Understanding Culture and Identity in Online Communities: Societal Dynamics and Implications

Theresa Henn-Latus
University of Bamberg

Bamberg 2025

Understanding Culture and Identity in Online Communities: Societal Dynamics and Implications

Dissertation zur Erlangung des akademischen Grades Doktor der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.) der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg vorgelegt von Theresa Henn-Latus im August 2025.

Erstgutachter: Prof. Dr. Oliver Posegga

Zweitgutachter: Prof. Dr. Andreas Jungherr

Mitglied der Promotionskommission: Prof. Dr. Ute Schmid

Tag der Disputation: 31.10.2025

URN: urn:nbn:de:bvb:473-irb-111437x
DOI: https://doi.org/10.20378/irb-111437

To everyone who believed in me when I thrived, and even more when I struggled.

Contents

Acknowledgements

This work marks a significant milestone in my life. Five years ago, I would never have imagined how much work, dedication, and endurance a dissertation would require. It has been a wild ride, filled with both rewards and challenges, and it has taught me more than I imagined. Yet, I could not have finished this ride without the many people who supported me along the way.

First and foremost, I want to thank my supervisor, Prof. Dr. Oliver Posegga, as well as my co-supervisors, Prof. Dr. Andreas Jungherr and Prof. Dr. Ute Schmid, for their invaluable guidance, thoughtful feedback, and ongoing support. Your input has played a key role in shaping this dissertation into what it is today.

Beyond academic mentorship, I am deeply grateful to my colleagues and friends who supported me through this journey. A very special thanks goes to Resi and Nina: without your friendship and support, this achievement would not have been possible. Thank you for all the laughter we shared, the support during challenging moments, unforgettable conference experiences, and, of course, the iconic Christel boat tours. I also want to thank Sarah and Julian, with whom I spent four years working on the same paper, an experience that taught me persistence and reminded me that even long academic marathons can be fun when run with the right people. Thank you for never doubting that this paper would eventually get published. Also, a heartfelt thanks to Gudrun Stilkerich, the heart and soul of the chair, whose extensive knowledge in all university-related matters has helped me more times than I can count. My sincere thanks also go to our copyeditor, Scott Cooper, who gave this work its final polish and who taught me, among other things, to appreciate the Oxford comma.

Last but by no means least, I want to thank my parents, Katja and Dieter, for your unconditional love and support throughout my life. I would not be who I am today without you. And most importantly, my deepest gratitude goes to my wife, Insa. Thank you for being by my side through the best and worst moments. You make me feel loved, appreciated, and accepted every single day. Without you, I would probably not have finished this dissertation.

To each and every one of you, thank you!

List of Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLM	Black Lives Matter
CLIP	Contrastive Language-Image Pre-Training
CNN	Convolutional Neural Networks
GPT	Generative Pre-Trained Transformer
IS	Information System
LLM	Large Language Model
OCCRP	Organized Crime and Corruption Reporting Project
SMO	Social Movement Organization
Tip-Adapter	Training-free adaption method for CLIP

Abstract

This dissertation investigates how online communities, particularly social movements, can become distinct and persistent actors in online environments characterized by low entry and exit barriers and fluid actor constellations. To approach this topic, I draw on the concept of collective identity, which enables collective actors to define a shared *we-ness*, distinguish themselves from other groups, and become recognizable to external audiences. I adopt a relational perspective on collective identity that encompasses two dimensions: the understanding that collective identity is formed in relation to other actors; and that it is shaped through the co-constitution of culture and structure. To specify the first dimension, I apply the concept of identity fields, analyzing collective identity formation across three actor categories: protagonist, antagonist, and audience. Each dimension is linked to a separate research question, with two questions formulated for the audience identity field, all of which contribute to answering the overarching research question. To substantiate the second dimension of the relational understanding, I draw on the concept of identity work, which helps specify the cultural and social expressions through which collective identity formation is examined in this dissertation. I outline the analytical potential offered by digital trace data, which, when analyzed with computational methods and qualitative approaches, enables a nuanced and large-scale empirical analysis of collective identity formation. Thus, this dissertation advances our understanding of how social movements online emerge as distinct and persistent actors through collective identity formation, offering valuable insights into digital protest and online organization.

Introduction

> [I]t is precisely because of the high level of individualization of our society as reflected in the personalized nature of social media communication that the construction of collective identity becomes so important and urgent. (Gerbaudo, 2015, p. 921)

In 2024, more than five billion people around the world were active on social media (Statista, 2025b), spending an average of 143 minutes per person per day on these platforms (Statista, 2025a). These figures reflect not only the widespread adoption of social media but also illustrate the central role platforms such as Reddit, X (formerly Twitter), Instagram, TikTok, and Facebook play in shaping how people connect and communicate. These platforms have fundamentally transformed social interaction (boyd, 2010), as a physical copresence and formal organizations are no longer a prerequisite for connecting like-minded individuals (Shirky, 2008). Instead, online communities, defined as large online collectives united by shared goals and interests (Sproull & Arriaga, 2007), represent a new form of social organizing, "bring[ing] together thousands of strangers across national, geographic, time zone, and organizational boundaries" (Johnson et al., 2015, p. 165).

Since these new forms of organizing are central to understanding digital collaboration, interaction, and user behavior, they have long been a subject of analysis in the information systems (IS) field. A vast body of literature examines various aspects of online communities, including their sustainability (Bock et al., 2015; Lasfer & Vaast, 2021), the language used within them (Danescu-Niculescu-Mizil et al., 2013; Fayard & DeSanctis, 2010; Moser et al., 2013; D. Zhang et al., 2024), characteristics of collaboration and member contributions (Faraj et al., 2011; Park & Zaggl, 2022; Ransbotham & Kane, 2011), drivers of member engagement (Chen et al., 2018; Connolly et al., 2023; Germonprez & Hovorka, 2013; S. Ray et al., 2014), and the role of leadership (Faraj et al., 2015; Johnson et al., 2015; Lu et al., 2022; Oh et al., 2016; Panteli & Sivunen, 2019).

In this dissertation, I focus on a specific type of online community that has attracted considerable attention in IS research and lies at the intersection of IS and the social sciences: *social movements online*. Social movements are defined as "a network of informal interactions between a plurality of individuals, groups and/or organizations, engaged in a political or cultural conflict, on the basis of a shared collective identity" (Diani, 1992, p. 13). These political or cultural conflicts are typically directed against actors perceived to be at the center of society, such as political authorities, who appear unwilling to implement the changes demanded by the movement (Rucht, 2023, p. 102 ff.). While this definition was originally grounded in the study of offline activism, the rise of social media and the emergence of movements such as Occupy Wall Street in the early 2010s (e.g., Kavada, 2015) have underscored the growing importance of the online sphere for social movements. Today, social media platforms play a central role in their mobilization, organization, and the articulation of demands (e.g., Gerbaudo, 2012; Milan, 2021).

Against this backdrop, social movements represent a distinct type of online community, as they move beyond merely sharing common interests to actively pursuing political and cultural change. At the same time, they share important organizational characteristics with other online communities, such as interest-based communities, centered on shared affinities like hobbies, music, and entertainment, and protest movements, which engage in public action to articulate political or societal demands (Rucht, 2003, p. 157). Their organizational similarities are rooted in their reliance on self-organization and peer production (Bennett et al., 2014), as all three types of communities must organize within online environments characterized by fluid boundaries and the absence of centralized leadership and formal hierarchies (e.g., Johnson et al., 2015). However, they differ in their stability and coordination: interest-based communities typically operate under more stable conditions, with clearly defined goals and outputs (Bennett et al., 2014, p. 236). In contrast, large-scale protest movements often lack such structuring conditions and are characterized by more fluid and dynamic organizational forms (Bennett et al., 2014). While social movements share some of these dynamic characteristics of protest movements, they tend to be more structurally stable, as they engage in long-term conflicts and rely on a shared collective identity (see definition). To better understand these organizational similarities and differences, I also include interest-based communities and protest movements as complementary cases, which enriches and contextualizes the analysis of social movements online that remains the central focus of this dissertation.

Better understanding these mechanisms of social movements' online organization and mobilization has become a central concern in IS research (e.g., Cardoso et al., 2019; Leong et al., 2019, 2020; Vaast et al., 2017; Young et al., 2019). Yet, while significant progress has been made in understanding various aspects of digital activism—such as the action repertoire of online activists (Selander & Jarvenpaa, 2016), factors contributing to movement continuity (Leong et al., 2019; Syed & Silva, 2023), and the role of social media in shaping protest cycles (Tarafdar & Kajal Ray, 2021)—questions remain about how social movements organize to become recognizable and distinct collective actors online, capable of distinguishing themselves from other actors and sustaining their persistence over time. Expanding our knowledge of these processes is essential, as social movements online increasingly shape public discourse, influence political agendas, and are capable of mobilizing large-scale collective action (e.g., Gerbaudo, 2012; Tufekci, 2017). These insights are, therefore, not only of academic interest but also hold practical value for policymakers, platform owners, and social movement organizations seeking to engage with, respond to, or leverage the digital potential of social movements online.

I contribute to this body of knowledge by employing *collective identity* as a concept to examine how social movements online emerge as distinct and persistent actors that can be demarcated from other online groups. As emphasized in the opening quote, in a world where social media platforms increasingly foster individualization, the concept of collective identity can help uncover the *we-ness* within these individualized communication spaces, enabling the identification of collective actors and the evaluation of their persistence over time. Building on this theoretical concept, this dissertation draws on the analytical opportunities offered by digital trace data and combines computational methods with qualitative analyses to enable a large-scale, yet nuanced examination of collective identity formation online.

This dissertation is structured as follows. The next section begins with a brief overview of the scholarly debate surrounding the organization and identification of social movements in online environments, with a particular focus on the role of collective identity in this context. Building on these insights, I formulate an overarching research question that guides the remainder of this dissertation. To answer this question, I adopt a relational understanding of collective identity, drawing on the seminal work of scholars such as Harrison White. This understanding comprises two conceptual dimensions: the notion that collective identity is constructed in relation to other actors; and that it is shaped by the co-constitution of culture and structure.

To substantiate the first dimension, I apply the concept of *identity fields*, which offers a multidimensional perspective for analyzing social movements' collective identity formation in relation to other actors. For each identity field, I formulate a specific research question, with two questions developed for the third field, each contributing to answering the overarching research question. To specify the second dimension, I apply the concept of *identity work* and outline the types of cultural and social expressions, as well as external identity attributions, that are analyzed to examine social movements' collective identity formation. Since these expressions manifest in digital trace data, I highlight the analytical opportunities such data offer for the analysis of collective identity formation.

Building on these conceptual insights, I introduce a framework that maps each article of this dissertation to the presented identity fields and distinguishes between different expressions of collective identity analyzed in each article. I then synthesize the main findings of each article in relation to the four underlying research questions to answer the overarching research question. This is followed by a brief executive summary of each individual article.

In the conclusion, I reflect on key insights, address limitations, and outline directions for future research. The dissertation closes with the appended research papers, a summary of the findings in German, and a list of the included publications.

Theoretical Background and Related Research

Social media platforms offer social movements new ways to organize and act collectively (e.g., Young et al., 2019), as they embody a fundamentally different form of *materiality* (e.g., Leonardi, 2012, p. 29) compared to offline environments in which collective action and movement formation occurred before the advent of social media and Web 2.0 technologies. To comprehend how social and material agency interact as mutually constitutive elements of one another, which is referred to as *sociomateriality* (Leonardi, 2012, p. 49), research commonly draws on the concept of affordances (Karahanna et al., 2018; Leonardi & Vaast, 2017; Strong et al., 2014; Treem & Leonardi, 2013; Vaast et al., 2017). Originating in ecological psychology (Gibson, 1977), an affordance is generally defined as "a possibility for action" (Bygstad et al., 2016, p. 84) that emerges "*from the relation between an object (e.g., an IT artifact) and a goal-oriented actor or actors*" (Volkoff & Strong, 2013, p. 823, emphasis in original). While the features that define an object's materiality exist independently of those who interact with it, affordances—as well as their actualization—depend on the actors' goals, competencies, and knowledge (Bygstad et al., 2016, pp. 87–88). Thus, the concept underscores the significance of technology's materiality in both enabling and constraining social action, without falling into a deterministic view that assigns sole responsibility to either technological or social factors (Ronzhyn et al., 2023, p. 3175; Vaast et al., 2017, p. 1182).

Leveraging these new possibilities for action enabled through platform affordances presents both opportunities and challenges for social movements. On the one hand, social media can empower movements (Milan, 2021) by enabling new forms of mobilization and organization characterized by horizontal and leaderless structures (Juris, 2008; Milan, 2021). This allows networked individuals (Rainee & Wellman, 2012) to coordinate collective action without relying on the formal organizational structures typically imposed by social movement organizations (SMOs), effectively enabling what Shirky (2008) calls "organizing without organizations." Social media thus allows for new modes of interaction and connection that do not require physical co-presence (e.g., Milan, 2015c). In addition to enabling these new forms of engagement, digital platforms also support the exchange of cultural artifacts, such as hashtags and memes (Milan, 2021, p. 301), that carry symbolic meaning and foster group cohesion. Together, these new forms of connection and exchange allow dispersed individuals to engage, unify around shared symbols, and form collective actors capable of coordinated action. This powerful mobilization potential has been demonstrated by both earlier and more recent movements, such as Occupy Wall Street (Kavada, 2015), Black Lives Matter (BLM) (Mundt et al., 2018), and MeToo (Reyes-Menendez et al., 2020), as well as protests during the Arab Spring (Tufekci, 2017). These examples show that it is possible to form persistent online collectives that are characterized by sustained engagement and the ongoing mobilization of supporters.

On the other hand, while social media can significantly "lower the cost of participation and mobilization" (Milan, 2021, p. 302), the same platform affordances that enable this rapid engagement also make it easy for users to join and leave groups and conversations with minimal effort, resulting in low entry and exit barriers for engaging in collective action (Massa & O'Mahony, 2021). This fluidity of actor constellations is further amplified by "the absence of either formal governance or managerial control" (Massa & O'Mahony, 2021, p. 1041), which means that social movements online must rely on self-organization, as the organizational structures typically provided by SMOs are no longer present. Yet, this self-organization depends on the voluntary contributions by users (Johnson et al., 2015), which often makes it difficult to establish lasting social ties and keep up long-term mobilization (Milan, 2021). As a result, online collectives often exhibit "fluid structures and shifting membership" (Johnson et al., 2015, p. 166; Faraj et al., 2011, 2015), which can impede the formation of collective actors that can be clearly distinguishable from other groups online, as social media often reduce sustained commitment (Milan, 2015a).

To fully understand the academic discourse on online mobilization, formation of collectives, and the enactment of (sustained) collective action by social movements online, it is essential to return to Bennett and Segerberg's seminal work *The Logic of Connective Action* (2012, 2013). In contrast to the traditional concept of *collective action*, the authors argue that *connective action* reflects more accurately how social movements organize and mobilize in online environments. While traditional collective action is characterized by formal structures imposed by SMOs, strong ideological commitment, and a high level of coordination among activists, connective action is a new form of collective engagement enabled via social media and organized through informal online networks. These networks do not require "strong ideological bands or collective identities" (Bennett & Segerberg, 2013, p. 50) among individuals and generally do not promote long-term commitment among protesters and formal organizations. In other words, as Bakardjieva (2015, p. 985) notes in her interpretation of Bennett and Segerberg's work, collective identity can be bypassed, as individuals no longer need to engage in the laborious process of negotiating shared meanings or constructing a collective *we* to participate in collective action. Instead, highly inclusive action frames enable spontaneous and large-scale mobilization without requiring participants to adopt "more demanding models regarding how to think and act" (Bennett & Segerberg, 2013, p. 197). As a result, the authors argue that adherents tend to be less committed, and social movements online often disperse quickly, failing to foster lasting connections between supporters and the movement.

While the concept of connective action has played a central role in shaping our understanding of social movement organization online, it has been increasingly revisited and expanded in recent years. For instance, Vaast et al. (2017) introduce the concept of *connective affordances* to better understand how social media enable new forms of engagement and how platform affordances, in combination with interdependent user roles, facilitate the enactment of connective action. By examining the empowerment mechanisms of social media in greater depth, Leong et al. (2019) demonstrate that understanding long-term commitment for social movements online requires differentiating between periods of rupture and abeyance. While also finding support for aspects of connective action, they show that social media can "retain[] the people (resources) needed to accumulate and maintain the power to carry on the movement" (p.191). This suggests that social media platforms are not only used for short-term mobilization but also enable the formation of persistent social movements, whose visibility shifts over time. Expanding on these insights, Leong et al. (2020) identify two types of emergent mechanisms fostered by social media—clustering and structuring, which enable movements to transition from spontaneous connective action to more sustained mobilization of supporters. Based on their analyses, they identify different organizational stages associated with different types of emergence, which ultimately lead to the formation of networked groups whose sustained activities can be described by "a logic of organized collective action" (p. 3). Diving deeper into the reasons for sustained engagement, Syed and Silva (2023) demonstrate that long-term connective action by social movements online is enabled through the sharing of messages featuring collective action frames that also foster the formation of a movement's collective identity.

Thus, in recent years, scholars have increasingly challenged Bennett and Segerberg's assumption that online collective action is primarily spontaneous and short-lived. Their research highlights how specific affordances, emergence mechanisms, and framing strategies can facilitate the formation of persistent collective actors capable of developing a collective identity. Real-world movements such as BLM (Mundt et al., 2018), MeToo (Reyes-Menendez et al., 2020), and Fridays for Future (Nasrin & Fisher, 2022) further demonstrate that social movements can indeed emerge as recognizable and persistent actors online, characterized by a distinct identity that distinguishes them from other groups. This aligns with established social movement research, which emphasizes collective identity as an essential "organizing principle" (della Porta & Diani, 2006, p. 93) and a prerequisite for creating persistent and sustained actors capable of coordinated action (see, e.g., Hunt & Benford, 2004; Polletta & Jasper, 2001). As Melucci (1995, p. 49) argues, "Collective identity ensures the continuity and permanence of movements over time; it establishes the limits of the actor with respect to its social environment." Given its central role in movement formation and persistence, Gamson (1991, p. 27) concluded that social movements "must make the construction of collective identity one of [their] most central tasks."

These findings stand in contrast to Bennett and Segerberg's logic of connective action, which describes online mobilization as primarily spontaneous and short-lived, lacking the need for the formation of a collective identity. To address this contradiction, I pose the following overarching research question that guides all further theoretical considerations, empirical analyses, and conclusions of this dissertation:

RQ: How do social movements online become distinct and persistent actors?

To answer this question and contribute to the ongoing debate about the role and form of collective identity online, I adopt a relational understanding of collective identity, which is introduced in sections that follow.

While related concepts, such as connective emotions (George & Leidner, 2019), framing (Syed & Silva, 2023), and solidarity (Stewart & Schultze, 2019), have also been used to explain the sustainability and formation of social movements online, collective identity is a more suitable concept for addressing the question. Rather than focusing solely on feelings of identification or being "interpretive packages that activists develop to mobilize potential adherents" (Polletta & Jasper, 2001, p. 291), collective identity captures the broader processes through which movements emerge, draw boundaries, and persist as distinct actors in digital environments (for more information on these concepts in relation to collective identity, see, e.g., Hunt & Benford, 2004, p. 439; Polletta & Jasper, 2001, p. 291; Stewart & Schultze, 2019, p. 3).

While using collective identity to examine long-term mobilization and to understand the organization of social movements online is not unprecedented (e.g., Barandiaran et al., 2020; Gerbaudo & Treré, 2015; Kavada, 2015; Milan, 2015a, 2021; Monterde et al., 2015; Treré, 2015), the following section outlines conceptual limitations in existing approaches and explains how I address them with a relational conceptualization of collective identity.

Collective Identity

Despite its conceptual significance, there is still "no consensual definition" (Snow, 2001, p. 2) of collective identity. This remains true even after decades of efforts by social movement scholars to define the concept, beginning with Blumer's (1939) emphasis on the need for an *esprit de corps* to foster unity and enable collective action, and continuing through numerous attempts by others (e.g., Flesher Fominaya, 2010; Hunt & Benford, 2004; Melucci, 1995, 1996; Polletta & Jasper, 2001; Taylor & Whittier, 1992).

I adopt David A. Snow's definition as a working definition of collective identity for this dissertation. Snow defines collective identity as

> a shared sense of "one-ness" or "we-ness" anchored in real or imagined attributes and experiences among those who comprise the collectivity and in relation or contrast to one or more actual or imagined sets of "others." Embedded within the shared sense of a "we" is a corresponding sense of "collective agency." (Snow, 2001, p. 2)

While the significance of Snow's definition has been acknowledged in the context of collective identity formation within social movements online (e.g., Stewart & Schultze, 2019), much of the literature on collective identity in online activism (e.g., Coretti & Pica, 2015; Gerbaudo & Treré, 2015; Kavada, 2009, 2015; Milan, 2015a, 2015b, 2015c; Treré, 2015) draws on Alberto Melucci's definition (1995, 1996), often described as "the obligatory entry point" (Monterde et al., 2015, p. 931). From Melucci's perspective, collective identity arises from the formation of a shared understanding forged during interaction processes (e.g. Melucci, 1995, p. 44). Hence, Melucci adopts a processual perspective on collective identity, emphasizing that it is not a fixed or static concept but rather "a recurrent process of activation of the relations that bind actors together" (Melucci, 1996, p. 70). Against this background, he defines collective identity along three dimensions: the sharing of "cognitive definitions concerning the ends, means, and field of action" (Melucci, 1995, p. 44); the formation of a "network of active relationships between the actors" (Melucci, 1995, p. 45); and the existence of "a certain degree of emotional investment, which enables individuals to feel like part of a common unity" (Melucci, 1995, p. 45).

Despite the widespread use of Melucci's understanding of collective identity, much of the research applying his framework falls short of assessing all three dimensions of identity formation, as researchers either rely on qualitative methods that focus on the evaluation of the cognitive dimension while overlooking the active relationships and emotional investments (e.g., Kavada, 2015), or conduct structural and dynamic network analyses that neglect the cognitive and emotional dimensions (e.g., Monterde et al., 2015). Yet, according to Melucci, identifying and analyzing all three dimensions is necessary for understanding collective identity (Melucci, 1995, pp. 44–45). The challenges of evaluating all three dimensions highlight the ongoing difficulty of translating Melucci's far-reaching conceptual considerations into a concrete operationalization. This issue has also been raised in previous research. Opp (2009, p. 211) notes that while Melucci identifies numerous aspects relevant for the emergence of collective identity, he offers little guidance on how these aspects interact and how they can be measured empirically. This conceptual vagueness extends to the definition of key terms, the relationship between his proposed dimensions, and the connection between micro- and macro-level processes (Opp, 2009, pp. 210–215). As a result, Melucci's work has been considered to offer "orienting hypotheses" (Opp, 2009, p. 211) rather than a coherent empirical approach.

These challenges of operationalizing Melucci's approach and disentangling its different analytical levels persist in both offline and online contexts. Online environments, however, present an additional difficulty that complicates the meaningful application of his understanding of collective identity: Melucci's approach relies heavily on the self-reflective assessment of identity by movement participants (e.g., Melucci, 1996, p. 73). This reflective aspect is inherently inaccessible when working solely with digital trace data (Howison et al., 2011), as such data capture only the observable outcomes of actions, offering no insights into actors' internal reflections, interpretations, and intentions.

Taken together, the challenges of translating Melucci's definition into empirical measures, the difficulties of observing the interplay between his analytical levels, and the reliance on self-assessments by movement participants make it difficult to apply his definition of collective identity to the empirical analysis of social movements in online environments. Nevertheless, his work remains a crucial point of departure for this dissertation, particularly because of its emphasis on the idea of identity as an emergent, processual phenomenon constructed through mutual recognition and active relationships among actors within a shared social field. Thus, although not all aspects of Melucci's conceptualization are fully grounded in a relational perspective, his approach incorporates key relational elements by emphasizing that a "collective actor cannot construct its identity independently of its recognition ... by other social and political actors" (Melucci, 1995, pp. 47–48). Instead, collective identity arises through mutual recognition within a social field of ongoing interaction, where "both actors and observers can give account of this field through unified, delimited, and static definitions of the 'we'" (Melucci, 1995, p. 50). This reciprocal process helps ensure movement continuity and permanence. Building on these insights, this dissertation adopts a fully *relational approach* to collective identity formation, drawing on insights from seminal authors such as Harrison White (1992, 2008).

Relational Understanding of Collective Identity Online

Adopting a relational approach to the analysis of collective identity in online environments addresses key limitations of Melucci's concept, as is briefly summarized here and detailed in the following section. First, unlike Melucci's reliance on actors' self-assessment, a relational perspective enables the empirical investigation of observable cultural and social traces users leave behind, which can be aggregated from the individual to the macro level. This not only bridges the gap between micro and macro levels from Melucci's approach, but also eliminates the need for participants' self-assessment. Second, these traces can be translated into empirical indicators, thereby addressing another major shortcoming of Melucci's approach. To fully understand the relational perspective, it is necessary to first outline the foundational assumptions of relational sociology, which underpin the theoretical approach adopted in this dissertation.

Relational sociology is grounded in the premise that social actors and cultural elements derive their meaning only "in and through their relatedness to others" (Mohr & White, 2008, p. 489). Accordingly, they are not defined by any "inherent attributes or properties," but by "their relational location within a field of relations," as "it is the relations (and the patterns of relations) that matter" (Mohr & White, 2008, p. 489). These relations are dynamic in nature (Emirbayer, 1997, p. 289), making the ongoing positional shifts and interactions between actors and meanings the focus of the analysis (Emirbayer, 1997, p. 287). The cultural structures and actor constellations that manifest in these processes can be mapped in networks since both are fundamentally "ordered in relational systems"

(Saussure 1959 as cited in Mohr & White, 2008, p. 489; see also, e.g., Basov et al., 2020; Basov & Brennecke, 2017; Breiger, 2010; Lee & Martin, 2018; Mohr, 2000; Mohr & Duquenne, 1997; Mohr et al., 2020; Powell & Dépelteau, 2013). In line with these assumptions, relational sociology views identity not as a fixed or inherent attribute, but rather as the outcome of ongoing processes of differentiation and identification (Powell & Dépelteau, 2013, p. 2). On this basis, the relational perspective offers a powerful approach for evaluating what Melucci describes as the "relational texture of social phenomena" (1996, p. 80), a key feature of his collective identity conceptualization.

I apply a relational perspective on collective identity formation, which comprises two dimensions: the notion that collective identity is constructed in relation to other actors situated within a social field; and the co-constitution of culture and social structures, as reflected in the digital traces users leave online. Together, these two dimensions enable a nuanced and multifaceted analysis of social movements' collective identity formation in online environments.

The first dimension has already been alluded to in Snow's (2001) definition introduced earlier, where he highlights that self-aware actors capable of collective action create a shared sense of *we-ness* in relation to an actual or imagined set of others. This differentiation is a fundamental component of a relational understanding of collective identity formation and is commonly referred to as *boundary work* (see, e.g., Flesher Fominaya, 2010; J. Gamson, 1995; Gieryn, 1983; Hunt & Benford, 2004; Taylor & Whittier, 1992). Boundary work involves emphasizing in-group communalities while simultaneously underscoring differences from the out-group (Snow, 2001, p. 7), thereby clearly demarcating the group from other actors online and expressing its collective identity (e.g., Hunt & Benford, 2004; Melucci, 1995). These boundaries are often expressed through the sharing of *boundary markers*, which are symbolic resources that reinforce these group distinctions (Snow, 2001, p. 7).[1] Since the concept of boundary work can be applied empirically to "almost any grouping or aggregation in a variety of contexts" (Snow, 2001, p. 3), it offers strong analytical utility by helping overcome the limitation of connecting different levels in Melucci's approach, providing a tangible way to analyze collective identity formation.

To elaborate on the second dimension of the relational perspective applied in this dissertation—namely, the co-constitution of culture and social structures—it is first necessary to define both terms. As with the concept of collective identity, there is no "consensual, unitary definition" (Mohr et al., 2020, p. 15) of culture. Rather, the literature offers a range of conceptualizations and definitions (e.g., Johnston & Klandermans, 1995; Swidler, 1986; Williams, 2013; Wuthnow, 1987). I adopt the definition proposed by Nina Eliasoph and

[1] The online expressions of these symbolic resources examined in this dissertation are discussed in greater detail in the "Identity Work: Expressions of Collective Identity Online" and "Classification of Articles" sections.

Paul Lichterman, who broadly describe culture as "a set of publicly shared codes or repertoires, building blocks that structure people's ability to think and to share ideas" (2003, p. 735). They further emphasize that the same cultural artifacts, such as symbols, vocabularies, and codes, "can take on different meanings in different contexts" as "[p]eople make meanings in specific social settings ... and they make those meanings in relation to each other as they perceive each other" (Eliasoph & Lichterman, 2003, p. 736). From this perspective, culture is not a static system of meaning but rather a dynamic process of meaning-making shaped through social interaction and context-dependent interpretation (Eliasoph & Lichterman, 2003, p. 782). This understanding of culture aligns closely with the relational perspective on collective identity introduced earlier in which collective identity is likewise seen as emerging through interaction and in relation to other groups. The chosen definition of culture, therefore, complements this relational perspective that constitutes the conceptual basis of this dissertation.

The importance of social structure in the analysis of social movements, particularly in relation to culture and collective identity, has been emphasized by Mario Diani. He argues that "[s]ocial ties discourage certain courses of action and facilitate others, which in turn affects how actors attribute meaning to their social linkages, thus (re)creating rules and arrangements perceived as relatively stable and 'structural'" (Diani, 2003, p. 303). Through ongoing interaction and processes of cultural production, "people create webs of ties that enable them to further act collectively and shape their future behaviour" (Diani, 2003, p. 303). In this way, Diani introduces a network-based perspective on the analysis of social movements, which also aligns with the definition of social movements adopted earlier in this dissertation. By emphasizing the role of social ties and relational structures in the context of social movements' cultural production, he highlights the importance of social structure in both enabling and constraining collective action, meaning-making processes, as well as shaping the formation of collective identity.

Having briefly defined both concepts, I now turn to Harrison White (1992, 2008), one of the most influential representatives of relational sociology. I use his work to explain how the co-constitution of culture and social structures informs this dissertation's conceptualization of collective identity formation. For White, networks and culture—his two fundamental conceptual building blocks for analyzing social life—converge in the concept of *netdoms*. On the one hand, this concept captures that actors constantly switch among "specialized fields of interactions" (Mische, 2003, p. 265) that inform their day-to-day lives. Each of these fields is characterized by its respective domain (the "dom" in "netdom"), defined as an "array of signals—including story sets, symbols, idioms, registers, grammatical patternings, and accompanying corporeal markers—that characterize a particular specialized field of interaction" (Mische & White, 1998, p. 702). Thus, the domain reflects the cultural dimension of social life. On the other hand, the concept incorporates the notion that the formal network topology (the "net" in "netdom") of relevant actors changes as they switch among these interaction fields. Neither of the two elements can be analyzed separately, as the network's edges themselves are "constituted and reconstructed through the course of communication" (White et al., 2007, p. 545), a process that draws on the ordering of cultural elements specific to the domain in which the ties evolve. These cultural elements, however, constitute expressions whose meaning is linked to particular social contexts (e.g., White, 2008, p. 337). Hence, White's theoretical approach builds on the idea of duality between "the world of action" and the "cultural world"(Mohr & Duquenne, 1997, p. 309), emphasizing that cultural and social dimensions are deeply intertwined and must be examined together (e.g., White, 2008, p. 337).

Building on this perspective, White's understanding of *identity formation* is grounded entirely in a relational understanding that incorporates the duality of social and cultural order (White, 2008). Neither can be analyzed independently, since networks and domains are conceptualized as "dual and co-constitutive" (Mische & White, 1998, p. 695). Against this conceptual background, identity emerges as soon as actors attribute meaning to an entity they perceive as a coherent source of action (White, 2008, p. 2)—which results from social interaction in a particular context, that is, in a netdom at a specific time. In other words, identities manifest as patterns of "*observable processes-in-relations*" (White, 1997, p. 60, emphasis in original), that is, as recurring social interactions among a set of actors tied to a distinct domain.[2] This focus on observable relational processes, grounded in the joint evaluation of cultural and social dimensions, can be meaningfully transferred to online environments, where they manifest through cultural and social traces users leave behind. By focusing on traces that are *observable*, this approach avoids the reliance on self-reflective assessments that characterize Melucci's framework. Furthermore, the visibility of these traces makes it possible to operationalize them into empirical indicators, thus addressing another limitation of Melucci's concept.

To briefly summarize, the relational perspective on social movements' collective identity formation adopted in this dissertation builds on two key dimensions. First, it is based on the understanding that identities emerge in relation to other actors, with particular emphasis on the differences between "us" and "them," as highlighted through the notion of boundary work (see Snow, 2001). Second, it incorporates the idea of the duality of the cultural and social dimensions, which together constitute collective identity when recurring patterns emerge across both dimensions (White, 2008). The importance of evaluating social movements' collective identity formation through these two perspectives[3] has also been acknowledged in previous literature, which notes that

> Collective identity is not strictly an individual attribute. Rather, it is a cultural representation, a set of shared meanings that are produced and reproduced, negotiated

[2] Article I includes a more detailed account of White's conceptualization of identity.
[3] Article I includes a comprehensive integration of Snow's and White's approaches to collective identity that provides both a conceptual framework and a concrete empirical operationalization.

> and renegotiated, in the interactions of individuals embedded in particular sociocultural contexts. (Hunt & Benford, 2004, p. 447)

> Research indicates that collective identity not only emerges from the interactions of activists within a specific movement or movement organization, but it is also produced from the relationships between allies, those opposed to the movement in some fashion, and bystander audiences such as print and electronic news media. (Einwohner 2002 and Meyer 2002, as cited in Hunt & Benford, 2004, p. 447; see also Polletta & Jasper, 2001, p. 298 for a similar interpretation)

In summary, the relational perspective is particularly well-suited for analyzing collective identity formation online based on digital trace data, as it focuses on the empirical observation of recurring patterns, both in relation to other actors and in the interplay between the cultural and social dimensions. These recurring patterns through which collective identity is expressed can be translated into empirical indicators, offering a tangible way to connect micro- and macro-level processes. When aggregated, such patterns form a recognizable signature through which collective actors in online environments can be identified and distinguished. Thus, applying a relational perspective offers a robust approach to both conceptualizing and empirically analyzing collective identity formation online, thereby providing a compelling alternative to existing approaches.

In the following, I introduce the concept of identity fields, which defines the set of actors that constitute the relational field within which social movements operate and against whom they position themselves when constructing their collective identity. Thus, this concept concretizes the first dimension of the relational perspective on collective identity formation as derived from Snow's definition.

Identity Fields: Relevant Actors for Collective Identity Formation

In 1994, Scott A. Hunt, Robert D. Benford, and David A. Snow (1994) introduced the concept of identity fields, bringing together two previously disconnected yet highly influential theoretical concepts in social movement research: framing and collective identity formation. By integrating these concepts, the authors aimed to enhance understanding of how social movement actors mobilize collective action and come to operate as unified actors. They argue that "identity constructions, whether intended or not, are inherent in all social movement framing activities" (Hunt et al., 1994, p. 185), thereby emphasizing the inseparability of framing and the collective identity formation and their dynamic and reciprocal relationship. More specifically, the authors argue that identities are established through framing processes "by situating or placing relevant sets of actors in time and space and by attributing characteristics to them that suggest specifiable relationships and lines of action" (Hunt et al., 1994, p. 185). This perspective emphasizes the relational understanding of collective identity formation, recognizing that social movements do not operate in isolation but within a "*multiorganizational field*" (Benford, 2013, p. 1, emphasis in original), where ongoing interactions with other actors continually shape their collective identity formation, thereby aligning with the relational perspective introduced before.

The concept of collective identity and its relational understanding has been discussed in detail above. I now briefly introduce the concept of framing to establish a shared understanding of the term and clarify how both concepts interact within the concept of identity fields.

Framing, originally introduced by Goffman (1974), refers to the use of "an interpretive schemata that simplifies and condenses the 'world out there' by selectively punctuating and encoding objects, situations, events, experiences, and sequences of actions within one's present or past environments" (Snow & Benford, 1992, p. 137). While the concept has gained considerable traction in fields such as media and communication studies (e.g., Entman, 1993; Gitlin, 2003), it has been particularly influential in social movement research, especially through the work of Snow and Benford (1988). In their article, the authors introduced the widely adopted typology of diagnostic, prognostic, and motivational framing, which social movements employ as part of their communication strategies. Diagnostic framing refers to how a movement defines a particular issue or problem and assigns responsibility or blame; prognostic framing builds on this by proposing possible solutions, including specific tactics, and targets of action; and motivational framing goes further by issuing a call for action that moves beyond problem identification and solution (Snow & Benford, 1988, pp. 200–202). Framing is, therefore, crucial to how social movements make sense of the world, as it not only shapes their "ideology or belief systems" but also functions as a set of "cognitive structures that guide collective action" (Hunt et al., 1994, pp. 191–192).

In the context of collective identity formation, a key function of framing is the attribution of "characteristics to relevant sets of actors within a movement's orbit of operation" (Hunt et al., 1994, p. 192). Based on previous social movement research, Hunt, Benford, and Snow (1994, p. 186) identified three sets of actors relevant for the collective identity formation of social movements: the protagonist, the antagonist, and the audience. The dynamic interplay among these "socially constructed movement actor categories" (Benford, 2013, p. 1) and the identities attributed to them through framing processes is conceptualized as identity fields (Benford, 2013; Hunt et al., 1994). While each identity field is important by itself, their combination offers a deep and multifaceted understanding of how various actors relevant to social movements, and the identities attributed to them, contribute to the construction of a movement's overall collective identity and its capacity for collective action. Yet, it is important to note that the boundaries between these fields are not fixed; instead, the identities attributed within each field frequently overlap and influence one another, especially when examined over time (Hunt et al., 1994, p. 186).

The *protagonist identity field* comprises collective identity attributions related to "individuals and collectives taken to be advocates of the movement cause" (Hunt et al., 1994, p. 193), focusing on the characteristics of the movement itself and its immediate allies. Within this field, all three previously introduced framing processes—diagnostic, prognostic, and motivational—typically come into play, as the movement defines its capacities and builds a sense of agency by situating its "organization and its views within a collective action field or context" (Hunt et al., 1994, p. 193). Hunt, Benford, and Snow (1994, pp. 193–194) also discuss the importance of distinguishing between in-groups and out-groups and the implementation of symbolic boundaries within this field, thereby addressing the concept of boundary work featured in Snow's collective identity definition. Yet, while initial boundary work begins here, particularly in establishing who belongs to the movement, the more explicit attribution of characteristics to opponents is central to the antagonist identity field, is addressed in the following section. Benford (2013) reinforces this by highlighting that the drawing of boundaries between "us" and "them" is a key feature of the antagonist identity field. Thus, the protagonist identity field focuses primarily on the social movement itself and the characteristics attributed to it as it develops self-awareness and a sense of agency. In light of my earlier discussion of collective identity, I interpret this process as the creation of a shared sense of *we-ness* as introduced by Snow that enables the formation of a unified actor capable of collective action. To gain a deeper understanding of this aspect of social movements' collective identity formation online, I pose the following research question:

RQ1: How do social movements online create and sustain a shared sense of we-ness?

To develop a collective identity, social movements must define not only "who they are" but also "who they are not." This is where the *antagonist identity field* becomes relevant. It is defined as "constellations of identity attributions about individuals and collectivities imputed to be opponents of movement causes" (Hunt et al., 1994, p. 197). The identity attributions within this field are typically negatively connoted, since opponents are framed as adversaries and positioned in direct contrast to protagonist identities (Hunt et al., 1994, p. 198). As suggested in the discussion of the protagonist identity field, these negatively associated identity traits attributed to opponents serve as boundary work, drawing clear boundaries between the in-group and out-group, which are often framed in terms of good versus evil (Benford, 2013). By positioning itself in contrast to its opponents, the movement elevates its own status by implicitly portraying its characteristics as morally superior. To better understand how these processes of demarcation from opponents contribute to the overall collective identity formation of social movements, and in light of the previously emphasized importance of boundary work, I pose the following research question:

RQ2: How do social movements online draw boundaries to differentiate themselves from other groups?

While scholarly attention has often centered on the boundary work between the protagonist and antagonist identity fields (Benford, 2013, p. 2), it is equally important to consider actors who observe these actions and the associated identity manifestations when evaluating the collective identity formation of social movements. The importance of external recognition in the construction and maintenance of collective identity has also been emphasized by Melucci, who argues that such identity formation requires "at least a minimal degree of reciprocity in social recognition between the actors" (Melucci, 1995, p. 48). Hunt et al. (1994, p. 199) refer to these observing actors as part of the *audience identity field*, defined as "constellations of identity attributions about individuals and collectivities imputed to be neutral or uncommitted observers who may react to or report on movement activities." Within this identity field, social movements must strategically frame both their protagonist and antagonist identities in ways that resonate with the audience, as audience reports and interpretations can significantly influence public support. In addition to anticipating how their portrayal will be received, movements must also respond to the identity attributions made by these observers. This includes interpreting, contesting, and reinforcing such attributions as part of their ongoing process of collective identity formation (Benford, 2013; Hunt et al., 1994).

The media,[4] a prominent example of actors within the audience identity field, have long played a central role in reporting on social movements, "hav[ing] historically provided material and symbolic support for a movement's elaboration of 'who we are'" (Milan, 2015a, p. 888). Given its influential role in framing social movements, this dissertation focuses on the media as a representative actor of the audience field and examines how it contributes to shaping a social movement's collective identity. However, to understand how movements position themselves in relation to the media and their attributed identity traits, as highlighted in the initial conceptualization of the audience identity field, it is first necessary to analyze how media assign these identity traits to social movements.

Expanding our knowledge of this process is essential, as prior research has demonstrated that media framing and identity attribution have a powerful influence on public opinion, impacting whether social movements are perceived as legitimate actors or considered to be disruptive and illegitimate (e.g., Brown & Mourão, 2021; Harlow & Brown, 2023; Masullo et al., 2024; McLeod & Hertog, 1992). Studies have found that media coverage often emphasizes negative or delegitimizing aspects of social movements rather than portraying them as credible actors with justified grievances and concrete demands, a pattern referred to as the *protest paradigm* (Chan & Lee, 1984; McLeod & Hertog, 1999). Such portrayals can, in turn, significantly hinder a movement's ability to gain public legitimacy and mobilize broader support (e.g., Harlow & Brown, 2023, p. 334). It is, therefore, crucial to examine

[4] Note that this dissertation exclusively analyzes online news articles shared on social media. Herein, media actors are, therefore, understood as digital extensions of traditional news outlets (e.g., *The New York Times*) as well as digital-born news outlets such as blogs. Article VII offers further information on the different types of media actors considered in this dissertation.

critically how media actors attribute identities to social movements, as media framing plays a central role in shaping how those movements' collective identity is perceived in the public sphere.

This dissertation does not examine how social movements position themselves in response to media identity attributions and framing within the audience identity field, but rather concentrates on how these attribution processes occur from the media's perspective. Given the complexity of these dynamics, many questions emerge. This dissertation addresses two of them:

RQ3a: How can emerging technologies support both the analysis and production of media reporting that attributes identity traits to social movements online?

RQ3b: How does the media's attribution of identity traits to social movements affect user attention?

These two questions address distinct but interconnected aspects of media's identity attribution to and framing of social movements: the role of technology in shaping and analyzing such attributions, and the impact of these portrayals on audience attention. Although these questions do not exhaust the entire landscape of media's role in assigning identity traits to social movements, they do provide valuable initial insights and were chosen for two key reasons. First, technological advances—particularly in computer vision, multimodal models, and large language models (LLMs)—offer unprecedented possibilities for systematically analyzing how identity traits are attributed and communicated in media coverage. Thereby, these technologies contribute to a more comprehensive understanding of how collective identity formation unfolds online. At the same time, these technologies are also increasingly embedded in the production of news itself, performing tasks such as translation, automated tagging, content recommendation, audience analytics, and summarization (Simon, 2024, p. 150). As such, it is crucial not only to use these technologies for analytical purposes but also to examine critically how they shape journalistic practices and can be used for assigning identity traits to social movements.

Second, we know that the media's portrayal of social movements can significantly shape public perception and determine the extent to which these movements receive public support. Understanding which types of identity attributions attract greater user attention can offer important insights into how visibility, resonance, and media framing interact in online environments (for an analysis of how images in this context impact visibility on social media, please, see Neumayer & Rossi, 2018). By examining how media representations affect user attention, we can better understand which portrayals resonate most strongly and how these, in turn, may shape broader public opinion and support for social movements.

Overall, the identity field approach enables the multifaceted analysis of social movements' collective identity formation in relation to multiple actors and perspectives, while integrating the widely recognized concept of framing. It thus moves beyond research focused solely on a singular dimension of collective identity formation, typically centered on the movement itself, acknowledging that collective identity emerges through a movement's relationship with other actors.

I now return to the second dimension of this dissertation's relational understanding of collective identity: the co-constitution of culture and structure. In the following section, I specify the expressions of collective identity within the previously defined identity fields that will be analyzed to assess social movements' collective identity formation online. To define these expressions precisely, I draw on the concept of identity work and external identity attributions, discussed in more detail below.

Identity Work: Expressions of Collective Identity Online

Broadly defined, identity work encompasses "anything people do, individually or collectively, to give meaning to themselves or others" (Schwalbe & Mason-Schrock, 1996, p. 115; see, also McAdam & Snow, 2000; Snow, 2001; for the relation between identity work and the construction of personal identities, please see Snow & Anderson, 1987). This meaning-making process is particularly shaped by the "generation, invocation, and maintenance of symbolic resources" (Snow, 2001, p. 7), which help establish the collective as a distinct actor while drawing boundaries toward other groups. Thus, the key element of identity work is the sharing of symbolic resources that define the group in relation to others—a cultural process that closely aligns with the definition of culture applied in this dissertation. These symbolic resources can take many forms, such as rituals, songs, names, and interpretative frameworks (Snow, 2001, p. 7), and serve to distinguish insiders from outsiders, thereby depicting a manifestation of the concept of boundary work introduced earlier. Schwalbe and Mason-Schrock describe this process of differentiation as a *semiotic bricolage* (1996, p. 119), which Snow interprets as an essential part of collective identity formation by giving "symbolic substance to the claimed distinctive 'we'" (Snow, 2001, p. 7). This symbolic expression is essential not only for expressing collective identity within social movements but also for communicating that identity to the broader public.

In its original sense, identity work captures the cultural dimension of collective identity formation through the expression and sharing of symbolic resources. In this dissertation, I examine two specific types of symbolic resources shared online, narratives and memes, which play a key role in shaping collective identity. Narratives are closely tied to framing processes and play a central role in collective identity formation by helping movements "create boundaries between themselves and outsiders" (Steward Jr. et al., 2002, p. 114). They thus contribute to the identification of out-groups and the definition of group membership, thereby expressing a form of boundary work. In this dissertation, narratives are understood as written accounts that articulate this distinction between in-groups and out-groups.

Memes, the second symbolic resource examined in this dissertation, have been identified in prior research as powerful identity-shaping symbols for online communities (e.g., DeCook, 2018; Gal et al., 2016; Nissenbaum & Shifman, 2017). They contribute to group identification and community building (e.g., DeCook, 2018; Nissenbaum & Shifman, 2017; Seiffert-Brockmann et al., 2018), serve as cues for group membership (Nissenbaum & Shifman, 2017), and reflect shared norms and ideologies (e.g., DeCook, 2018). Analyzing how social movements online use memes in their communication thus offers valuable insights into how they express collective identity on a cultural dimension. Throughout this dissertation, I focus on memes in the form of image macros, defined as "a repeating image overlaid with text and usually deal[ing] with a specific topic or situation" (Nissenbaum & Shifman, 2018, p. 298), as they represent one of the most widely shared meme formats online (see, e.g., Nissenbaum & Shifman, 2018).

Taken together, I analyze narratives and memes as concrete symbolic resources of identity work to examine collective identity formation on a cultural level. However, to fully capture the co-constitutive nature of collective identity, it is necessary to broaden the concept of identity work to also include a social dimension examined through network structures. Such structures arise from user interactions like following, reposting, or commenting on others' content (see, e.g., Karahanna et al., 2018 for an overview of platform affordances that enable or constrain such actions). Accordingly, this dissertation conceptualizes identity work as comprising both symbolic resources as representations of the cultural dimension, expressed through narratives and memes, and network structures as a manifestation of the social dimension.

This understanding of identity work reflects the *internal* perspective of collective identity formation, capturing how movements construct and express their collective identity from their own viewpoint. However, scholars such as Melucci (e.g., 1995, p. 48) have also emphasized that external recognition is an essential component of collective identity formation. To account for this, I add an *external* perspective by analyzing how media actors attribute identity traits to social movements. These identity attributions are analyzed through frames, which, as previously outlined, involve the selective emphasis of specific characteristics of social movements. The application of these frames within news reporting is increasingly shaped by the technological tools available to journalists. In particular, the adoption of artificial intelligence (AI)-based tools in newsrooms plays a growing role in shaping how content is created, curated, and presented. Accordingly, this dissertation's external perspective not only considers frames as concrete manifestations of how identity is attributed by the media but also examines the conditions under which AI tools are integrated into journalistic practice and thus influence the production of news and the construction of collective identity.

To analyze the formation of social movements' collective identity online, I draw on digital trace data. The following section critically examines the analytical potential of such data for studying collective identity formation, especially when combined with computational methods and qualitative approaches that together enable both nuanced and large-scale analyses.

Digital Trace Data and Collective Identity

Digital trace data are defined as "records of activity (trace data) undertaken through an online information system (thus, digital)" (Howison et al., 2011, p. 769). This type of data offers many analytical advantages for studying online phenomena, such as collective identity formation. First and most notably, digital trace data provide reliable and longitudinal evidence of users' past actions, capturing observable traces of their opinions and interactions and thus enabling the analysis of dynamic phenomena and their evolution over time (Howison et al., 2011, pp. 769–771; see, also Franzoi et al., 2023). Unlike self-reported data, digital traces are not subject to the biases that arise from individuals' retrospective accounts. In identity research, this is a valuable complement to more traditional forms of data, such as surveys or interviews (often referred to as "summary-based data"), which rely on individuals' retrospective self-reports (Howison et al., 2011). This is particularly important in politically sensitive contexts such as protest participation, where analyzing actual online behavior can yield more nuanced and potentially less-biased understanding of how collective identities are formed, offering a perspective grounded in observable interactions rather than post-hoc reflections.

Second, and relatedly, digital trace data help mitigate biases typically found in traditional network research, which often relies on self-reported data prone to inaccuracies resulting from forgetting, misreporting, or selectively naming network ties (Marsden, 1990). In contrast, digital traces capture users' actual interactions, enabling the construction of networks based on observable behavior rather than subjective accounts.

Third, another key strength of digital trace data lies in their capacity to explain emergent "macro-phenomena based on micro-behavior" (Jungherr, 2015, p. 26). This characteristic directly addresses a central limitation of Melucci's approach, as the ability to analytically distinguish and connect micro- and macro-level processes is essential for understanding how collective identities of social movements emerge online.

Fourth, the abundance of diverse data types available within a single environment, such as images, text, and network data, makes digital trace data a uniquely rich data source for analyzing complex phenomena from multiple perspectives. This data diversity enables a nuanced and longitudinal examination of how social movements construct and sustain collective identity by allowing researchers to trace various cultural and social expressions as manifestations of collective identity over time. Such integrated and multifaceted analysis was difficult, if not impossible, in pre-digital contexts, where relevant data were often fragmented, scarce, or unavailable altogether. Thus, due to their diversity, scalability, observability, and ability to capture temporal dynamics, digital trace data provide a highly suitable database for analyzing how collective identity emerges and evolves within social movements online.

Despite their multiple advantages, however, digital trace data are not "produced by a designed research instrument," but are instead "found data"(Howison et al., 2011, p. 769). As such, they capture only certain facets of social phenomena that are mediated by digital platforms (Jungherr, 2015, p. 26). Researchers must, therefore, approach digital trace data with caution and remain critically aware of biases and limitations, particularly when analyzing phenomena such as collective identity. These limitations manifest in several ways. First, if the focus is on the identification and interpretation processes of individual actors, digital trace data offer only limited insights into "individual-level determinants of human behavior" (Stier et al., 2020, p. 504). In other words, we can observe the outcomes of behavior mediated through the platform, but not the underlying motivations or reasons behind it. This limitation is particularly relevant when analyzing collective identity formation, as digital trace data offer no insights into the self-reflective processes that influence participants' expressions of identity, an aspect central to Melucci's understanding of collective identity.

Second, and in a similar vein, while digital trace data allow us to observe user behavior, studies show that most social media users primarily consume content without actively contributing or leaving visible traces (Jones, 2023). As a result, we cannot draw conclusions about these passive users, even though they represent the largest share of social media participants.

Third, since digital traces merely provide "evidence of the 'raw material' of social relationships" (Howison et al., 2011, p. 770), they are often challenging to interpret, making it difficult at times to draw meaningful conclusions. Therefore, they must "be adapted for research purposes" (Howison et al., 2011, p. 769). This underscores the need for a solid theoretical approach to guide the interpretation of these observable user interactions, which occur naturally within an online environment rather than an artificial setting constructed by researchers.

I address the first limitation through a relational conceptualization of collective identity, which is grounded in the analysis of *observable* social interactions and does not require self-assessments of movement participants central to Melucci's definition. While I cannot make inferences about the users who do not leave visible traces (second limitation), I respond to the third limitation by providing a well-defined theoretical conceptualization that allows for a meaningful interpretation of digital trace data as social and cultural expressions left by social movements that reflect collective identity.

To complement this theoretical foundation, I also respond to the interpretative limitations of digital trace data (third limitation) with a robust methodological approach. Specifically, I combine computational methods, including applications of LLMs such as OpenAI's Chat Generative Pre-trained Transformer (ChatGPT), with qualitative approaches, such as qualitative content analysis, for analyzing collective identity formation. This mixed-methods design enables the analysis of the multimodal nature of digital trace data while also ensuring the reliability and validity of automated methods through complementary manual coding, benchmarking, and qualitative interpretation. Combined with the robust theoretical conceptualization, this approach provides both the scale and analytical depth required to analyze the formation of social movements' collective identity across social and cultural dimensions.

While this dissertation offers a novel and well-suited approach to analyzing collective identity online, the use of digital trace data to examine the social and cultural traces left by users in the context of collective identity formation is not without precedent. When analyzing these traces, however, many studies tend to focus exclusively on either the cultural or social dimension, rather than addressing their co-constitution. For example, some studies have explored cultural traces through qualitative discourse analysis of online forum content (Fayard & DeSanctis, 2008, 2010), conducted qualitative case studies evaluating visual and textual elements shared on social media (Khazraee & Novak, 2018), or combined social media text data with interview data (Kavada, 2009, 2015). Others have analyzed avatars and profile pictures (Gerbaudo, 2015) or applied qualitative content analysis to tweets from various social movements (Maghrabi & Salam, 2011; D. Ray & Tarafdar, 2017). To investigate social traces, researchers have commonly used network analysis based on user interactions, examining networks' structural features, connectedness, and communities (Barandiaran et al., 2020). Some studies have triangulated network data with surveys and participant observations, applying standard network metrics such as the evaluation of components, modularity, and k-core decomposition (Monterde et al., 2015). Although a few studies have considered both cultural and social traces (Bhattacharya et al., 2024; Coretti & Pica, 2015), they often lack a theoretical and empirical approach that accounts for the co-constitution of these dimensions, which is, however, an essential requirement for analyzing collective identity from a dual perspective.[5] By adopting a relational understanding of collective identity that accounts for the co-constitution of culture and structure and builds on the concepts of identity fields and identity work, this dissertation addresses these

[5] Article I provides a more detailed overview on shortcomings related to the previous evaluation of cultural and social traces in the context of collective identity online.

shortcomings and offers a multidimensional approach that expands and enriches previous research in this field.

Classification of Articles

The four research questions, derived from the concept of identity fields, guide the presentation of the subsequent results and collectively contribute to answering the overarching research question: *How do social movements online become distinct and persistent actors?*

- Articles I, II, and III address the protagonist identity field by examining the shared cultural and social expressions of social movements, with a particular focus on memes and narrative as symbolic resources, as well as network structures. These articles provide insights into how social movements create and sustain a shared sense of *we-ness*, thereby addressing the internal perspective of collective identity formation and answering RQ1.
- Articles I and IV adopt an internal perspective on collective identity formation and contribute to the antagonist identity field, offering insights into processes such as boundary work and social movements' demarcation toward other collectives. Drawing on insights obtained through the analysis of memes and narratives, these articles contribute to a deeper understanding of how social movements establish boundaries to distinguish themselves from others, thereby becoming distinct actors in the online space. They thus address RQ2.
- Articles V, VI, and VII adopt an external perspective on collective identity formation to examine the audience identity field by focusing on how media actors attribute identity traits to social movements. Within this identity field, I place particular emphasis on the role of emerging technologies, both in uncovering how media reporting assigns identity traits to social movements through framing (Article V) and in exploring how such AI-based tools should be designed to enable their effective adoption in journalistic production processes (Article VI). Together, the two articles contribute to this identity field from complementary perspectives, addressing RQ3a. Article VII extends this line of research by investigating how the media's framing of social movements affects user attention, addressing RQ3b.

In summary, Articles I through IV analyze social expressions, cultural expressions, or a combination of both as manifestations of identity work, thus capturing the internal perspective of collective identity formation. Articles V, VI, and VII focus on the media's identity attributions through framing and examine the conditions under which AI-based technologies are adopted in journalistic practice, thereby covering the external perspective that is essential for the recognition of collective identity. Each article contributes individually to one or more of the four derived research questions; together, their findings offer a comprehensive response to the overarching research question of this dissertation.

Table 1 is an overview of each article, including the title, contributing authors, publication venue, and the specific research question(s) addressed.

Article No.	Article	RQ
I	Henn-Latus, T., Tell, S., Polenz, P., Kern, T., Posegga, O. 2025. In Search of a "Style:" Capturing the Collective Identity of Social Movements Based on Digital Trace Data. Journal of the Association for Information Systems, 26(6), 1566-1629.	1 & 2
II	Henn, T., Posegga, O. 2023. What Do They Meme? Exploring the Role of Memes as Cultural Symbols of Online Communities. International Conference on Information Systems (ICIS). Hyderabad, India.	1
III	Henn, T. 2024. Follow the Memeing: Analyzing the Cultural Diffusion between Mainstream and Alt-Right Communities based on Shared Memes. European Conference on Information Systems (ECIS). Paphos, Cyprus.	1
IV	Henn, T., Posegga, O. 2024. Peeking behind the Memes: Evaluating the Boundary Work of Online Communities through Shared Memes. International Conference on Information Systems (ICIS). Bangkok, Thailand.	2
V	Henn-Latus, T., Posegga, O. Submitted. Classifying Protest Frames in Online News Reporting Using Multimodal Models with Zero-Shot and Few-Shot Learning. Journal of Computational Social Science.	3a
VI	Cordero, J.M., Henn, T., Holtel, F., Sánchez Gómez, J.Á., Arenas D., Šipka, A., Vollmer, S. 2025. Developing Effective and Value-Aligned AI Tools for Journalists: 12 Critical Questions to Reflect upon. Journalism Practice, 1-20.	3a
VII	Henn, T., Posegga, O. 2023. Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit. PLOS One, (18)8, e0288962.	3b

Table 1. Overview of Articles

As established above, I conceptualize collective identity as a relational concept grounded in two dimensions: the idea that collective identity is formed in relation to other actors online, as captured by the concept of identity fields; and the understanding that collective identity emerges through cultural and social expressions, as outlined by the concept of identity work (internal perspective), as well as through identity attributions (external perspective). Note that while the primary focus of this dissertation is on the internal perspective, I also integrate external recognition as an additional analytical dimension, given its widely acknowledged importance in the literature. Although this external perspective could likewise be differentiated into symbolic resources and social structures, this dissertation adopts a more high-level view as a first important step toward developing a multi-perspective analysis of how social movements form collective identities online.

Table 2, an overview of the specific expressions of collective identity examined in each article and the corresponding identity field it addresses, illustrates how the articles relate to both dimensions and thereby offer distinct insights into social movements' collective identity formation online.

Expressions of Collective Identity			Identity Fields		
			Protagonist	Antagonist	Audience
Internal Perspective	Identity Work	Symbolic Resources	• Narratives (Article I) • Memes (Articles II, III)	• Narratives (Article I) • Memes (Article IV)	• *Not applicable*
		Social Structures	• Networks (Articles I, II)		• *Not applicable*
External Perspective	Identity Attribution		• *Not applicable*	• *Not applicable*	• Frames (Articles V, VII) • Conditions for AI Adoption (Article VI)

Table 2. Classification of Articles According to Expressions of Collective Identity and Identity Fields

The following section presents the answers to each individual research question by summarizing the key findings from the respective articles.

Contributions

Research Question 1

This section focuses on the protagonist identity field and includes the contributions of Articles I, II, and III, which collectively address RQ1: *How do social movements online create and sustain a shared sense of we-ness?* The articles offer an internal perspective on collective identity formation, examining how narratives, memes, and networks serve as expressions of collective identity.

In Article I, we introduced a novel conceptualization and operationalization of collective identity for social movements online that accounts for the peculiarities of digital trace data. Grounded in Harrison White's (1992, 2008) relational approach and his concept of *style*, our approach emphasizes the co-constitution of structure and culture in the formation of collective identity. We applied this conceptualization empirically to a case study of the social movement Querdenken (*lateral thinker*) on the platform X.

When examining the narratives Querdenken used to define itself, we observed that the movement initially remained notably ambiguous regarding which actor groups belonged to it, shifting only gradually toward a clearer self-definition of a protest movement. We interpreted this ambiguity through the concept of *tolerant identities* (della Porta, 2005), which offer the advantage of being ideologically flexible and inclusive. Such identities enable broad identification without explicitly excluding individuals, thereby supporting wider mobilization. At the same time, despite this vagueness, Querdenken effectively attributed clearly delineated positive character traits to itself, in line with Alexander's (2006, pp. 57–59) set of binary codes, which helped to establish a shared sense of *we-ness* within the movement.

In terms of thematic focus, Querdenken succeeded in developing unifying narratives that expressed a shared sense of *we-ness*. We found that certain themes, such as criticism of political authorities, were consistently prominent; over time, the movement increasingly shifted from neutral to a more politicized content. In parallel, Querdenken became more aware of its collective agency, as reflected in the growing emphasis on concrete actions against the authorities it targeted in its narratives. This increasing self-awareness as a collective actor capable of action represented a further step in the movement's evolving understanding of its own identity and "who we are."

Our analysis of social structures based on network patterns revealed evidence of emergent leadership within the movement, consistent with prior research on online leadership (e.g., Faraj et al., 2015; Johnson et al., 2015). While the specific individuals occupying leadership roles changed over time, the overall structure remained stable, following a persistent power-law distribution of users considered influential by the movement. This suggests that leadership roles are structurally consistent over time but not dependent on specific actors. When examined alongside the persistent cultural patterns discussed earlier, these findings indicate that both social and cultural patterns can become autonomous from individual users. Thus, sustaining a shared sense of *we-ness* does not require the continuity of specific actors, but rather the recurrence of recognizable patterns at the social and cultural levels.

Finally, we drew on the distinction between *weak cultures* and *strong cultures* (Schultz & Breiger, 2010) to interpret the evolution of Querdenken's social and cultural patterns over time. While both dimensions were initially characterized by ambiguity, we observed a gradual consolidation: the movement became increasingly aware of its unifying characteristics, collective agency, and important topics, while also exhibiting recurring structural patterns, such as those related to leadership roles. Interpreted through the concepts of weak and strong cultures, these developments suggest that Querdenken transitioned from a weak to a strong culture. This shift reinforces the movement's shared sense of *we-ness*, as an increasing cultural and structural coherence reflects the emergence of a more distinct and unified collective actor.

Articles II and III examine the role of memes as symbolic resources in expressing the collective identities of online communities in terms of *we-ness*. While both articles focus on the platform Reddit, they analyze the meme-sharing of different types of communities. Article II examines interest-based online communities by comparing how meme sharing differs between clearly defined subreddit communities—with specific rules and moderation policies—and communities based on user interaction patterns that span across subreddits and platform-defined boundaries. Article III focuses on subreddits associated with the alt-right movement,[6] investigating how their meme-sharing differs from or resembles that of mainstream meme subreddits, serving as an example of interest-based communities. To systematically analyze meme sharing as an expression of collective identity, both articles draw on the newly introduced concept of *meme languages*, which capture the distinctive patterns of shared memes specific to communities and build on Stuart Hall's (1997) theory of the two systems of representation.

When assessing the meme languages of interaction-based and subreddit communities in Article II, we found that the two community types differ significantly in their meme languages, with only a small number of memes appearing across multiple meme languages. This indicates that meme languages reflect a recognizable cultural signature and serve as a distinct expression of collective identity for individual communities. We were surprised to find, however, that subreddit communities—despite their stricter moderation policies—did not exhibit more homogenous meme languages than interaction-based communities. This finding challenges the assumption that moderation policies foster greater internal cultural coherence. It implies instead that a shared sense of *we-ness* based on memes can be expressed through heterogeneous cultural symbols, remaining unaffected by community moderation.

Article III expands on these findings and demonstrates that a community's purpose, self-perception, and ideological boundaries influence meme-sharing practices and its expression of a shared sense of *we-ness*. Analyzing the composition of meme languages across mainstream meme and alt-right related subreddits revealed that communities with a

[6] Note that I classify the alt-right movement as a social movement in this dissertation, as it is defined as "a right-wing, primarily online political movement or grouping based in the U.S. whose members reject mainstream conservative politics and espouse extremist beliefs and policies typically centered in the idea if white nationalism" (Merriam-Webster, 2024). This definitions aligns with Diani's (1992) definition of social movements introduced earlier.

clearly defined purpose and strong ideological boundaries shared more visually similar memes, resulting in a more homogenous and coherent expression of collective identity. For instance, the subreddit r/wholesomememes featured a high degree of visually similar memes, which can be attributed to its explicit community purpose and self-perception centered around sharing wholesome animal content (Chabot & Chen, 2020). Similarly, r/The_Donald, formerly the most prominent alt-right subreddit before its ban (Gaudette et al., 2021; Trujillo & Cresci, 2022), also exhibited a high degree of visually similar memes. This visual coherence can be attributed to the community's clearly defined boundaries and a shared consensus on its collective identity, depicting a common understanding of "who we are."

This finding is further reinforced when comparing the meme languages between mainstream and alt-right communities. The analysis revealed that r/The_Donald and the mainstream meme subreddits exhibit greater similarity in their meme languages than any other combination of subreddits examined. To explain this result, I returned to the role of community purpose, self-perception, and ideological boundaries. Mainstream meme subreddits are primarily oriented toward the creation and dissemination of novel, viral content that spreads widely across the platform and beyond. In contrast, ideological communities such as r/The_Donald are defined by strong internal boundaries and a clear sense of group identity. When a meme enters such a community, it is quickly evaluated by users based on how well it aligns with the community's values. Memes that resonate with the collective identity are frequently remixed and reshared, while others are ignored and fail to gain traction.

Overall, these findings underscore that memes serve as symbolic expressions of a community's identity work, reinforcing a shared sense of *we-ness* within the community. However, meme-sharing practices are shaped by a community's boundaries, purpose, and self-perception. Communities linked to social movements, such as those associated with the alt-right, exhibit high levels of visual similarity in their meme content due to their strong ideological boundaries. Interest-based communities can also express their identity through visually similar memes, especially when their purpose is directed toward specific topics. However, in the absence of such clarity, their meme languages tend to be visually diverse, resulting in a more heterogeneous expression of *we-ness*. Moreover, these findings also demonstrate that visually similar memes can cross the cultural boundaries between different types of communities, carrying different meanings depending on the community's context and thus aligning with the earlier definition of culture introduced in this dissertation.

Research Question 2

This section focuses on the antagonist identity field and includes the contributions of Articles I and IV, which collectively address RQ2: *How do social movements online draw* *boundaries to differentiate themselves from other groups?* The articles offer an internal perspective on collective identity formation, examining how narratives and memes serve as expressions of collective identity.

In Article I, we also examined how Querdenken used narratives to distinguish itself from other groups online. Our analysis revealed that the movement consistently identified political authorities as its primary adversaries, a common pattern in the collective identity formation of social movements (della Porta & Diani, 2006, p. 112). In addition to political authorities, scientific experts were also frequently targeted, reflecting the specific context of the movement during the COVID-19 pandemic. In line with Alexander's (2006, pp. 57–59) set of binary codes, these opponents were associated with negative traits such as incompetence, abuse of power, and conspiratorial behavior. These findings stand in sharp contrast to how Querdenken portrayed itself: while being somewhat ambiguous about who exactly belonged, the movement consistently attributed morally desirable traits to itself. In doing so, Querdenken positioned itself as morally superior to its opponents, clearly drawing boundaries and distinguishing itself from its adversaries.

Beyond a clear consensus on the adversaries and the morally negative characteristics ascribed to them, references to the demarcation towards opponents were also the most frequently occurring themes in the dataset and exhibited the most significant increase over time. These findings support existing research suggesting that boundary-drawing against perceived opponents is an early and essential step in collective identity formation (della Porta & Diani, 2006, pp. 93–94; Tilly, 2003, p. 608). They also suggest that it is often easier for social movements to define "who they are not" than to reach consensus on "who they are." This process of boundary work plays a central role in shaping a movement's collective identity and offers insights into how they demarcate themselves from other groups online.

In Article IV, we investigate how social movements use memes to draw boundaries and distinguish themselves from other actors online. To analyze this boundary work, the article draws on Tajfel and Turner's (2004) theory of social identity, which explains how individuals form a sense of identity through membership in in-groups and differentiation from out-groups. This theory was used to develop a coding scheme that categorizes different types of boundaries featured in memes. To enrich the analysis, this coding scheme was further complemented by themes and roles. We applied the coding scheme to memes shared in the subreddit r/The_Donald on the platform Reddit before and after the 2016 U.S. presidential election, allowing us to assess whether major offline events corresponded with shifts in the boundaries articulated through meme content.

Our results showed that the boundaries and themes expressed in memes followed recurring patterns. Specifically, the three boundary types derived from social identity theory were consistently associated with two recurring themes each, resulting in six key themes that more concretely characterized the boundary work observed in memes shared in r/The_Donald. These themes were, for instance, caring—where memes express solidarity and support among in-group members—and victimization, where the in-group presents itself as a victim of the out-group. We also identified a set of recurring roles frequently featured in the memes— heroes, persecutors, victims, and fools—that further enriched our framework of boundaries and themes. Taken together, the recurring boundaries, themes, and roles offered deeper insights into how users of r/The_Donald engage in boundary work. These findings underscored the complex interplay of these three dimensions in expressing collective identity and in demarcating the community from other groups in the online space.

Moreover, we found evidence that the 2016 U.S. presidential election correlated with notable shifts in the types of boundaries, themes, and roles featured in memes. The findings aligned with Sherif's (1966) realistic group conflict theory, which suggests a heightened hostility toward the out-group in situations involving conflicting goals or competition over resources. Our analysis revealed that prior to the election, memes more frequently featured boundaries characterized by hostility, along with a dominant antagonistic dynamic between heroes and persecutors. Following the election, however, we observed a shift toward boundaries that emphasized direct comparisons between in-group and out-group members, along with an increased portrayal of the in-group as victims of the out-group. These patterns underscore the reciprocal relationship between offline events and online identity work, suggesting that changes in the boundary work of social movements online are shaped by developments in the offline world.

Research Question 3a

This section focuses on the audience identity field and includes the contributions of Articles V and VI, which collectively address RQ3a: *How can emerging technologies support both the analysis and production of news content that attributes identity traits to social movements online?* The articles offer an external perspective on collective identity formation by examining how media actors use frames to attribute identity traits to social movements and how conditions for AI adoption in news production influence how these attributions are created.

In Article V, we examined the performance of two different multimodal models in classifying news frames in online reporting on protests. The article focused particularly on frames associated with the protest paradigm (Chan & Lee, 1984; McLeod & Hertog, 1999) and also classified news posts based on whether the protest was portrayed as violent and/or peaceful. The study used Instagram posts from major U.S. news outlets, with both the image and caption serving as units of analysis. After manually labeling the data, the models were applied in zero-shot and few-shot learning settings to detect these labels for images and image and captions, thereby assessing how well these technologies can identify complex phenomena such as frames.

When comparing the performance of our two models, Contrastive Language-Image Pre-Training (CLIP) and GPT4.1, we found that both achieved promising classification results. However, a closer evaluation revealed important nuances. Notably, we observed considerable variation in classification performance across individual labels, reflecting the inherent complexity of certain frames, which are known to also challenge human coders. Moreover, GPT4.1 outperformed CLIP architectures across nearly all performance metrics. This is unsurprising given that GPT4.1 is not only a larger model but also trained on a significantly more diverse and extensive dataset. However, its use comes with trade-offs: the model architecture is less transparent, offers limited fine-tuning options, and applies token-based fees for classification. In contrast, CLIP models accessed through OpenCLIP (Ilharco et al., 2021) are open-source, more transparent, and free of token-based costs, making them a compelling alternative despite their comparatively lower performance. In addition, CLIP models showed notable improvements when moving from zero-shot to few-shot learning, highlighting another of the architecture's advantages. Finally, our results demonstrated that combining textual and visual features improved classification performance compared to using image data alone, emphasizing the importance of multimodal approaches in understanding media framing online.

Overall, the results showed that models capable of processing multimodal input perform considerably well in classifying frames associated with the protest paradigm. A key advantage of these models over earlier approaches is their ability to achieve strong performance without relying on large, curated training datasets. In many cases, just a few labeled examples are sufficient to improve results, underscoring their adaptability to new, previously unseen contexts. As such, these technologies are powerful tools for uncovering the identity traits that media actors attribute to social movements, enabling the large-scale analysis of media framing.

Beyond assessing the technological performance of the models, Article V also provided important insights into the distribution of frames, offering a deeper understanding of how media actors attribute identity to social movements. The distribution of frames within our manually labeled ground truth revealed that the debate, spectacle, and victimization frames appeared most frequently in both image and caption data. The prominence of the debate and spectacle frames aligns with previous research (e.g., Harlow et al., 2017; Harlow & Brown, 2021) and likely reflects the competitive nature of social media, where emotionally charged and visually striking content attracts greater attention. The prominence of the victimization frame can be attributed to the inclusion of protest events organized by the BLM movement, which often feature police violence and were, therefore, coded under this frame. These findings indicate that both the logic of social media platforms and the nature of the protest events analyzed influence which frames are most prominently featured in media coverage.

Article VI focused on how AI-based technological tools should be designed to ensure their effective adoption by the journalistic community. Drawing on existing literature, we identified two challenges that currently hinder widespread adoption of these technologies in newsrooms: effectiveness and value alignment. These challenges were examined in practice through a collaboration with the Organized Crime and Corruption Reporting Project (OCCRP), one of the world's leading investigative journalism networks. Together, we developed an AI-based prototype capable of classifying document types within larger datasets and prioritizing pages based on their relevance to journalists. Through the design and implementation of the prototype, alongside our exploration of the identified challenges, we derived twelve guiding questions that we consider critical for the future design of AI tools in journalism. They address issues related to the two core challenges identified in the literature.

Our findings offered several key insights regarding the effectiveness challenge. First, issues related to generalizability can be mitigated by designing models that are easily adaptable to new classification contexts and domains, for example, through the use of dynamic and adjustable thresholds (Shu et al., 2017). Second, cost-related concerns can be addressed by leveraging publicly available datasets for model pre-training, alongside methodological techniques such as image augmentation to enhance dataset diversity. Third, issues related to user-friendliness can be improved by integrating new functionalities into the existing technological infrastructures of newsrooms and organizations, ensuring a seamless integration into the already existing workflow of journalists. Finally, accuracy concerns can be tackled through the implementation of technical safeguards aimed at reducing erroneous outputs.

Our findings also provided helpful insights into the value alignment challenge. Our close collaboration with OCCRP in developing the prototype allowed for an active exchange of ideas with OCCRP staff and journalists, thereby ensuring that journalistic values are reflected within the prototype. This collaboration highlights the importance of involving journalists and media organizations directly in the development of AI-based tools. In addition, by partnering with a non-profit investigative journalism organization rather than a commercial tech company, we addressed concerns related to independence and reduced reliance on major technology providers in the development of AI tools for journalism. Further, our prototype is open source and available on GitHub,[7] which addresses concerns around transparency that are often raised with proprietary systems. Finally, with respect to data privacy, our tool can be run locally, ensuring that sensitive data is not shared with third parties.

Article VI highlights the crucial role non-profit organizations such as OCCRP can play in creating AI-based tools that are both effective and align with journalistic values. Our

[7] https://github.com/DSSGxDFKI/occrp-document-classifier?tab=readme-ov-file

collaboration with OCCRP demonstrates the potential of partnership between researchers and practitioners to develop new tools and technologies. These insights contribute to a better understanding of how AI-based tools can be developed to support their effective adoption into journalistic workflows, ultimately shaping the creation of news content.

Research Question 3b

This section focuses on the audience identity field and includes the contributions of Article VII, which addresses RQ3b: *How does the media's attribution of identity traits to social movements affect user attention?* The article offers an external perspective on collective identity formation, examining how media actors use frames to attribute identity traits to social movements.

Article VII investigated how the media framing of the BLM movement online affects user attention. Grounded in the concept of the negativity bias (e.g., Rozin & Royzman, 2001; Soroka et al., 2019) and the picture superiority effect (e.g., Stenberg, 2006), the article focused on how images featuring violence and negatively framed headlines impact user attention on Reddit. The article also examined how other factors related to the news provider, such as the level of factual reporting, as well as community rules, such as the flagging of submissions by a moderator, affect user attention.

Our results showed that while violent visual framing of the BLM movement did not significantly influence user attention, negatively framed headlines had a significant impact. This finding supported our assumptions based on the negativity bias, indicating that the negative framing in headlines can effectively attract user attention. In contrast, the picture superiority effect did not hold in this context, as headlines had a stronger and more significant influence on user attention than images. Beyond framing effects, our analysis also underscored the importance of attributes related to community rules and the news outlet in affecting user attention. For instance, Reddit-specific variables, such as the subreddit, the number of cross-postings, and whether the post was flagged by the moderator, significantly influenced user attention.

These findings illustrate the complexity of social media environments, where multiple factors affect user attention. While the violent visual depiction of the BLM movement did not affect user attention, negatively framed headlines had a significant impact. This suggests that, in this context, headlines and their framing play a more influential role than images in attracting user attention to news about social movements.

The following section presents executive summaries of each article to offer a more detailed overview. For additional information, please see the full articles included at the end of this dissertation.

Executive Summaries

Article I

Article I is the cornerstone of this dissertation, addressing the ongoing scholarly debate—introduced above—over whether and how social movements can construct collective identity within online environments. In response to this controversy, the article proposes a novel conceptualization and operationalization of collective identity for social movements online by drawing on Harrison White's (1992, 2008) concept of style. This concept offers a relational perspective that accounts for the co-constitution of culture and structure in collective identity formation. To translate the concept of style to online environments, we derive several subdimensions of cultural and social patterns. For cultural patterns, we analyze the emergence of collective agency, sameness, and distinctiveness, drawing on Snow's (2001) definition of collective identity. For the social patterns, we analyze social cohesion and social positions as key subdimensions, based on the work of Diani (2002) and Mustafa Emirbayer and Jeff Goodwin (1994). These dimensions are translated into measurable empirical indicators and applied to a case study of the German protest movement Querdenken (*lateral thinker*) on the platform X.

To analyze the style of Querdenken, we apply a socio-semantic network analysis that integrates quantitative network metrics with qualitative content analysis. The dataset for this analysis includes tweets from Querdenken-affiliated users posted throughout 2020 that contain at least one hashtag and were retweeted at least once. To capture temporal dynamics, these tweets are segmented into three time periods aligned with key offline events. The analysis unfolds in three stages. First, bipartite user–hashtag networks are constructed for each period and qualitatively examined to identify key topics and assess their level of politicization. Second, bipartite user–retweet networks are transformed into statistically validated monopartite Querdenken user networks using entropy-based null models; these networks are used to assess social cohesion through network metrics and to evaluate user positions via a leadership score. Third, to investigate the cultural expressions of collective agency, sameness, and distinctiveness, we construct monopartite hashtag co-occurrence networks and conduct a qualitative content analysis of a subsample of tweets. This analysis combines inductive coding with a deductive approach using Alexander's (2006, pp. 57–59) set of binary codes. The coding was carried out by multiple coders and refined through several rounds of open coding.

Our results show that Querdenken can indeed develop a recognizable style through recurring social and cultural patterns. Based on these findings, we formulate nine propositions that explain how specific user behavior contributes to the manifestation of the movement's collective identity online. These insights offer a process-oriented interpretation of collective identity formation, grounded in the co-constitution of social structures and culture. The resulting propositions provide a foundation for future research to further investigate collective identity formation processes online.

The article is published in the Journal of the Association for Information Systems.

Article I. Henn-Latus, T., Tell, S., Polenz, P., Kern, T., Posegga, O. 2025. In Search of a "Style:" Capturing the Collective Identity of Social Movements Based on Digital Trace Data. Journal of the Association for Information Systems, 26(6), 1566-1629.

Articles II and III

Research has repeatedly demonstrated that memes serve as powerful cultural symbols, expressing the identity and ideology of online communities (e.g., DeCook, 2018; Gal et al., 2016; Nissenbaum & Shifman, 2017). However, existing research on meme sharing within online communities reveals two somewhat conflicting perspectives. On the one hand, memes are understood as symbolic and identity-shaping symbols tied to specific online communities that express a community's collective identity (e.g., DeCook, 2018; Nissenbaum & Shifman, 2017). On the other hand, research highlights the inherently fluid and evolving nature of memes, which often diffuse across community boundaries (e.g., Bauckhage, 2011; Morina & Bernstein, 2022; Zannettou et al., 2018). Articles II and III address this apparent contradiction by examining the meme-sharing of different communities on the platform Reddit. Given the articles' shared conceptual and methodological foundation, they are presented together. Article II examines the meme-sharing of interest-based communities defined by subreddit affiliation and interaction patterns. Article III focuses on communities associated with the alt-right movement, analyzing how they use memes to express their collective identity and how their meme-sharing compares to that of mainstream meme subreddits.

To systematically assess how these communities differ or resemble each other in their use of memes to express their collective identity, both articles apply the newly introduced concept of meme languages. This concept captures the distinctive patterns of shared memes specific to communities and is grounded in Stuart Hall's (1997) theory of the two systems of representation. Article III further draws on the concept of *memetics* (e.g., Lugea, 2019) to better understand the cultural diffusion between alt-right and mainstream meme communities. To empirically detect meme languages, both studies apply a visual feature-matching algorithm developed by Courtois and Frissen (2022), which calculates visual similarity based on the spatial distances of visual features. This allows for the detection of memes and their visual variants by identifying memes that are based on the same initial template. Through these visual similarities, networks are created in which visually similar memes are connected within individual components. The meme language of each community is defined as the aggregation of these components. To compare meme languages across communities and assess which memes appear in multiple communities, we manually code each component by assigning a label corresponding to the meme it represents.

To create these meme languages, we use data from Reddit, downloading submission and comment dumps via the Pushshift API (Baumgartner et al., 2020). After identifying the relevant subreddits for analysis, we download all associated images and detect memes within them. In Article II, this detection is automated using a self-trained convolutional neural network (CNN) for meme classification; Article III, in contrast, relies on manual annotation to determine whether an image follows the meme definition. Also, we create interaction-based communities in Article II by connecting users with a tie if they authored a post or comment in the dataset and interacted with another user's submission or comment. We then apply the widely used Louvain algorithm (Blondel et al., 2008) for community detection to identify communities based on these interaction patterns.

The findings from Article II show that subreddits develop unique meme languages that characterize them as distinct actors. Moderation policies, though, appear to have little effect on the internal homogeneity of these languages. Article III finds that alt-right and mainstream meme subreddits share notable similarities in their meme languages, suggesting that memes can diffuse across cultural boundaries, change their meaning according to the community context, and serve as expressions for the collective identity of multiple communities. Article III also emphasizes the importance of understanding a community's purpose and ideological boundaries to explain how and why certain memes survive a competitive selection process and become popular within communities. Thus, by examining how different types of communities with varying levels of ideological boundaries, moderation policies, and purposes differ from or resemble one another in their meme-sharing practices, both articles make an important contribution to understanding how memes function as cultural symbols that are expressions of communities' collective identity.

Article II is a short paper published in the Proceedings of the International Conference on Information Systems (ICIS), while Article III is a full paper published in the Proceedings of the European Conference on Information Systems (ECIS).

Article II. Henn, T., Posegga, O. 2023. What Do They Meme? Exploring the Role of Memes as Cultural Symbols of Online Communities. International Conference on Information Systems (ICIS). Hyderabad, India.

Article III. Henn, T. 2024. Follow the Memeing: Analyzing the Cultural Diffusion between Mainstream and Alt-Right Communities based on Shared Memes. European Conference on Information Systems (ECIS). Paphos, Cyprus.

Article IV

While the cultural and identity-shaping value of memes has been thoroughly demonstrated in the literature (e.g., DeCook, 2018; Gal et al., 2016; Nissenbaum & Shifman, 2017), it remains unclear how memes are used to create and sustain boundaries between online communities, particularly those related to social movements. Article IV, therefore, explores how social movements use memes to construct boundaries and differentiate themselves from other actors online. Drawing these boundaries represents a form of boundary work that plays a key role in collective identity formation. The analysis focuses on memes shared within the subreddit r/The_Donald, which was widely regarded as the most important community for the alt-right movement before its ban (Gaudette et al., 2021; Trujillo & Cresci, 2022). To explore how offline events might shape this boundary work featured in memes, the article compares memes shared before and after the 2016 U.S. presidential election.

To conceptualize the boundaries featured in memes, the article draws on Tajfel and Turner's (2004) theory of social identity, which explains how individuals form a sense of identity through membership in in-groups and differentiation from out-groups. This theory is used to develop a coding scheme that categorizes different types of boundaries expressed in memes. To assess whether offline events correlate with shifts in the boundaries featured in memes, we draw on realistic group conflict theory (Sherif, 1966), which posits a heightened hostility toward the out-group in situations involving conflicting goals or competition over resources. The 2016 U.S. presidential election, characterized by an intense competition over resources such as power, is treated as a particularly relevant event for examining potential changes in the boundary work featured in memes.

We collect meme data from r/The_Donald using the Pushshift API (Baumgartner et al., 2020) by downloading the submission dump of November 2016. Wherever possible, we retrieved the image attached to each submission and manually review all images to identify those that match our definition of memes. To identify the most popular memes within this sample before and after the election, we use the Reddit score, a metric that reflects the difference between upvotes and downvotes, indicating how well each submission resonates with the community.

The boundaries featured in memes are operationalized through three boundary types that we derive from social identity theory: in-group favoritism, where the in-group is portrayed positively without direct reference to the out-group; in-group enhancement, where the in-group is depicted as superior in explicit comparison to an out-group; and out-group degradation, where the out-group is portrayed negatively without reference to the in-group. Beyond identifying the boundary type, we also inductively code a meme's featured theme, referring to the specific manifestation of the detected boundary. We also deductively code the depicted roles of actors within memes using the typology introduced by De Saint

Laurent et al. (2021), which distinguishes among heroes, victims, persecutors, and fools. All coding categories are developed collaboratively through multiple rounds of open coding between two researchers. After coding all memes along these three dimensions—boundary type, theme, and roles—the distribution of these dimensions is analyzed before and after the 2016 election. This analysis not only helps deepen the understanding of how boundary work is constructed within memes but also assess whether these patterns shift in response to offline events.

Overall, the findings demonstrate that memes serve as powerful boundary markers, offering valuable insights into the boundary work of one of the formerly most important communities associated with the alt-right movement. The results highlight that a nuanced conceptualization of boundary work allows for a deeper understanding of how groups online differentiate themselves from others. Moreover, our analysis shows that important offline events can influence online boundary work, as evidenced by notable shifts in the types of boundaries, themes, and roles featured in memes before and after the election. This underscores the reciprocal relationship between the online and offline world.

The article is a full paper published in the Proceedings of the International Conference on Information Systems (ICIS).

Article IV. Henn, T., Posegga, O. 2024. Peeking behind the Memes: Evaluating the Boundary Work of Online Communities through Shared Memes. International Conference on Information Systems (ICIS). Bangkok, Thailand.

Article V

Given the influential role of media framing in shaping public perception and discourse around protest movements (e.g., Brown & Mourão, 2021; McCurdy, 2012; McLeod & Detenber, 1999; McLeod & Hertog, 1992), Article V examines the extent to which recent multimodal models can accurately classify these frames automatically. In doing so, the article helps to gain deeper insights into how media actors attribute identity traits to protest movements on social media platforms. Specifically, the article compares the performances of OpenAI's CLIP and GPT4.1 models, distinguishing between zero-shot and few-shot learning. The models are evaluated based on their ability to classify frames using images and images and captions as input data.

The article focuses on frames associated with the protest paradigm (Chan & Lee, 1984; McLeod & Hertog, 1999), which emphasizes the deviant and disruptive behavior of protest movements rather than their grievances, historical context, or demands. In this context, four frames are commonly used in media coverage: riot, confrontation, spectacle, and debate (e.g., Brown et al., 2018; Harlow & Brown, 2023). The first three—riot, confrontation, and spectacle—are typically considered delegitimizing because they portray protestors as disruptive or violent, whereas the debate frame is considered legitimizing as it presents protestors' grievances and demands. While protest-led violence plays a central role in these frames, scholars have pointed out that violence directed *against* protestors, especially when committed by state authorities, has received insufficient attention in prior research (Harlow & Brown, 2023). To address this shortcoming, the article introduces a fifth frame: the victimization frame, which focuses on physical violence committed against protestors by state actors. In addition to analyzing these frames as expressions of identity attribution by the media, the article also classifies news posts based on whether the protest is portrayed as violent and/or peaceful.

To assess the performance of different models in classifying news frames in the context of protest, we collect data from sixteen major U.S. news outlets on the platform Instagram using the platform CrowdTangle. We filter the dataset to include only posts that contain images and appear to depict protests. However, due to the ambiguity of some keywords used for filtering, the dataset also includes content unrelated to protests. To establish a reliable ground truth, we conduct a two-step qualitative coding process: first identifying whether each post depicts a protest by a movement or its supporters, and then coding these posts according to the five pre-defined frames and assessing whether the protest was portrayed as peaceful and/or violent. Labeling involved multiple rounds of open coding and team discussions to reach consensus. In line with related research (e.g., Harlow et al., 2017; Harlow & Brown, 2021), frames are not mutually exclusive.

We implement different CLIP architectures using OpenCLIP models (Ilharco et al., 2021) to automatically classify protest-related frames and identify whether protests were portrayed as violent or peaceful. These models are first evaluated in a zero-shot setting, where we used grid search to fine-tune parameters and use different prompts that vary in their level of detail. Classifications are conducted using two input types: image-only and a combination of image and caption. To benchmark the performance of these CLIP models, we evaluate the performance of GPT4.1, also varying between different prompts.

For the few-shot setting, we employ Tip-Adapter-F, a fine-tuned version of the Training-free Adaptation Method for CLIP (Tip-Adapter) introduced by Zhang et al. (2022). We train this adapter with a few samples per class and attach it to the frozen CLIP backbone. Hyperparameters of Tip-Adapter-F are again optimized via grid search. For the few-shot GPT-4.1 setup, we embed a small number of labeled examples directly into the prompts. As in the zero-shot setting, both image-only and image and caption inputs are tested for all model variations.

Our findings demonstrate that automated methods yield promising results in classifying complex frames; however, performance varies depending on frame type, model architecture, and learning approach. Overall, this article makes a significant contribution by demonstrating how AI-based methodological advances can be used to uncover how media reporting attributes identity traits to social movements at scale.

The article has been submitted to the Journal of Computational Social Science.

Article V. Henn-Latus, T., Posegga, O. Submitted. Classifying Protest Frames in Online News Reporting Using Multimodal Models with Zero-Shot and Few-Shot Learning. Journal of Computational Social Science.

Article VI

AI-based technologies offer unprecedented opportunities for journalistic practice and have fundamentally reshaped many aspects of journalistic work (e.g., de Haan et al., 2022; Helberger et al., 2022). Despite this potential, however, research shows that some journalists and media organizations remain hesitant to adopt advanced AI-based tools (Cools & Diakopoulos, 2024; de Haan et al., 2022; Nishal & Diakopoulos, 2024; Stray, 2019; Van Dalen, 2024). Drawing on existing literature, Article VI identifies two challenges—effectiveness and value alignment—that hinder broader adoption of these technologies in newsrooms and investigates how AI-based technologies can be designed to better meet the needs of the journalist community.

In order to gain a more nuanced and practical understanding of these challenges to AI-based tool adoption in journalism, we cooperate with OCCRP and develop an AI-based prototype. This tool is capable of classifying document types within larger datasets and prioritizing pages based on their relevance to journalists. The prototype is developed using a model pipeline that integrates multiple CNN architectures to classify documents both by type and their information value through image-based analysis. The models are trained using labeled data provided by OCCRP, covering nine document types relevant to journalists. To ensure the validity of these labels, we develop a detailed codebook through several rounds of open coding, refining definitions of each document type. We proceed similarly for information-rich and information-poor pages. Model performance is improved by pre-training the CNN architectures on publicly available datasets and augmenting the dataset with image augmentation techniques. To enhance the models' ability to handle unseen data and perform more reliably across new settings, we implement a dynamic classification threshold based on the approach by Shu et al. (2017). The final prototype is openly accessible on GitHub.

Drawing on the design and implementation process of the prototype, we reflect on the identified challenges that potentially hinder AI-based tool adoption in journalism. From these insights, we derive twelve questions that can guide the design process of such tools and help address these challenges. The article thus advances our understanding of how AI-based technologies should be designed and implemented to ensure their effective adoption into journalistic practice. It further emphasizes the role non-profit organizations like OCCRP, in collaboration with researchers, can play in advancing this process.

The article is published in the journal Journalism Practice.

Article VI. Cordero, J.M., Henn, T., Holtel, F., Sánchez Gómez, J.Á., Arenas D., Šipka, A., Vollmer, S. 2025. Developing Effective and Value-Aligned AI Tools for Journalists: 12 Critical Questions to Reflect upon. Journalism Practice, 1-20.

Article VII

Research shows that media reporting plays a significant role in influencing how movements are perceived by the public and whether they are supported (e.g., Brown & Mourão, 2021; Masullo et al., 2024; McLeod & Hertog, 1992). As a result, it is becoming increasingly important to understand which types of news attract user attention and what factors drive this engagement. Article VII addresses these questions by analyzing how images featuring violence and negatively framed headlines in the news coverage of the BLM movements on Reddit affect user attention. The article also investigates the influence of other factors such as the political leaning of news outlets, their level of factual reporting, and platform-specific factors such as cross-postings and flair tags.

This article draws on the concept of negativity bias (e.g., Rozin & Royzman, 2001; Soroka et al., 2019), which suggests that people are more likely to pay attention to negative news. Building on this theoretical foundation, the study focuses in particular on the role of violence depicted in news reporting and the effect of negatively framed headlines on user attention. Special emphasis is placed on images featuring violence due to the so-called picture superiority effect (e.g., Stenberg, 2006), which posits that images have a stronger impact on human cognition than text, a finding that has become increasingly relevant as visual content plays a growing role in social media communication (e.g., Marwick, 2015). In addition to analyzing the effects of negative headlines and images featuring violence on user attention, the article considers other factors identified in previous research that may influence user engagement. These factors are grouped into three categories: attributes related to the news item; characteristics of the news provider; and platform-specific community rules.

For our analysis, we collect data from Reddit using the Pushshift API (Baumgartner et al., 2020), downloading submission dumps from May 2020 to May 2021. This timeframe captures the year following the murder of George Floyd, which reignited the BLM movement and sparked global protests. Within this dataset, we identify important news subreddits and download submissions related to the BLM movement. To supplement missing information about news outlets not available on Reddit but necessary for our analysis, we use the website mediabiasfactcheck.com. To detect violence in images, we train and test several

CNN architectures, selecting the best-performing model for our analysis. For sentiment analysis of headlines, we fine-tune a Bidirectional Encoder Representations from Transformers (BERT) model. To analyze the statistical relationship between all independent variables and our dependent variable, user attention, we conduct multiple negative binomial regressions.

Our findings indicate that while images featuring violence do not significantly affect user attention, other factors, particularly negative headlines, have a notable impact. By analyzing these dynamics, the article contributes to a deeper understanding of how media frames shape online attention in the context of news reporting.

The article is published in the journal PLOS One.

Article VII. Henn, T., Posegga, O. 2023. Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit. PLOS One, (18)8, e0288962.

Conclusion

The goal of this dissertation was to resolve the ongoing debate in the literature regarding whether social movements can establish a collective identity in online environments, and if so, how this identity manifests. It thus answers the overarching research question: *How do social movements online become distinct and persistent actors?*

To address this question and contribute to the debate, I adopted a relational understanding of collective identity, grounded in the notion that social movements' collective identities are formed in relation to other actors and shaped by the co-constitution of culture and structure. This understanding led to the conceptual division of collective identity into the protagonist, antagonist, and audience identity field, each capturing a different perspective of collective identity formation of social movements. Based on these identity fields, I derived four subordinate questions, each of which is addressed through one or more of the seven articles included in this dissertation. These four questions are:

RQ1: How do social movements online create and sustain a shared sense of we-ness?

RQ2: How do social movements online draw boundaries to differentiate themselves from other groups?

RQ3a: How can emerging technologies support both the analysis and production of news content that attributes identity traits to social movements online?

RQ3b: How does the media's attribution of identity traits to social movements affect user attention?

I applied the concept of identity work to specify the different expressions of collective identity analyzed within the respective identity fields. In its original form, this concept captures the cultural dimension through the sharing of symbolic resources, which in this dissertation are represented by narratives and memes. To reflect the dual nature of collective identity, I extended the concept to include a social dimension that expresses social structures through networks. Both dimensions represent the internal perspective of collective identity formation, as they capture how movements themselves construct and negotiate their identity. To account as well for the external perspective—the significance of which is widely emphasized in the literature—I extended the expressions of collective identity to also include identity attributions made by external actors, specifically by the media.

While each of the four research questions has been addressed individually through the results of the respective articles, I now briefly summarize key insights that emerge from the overall findings of this dissertation.

First, by highlighting the conceptual and empirical limitations of previous approaches, this dissertation demonstrates that a relational understanding offers a particularly suitable approach for analyzing collective identity formation in online environments. While alternative conceptualizations capture important aspects of collective identity, I demonstrate that a relational perspective is especially well-suited for working with digital trace data. By integrating the concepts of identity fields and identity work into a comprehensive analytical framework, the dissertation shows that collective identity can be effectively conceptualized and operationalized to account for the specific characteristics of online environments, while also enabling a multifaceted analysis of collective identity formation.

Second, by applying the concept of identity fields, this dissertation analyzes social movements' collective identity formation from multiple perspectives, taking into account the most relevant sets of actors in this process. This multi-perspective approach highlights that collective identity is not merely something social movements construct in isolation; rather, it emerges through the dynamic interplay of multiple key actors situated within the same social field.

Third, while the primary focus of this dissertation is on the collective identity formation of social movements online, it also examines interest-based communities (Articles II, III) and protest movements online (Article V) to enrich the findings. The results show that the sharing of cultural symbols used to express collective identity, such as memes, is influenced less by platform regulations and more by how communities define and enforce their boundaries, as well as by their underlying purpose and self-perception. Communities with strong ideological boundaries, such as those associated with social movements, tend to share more visually homogenous memes, as content is closely scrutinized and discarded if it does not align with the community's norms and values. Interest-based communities can also express a shared identity through the recurrence of similar memes when they have a clearly defined purpose centering on specific topics; in contrast, communities lacking such clarity exhibit more heterogeneous meme-sharing. Lastly, the findings show that memes based on the same template are often adopted by multiple communities but are modified and remixed to reflect each group's particular values, norms, and identity. This suggests that the meaning of memes as cultural symbols is not fixed but shaped by the communities that use them.

The findings on media framing of protest movements show that identity traits are most commonly attributed through the debate, spectacle, and victimization frames. While this aligns with previous research and may be influenced by both platform dynamics and the specific protest movements examined, it diverges from the expectations based on the protest paradigm, which predicts a dominance of delegitimizing frames like riot and confrontation. Future research could explore whether the media framing of social movements online follows a similar frame distribution and investigate the broader mechanisms that shape these patterns.

Fourth, this dissertation demonstrates the analytic potential of digital trace data in combination with a mixed-methods design to examine collective identity formation both at scale and in depth. In addition to accounting for the multimodal nature of digital trace data and examining collective identity from different perspectives, the combination of methods offers a major advantage: it helps compensate for the limitations inherent in each individual approach. For example, in Article V, we evaluated the performance of multimodal models to detect frames in online news reporting based on images and captions. While the results were promising, particularly for certain frame types, manual qualitative coding remained essential for establishing a reliable ground truth and revealing the limitations of automated methods. Benchmarking model performance through manual validation is therefore essential, not only to improve these methods but also to ensure meaningful interpretation of their results. As all articles in this dissertation include a manual validation of automated methods, the findings offer a robust basis for analyzing collective identity formation online.

In sum, this dissertation demonstrates that social movements can indeed establish a collective identity online. Movements achieve this by developing a shared sense of *we-ness*, expressed through the use of symbolic resources such as narratives and memes, as well as through the formation of consolidating network structures; by demarcating themselves from antagonists through the assignment of negative traits and the drawing of clear boundaries; and by being recognized and attributed with specific identity traits through the framing of media actors who communicate these characteristics to the public. While the expression of collective identity within these identity fields is shaped by multiple factors, such as a community's boundaries and purpose, the findings show that social movements are capable of forming a coherent collective identity that enables them to become recognizable, persistent actors within the fluid and dynamic nature of online environments, thereby answering the overarching research question. These findings challenge Bennett and Segerberg's claim that collective identity plays a diminishing role online and contribute to a deeper understanding of online organizing in the context of digital activism.

While this dissertation offers important insights into how collective identity is formed and expressed online, it also opens up several promising avenues for future research. First, it would be valuable to further explore the reciprocal relationships between the identity fields. For example, future work could examine how the framing of a protagonist identity influences the formation of an antagonist identity and vice versa. Similarly, an important next step would be to investigate how social movements respond to external identity attribution made by actors such as the media, as originally conceptualized within the audience identity field. Expanding this perspective to include other external actors, such as political stakeholders, and analyzing how movements react to their identity attributions, would also offer valuable insights. In sum, a dynamic analysis of the interplay between all three identity fields and how these interrelations affect collective identity formation represents a promising direction for future research.

Second, when analyzing identity attributions from an audience perspective, it would be valuable to differentiate between symbolic resources and social structures, mirroring the conceptualization of identity work in this dissertation. Doing so would allow the dual understanding of collective identity, grounded in both culture and structure, to be incorporated into the external perspective of identity formation as well. Applying this co-constitution of culture and structure across all identity fields would enable a more nuanced analysis of both dimensions, thereby deepening the insights of collective identity formation online.

Third, while Article I examines collective identity formation from a temporal perspective by analyzing the Querdenken movement across different time periods, it would be valuable to further develop this processual and dynamic approach, as emphasized by Melucci. Future research could, for example, investigate how the use of memes changes over time within specific social movements. Such an analysis would provide a deeper understanding of how cultural collective identity expressions evolve, while also offering valuable insights into how movements adapt and transform their identity work over time.

References

Alexander, J. C. (2006). *The Civil Sphere*. Oxford University Press.

Bakardjieva, M. (2015). Do clouds have politics? Collective actors in social media land. *Information Communication and Society*, *18*(8), 983–990. https://doi.org/10.1080/1369118X.2015.1043320

Barandiaran, X. E., Calleja-López, A., & Cozzo, E. (2020). Defining Collective Identities in Technopolitical Interaction Networks. *Frontiers in Psychology*, *11*, 1549. https://doi.org/10.3389/fpsyg.2020.01549

Basov, N., Breiger, R., & Hellsten, I. (2020). Socio-semantic and other dualities. *Poetics*, *78*, 101433. https://doi.org/10.1016/j.poetic.2020.101433

Basov, N., & Brennecke, J. (2017). Duality beyond dyads: Multiplex patterning of social ties and cultural meanings. *Research in the Sociology of Organizations*, *53*, 87–112. https://doi.org/10.1108/S0733-558X20170000053005

Bauckhage, C. (2011). Insights into internet memes. *Proceedings of the 5th International Conference on Web and Social Media*, 42–49. https://doi.org/10.1609/icwsm.v5i1.14097

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *Proceedings of the 14th International Conference on Web and Social Media*, 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

Benford, R. D. (2013). Identity Fields. In D. A. Snow, D. della Porta, B. Klandermans, & D. McAdam (Eds.), *The Wiley-Blackwell Encyclopedia of Social and Political Movements* (1st ed., pp. 1–2). Blackwell Publishing.

Bennett, W. L., & Segerberg, A. (2012). The logic of connective action: Digital media and the personalization of contentious politics. *Information Communication and Society*, *15*(5), 739–768. https://doi.org/10.1080/1369118X.2012.670661

Bennett, W. L., & Segerberg, A. (2013). *The logic of connective action: Digital media and the personalization of contentious politics*. Cambridge University Press.

Bennett, W. L., Segerberg, A., & Walker, S. (2014). Organization in the crowd: Peer production in large-scale networked protests. *Information Communication and Society*, *17*(2), 232–260. https://doi.org/10.1080/1369118X.2013.870379

Bhattacharya, S., Spann, B., & Agarwal, N. (2024). Solidarity to storming: Assessing the socio-technical factors behind modern social movements. *Proceedings of the 32nd European Conference on Information Systems*.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Blumer, H. G. (1939). Collective behavior. In R. E. Park (Ed.), *An outline of the principle of sociology* (pp. 221–280). Barnes & Noble.

Bock, G.-W., Ahuja, M. K., Suh, A., & Yap, L. X. (2015). Sustainability of a virtual community: Integrating individual and structural dynamics. *Journal of the Association for Information Systems*, *16*(6), 418–447. https://doi.org/10.17705/1jais.00400

boyd, D. (2010). Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. In Z. Papacharissi (Ed.), *Networked Self: Identity, Community, and Culture on Social Network Sites* (1st ed., pp. 47–66). Routledge.

Breiger, R. L. (2010). Dualities of Culture and Structure: Seeing Through Cultural Holes. In J. Fuhse & S. Mützel (Eds.), *Relationale Soziologie* (pp. 37–47). VS Verlag für Sozialwissenschaften.

Brown, D. K., Harlow, S., García-Perdomo, V., & Salaverría, R. (2018). From #Ferguson to #Ayotzinapa: Analyzing differences in domestic and foreign protest news shared on social media. *Mass Communication and Society, 21*(5), 606–630. https://doi.org/10.1080/15205436.2018.1469773

Brown, D. K., & Mourão, R. R. (2021). Protest Coverage Matters: How Media Framing and Visual Communication Affects Support for Black Civil Rights Protests. *Mass Communication and Society*, *24*(4), 576–596. https://doi.org/10.1080/15205436.2021.1884724

Bygstad, B., Munkvold, B. E., & Volkoff, O. (2016). Identifying generative mechanisms through affordances: A framework for critical realist data analysis. *Journal of Information Technology*, *31*(1), 83–96. https://doi.org/10.1057/jit.2015.1

Cardoso, A., Boudreau, M.-C., & Carvalho, J. Á. (2019). Organizing collective action: Does information and communication technology matter? *Information and Organization*, *29*(3), 100256. https://doi.org/10.1016/j.infoandorg.2019.100256

Chabot, R., & Chen, Y. (2020). Living your best life and radiating positivity: Exploratory conceptions of wholesome memes as the new sincerity. *Proceedings of the Annual Conference of CAIS*, 1–7. https://doi.org/10.29173/cais1172

Chan, J. M., & Lee, C.-C. (1984). The journalistic paradigm on civil protests: A case study of Hong Kong. In A. Arno & W. Dissanayake (Eds.), *The news media in national and international conflict* (pp. 183–202). Routledge.

Chen, W., Wei, X., & Zhu, K. X. (2018). Engaging voluntary contributions in online communities. *Mis Quarterly*, *42*(1), 83–100. https://doi.org/10.25300/MISQ/2018/14196

Connolly, R., Sanchez, O. P., Compeau, D., & Tacco, F. M. de S. (2023). Understanding engagement in online health communities: A trust-based perspective. *Journal of the Association for Information Systems, 24*(2), 345–378. https://doi.org/10.17705/1jais.00785

Cools, H., & Diakopoulos, N. (2024). Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities. *Journalism Practice*, 1–19. https://doi.org/10.1080/17512786.2024.2394558

Coretti, L., & Pica, D. (2015). The rise and fall of collective identity in networked movements: Communication protocols, Facebook, and the anti-Berlusconi protest. *Information Communication and Society, 18*(8), 951–967. https://doi.org/10.1080/1369118X.2015.1043317

Courtois, C., & Frissen, T. (2022). Computer Vision and Internet Meme Genealogy: An Evaluation of Image Feature Matching as a Technique for Pattern Detection. *Communication Methods and Measures, 17*(1), 17–39. https://doi.org/10.1080/19312458.2022.2122423

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd International Conference on World Wide Web*, 307–318. https://doi.org/10.1145/2488388.2488416

de Haan, Y., van den Berg, E., Goutier, N., Kruikemeier, S., & Lecheler, S. (2022). Invisible Friend or Foe?How Journalists Use and Perceive Algorithmic-Driven Tools in Their

Research Process. *Digital Journalism, 10*(10), 1775–1793. https://doi.org/10.1080/21670811.2022.2027798
De Saint Laurent, C., Glăveanu, V. P., & Literat, I. (2021). Internet Memes as Partial Stories: Identifying Political Narratives in Coronavirus Memes. *Social Media + Society, 7*(1), 205630512198893. https://doi.org/10.1177/2056305121988932
DeCook, J. R. (2018). Memes and symbolic violence: #proudboys and the use of memes for propaganda and the construction of collective identity. *Learning, Media and Technology, 43*(4), 485–504. https://doi.org/10.1080/17439884.2018.1544149
della Porta, D. (2005). Multiple belongings, tolerant identities, and the "'construction of another politics'": Between the European Social Forum to local Social Fora. In D. della Po & S. Tarrow (Eds.), *Transnational protest and global activism* (pp. 175–202). Rowman & Littlefield.
della Porta, D., & Diani, M. (2006). *Social Movements. An Introduction*. Blackwell.
Diani, M. (1992). The concept of social movement. *The Sociological Review, 40*(1), 1–25. https://doi.org/10.1111/j.1467-954X.1992.tb02943.x
Diani, M. (2002). Network analysis. In B. Klandermans & S. Staggenborg (Eds.), *Methods of social movement research* (pp. 173–200). University of Minnesota Press.
Diani, M. (2003). Networks and Social Movements: A Research Programme. In M. Diani & D. McAdam (Eds.), *Social Movements and Networks: Relational Approaches to Collective Action* (pp. 299–319). Oxford University Press.
Eliasoph, N., & Lichterman, P. (2003). Culture in Interaction. *American Journal of Sociology, 108*(4), 735–794. https://doi.org/10.1086/367920
Emirbayer, M. (1997). Manifesto for a relational sociology. *American Journal of Sociology, 103*(2), 281–317. https://doi.org/10.1086/231209
Emirbayer, M., & Goodwin, J. (1994). Network analysis, culture, and the problem of agency. *American Journal of Sociology, 99*(6), 1411–1454. https://doi.org/10.1086/230450
Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication, 43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x
Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge Collaboration in Online Communities. *Organization Science, 22*(5), 1224–1239. https://doi.org/10.1287/orsc.1100.0614
Faraj, S., Kudaravalli, S., & Wasko, M. (2015). Leading collaboration in online communities. *MIS Quarterly, 39*(2), 393–412.
Fayard, A.-L., & DeSanctis, G. (2008). Kiosks, clubs and neighborhoods: The language games of online forums. *Journal of the Association for Information Systems, 9*(10), 677–705. https://doi.org/10.17705/1jais.00173
Fayard, A.-L., & DeSanctis, G. (2010). Enacting language games: The development of a sense of 'we-ness' in online forums. *Information Systems Journal, 20*(4), 383–416. https://doi.org/10.1111/j.1365-2575.2009.00335.x
Flesher Fominaya, C. (2010). Collective Identity in Social Movements: Central Concepts and Debates. *Sociology Compass, 4*(6), 393–404. https://doi.org/10.1111/j.1751-9020.2010.00287.x
Franzoi, S., Grisold, T., & vom Brocke, J. (2023). Studying dynamics and Change with Digital Trace Data: A Systematic literature Review. *Proceedings of the 31st European Conference on Information Systems*.
Gal, N., Shifman, L., & Kampf, Z. (2016). "It Gets Better": Internet memes and the construction of collective identity. *New Media & Society, 18*(8), 1698–1714. https://doi.org/10.1177/1461444814568784

Gamson, J. (1995). Must Identity Movements Self-Destruct? A Queer Dilemma. *Social Problems, 42*(3), 390–407. https://doi.org/10.1525/sp.1995.42.3.03x0104z

Gamson, W. A. (1991). Commitment and agency in social movements. *Sociological Forum, 6*(1), 27–50. https://doi.org/10.1007/BF01112726

Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2021). Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society, 23*(12), 3491–3508. https://doi.org/10.1177/1461444820958123

George, J. J., & Leidner, D. E. (2019). From clicktivism to hacktivism: Understanding digital activism. *Information and Organization, 29*(3), 100249. https://doi.org/10.1016/j.infoandorg.2019.04.001

Gerbaudo, P. (2012). *Tweets and the Streets: Social Media and Contemporary Activism*. Pluto Press.

Gerbaudo, P. (2015). Protest avatars as memetic signifiers: Political profile pictures and the construction of collective identity on social media in the 2011 protest wave. *Information Communication and Society, 18*(8), 916–929. https://doi.org/10.1080/1369118X.2015.1043316

Gerbaudo, P., & Treré, E. (2015). In search of the 'we' of social media activism: Introduction to the special issue on social media and protest identities. *Information Communication and Society, 18*(8), 865–871. https://doi.org/10.1080/1369118X.2015.1043319

Germonprez, M., & Hovorka, D. S. (2013). Member engagement within digitally enabled social network communities: New methodological considerations. *Information Systems Journal, 23*(6), 525–549. https://doi.org/10.1111/isj.12021

Gibson, J. J. (1977). The Theory of Affornaces. In R. Shaw & J. Bransford (Eds.), *The senses considered as perceptual systems* (pp. 67–82). Erlbaum.

Gieryn, T. F. (1983). Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review, 48*(6), 781–795. https://doi.org/10.2307/2095325

Gitlin, T. (2003). *The whole world is watching: Mass media in the making and unmaking of the new left*. University of California Press.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harvard University Press.

Hall, S. (1997). The Work of Representation. In S. Hall (Ed.), *Representation: Cultural Representations and Signifying Practices* (2nd ed., pp. 13–58). SAGE.

Harlow, S., & Brown, D. K. (2021). Protest News and Facebook Engagement: How the Hierarchy of Social Struggle Is Rebuilt on Social Media. *Journalism & Mass Communication Quarterly, 98*(3), 665–691. https://doi.org/10.1177/10776990211017243

Harlow, S., & Brown, D. K. (2023). A New Protest Paradigm: Toward a Critical Approach to Protest News Analyses. *The International Journal of Press/Politics, 28*(2), 333–343. https://doi.org/10.1177/19401612231153377

Harlow, S., Salaverría, R., Kilgo, D. K., & García-Perdomo, V. (2017). Protest paradigm in multimedia: Social media sharing of coverage about the crime of Ayotzinapa, Mexico. *Journal of Communication, 67*(3), 328–349. https://doi.org/10.1111/jcom.12296

Helberger, N., van Drunen, M., Moeller, J., Vrijenhoek, S., & Eskens, S. (2022). Towards a normative perspective on journalistic AI: Embracing the messy reality of normative ideals. *Digital Journalism, 10*(10), 1605–1626. https://doi.org/10.1080/21670811.2022.2152195

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems, 12*(12), 767–797. https://doi.org/10.17705/1jais.00282

Hunt, S. A., & Benford, R. D. (2004). Collective identity, solidarity, and commitment. In D. A. Snow, S. A. Soule, & H. Kriesi (Eds.), *The Blackwell Companion to social movements* (pp. 433–457). Blackwell.

Hunt, S. A., Benford, R. D., & Snow, D. A. (1994). Identity Fields: Framing Processes and the Social Construction of Movement Identities. In E. Laraña, H. Johnston, & J. R. Gusfield (Eds.), *New Social Movements: From Ideology to Identity* (pp. 185–208). Temple University Press.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L., & Shankar, V. (2021). *OpenCLIP* (Version 0.1) [Computer software]. https://github.com/mlfoundations/open_clip/tree/main

Johnson, S. L., Safadi, H., & Faraj, S. (2015). The emergence of online community leadership. *Information Systems Research, 26*(1), 165–187. https://doi.org/10.1287/isre.2014.0562

Johnston, H., & Klandermans, B. (1995). The cultural analysis of social movements. In H. Johnston & B. Klandermans (Eds.), *Social Movements and Culture* (4th ed., pp. 3–24). University of Minnesota Press.

Jones, J. M. (2023). *Social Media Users More Inclined to Browse Than Post Content*. Gallup.Com. https://news.gallup.com/poll/467792/social-media-users-inclined-browse-post-content.aspx

Jungherr, A. (2015). *Analyzing Political Communication with Digital Trace Data. The role of Twitter Messages in Social Science Research*. Springer.

Juris, J. S. (2008). *Networking futures: The movements against corporate globalization*. Duke University Press.

Karahanna, E., Xin Xu, S., Xu, Y., & Zhang, N. (Andy). (2018). The needs-affordances-features perspective for the use of social media. *MIS Quarterly, 42*(3), 737–756. https://doi.org/10.25300/MISQ/2018/11492

Kavada, A. (2009). Email lists and the construction of an open and multifaceted identity. *Information, Communication & Society, 12*(6), 817–839. https://doi.org/10.1080/13691180802304854

Kavada, A. (2015). Creating the collective: Social media, the Occupy Movement and its constitution as a collective actor. *Information Communication and Society, 18*(8), 872–886. https://doi.org/10.1080/1369118X.2015.1043318

Khazraee, E., & Novak, A. N. (2018). Digitally Mediated Protest: Social Media Affordances for Collective Identity Construction. *Social Media and Society, 4*(1), 2056305118765740. https://doi.org/10.1177/2056305118765740

Lasfer, A., & Vaast, E. (2021). Finding Simplicity in Complexity-Towards an Integrative Understanding of Online Community Sustainability. *Proceedings of the 42nd International Conference on Information Systems*.

Lee, M., & Martin, J. L. (2018). Doorway to the dharma of duality. *Poetics, 68*, 18–30. https://doi.org/10.1016/j.poetic.2018.01.001

Leonardi, P. M. (2012). Materiality, sociomateriality, and socio-technical systems: What do these terms mean? How are they different? Do we need them? In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), *Materiality and organizing: Social interaction in a technological world* (pp. 24–48). Oxford University Press.

Leonardi, P. M., & Vaast, E. (2017). Social media and their affordances for organizing: A review and agenda for research. *Academy of Management Annals, 11*(1), 150–188. https://doi.org/10.5465/annals.2015.0144

Leong, C., Faik, I., Tan, F. T. C., Tan, B., & Khoo, Y. H. (2020). Digital organizing of a global social movement: From connective to collective action. *Information and Organization, 30*(4), 100324. https://doi.org/10.1016/j.infoandorg.2020.100324

Leong, C., Pan, S. L., Bahri, S., & Fauzi, A. (2019). Social media empowerment in social movements: Power activation and power accrual in digital activism. *European Journal of Information Systems, 28*(2), 173–204. https://doi.org/10.1080/0960085X.2018.1512944

Lu, X., Jiang, J., Head, M., & Yang, J. (2022). The impact of linguistic complexity on leadership in online q&a communities: Comparing knowledge shaping and knowledge adding. *Information & Management, 59*(6), 103675. https://doi.org/10.1016/j.im.2022.103675

Lugea, J. (2019). The pragma-stylistics of 'image macro'internet memes. In H. Ringrow & S. Pihlaja (Eds.), *Contemporary media stylistics* (pp. 81–106). Bloomsbury Publishing.

Maghrabi, R., & Salam, A. F. (2011). Social Media, Social Movement and Political Change: The Case of 2011 Cairo Revolt. *Proceedings of the 32nd International Conference on Information Systems.*

Marsden, P. V. (1990). Network Data and Measurement. *Annual Review of Sociology, 16*(1), 435–463. https://doi.org/10.1146/annurev.so.16.080190.002251

Marwick, A. E. (2015). Instafame: Luxury selfies in the attention economy. *Public Culture, 27*(1), 137–160. https://doi.org/10.1215/08992363-2798379

Massa, F. G., & O'Mahony, S. (2021). Order from chaos: How networked activists self-organize by creating a participation architecture. *Administrative Science Quarterly, 66*(4), 1037–1083. https://doi.org/10.1177/00018392211008880

Masullo, G. M., Brown, D. K., & Harlow, S. (2024). Shifting the protest paradigm? Legitimizing and humanizing protest coverage lead to more positive attitudes toward protest, mixed results on news credibility. *Journalism, 25*(6), 1230–1251. https://doi.org/10.1177/14648849231200135

McAdam, D., & Snow, D. (2000). Identity Work Processes in the Context of Social Movements: Clarifying the Identity/Movement Nexus. In S. Stryker, T. J. Owens, & R. W. White (Eds.), *Self, Identity and Social Movements* (Vol. 13, pp. 41–67). University of Minnesota Press.

McCurdy, P. (2012). Social movements, protest and mainstream media. *Sociology Compass, 6*(3), 244–255. https://doi.org/10.1111/j.1751-9020.2011.00448.x

McLeod, D. M., & Detenber, B. H. (1999). Framing effects of television news coverage of social protest. *Journal of Communication, 49*(3), 3–23. https://doi.org/10.1111/j.1460-2466.1999.tb02802.x

McLeod, D. M., & Hertog, J. K. (1992). The Manufacture of ` Public Opinion' by Reporters: Informal Cues for Public Perceptions of Protest Groups. *Discourse & Society, 3*(3), 259–275. https://doi.org/10.1177/0957926592003003001

McLeod, D. M., & Hertog, J. K. (1999). Social control, social change and the mass media's role in the regulation of protest groups. In D. Demers & K. Viswanath (Eds.), *Mass media, social control, and social change: A macrosocial perspective* (pp. 305–332). Iowa State University Press.

Melucci, A. (1995). The Process of Collective Identity. In H. Johnston & B. Klandermans (Eds.), *Social Movements and Culture* (4th ed., pp. 41–63). University of Minnesota Press.

Melucci, A. (1996). *Challenging Codes: Collective action in the information age*. Cambridge University Press.

Merriam-Webster. (2024). *Definition of ALT-RIGHT*. https://www.merriam-webster.com/dictionary/alt-right

Milan, S. (2015a). From social movements to cloud protesting: The evolution of collective identity. *Information Communication and Society*, *18*(8), 887–900. https://doi.org/10.1080/1369118X.2015.1043135

Milan, S. (2015b). Mobilizing in Times of Social Media: From Politics of Identity to a Politics of Visibility. In L. Dencik & O. Leistert (Eds.), *Critical Perspectives on Social Media and Protest* (pp. 53–71). Rowman & Littlefield.

Milan, S. (2015c). When Algorithms Shape Collective Action: Social Media and the Dynamics of Cloud Protesting. *Social Media + Society*, *1*(2), 2056305115622481. https://doi.org/10.1177/2056305115622481

Milan, S. (2021). Peer Production and Collective Action. In M. O'Neil, C. Pentzold, & S. Toupin (Eds.), *The Handbook of Peer Production* (1st ed., pp. 299–310). John Wiley & Sons.

Mische, A. (2003). Cross-Talk in Movements: Reconceiving the Culture-Network Link. In M. Diani & D. McAdam (Eds.), *Social Movements and Networks: Relational Approaches to Collective Action* (pp. 258–280). https://doi.org/10.1093/0199251789.003.0011

Mische, A., & White, H. C. (1998). Between conversation and situation: Public switching dynamics across network domains. *Social Research*, *65*(3), 695–724.

Mohr, J. W. (2000). Introduction: Structures, institutions, and cultural analysis. *Poetics*, *27*(2–3), 57–68. https://doi.org/10.1016/S0304-422X(00)00002-4

Mohr, J. W., Bail, C., Frye, M., Lena, J., Lizardo, O., McDonnell, T., Mische, A., Tavory, I., & Wherry, F. F. (2020). *Measuring Culture*. Columbia University Press.

Mohr, J. W., & Duquenne, V. (1997). The duality of culture and practice: Poverty relief in New York City, 1888-1917. *Theory and Society*, *26*(2–3), 305–356. https://doi.org/10.1023/A:1006896022092

Mohr, J. W., & White, H. C. (2008). How to model an institution. *Theory and Society*, *37*(5), 485–512. https://doi.org/10.1007/s11186-008-9066-0

Monterde, A., Calleja-López, A., Aguilera, M., Barandiaran, X. E., & Postill, J. (2015). Multitudinous identities: A qualitative and network analysis of the 15M collective identity. *Information Communication and Society*, *18*(8), 930–950. https://doi.org/10.1080/1369118X.2015.1043315

Morina, D., & Bernstein, M. S. (2022). A Web-Scale Analysis of the Community Origins of Image Memes. *Conference on Human-Computer Interaction*, *6*, 74:1-74:25. https://doi.org/10.1145/3512921

Moser, C., Ganley, D., & Groenewegen, P. (2013). Communicative genres as organising structures in online communities – of team players and storytellers. *Information Systems Journal*, *23*(6), 551–567. https://doi.org/10.1111/isj.12022

Mundt, M., Ross, K., & Burnett, C. M. (2018). Scaling Social Movements Through Social Media: The Case of Black Lives Matter. *Social Media + Society*, *4*(4), 2056305118807911. https://doi.org/10.1177/2056305118807911

Nasrin, S., & Fisher, D. R. (2022). Understanding collective identity in virtual spaces: A study of the youth climate movement. *American Behavioral Scientist, 66*(9), 1286–1308. https://doi.org/10.1177/00027642211056257
Neumayer, C., & Rossi, L. (2018). Images of protest in social media: Struggle over visibility and visual narratives. *New Media and Society, 20*(11), 4293–4310. https://doi.org/10.1177/1461444818770602
Nishal, S., & Diakopoulos, N. (2024). *Envisioning the Applications and Implications of Generative AI for News Media* (arXiv:2402.18835). arXiv. https://doi.org/10.48550/arXiv.2402.18835
Nissenbaum, A., & Shifman, L. (2017). Internet memes as contested cultural capital: The case of 4chan's /b/ board. *New Media & Society, 19*(4), 483–501. https://doi.org/10.1177/1461444815609313
Nissenbaum, A., & Shifman, L. (2018). Meme Templates as Expressive Repertoires in a Globalizing World: A Cross-Linguistic Study. *Journal of Computer-Mediated Communication, 23*(5), 294–310. https://doi.org/10.1093/jcmc/zmy016
Oh, W., Moon, J. Y., Hahn, J., & Kim, T. (2016). Research Note—Leader Influence on Sustained Participation in Online Collaborative Work Communities: A Simulation-Based Approach. *Information Systems Research, 27*(2), 383–402. https://doi.org/10.1287/isre.2016.0632
Opp, K.-D. (2009). Collective identity and social movement activity. In K.-D. Opp (Ed.), *Theories of political protest and social movements: A multidisciplinary introduction, critique, and synthesis* (pp. 204–233). Routledge.
Panteli, N., & Sivunen, A. (2019). "I Am Your Fan; Bookmarked!" Members' identification development in founder-led online communities. *Journal of the Association for Information Systems, 20*(6), 824–841. https://doi.org/10.17705/1.jais.00552
Park, H., & Zaggl, M. A. (2022). When to signal? Contingencies for career-motivated contributions in online collaboration communities. *Journal of the Association for Information Systems, 23*(6), 1386–1419. https://doi.org/10.17705/1jais.00765
Polletta, F., & Jasper, J. M. (2001). Collective identity and social movements. *Annual Review of Sociology, 27*(1), 283–305. https://doi.org/10.1146/annurev.soc.27.1.283
Powell, C., & Dépelteau, F. (2013). Introduction. In C. Powell & F. Dépelteau (Eds.), *Conceptualizing relational sociology: Ontological and theoretical issues* (pp. 1–12). Palgrave Macmillan. https://doi.org/10.1057/9781137342652
Rainee, L., & Wellman, B. (2012). *Networked: The new social operating system.* MIT Press.
Ransbotham, S., & Kane, G. C. (2011). Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *MIS Quarterly, 35*(3), 613–627. https://doi.org/10.2307/23042799
Ray, D., & Tarafdar, M. (2017). How does Twitter influence a social movement? *Proceedings of the 25th European Conference on Information Systems.*
Ray, S., Kim, S. S., & Morris, J. G. (2014). The Central Role of Engagement in Online Communities. *Information Systems Research, 25*(3), 528–546. https://doi.org/10.1287/isre.2014.0525
Reyes-Menendez, A., Saura, J. R., & Thomas, S. B. (2020). Exploring key indicators of social identity in the #MeToo era: Using discourse analysis in UGC. *International Journal of Information Management, 54*, 102129.
Ronzhyn, A., Cardenal, A. S., & Rubio, A. B. (2023). Defining affordances in social media research: A literature review. *New Media & Society, 25*(11), 3165–3188. https://doi.org/10.1177/14614448221135187

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Rucht, D. (2003). The changing role of political protest movements. *West European Politics, 26*(4), 153–176. https://doi.org/10.1080/01402380312331280728

Rucht, D. (2023). *Social movements. A theoretical approach*. Oxford University Press.

Schultz, J., & Breiger, R. L. (2010). The strength of weak culture. *Poetics, 38*(6), 610–624. https://doi.org/10.1016/j.poetic.2010.09.002

Schwalbe, M. L., & Mason-Schrock, D. (1996). Identity work as group process. *Advances in Group Processes, 13*, 113–147.

Seiffert-Brockmann, J., Diehl, T., & Dobusch, L. (2018). Memes as games: The evolution of a digital discourse online. *New Media & Society, 20*(8), 2862–2879. https://doi.org/10.1177/1461444817735334

Selander, L., & Jarvenpaa, S. L. (2016). Digital action repertoires and transforming a social movement organization. *MIS Quarterly, 40*(2), 331–352. https://doi.org/10.25300/MISQ/2016/40.2.03

Sherif, M. (1966). *Group conflict and co-operation: Their social psychology*. Routledge & Kegan Paul.

Shirky, C. (2008). *Here Comes Everybody: The Power of Organizing Without Organizations*. Penguin Books.

Shu, L., Xu, H., & Liu, B. (2017). DOC: Deep Open Classification of Text Documents. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2911–2916.

Simon, F. M. (2024). Escape Me If You Can: How AI Reshapes News Organisations' Dependency on Platform Companies. *Digital Journalism, 12*, 149–170. https://doi.org/10.1080/21670811.2023.2287464

Snow, D. A. (2001). *Collective Identity and Expressive Forms* [Working Paper]. Center for the Study of Democracy.

Snow, D. A., & Anderson, L. (1987). Identity Work Among the Homeless: The Verbal Construction and Avowal of Personal Identities. *American Journal of Sociology, 92*(6), 1336–1371. https://doi.org/10.1086/228668

Snow, D. A., & Benford, R. D. (1988). Ideology, Frame Resonance, and Participant Mobilization. In B. Klandermans, H. Kriesi, & S. G. Tarrow (Eds.), *International Social Movement Research* (pp. 197–217). JAI Press.

Snow, D. A., & Benford, R. D. (1992). Master frames and cycles of protest. In A. D. Morris & C. McClurg Mueller (Eds.), *Frontiers in social movement theory* (pp. 133–155). Yale University Press.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences of the United States of America*.

Sproull, L., & Arriaga, M. (2007). Online communities. In H. Bidogli (Ed.), *Handbook of Computer Networks* (Vol. 3, pp. 898–914). John Wiley & Sons.

Statista. (2025a). *Global daily social media usage 2024*. Statista. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Statista. (2025b). *Number of worldwide social network users 2028*. Statista. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology*, *18*(6), 813–847. https://doi.org/10.1080/09541440500412361
Steward Jr., G. A., Shriver, T. E., & Chasteen, A. L. (2002). Participant narratives and collective identity in a metaphysical movement. *Sociological Spectrum*, *22*(1), 107–135. https://doi.org/10.1080/027321701753284305
Stewart, M., & Schultze, U. (2019). Producing solidarity in social media activism: The case of My Stealthy Freedom. *Information and Organization*, *29*(3), 100251. https://doi.org/10.1016/j.infoandorg.2019.04.003
Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669
Stray, J. (2019). Making artificial intelligence work for investigative journalism. *Digital Journalism*, *7*(8), 1076–1097. https://doi.org/10.1080/21670811.2019.1630289
Strong, D. M., Volkoff, O., Johnson, S. A., Pelletier, L. R., & Tulu, B. (2014). A theory of organization: EHR affordance actualization. *Journal of the Association for Information Systems*, *15*(2), 53–85. https://doi.org/10.17705/1jais.00353
Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review*, *51*(2), 273–286. https://doi.org/10.2307/2095521
Syed, R., & Silva, L. (2023). Social Movement Sustainability on Social Media: An Analysis of the Women's March Movement on Twitter. *Journal of the Association for Information Systems*, *24*(1), 249–293. https://doi.org/10.17705/1jais.00776
Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Eds.), *Political psychology* (pp. 276–293). Psychology Press.
Tarafdar, M., & Kajal Ray, D. (2021). Role of Social Media in Social Protest Cycles: A Sociomaterial Examination. *Information Systems Research*, *32*(3), 1066–1090. https://doi.org/10.1287/isre.2021.1013
Taylor, V., & Whittier, N. E. (1992). Collective Identity in Social Movement Communities: Lesbian Feminist Mobilization. In A. D. Morris & C. M. C. Mueller (Eds.), *Frontiers in Social Movement Theory* (pp. 104–129). Yale University Press.
Tilly, C. (2003). Political identities in changing polities. *Social Research*, *70*(2), 605–619. https://doi.org/10.1353/sor.2003.0021
Treem, J. W., & Leonardi, P. M. (2013). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association*, *36*(1), 143–189. https://doi.org/10.1080/23808985.2013.11679130
Treré, E. (2015). Reclaiming, proclaiming, and maintaining collective identity in the #YoSoy132 movement in Mexico: An examination of digital frontstage and backstage activism through social media and instant messaging platforms. *Information Communication and Society*, *18*(8), 901–915. https://doi.org/10.1080/1369118X.2015.1043744
Trujillo, A., & Cresci, S. (2022). Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald. *Proceedings of the Conference on Human-Computer Interaction*, *6*, 1–28.
Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Vaast, E., Safadi, H., Lapointe, L., & Negoita, B. (2017). Social Media Affordances for Connective Action. *MIS Quarterly*, *41*(4), 1179–1206. https://doi.org/10.25300/MISQ/2017/41.4.08

Van Dalen, A. (2024). Revisiting the Algorithms Behind the Headlines. How Journalists Respond to Professional Competition of Generative AI. *Journalism Practice*, 1–18. https://doi.org/10.1080/17512786.2024.2389209

Volkoff, O., & Strong, D. M. (2013). Critical realism and affordances: Theorizing IT-associated organizational change processes. *MIS Quarterly*, *37*(3), 819–834.

White, H. C. (1992). *Identity and Control a structural theory of social action*. Princeton University Press.

White, H. C. (1997). Can mathematics be social? Flexible representations for interaction process and its sociocultural constructions. *Sociological Forum*, *12*(1), 53–71. https://doi.org/10.1023/A:1024604606367

White, H. C. (2008). *Identity and Control: How Social Formations Emerge*. Princeton University Press.

White, H. C., Fuhse, J., Thiemann, M., & Buchholz, L. (2007). Networks and Meaning: Styles and Switchings. *Soziale Systeme*, *13*(1–2). https://doi.org/10.1515/sosys-2007-1-246

Williams, R. H. (2013). Culture and Social Movements. In D. A. Snow, D. Della Porta, D. McAdam, & B. Klandermans (Eds.), *The Wiley-Blackwell Encyclopedia of Social and Political Movements*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470674871.wbespm429

Wuthnow, R. (1987). *Meaning and moral order: Explorations in cultural analysis*. University of California Press.

Young, A., Selander, L., & Vaast, E. (2019). Digital organizing for social impact: Current insights and future research avenues on collective action, social movements, and digital technologies. *Information and Organization*, *29*(3), 100257. https://doi.org/10.1016/j.infoandorg.2019.100257

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Suarez-Tangil, G. (2018). On the origins of memes by means of fringe web communities. *Proceedings of the Internet Measurement Conference*, 188–202.

Zhang, D., Zhang, K., & Yang, Y. (2024). TM-OKC: AN UNSUPERVISED TOPIC MODEL FOR TEXT IN ONLINE KNOWLEDGE COMMUNITIES. *MIS Quarterly*, *48*(3). https://doi.org/10.25300/MISQ/2023/17885

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., & Li, H. (2022). Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Proceedings of Computer Vision – ECCV 2022* (Vol. 13695, pp. 493–510). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19833-5_29

Appendix

Appendix A.1: Paper I

Bibliographic data	Henn-Latus, T., Tell, S., Polenz, J., Kern, T., Posegga, O. 2025. In Search of a "Style:" Capturing the Collective Identity of Social Movements Based on Digital Trace Data. Journal of the Association for Information Systems, 26(6), 1566-1629.
Abstract	The emergence of social media platforms such as Twitter and Facebook is widely acknowledged to have fundamentally changed digital activism. Scholars remain divided, however, on the role of collective identity for social movements online, particularly in terms of the debate between connective and collective action. This paper contributes to this debate by demonstrating that collective identity formation online is possible and can create distinguishable social actors that can be sustained over time. When evaluating collective identity formation online, previous research has focused either on cultural references shared by platform users or on the interaction patterns among them. Foundational theories on collective identity, however, argue that both cultural and social dynamics must be examined together. This paper fills this research gap by drawing on the notion of "style" as introduced by network sociologist Harrison White, which inherently captures collective identity's duality of culture and structure. We operationalized style based on socio-semantic network analysis, enabling the empirical assessment of collective identity formation online. We applied our style conceptualization to the Querdenken (German for "thinking outside the box") movement—Germany's most successful movement mobilizing against COVID-19 measures—analyzing its activity on Twitter. Our findings revealed that Querdenken's collective identity online materialized through recurring social and cultural patterns that persisted independently of individual users. These patterns were sustained over time by a multitude of users, with some temporarily assuming emergent leadership roles, thereby significantly shaping Querdenken's collective identity. We theorize our findings in relation to the existing literature, offering nine propositions to guide future research on the collective identity formation of social movements online.
DOI	10.17705/1jais.00951

Journal of the Association for Information Systems (2025) **26**(6), 1566-1629
doi: 10.17705/1jais.00951

RESEARCH ARTICLE

ISSN 1536-9323

In Search of a "Style:" Capturing the Collective Identity of Social Movements Based on Digital Trace Data

Theresa Henn-Latus,[1] Sarah Tell,[2] Julian Polenz,[3] Thomas Kern,[4] Oliver Posegga[5]
[1]University of Bamberg, Germany, theresa.henn@uni-bamberg.de
[2]University of Bremen, Germany, sarah.tell@uni-bremen.de
[3]University of Bamberg, Germany, julian.polenz@uni-bamberg.de
[4]University of Bremen, Germany, thomas.kern@uni-bremen.de
[5]University of Bamberg, Germany, oliver.posegga@uni-bamberg.de

Abstract

The emergence of social media platforms such as Twitter and Facebook is widely acknowledged to have fundamentally changed digital activism. Scholars remain divided, however, on the role of collective identity for social movements online, particularly in terms of the debate between connective and collective action. This paper contributes to this debate by demonstrating that collective identity formation online is possible and can create distinguishable social actors that can be sustained over time. When evaluating collective identity formation online, previous research has focused either on cultural references shared by platform users or on the interaction patterns among them. Foundational theories on collective identity, however, argue that both cultural and social dynamics must be examined together. This paper fills this research gap by drawing on the notion of "style" as introduced by network sociologist Harrison White, which inherently captures collective identity's duality of culture and structure. We operationalized style based on socio-semantic network analysis, enabling the empirical assessment of collective identity formation online. We applied our style conceptualization to the Querdenken (German for "thinking outside the box") movement—Germany's most successful movement mobilizing against COVID-19 measures—analyzing its activity on Twitter. Our findings revealed that Querdenken's collective identity online materialized through recurring social and cultural patterns that persisted independently of individual users. These patterns were sustained over time by a multitude of users, with some temporarily assuming emergent leadership roles, thereby significantly shaping Querdenken's collective identity. We theorize our findings in relation to the existing literature, offering nine propositions to guide future research on the collective identity formation of social movements online.

Keywords: Collective Identity Online, Connective Action, Collective Action, Social Media, Social Movements, Socio-Semantic Networks, Digital Trace Data, Online Collectives, Leadership

Yulin Fang was the accepting senior editor. This research article was submitted on April 19, 2023, and underwent three revisions.

1 Introduction

Political activism has undergone tremendous changes over the last few decades. The rise of social media platforms such as X and Facebook has provided social movements with unprecedented ways to mobilize and interact with adherents (e.g., Earl, 2018; George & Leidner, 2019; Syed & Silva, 2023; Tarafdar & Ray, 2021; Tufekci, 2017; Young et al., 2019).[1] This transformative influence of social media on activism has intrigued information systems (IS) scholars, who have increasingly been evaluating these new forms of digital engagement and mobilization (e.g., Braccini et al., 2019; Ghobadi & Clegg, 2015; Leonardi & Vaast, 2017; Leong et al., 2020; Venkatesan et al., 2021; Young et al., 2019). Recent papers have explored reasons for sustained protestor engagement (Syed & Silva, 2023) and investigated shared solidarity between movement supporters (Stewart & Schultze, 2019), seeking to better understand the evolving action repertoires of activists (Selander & Jarvenpaa, 2016). In addition, scholars have shed light on new forms of empowerment (Leong et al., 2019; Miranda et al., 2016) and have evaluated novel approaches to value creation (Chamakiotis et al., 2021) and collective sensemaking within social movements (Oh et al., 2015).

Despite this growing interest in digital activism within the IS field in recent years, however, George and Leidner (2019) describe this area of research as still a "raucous teenager" (p. 16), emphasizing its early stage and numerous questions that remain unanswered. These include "what it means to be organized" (Young et al., 2019, p. 1), which has emerged as a focal point for scholars (e.g., Cardoso et al., 2019; Leong et al., 2020; Levina & Arriaga, 2014; Massa & O'Mahony, 2021; Vaast et al., 2017; Young et al., 2019). This field of study is rooted in Bennett and Segerberg's seminal work "The Logic of Connective Action" (2013), which introduced the concept of "connective action" as the primary mode of organization for online collectives, diverging from the traditional form of "collective action." Whereas collective action relies on social movement organizations (SMOs) to raise a movement's profile, gather resources, formulate its strategy, and ultimately ensure its continuity by contributing to the creation and maintenance of its identity (for an overview, see Earl, 2015, pp. 37-38; generally, see Bennett & Segerberg, 2013), connective action is fueled by online sharing and adapting personalized versions of highly inclusive action frames, allowing adherents to "specify their own connection to an issue rather than adopt more demanding models regarding how to think and act" (Bennett & Segerberg, 2013, p. 197). Thus, unlike collective action, connective action requires neither shared ideological beliefs (Bennett & Segerberg, 2013, p. 50) nor centers on the "symbolic construction of a united 'we'" (Bennett & Segerberg, 2012, p. 748). Consequently, it imposes minimal entry and exit barriers, facilitating informal and spontaneous self-organization, which may scale up quickly (Bennett & Segerberg, 2013; Vaast et al., 2017, p. 1180), but lacks a strong attachment to the collective. Bennett and Segerberg (2013, pp. 28, 50) thus declared the "demise of collective identity in contemporary Internet-enabled movements" (Stewart & Schultze, 2017, p. 4; for similar positions, see also Gladwell, 2010; Morozov, 2011).

In sharp contrast with Bennett and Segerberg, traditional social movement theory posits that developing a collective identity is crucial for social movements (especially Melucci, 1995, 1996), with social movement scholars arguing that social movements "must make the construction of collective identity one of [their] most central tasks" (Gamson, 1991, p. 27). Generally, collective identity denotes

> *a shared sense of "one-ness" or "we-ness" anchored in real or imagined shared attributes and experiences among those who comprise the collectivity and in relation or contrast to one or more actual or imagined sets of "others." Embedded within the shared sense of "we" is a corresponding sense of "collective agency"* (Snow, 2001, p. 2).[2]

Thus, it allows composite actors to "act as unified and delimited subjects" (Melucci, 1996, p. 72), that is, it enables them to recognize their own actions and attribute the outcomes of those actions to themselves, thereby becoming emergent social entities that endure over time (Melucci, 1996, p. 72; see also Syed & Silva, 2023, p. 267). This also holds true for composite actors other than social movements, such as organizations (e.g., Albert & Whetten, 1985; Gioia et al., 2000, 2013). Without a collective identity, connective action is therefore expected to "expand haphazardly" (Massa & O'Mahony, 2021, p. 1056) and disperse quickly thereafter due to its lack of emphasis on sustainably integrating its adherents. Yet, contemporary movements such as MeToo and Black Lives Matter have demonstrated that sustained integration of adherents and coordinated mobilization are indeed achievable online (for Black Lives Matter, see, for instance, Mundt et al., 2018; for MeToo, see, for instance, Reyes-Menendez et al., 2020), indicating the successful formation of a collective identity that can enable them to become discernable and enduring

[1] By "social media," we mean an internet-based "digital space" (Kapoor et al., 2018) consisting of "user-driven platforms that facilitate diffusion of compelling content, dialogue creation, and communication to a broader audience" (p. 536). Following Diani (1992), we define a "social movement" as "a network of informal interactions between a plurality of individuals, groups and/or organizations, engaged in a political or cultural conflict, on the basis of a shared collective identity" (p. 13).

[2] While there is "no consensual definition" (Snow, 2001, p. 2) of collective identity, there is broad agreement on the concept's core dimensions, as reflected in the definition provided by Snow: a sense of collective agency, sameness, and distinctiveness. We explain our choice of this definition in more detail in Section 2.1.

actors online. Against this backdrop, claiming that connective action lacks identity formation calls into question how these movements manage to sustain their support and maintain their existence online.

Summarizing the positions presented, there are two positions at odds. On the one hand, the scholarship on connective action posits that collective identity plays no role in online social movements, even though contemporary online movements demonstrate a remarkable capacity to bind and mobilize supporters. Traditional social movement research, on the other hand, argues that without collective identity, there is neither meaningful collective action nor the persistence of a movement online or offline. Our goal is to clarify whether and in what form collective identity plays a role online. This inquiry is not unprecedented. Faced with this contradiction, IS and social movement studies have increasingly challenged Bennett and Segerberg's (2013) assertion through empirical testing (e.g., Gerbaudo, 2015; Leong et al., 2019, 2020; Monterde et al., 2015; Syed & Silva, 2023; Treré, 2015; see also Section 2.2)—a step that was notably absent in the authors' original work.

Studies that empirically investigate collective identity online, however, generally either focus solely on cultural references shared by platform users (e.g., the use of shared hashtags and pronouns such as "us" and "we") or on interaction patterns among platform users. Foundational theoretical work on collective identity, however, has long posited that both cultural and social dynamics must be examined in tandem to fully elucidate the concept of collective identity (e.g., Melucci, 1996, p. 15). Only by adopting such an integrated approach can we eventually determine whether particular cultural elements sustainably integrate a particular set of actors, with the cultural and actor levels mutually defining and demarcating one another. Contemporary studies frequently overlook this multilayered complexity, leading to an incomplete understanding of collective identity as originally theorized. At the same time, this prevailing disjunction between the actor level and the cultural level means that the apparent contradiction derived above cannot be conclusively resolved, as this would require both levels of analysis to be examined in their co-constitution. Thus, we conclude that an appropriate conceptualization and operationalization of collective identity online is still lacking. Therefore, our first research question is:

RQ1: How can collective identity online be conceptualized and determined empirically?

Thus far, we have discussed collective identity as a "product" (Flesher Fominaya, 2010, p. 396), a "social object to which the movement's protagonists, adversaries, and audience(s) respond" (Snow, 2001, p. 4), without addressing the process that generates and sustains it and thereby provides the collective with temporal continuity. Social movement theory suggests that neither collective identity nor its continuity should be regarded as given (Melucci, 1996, p. 71-73). Therefore, it is essential to analyze how a movement's sense of we-ness—that is, collective identity as a product—is actually created and sustained over time (e.g., Melucci, 1996, p. 46). To explain this process of collective identity formation and its manifestation online, we theorize about how social media facilitates user behavior and the emergence of specific structures that shape the formation of social movements' collective identity online. Based on these theoretical advances, we contribute to research that "theorize[s] emerging phenomena from an IS perspective," shedding additional light on "the interplay between people and technologies" (Leidner & Gregory, 2024, p. 501) and also outline avenues for future research. Accordingly, our second research question is:

RQ2: How can the formation of collective identity online be explained?

To answer our first research question, we adopted network sociologist Harrison White's relational understanding of identity and derived a dual conceptualization of collective identity and operationalization. This approach enabled us to determine collective identity online based on digital trace data via socio-semantic network analysis and additional in-depth qualitative analyses. This approach also situates our examination of collective identity online within the broader debate on the duality of culture and social networks (for an overview, see, e.g., Basov et al., 2020; Lee & Martin, 2018; Pachucki & Breiger, 2010) and acknowledges previous calls for integrating these levels of analysis to fully leverage the potential of digital trace data (Barberá et al., 2015; Borgatti & Foster, 2003; Kane et al., 2014; Syed & Silva, 2020; Zhang et al., 2016). We illustrated our approach through a case study on the German movement Querdenken ("thinking outside the box") on Twitter.[3] Based on the empirical findings, we derived nine propositions on the conditions required for identity formation online, informed by our dual conceptualization and thus answering our second research question.

[3] The collected digital trace data we use originated from the period when Twitter was the name of the platform now known as X. Therefore, we refer to the former name.

The remainder of this paper is structured as follows. In Section 2, we outline our general understanding of collective identity in more detail and summarize previous attempts to empirically detect collective identity online, identifying existing gaps in the literature. In Section 3, we derive the dual conceptualization of collective identity based on White's notion of style as a tool to bridge the identified gaps in previous literature. We adapt this conceptualization to the context of social movements and delineate, more specifically, how a movement's style manifests online, resulting in a movement-specific framework for conceptualizing style in digital environments. In Section 4, we introduce Querdenken, the data collection process, and our operationalization and empirical analysis of style. In Section 5, we present the results of the case study, offering empirical insights into the manifestation of collective identity through the lens of style. In Section 6, we outline the propositions derived from our findings. Section 7 discusses the theoretical and practical implications of our study and suggests directions for future research.

Overall, our findings show that Querdenken was capable of creating a unique style, demarcating the movement as a distinct social actor. This expression of collective identity online became particularly tangible through the consolidation of recurring cultural and social patterns, which are a structurally autonomous phenomenon, as they are continuously reproduced independently of specific users.

2 Background Literature

2.1 Defining Collective Identity

Collective identity is a multidimensional concept originating from social movement studies that has been analyzed by numerous authors over a number of decades (e.g., Eder, 2009; Hunt & Benford, 2004; Melucci, 1995, 1996; Snow, 2001; Taylor & Whittier, 1992; Touraine, 1985), but, as yet, this has resulted in "no consensual definition" (Snow, 2001, p. 2). In the existing literature, the term is often left undefined or is explained using a variety of definitions (see Appendix A for details).

As noted in the introduction, we adopted Snow's (2001) definition of collective identity, which we chose for its comprehensive approach to the concept's multidimensional nature: It incorporates the three key dimensions associated with the concept in relevant definitional approaches, as delineated by Stewart and Schultze (2019, pp. 20-21): a sense of *collective agency*—that is, the conceived capability of acting as a unified social entity—that creates for activists the perception that they are, collectively, "subjects of history, not merely objects" (Gamson, 1991, p. 37; see also Snow, 2001, p. 2); *sameness*, as indicated in Snow's definition by a collective's shared attributes and experiences; and *distinctiveness*, as indicated by the demarcation toward sets of "others" (Snow, 2001, p. 2). Establishing sameness and distinctiveness—and thereby answering the questions "Who are we?" and "Who are they?"—inherently involves the creation of a boundary demarcating "us" from "them" (e.g., Gamson, 1991, p. 42; Tilly, 2003, p. 608). This process, known as "boundary work," is integral to collective identity formation (e.g., Flesher Fominaya, 2019, pp. 435-436; Hunt & Benford, 2004, pp. 442-444; Snow, 2001, p. 8; Taylor & Whittier, 1992, pp. 111-112; generally, see also Gieryn, 1983; Lamont & Molnár, 2002). As emphasized earlier, sustaining these conceptual dimensions is a "laborious process" (Melucci, 1996, p. 76) because the collective's agency, sameness, and distinctiveness must be continually reaffirmed to create and maintain the impression of a coherent social actor with "a certain stability and permanence over time" (Melucci, 1996, p. 71; for the significance of temporal continuity in relation to collective identity, see also Albert & Whetten, 1985). Consequently, if we seek to determine collective identity, it must be consistently recognizable over time; otherwise, the criterion of temporal continuity is not satisfied. Overall, we conclude that Snow's definition outlines the three dimensions that *must stabilize over time* to classify an empirical phenomenon as collective identity.

To address movement sustainability over time, IS research has also evaluated related concepts such as "solidarity" (Stewart & Schultze, 2019) and "connective emotions" (George & Leidner, 2019). These concepts, however, do not address the process by which a collective actually establishes a shared "we-ness" as a distinct social entity. Thus, while such emotionally focused concepts may be essential for understanding sustained engagement within a certain collective (e.g., Stewart & Schultze, 2019), they are insufficient for comprehending the construction process of a distinct composite actor in the first place. We, therefore, prioritized the concept of collective identity.[4]

In the following, we provide a brief overview of earlier attempts in the IS and social movement literature to empirically examine collective identity online.

[4] Other prominent approaches originating from social movement studies, such as resource mobilization and political process theory, accentuate material and socio-structural factors that influence mobilization (see also, e.g., Melucci, 1995, p. 42; Polletta & Jasper, 2001, pp. 283-285). Our focus, however, is on the formation of collectives. Therefore, we accentuate the "sociocultural … level" (Gamson, 1991, p. 41) of analysis and focus exclusively on the concept of collective identity. That said, there are several other concepts closely related to collective identity, particularly within social movement research, such as personal identity, social identity, and group consciousness (for a differentiation between these concepts, see, e.g., Gamson, 1991, p. 45; Hunt & Benford, 2004, p. 439; Klandermans, 2014, p. 3; Rucht, 2023a, p. 22).

2.2 Empirically Determining Collective Identity Online

The evaluation of collective identity formation online has gained prominence in IS research and related fields in recent years (e.g., Bhattacharya et al., 2024; Jarvenpaa & Selander, 2023; Priante, 2021; Priante et al., 2018; Ray & Tarafdar, 2017; Stewart & Schultze, 2017, 2019), highlighting the emergence of new forms of identity formation, collective engagement, and collectivity (Bakardjieva, 2015; Chamakiotis et al., 2021; George & Leidner, 2019; Gerbaudo, 2014; Gerbaudo & Treré, 2015; Langer et al., 2020; Stewart & Schultze, 2019). Based on their empirical approach, these studies can be divided into two main strands of research: one that reconstructs movements' collective identity based on cultural references shared by platform users (e.g., Fayard & DeSanctis, 2010; Khazraee & Novak, 2018; Maghrabi & Salam, 2011; Stewart & Schultze, 2017) and another that focuses solely on interaction patterns or social networks among platform users (e.g., Ackland & O'Neil, 2011; Barandiaran et al., 2020; Monterde et al., 2015; Priante, 2021). Table 1 compares both approaches in terms of their conceptual understanding of collective identity and their empirical measurement of the concept (see Appendix A for a detailed literature overview).

The foundational conceptual reflections that continue to shape our understanding of collective identity formation (see especially Melucci, 1995, 1996), however, emphasize the need to consider the interrelatedness of culture and structure and their reciprocal influence in this process. Accordingly, "analysis cannot simply identify action with that which the actors report about themselves, without taking into account the system of relationships in which goals, values, frames, and discourses are produced" (Melucci, 1996, p. 15). The existing literature on collective identity online today disregards this essential duality. Nevertheless, both the IS literature (Priante, 2021; Priante et al., 2018, p. 2663; Syed & Silva, 2023, p. 269) and social movement studies (Barandiaran et al., 2020, p. 19; Monterde et al., 2015, p. 944) have already emphasized the critical need to establish a connection between these two levels of analysis in the study of digital activism and collective identity. Developing such a conceptualization allows for a more comprehensive understanding of an online collective's formation, organization, and continuance, and contributes to the resolution of the contradiction between connective action and traditional social movement theory (see Table 1).

While there are some exceptions, with studies that consider both actors' interaction networks and culture (Bhattacharya et al., 2024; Coretti & Pica, 2015; see Table 1), they fall short in accounting for the co-constitution of these two levels both in their conceptualization and empirical implementation of collective identity, thus hindering a theoretically guided interpretation of their findings. We know of no study that has addressed the dual nature of collective identity by analyzing the reciprocal relationship between its actor and cultural level in the context of a comprehensive empirical framework. We addressed this research gap by capturing the duality inherent in collective identity to thoroughly determine this concept.

3 Dual Conceptualization of Collective Identity

Comprehensively capturing the dual nature of collective identity requires an integrated framework that acknowledges the interplay between the cultural level and the level of actors' interactions. To derive such a conceptualization, we drew on White's relational approach and integrated his perspective with the three dimensions of collective identity outlined in Section 2.1.

White's relational approach is particularly well-suited for this purpose for three reasons. First, it represents the foundational work on the duality of social and cultural structures in the context of identity (e.g., Emirbayer & Goodwin, 1994, pp. 1436-1437). His approach is based on the fundamental assumption characterizing duality, which posits that "social actors and cultural elements alike are ontologically defined by their relative structural position *both* within *and* across the two orders, the social and the cultural" (Basov et al., 2020, p. 1, emphasis in original). Second, White adopts a processual ontology, viewing social life as "essentially fluid and indeterminate" (Azarian, 2005, p. 58). This approach also allows for conceptually incorporating the process of identity formation from a dual perspective. Third, White is a relational sociologist and, as such, posits that the cultural structures and actor constellations are "ordered in relational systems" (Mohr & White, 2008, p. 489), allowing us to map these relational systems as socio-semantic networks—a methodical innovation corresponding to the needs of a dual perspective by linking the cultural level with the level of actors' interaction while accounting for their co-constitution (e.g., Basov et al., 2020, pp. 1-2).

We focused on White's understanding of collective identity as a *style*, which we address in the following section.

Table 1. Existing Culture-Based and Network-Based Approaches of Collective Identity in Comparison to Style

	Culture-based approaches	Networked-based approaches	Approaches considering culture and networks	Dual approach ("Style")
Conceptual understanding of collective identity	Collective identity manifests within the cultural references shared by platform users. **Limitation:** Neglects the role of interactions among platform users and their mutual co-constitution with the cultural references.	Collective identity manifests within the interaction among platform users. **Limitation:** Neglects the role of the cultural references shared by platform users and their mutual co-constitution with interactions.	Collective identity manifests within the interactions among platform users and their shared content. **Limitation:** Neglects the mutual co-constitution of the actor and cultural levels by analyzing them separately.	Collective identity manifests within the cultural- and network-based patterns that mutually co-constitute one another, shaped by the social interactions among platform users. **Solution:** An inherently dual approach grounded in the conceptual understanding that the actor and cultural levels are inseparably linked and mutually co-constitutive.
Empirical measurement of collective identity	Analysis of the cultural references shared by platform users such as through qualitative content analysis, discourse analysis, and the evaluation of user avatars and images. **Limitation:** Neglects underlying network structures and relational dynamics from which cultural references originate.	Analysis of network patterns, such as network density, centrality, significant interactions between users, modularity, *k*-core decompositions, and component structures. **Limitation:** Neglects cultural references that provide meaning to the detected network structures.	Analysis of network patterns, such as network modularity and tie strengths, and cultural references, such as shared topics and comments. **Limitation:** Neglects the mutual co-constitution of both dimensions.	Socio-semantic network analysis of social and cultural patterns, and a qualitative content analysis of the shared content structured according to users' interaction networks. **Solution:** Combined analysis of social and cultural patterns that co-constitute each other, grounded in a theoretical background that integrates both dimensions.

3.1 Collective Identity as a Style

White's concept of identity draws on this notion of processuality and duality as well as on insights from both ethnomethodology and sociolinguistics. Within the initially fluid and indeterminate "turbulent context" (White, 2008, p. 1) that actors face as they navigate their lives, identity serves as a reference point to which information can be attributed. White views identity abstractly as "any source of action, or any entity to which observers can attribute meaning not explicable from biophysical regularities" (White, 2008, p. 2), providing participants in interaction with "control amid contingencies" (White, 2008, p. 1). In essence, White conceptualizes identity as a cognitive framework that enables actors to interpret their actions and social environment, thereby offering a sense of order and predictability within an otherwise contingent, unpredictable world.

White specifies this general notion by distinguishing five different senses of identity that vary in their degree of complexity and persistence (for a detailed overview, see, e.g., White, 2008, pp. 10-18). White's *first sense* of identity involves an actor's performative effort to establish "footing" (Goffman, 1979) among other identities in a particular interaction situation (related to "self-presentation"; see Goffman, 1969). The *second sense* of identity emerges when individual actors establish their "role" (White et al., 2007, p. 186) within a particular social group (related to "role identity"; see McCall & Simmons, 1978). The *third sense* of identity incorporates temporality, reflecting an actor's trajectory from one social context to another over time (White, 2008, p. 17). In this sense, identity emerges from the intersection of multiple "social circles" (Simmel, 1955) as individuals navigate and interweave various social affiliations in their daily experience, forming a "networked identity" (Rainie & Wellman, 2012). White's *fourth sense* addresses personal identity, or "what a person perceives as his or her self" (White, 2008, p. 17), which is understood through a "biographical" introspection on the third sense (related to "reflexive identity"; see Giddens, 1991). Finally, his *fifth sense* of identity addresses "styles" (Fontdevila, 2018, p. 255), defined as stable "network-based patterns of social interaction" (Mische, 2008, p. 38) that consistently involve specific cultural codes and expressions (Godart & White, 2010, p. 578). This may encompass a range of recurring cultural elements, such as language, communication, clothing, and other means of expression.

Style in White's conception is, thus, inherently dual, encompassing both cultural patterns and social, i.e., network-based, patterns manifesting in social interaction. The term "pattern" in this context emphasizes that a style becomes established only through "continued reenactments" (White, 2008, p. 141), indicating temporal continuity. As a "meta-sense of identity" (Fontdevila, 2018, p. 255), style manifests in all social interactions linked to the first four senses of identity, as it synthesizes characteristics that emerge from these senses into a more abstract, emergent pattern. Consequently, style contributes to the establishment of social order and coherence.

In this context, White (2008) introduces the concept of "free scalability," emphasizing that styles can apply to a wide range of social entities. They are not confined to individual actors but extend across different levels of social reality, from smaller units such as personal conversations to broader, more encompassing structures such as communities (p. 134; see also White, 2008, p. 113). Thus, any social entity—ranging from individual interaction patterns in pub conversations, dance performances, and barbecue gatherings to larger social units such as social movements, institutional fields, and organizations—must develop a distinctive style to establish recognizability and distinguish itself from others.

Building on his dual notion of social life, White posits that identities settle in *netdoms* (e.g., White, 2008, pp. 7-9), a neologism combining *network*—a "social space … with broader temporal relational extension … than any smaller abstraction of momentary dyads" (Fontdevila, 2018, p. 232)—and a concurrent *domain*—a specific set of cultural structures characterizing the netdom's "phenomenological contents and horizons" (Fontdevila, 2018, p. 232; Mische & White, 1998, p. 702). A netdom thus marks a specific social context of actors' day-to-day experiences (White, 2008, pp. 2, 7). White illustrates the concept using an internet forum, where an actor interacts with a set of other users on a particular topic, switching netdoms by logging out of one forum and joining another (White, 2008, p. 2). As such, netdoms provide the social context for the dual manifestation of style.

To sum up, by distinguishing five senses of identity, White clarifies what we actually observe when identifying actors based on their empirical (digital) traces: a socially and culturally patterned style. Since style is freely scalable, it also provides a suitable framework for determining collective identity. Building on White's notion of style—manifesting within netdoms—we developed a dual conceptualization of collective identity. Our approach assumes that recurring cultural and social patterns, when consistently observed, indicate collective identity. In other words, as style emerges from interactions, it serves as an empirically detectable social signature of collective identity (Fontdevila, 2018, p. 241; Godart & White, 2010, p. 578; Mische, 2008, p. 38; White, 2008, p. 119).

However, given the abstract level of White's considerations, it remains necessary to identify the recurring social and cultural patterns within netdoms that distinctively characterize the style of collectives. Furthermore, social movements, as a more specific type of collective, exhibit a particular manifestation of style, which we describe in the following.

3.2 The Style of Social Movements

Social movements are a central actor in social conflicts in modernity (della Porta & Diani, 2006, pp. 20-21; Rucht, 2023a, pp. 18-19). Generally, movements emerge when awareness of shared grievances grows (Klandermans, 2014, p. 4) and those grievances become *politicized*, that is, they evolve into focal points of conflict that are socially relevant and involve explicitly named adversaries (Simon & Klandermans, 2001, p. 324). Through White's lens, movements and their style thus solidify within *politicized netdoms*, as they provide a fertile ground for protest and the formation of that distinctive style.

The *social patterns* of protest movements typically take the form of a decentralized network of movement organizations (SMOs)—which coordinate collective action—and a broader base of adherents who provide support without necessarily engaging in direct activism (Edwards & McCarthy, 2004; see also: Bennett & Segerberg, 2013; Marwell et al., 1988). Together, these two elements shape the characteristic relationship between leaders and adopters, which plays a critical role in the diffusion "of collective action and the spread of protest symbols and tactics" (Strang & Soule, 1998, p. 268; see also: Soule, 2004). Emirbayer and Goodwin (1994) outline two fundamental and complementary approaches that network analysts employ to understand these social patterns (for an application in the context of social movement research, see Diani, 2002, pp. 189-194). The first one focuses on a network's overall structure to derive conclusions regarding an actor's *social cohesion* (Diani, 2002, pp. 190-191; Emirbayer & Goodwin, 1994, p. 1419). The second one, "positional analysis" (Emirbayer & Goodwin, 1994, p. 1422), examines the structure of *social positions* within the network (Diani, 2002, pp. 191-194; Emirbayer & Goodwin, 1994, p. 1422), thereby providing insights into how actors organize and interact.

Building on Section 2.1, we assume that a sense of *collective agency*, *sameness*, and *distinctiveness* are discrete dimensions of the collective identity of any composite actor. Examining these three dimensions derived from Snow (2001) through White's style perspective—and thus integrating the two approaches—suggests that a collective's sense of agency, sameness, and distinctiveness should give rise to specific *cultural patterns* that emerge from the interactions among the set of actors forming the collective. Cultural patterns related to the collective's sense of agency arise, for example, when a specific repertoire of collective action is consistently referenced. Cultural patterns regarding sameness arise, for instance, when certain characteristics attributed to the collective are expressed repeatedly. Patterns of distinctiveness emerge, for example, when characteristics attributed to the collective's social environment—its "others"—are consistently expressed. For social movements, we can further specify these expected characteristics within the three dimensions based on the existing literature, thereby distinguishing a social movement's style from styles exhibited by other composite actors.

To begin with, the core cultural feature of a social movement's style—in contrast to other composite actors—lies in the specific communicative form of protest. As Niklas Luhmann (1997/2013) pointed out, the unity of "a protest movement arises from its form, from protest. … It expresses itself from a sense of responsibility *for* society but *against* it" (p. 157, emphasis in original). Social movements, therefore, typically adopt a perspective of criticizing society as if from the outside while being inherently a part of it. As a consequence, the distinctive communication style of social movements at the cultural level inevitably divides society into two opposing parts: the protesters and the targets of their protest. Their sameness is rooted in their self-perceived position at the periphery of society, enduring grievances, while their distinctiveness stems from their protest against those they perceive as positioned at the societal center—political authorities, big corporations, mass media, and the like (Rucht, 2023b, p. 103)—that is, individuals or groups that movement adherents believe are capable but unwilling to implement the social changes the movement advocates. A movement's agency, in turn, focuses on its supporters' capacity to influence the societal center through a particular "repertoire of collective action" (e.g., Tilly, 1978, p. 151). We consider patterns arising from recurring references concerning these three dimensions as indicative of the style of social movements at the cultural level.

Overall, since style manifests in discernible patterns produced through social interaction, it can be extracted from empirical traces of actors' past interactions. This insight is the key added value of conceptualizing collective identity based on style. This can be applied to analyze the patterns left by social movement adherents both *online* and *offline*. In the following section, we outline the online-specific expression of social movements' style based on digital traces left by users of social media platforms, concluding with our final conceptualization of style online.

3.3 Social Movements' Style Online

When adapting the concept of style specifically for the online context, it is important to consider that a style's manifestation online depends on the characteristics of online environments facilitating particular user behavior (e.g., Johnson et al., 2015; Leong et al., 2020), which results in specific social and cultural patterns.

In online environments, social patterns shaping a movement's style emerge through the unique "potential for action" (Leonardi & Vaast, 2017, p. 152) that platforms offer their users, such as friending, following, and joining online groups (e.g., Karahanna et al., 2018). Due to low entry and exit barriers and "the absence of either formal governance or managerial control" (Massa & O'Mahony, 2021, p. 1041), the resultant actor networks tend to exhibit "fluid structures and shifting membership" (Johnson et al., 2015, p. 166; see also, e.g., Faraj et al., 2011, 2015). Organizational coordination by SMOs is rendered unnecessary (see Section 1); instead, self-organization and relationship formation are sustained solely through users' voluntary contribution efforts (Johnson et al., 2015, p. 166). Hence, we anticipate that the social patterns of a movement's style will reflect this fluidity when social cohesion is assessed.

This fluidity extends to users' social positions, which in online contexts are neither predetermined nor consistently tied to specific users over time (Faraj et al., 2011, p. 1231; Johnson et al., 2015, pp. 166-167; Majchrzak et al., 2013; Vaast et al., 2017, p. 1183). Rather, positions are understood as emergent phenomena arising from interactional patterns, with users transitioning between positions (Faraj et al., 2011, p. 1231; Majchrzak et al., 2013).

In this context, the emergence of "leaders" is particularly significant for shaping a collective's style. Existing literature on the role of leaders in self-organized online collectives (e.g., Faraj et al., 2015; Harysi et al., 2019; Johnson et al., 2015; Lee et al., 2018) describes leadership online as "a local designation … emerging from observable interactions" (Johnson et al., 2015, p. 166). Leadership status arises from users' past contributions and the recognition received from others based on these contributions, rather than through formal, explicit acknowledgement (Johnson et al., 2015, p. 168; also, see Harysi et al., 2019; Faraj et al., 2015; van Haperen & Uitermark, 2023). Consequently, online leaders may be occasionally replaced by others who more accurately anticipate and align with the collective's preferences (Faraj et al., 2011; Harysi et al., 2019; Johnson et al., 2015). This form of leadership "will not generate domination, if by that we mean actors' capacity to impose sanctions over others … but rather varying degrees of influence" (Diani, 2003, p. 106; see also Johnson et al., 2015, p. 168). Emergent leaders, though, still fulfill classic leadership functions for social movements, effectively becoming their temporary de facto coordinators, decision makers, and spokespersons (della Porta & Diani, 2006, p. 143).

Leadership is crucial for assessing a movement's style for two key reasons. First, despite its transience, it shapes the network's topology. It creates social patterns that contribute to a collective's style, with leadership ranging from being evenly distributed across the network to highly concentrated in the hands of a few (e.g., Bennett & Segerberg, 2013, p. 152). Second, in light of a dual approach, users who attain leadership also play a pivotal role in defining a collective's style at the cultural level through the content they produce. Their particular influence is driven by two mechanisms. On the one hand, platforms amplify their influence through algorithmic curation, increasing the likelihood that prominent users and their content gain additional visibility (e.g., Ciampaglia et al., 2018). As a result, leaders typically benefit from a "cumulative advantage" (Merton, 1988), where initial recognition compounds over time, further solidifying their leadership position. This reinforcement can be explained through the structuring mechanism of preferential attachment (e.g., Barabási, 2012; Rainie & Wellman, 2012, pp. 48-49). On the other hand, these users shape the cultural level by setting precedents for others to follow. By observing present leaders, others may "learn the cumulative past preferences of existing network participants" (Johnson et al., 2014, p. 798; see also, e.g., Vaast et al., 2013, p. 1087) and emulate the cultural patterns found in the content of influential users (Vaast et al., 2013, p. 1082), thereby stabilizing and perpetuating the collective's style overall via "mimetic isomorphism" (DiMaggio & Powell, 1983). Similar dynamics have been observed in the context of organizational identity formation, where scholars have demonstrated that leaders play a crucial role in expressing and asserting key identity claims, providing direction in the process of identity formation (e.g., Gioia et al., 2013, p. 182; Massa & O'Mahony, 2021, p. 1058; Wry et al., 2011, pp. 456-457).

A movement style's manifestation at the cultural level—that is, the expression of collective agency, sameness, and distinctiveness over time—is also both restricted and enabled by the "potential for action" (Leonardi & Vaast, 2017, p. 152) that platforms provide to their users, such as posting texts, images, video clips, and other forms of expression (e.g., Karahanna et al., 2018). This potential for action not only facilitates the production of content that reflects conceptions of agency, sameness, or distinctiveness but also enables new internet-based forms of action to express collective agency, such as digital petitioning and email bombing (Selander & Jarvenpaa, 2016, p. 332; Theocharis et al., 2015, p. 204; Van Laer & Van Aelst, 2010, pp. 1148-1151). As these forms of expression leave digital traces, recurring patterns related to collective agency, sameness, and distinctiveness can be extracted based on actors' interactions online.

Building on the considerations above, we define *a movement's style online* as the *distinctive combination of recurring social and cultural patterns manifesting in actors' interactions occurring within politicized digital network domains (netdoms), which collectively demarcate the movement as a distinct social actor.* The recurring social patterns represent a movement's organizational structure based on the social networks emerging from the interactions between its supporters, especially regarding their connectedness (social cohesion) and the roles assumed by them (social positions). The cultural patterns represent a movement's creation of a common "we-ness" by establishing its capacity for action (collective agency) and the boundary work towards the societal center through the differentiation between "us" (sameness) versus "them" (distinctiveness), based on the perception that the societal center either hinders or promotes the social change the movement seeks to achieve or resist. Together, these patterns depict the *empirically detectable social signature* of the movement's collective identity online.

Figure 3 summarizes this conceptualization of style online, covering its manifestation within netdoms, its cultural and social patterns, along with their respective subdimensions, and its empirical indicators.

The ability to meaningfully interpret recurring patterns within actors' past interactions renders a conceptualization based on White suitable for examining collective identity online based on digital trace data, that is, records of past activities online (Howison et al., 2011, p. 769). In Sections 4.2.1 to 4.2.3, we operationalize the conceptual dimensions derived above by introducing specific empirical indicators to evaluate digital traces.

4 Methodology

As discussed in Section 3.1, style emerges within netdoms. To translate this theoretical idea into empirical indicators, we began our empirical analysis by reconstructing netdoms—specifically, by modeling dual network structures that accurately represent netdoms' social and cultural levels. We achieved this by constructing socio-semantic networks and then examining the emerging social and cultural patterns left by users over time within them, pertaining to the conceptualization of style derived above. This approach is particularly effective, as it inherently captures netdoms' and styles' duality (Basov et al., 2020, p. 1), a critical aspect overlooked in previous research (see Section 2.2).

To empirically illustrate our conceptualization, we adopted a single case study research design focused on the Querdenken movement on Twitter. Generally, a case study is defined as "*an intensive analysis of a single unit for the purpose of understanding a larger class of (similar) units*" (Gerring, 2004, p. 342, emphasis in original). Accordingly, the strengths of a case study lie not only in the descriptive depth possible with a specific case but also in its suitability for generating novel theoretical propositions that extend beyond purely idiographic analysis (e.g., Byrne, 2009; Eisenhardt, 1989; George & Bennett, 2005, p. 20; Gregor, 2006, p. 624). This involves identifying patterns and mechanisms that explain how and why the interrelationship among factors in a specific case produce a particular phenomenon (e.g., Byrne, 2009; Eisenhardt & Graebner, 2007, pp. 26-27; George & Bennett, 2005, p. 20; Gerring, 2004, p. 348; Gregor, 2006, p. 624), reflecting a nuanced view of causality as conjunctural and configurational rather than universal and linear (e.g., Byrne, 2009; Ragin, 1987, p. 25). Because case study-based theorizing is closely tied to its foundational empirical insights, the resultant theories are typically considered not only novel but also testable and empirically robust (Eisenhardt, 1989, p. 548). This explains why single case studies are widely used in the IS field, particularly in the context of digital activism (Leong et al., 2019; Miranda et al., 2016; Sæbø et al., 2020; Selander & Jarvenpaa, 2016; Syed & Silva, 2023; Tarafdar & Ray, 2021; Venkatesan et al., 2021).

Building on the strengths of this research design, the following case study not only illustrates the conceptualization outlined above but also allows for deriving general characteristics of the social and cultural patterns that explain the formation of social movements' style online. This is essential for addressing RQ2.

In the following, we introduce Querdenken, the data collection process, and the operationalization of style and its dimensions introduced above.

4.1 Case Introduction: Querdenken on Twitter

Querdenken emerged in Germany at the onset of the COVID-19 pandemic to mobilize protests against government public health directives such as lockdowns, mandatory face masks, and vaccination (Holzer et al., 2021). Beginning with protests in April 2020, the movement quickly developed an increasingly professionalized network of local chapters, organizing several large-scale demonstrations across Germany (Holzer et al., 2021, pp. 11-14). An online initiative directory lists, as of this writing, 47 local Querdenken chapters.[5] The movement was highly heterogeneous, drawing members from diverse social backgrounds (Grande et al., 2021, pp. 11-12; Koos, 2021b; Nachtwey et al., 2020) and spanning the political spectrum (Frei et al., 2021; Koos, 2021b). Adherents from various groupings, including right-wing extremists, QAnon followers, esotericists, and anti-vaxxers (e.g., Jarynowskyi et al., 2020, p. 525; Koos, 2021b, p. 67), united under the banner of Querdenken, using the protests against public health directives as a mobilizing infrastructure. Prominent politicians from the right-wing German party Alternative für Deutschland (Alternative for Germany, AfD), which includes a significant extremist element, also showed support for the movement (Weisskircher, 2024, pp. 165-167).

We selected Querdenken as a suitable case for our research for three reasons. First, Querdenken is an example of a contemporary movement that blended online engagement and offline protest events (e.g., Castells, 2015, pp. 249-250; Earl, 2018, p. 300; Tufekci, 2017, p. 6), leveraging social media as "a regular part of the activist tool kit" (Caren et al., 2020, p. 456) to mobilize supporters, amplify their message, and sustain protest (e.g., Caren et al., 2020, p. 448). Querdenken's reliance on social media—partly due to curfew regulations (e.g., Jost & Dogruel, 2023, p. 1; Koos, 2021a, p. 82)—makes it a compelling case for analyzing a social movement's style online. Second, while pinpointing the emergence of a new social movement can be challenging because its origin is often "fuzzy and contested" (Blee, 2013, p. 661), Querdenken's emergence clearly coincided with the outbreak of the COVID-19 pandemic, making it straightforward to determine when the movement was founded and thus offering a rare opportunity to study the formation of a movement's style from its inception. Third, the movement's internal heterogeneity raises questions about whether the movement was able to develop a collective identity and how this putative identity manifested (Koos, 2021b, pp. 82-84), making Querdenken an intriguing case for analyzing style formation.

To analyze Querdenken's style online, we relied on the social media platform Twitter. Through Twitter, "digital content can be delivered to (and redistributed by) … users in seconds" (Aladwani, 2015, p. 15), making it a highly popular arena for discussing political topics (e.g., Jungherr, 2015; Theocharis et al., 2015, p. 205). Consequently, it has been used multiple times in IS research to analyze social movements (see, e.g., Miranda et al., 2016; Syed & Silva, 2023; Tarafdar & Ray, 2021; Venkatesan et al., 2021). For Querdenken, Twitter served as a vital communication channel, especially during the early months of pandemic-skeptic protests, when it played a crucial role in facilitating political discussions about COVID-19 (e.g., Winter et al., 2021). Given Twitter's significance for Querdenken, this platform serves as a suitable field of study for analyzing whether a movement can develop a cohesive style online.

[5] https://app.querdenken-711.de/initiatives-directory (last access: August 1, 2024).

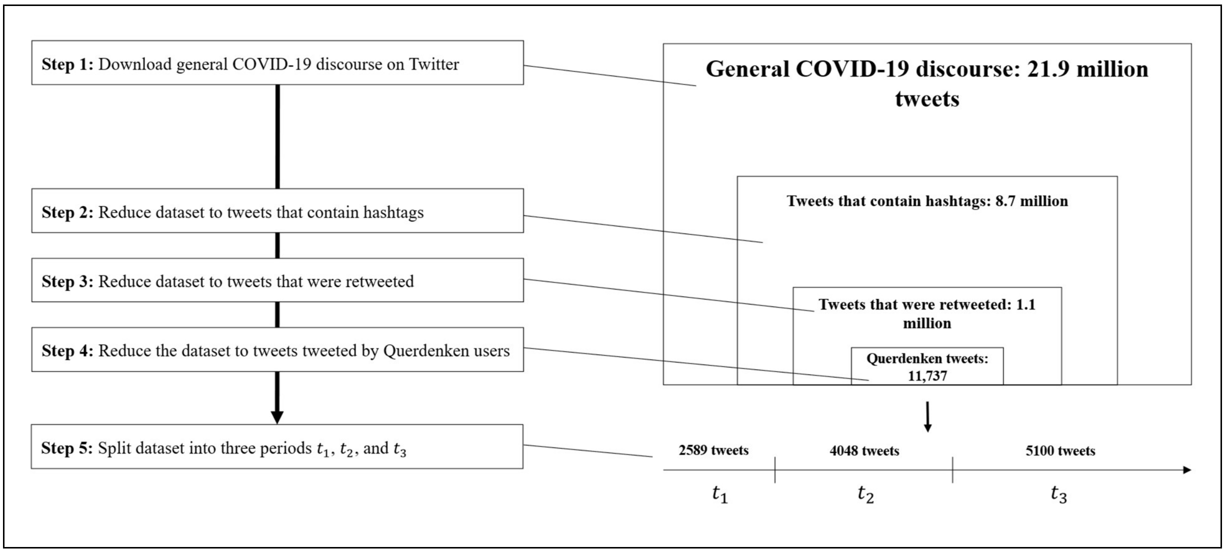


Figure 1. Data Collection, Reduction, and Temporal Segmentation

4.2 Data Collection and Analysis

To receive the data for constructing netdoms and detecting the style of Querdenken on Twitter, we applied the five steps of data collection, reduction, and temporal segmentation in Figure 1.

Using the academic Twitter API, we collected all German-language tweets containing a neutral set of keywords related to COVID-19,[6] along with their retweets, that were published between January 1, 2020, and December 21, 2020 (Step 1). This dataset captured the general COVID-19 discourse on Twitter.

Next, we reduced the dataset to tweets containing at least one hashtag (Step 2), since hashtags provide central thematic markers (Bennett & Segerberg, 2013, pp. 90-91; Jungherr, 2015, p. 69) essential for deducing the domain side of the netdom and the evaluation of cultural patterns. Misspelled or differently written yet synonymous hashtags were identified using Levenshtein distance and were harmonized manually when necessary.

We then reduced the dataset to tweets that were retweeted at least once but not commented on—meaning users simply redistributed another user's content without adding further comments or altering the original tweet (Step 3). Hereafter, we refer to this type of tweet as a "direct retweet." Direct retweets have a special status (Marsili, 2021, pp. 10461-10462): "Through their [direct] retweeting, retweeters align with the original stances and broadcast them to their followers" (Gruber, 2017, p. 7). Thus, ties formed by direct retweets arise from recognition of commonality, as the tweet is retweeted in its original form without (critical) comments on its content. As will be discussed later, direct retweets formed the basis of both the netdoms' network side and the social patterns indicating style.

Next, we narrowed the dataset to the tweets and retweets of Querdenken-related users (Step 4). To achieve this, we identified local Querdenken chapters with active Twitter accounts based on the movement's official website as of May 31, 2021, resulting in 39 accounts. To ensure a more inclusive representation of Querdenken-related users, we examined their followees—the accounts followed by the local chapter accounts—under the assumption that followees were deemed relevant by the chapters. After qualitatively categorizing the followees as Querdenken-related based on their tweets' content (see Appendix B1 for details), we obtained a total of 272 unique users for our analysis, referred to as "Querdenken users." Finally, we filtered the dataset to include only tweets posted by these 272 Querdenken users, resulting in a final dataset of 11,737 unique tweets and their respective retweets—the foundation of the following analysis.

To determine whether a style solidified within the Querdenken netdom over time, we divided the dataset into three timeframes: t_1, t_2, and t_3 (Step 5). t_1 covers the initial pandemic period in Germany from January 1, 2020, to April 17, 2020, ending the day before the

[6] Original search query: "(SARS-CoV-2 OR coronavirus OR nCoV OR corona OR covid OR covid19 OR lockdown OR pandemie OR maskenpflicht OR neuinfektionen OR coronakrise OR impfstoff OR sarscov2 OR quarantäne OR mutation) - lang:de"; Translated search query: "(SARS-CoV-2 OR coronavirus OR nCoV OR corona OR covid OR covid19 OR lockdown OR pandemic OR obligatory masks OR new infections OR corona crisis OR vaccine OR sarscov2 OR quarantine OR mutation)."

first Querdenken protest event took place (Hippert & Saul, 2021). t_2 includes the period from April 18, 2020—the date of the first Querdenken protest event (Hippert & Saul, 2021)—to August 31, 2020, which marks a period of significant mobilization success for the movement (Holzer et al., 2021, p. 13). t_3 begins on September 1, 2020, just after the end of this steady growth of protests (Holzer et al., 2021, p. 13), extends through the final stage of vaccine development and corresponding public discussion, and concludes on December 21, 2020, with the authorization of Comirnaty, the first official COVID-19 vaccine in Germany. The three final datasets, divided according to t_1, t_2, and t_3, enabled us to construct three netdoms in which Querdenken's style evolved over time.

In the following, we outline the process of constructing netdoms as well as identifying the social and cultural patterns that represent the movement's style. As mentioned earlier, we employed socio-semantic network analysis, using the final dataset outlined above. For this approach, we drew on the work of Radicioni et al. (2021a, 2021b), who used socio-semantic network analysis to evaluate the Italian social media landscape prior to the 2018 federal elections. Figure 2 visually outlines the different steps of our socio-semantic network construction and analysis.

4.2.1 Construction and Analysis of Netdoms

To model netdoms, we merged social and cultural information in our datasets by constructing three bipartite user-by-hashtag networks for each of the three timeframes t. Each network for a given timeframe t consisted of two distinct sets of nodes: $Q^t = \{1, \ldots, m\}$ representing Querdenken users (social information), and $H^t = \{1, \ldots, n\}$ representing hashtags (cultural information). We modeled relations between, but not within, both sets (bipartite networks). We then translated each of these networks into a matrix A^t, with the dimensions $m \times n$, where m corresponds to the number of Querdenken users within the timeframe t and n relates to the total number of hashtags posted within the same timeframe. In this matrix, the matrix element a^t_{qh} equals 1 if Querdenken user $q \in Q^t$ tweeted the hashtag $h \in H^t$ at least once during timeframe t. If Querdenken user q did not tweet hashtag h on any day during timeframe t, the matrix element a^t_{qh} equals 0. In sum, these bipartite networks depict how Querdenken users deployed central thematic markers, serving as the empirical manifestation of the Querdenken-specific netdom on Twitter, based on digital trace data.

To assess these netdoms, we analyzed their network and domain level in more detail. On the network level, we evaluated the basic network structure, which we present in Section 5.1. On the domain level, we examined the domain's central topics and its level of politicization.

Identifying the domains' characteristic topics based on qualitative coding required an inductive approach to category development (see Syed & Silva, 2023 for a similar approach). We developed a codebook based on multiple rounds of open coding, followed by consensus-building team meetings to review coding decisions (see Appendix B2 for details). To identify the most significant topics shaping the Querdenken-related domains, we focused on those that appeared above average in each timeframe.

To determine the domains' level of politicization, we derived a coding scheme (see Appendix B3 for details) to assess whether the content of a tweet was politicized, that is, if it addressed a subject of political conflict. Politicization is indicated by either the demarcation against a particular idea or collective shaping society at large, or the endorsement of or alignment with such an idea or collective (Simon & Klandermans, 2001, p. 324). Both forms of politicization are represented by a value of "1," indicating that the tweet took a clear political stance. If a tweet's content did not exhibit any signs of politicization, it was assigned a value of "0." To assess the domains' level of politicization, we analyzed frequency counts and mapped these across the three timeframes. The results of this descriptive analysis are detailed in Section 5.1. The tweet sample to which the coding scheme is applied is described in Section 4.2.3.

4.2.2 Analysis of Social Patterns

For the analysis of the social patterns of style, based on our conceptualization derived in Section 3.3, we concentrated on specific aspects of these overarching netdoms. This required constructing two types of networks: bipartite user-by-user networks and monopartite Querdenken user networks (see Figure 2), each tailored to a distinct analytical purpose. We detail the construction of these networks in the following section.

To construct the basis for receiving *significant* social patterns, we first modeled bipartite user-by-user networks for each timeframe t with the two node sets $Q^t = \{1, \ldots, m\}$ and $R^t = \{1, \ldots, n\}$. The matrices related to these networks can be expressed by B^t with the dimensions $m \times n$, where m corresponds to the total number of Querdenken users within t and n relates to the total number of retweeting users within the same timeframe. Querdenken user $q \in Q^t$ is linked with retweeting user $r \in R^t$ if retweeting user r retweeted Querdenken user q at least once during the respective timeframe, t. In this case, the element b^t_{qr} equals 1, otherwise 0. These bipartite user-by-user networks constituted the basis for receiving significant Querdenken user networks, which we describe in the following.

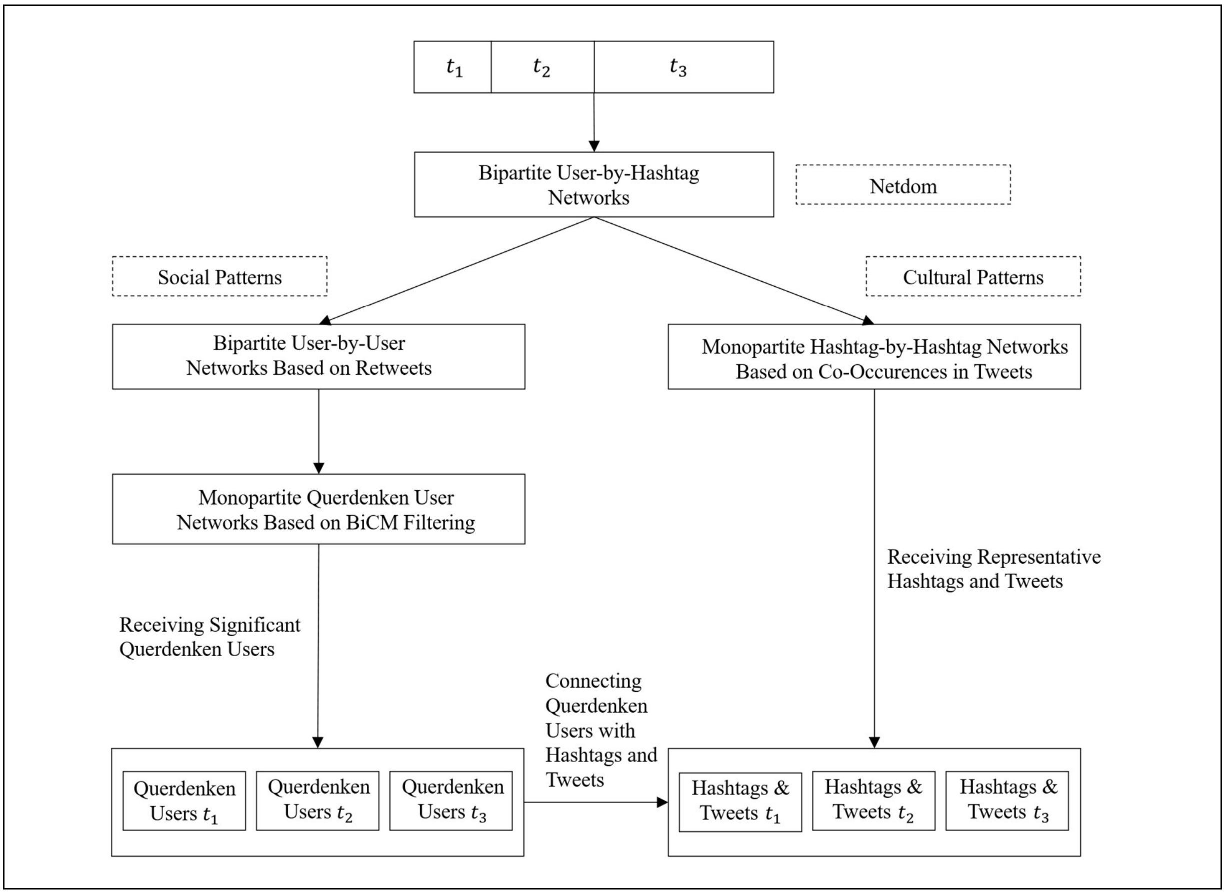


Figure 2. Overview of Socio-Semantic Network Construction and Analysis

To construct *significant* Querdenken user networks, we constructed monopartite projected networks of these bipartite user-by-user networks, referred to as Querdenken user networks (see Figure 2). One of the most common methods to construct monopartite networks from such a bipartite network is to link two nodes "if their number of common neighbors is *positive*" (Radicioni et al., 2021a, p. 3, emphasis in original). Thus, a link between two nodes is established if both share at least one common neighbor. However, entropy-based null models are a more sophisticated method for obtaining a statistically significant evaluation of the linking between two nodes of the same layer (Radicioni et al., 2021a; Saracco et al., 2017). Applying this additional filtering step guaranteed the construction of monopartite networks with *significant* edges among Querdenken users only (see Appendix B4 for details). This filtering step was conducted by applying a bipartite configuration model (BiCM), a commonly used entropy-based null model (Saracco et al., 2015, 2017). This filtering of the bipartite user-by-user networks resulted in monopartite Querdenken user networks that indicated which Querdenken users had a similar audience, that is, which users received recognition from the same sources.

These Querdenken user networks depict our final networks for evaluating the social patterns of Querdenken's style regarding *social cohesion* and for deriving the *social positions* of leaders via positional analysis.

For evaluating social cohesion, we used three different measurement techniques. First, we applied the community detection algorithm Louvain (Blondel et al., 2008) to our Querdenken user networks to test whether these networks exhibited structural signs of a unified actor. We chose this algorithm because it is a modularity-maximizing algorithm known for its scalability and convincing results (Lancichinetti & Fortunato, 2009).[7] Fewer communities, both at a given point in time and over time, would indicate that Querdenken was a relatively unified actor, whereas many communities would be a sign of fragmentation.

[7] We validated these findings by also running the Infomap algorithm, which yielded similar results. Also, we validated these findings by calculating a polarization score (see Appendix B5 for details).

Second, we evaluated the Querdenken user networks regarding their density to measure social cohesion. Density can be defined as "the relative fraction of (network) links that are present" (Jackson, 2010, p. 29). Thus, the density value of a network offers insights into the network's overall level of connectedness—measured in existing edges—compared to a fully connected network.

Third, we analyzed the Querdenken user networks regarding their transitive triplets using the average clustering coefficient (for more information, see Jackson, 2010), which measures the average likelihood of two nodes being connected within a network once they have a common neighbor, offering a more nuanced understanding of social cohesion. Thus, both measures, *density* on a macrolevel and *transitive triplets* on a microlevel, illustrated the network's cohesion at a given point in time, which is an expression of group unity. While larger and stabilizing values over time would suggest a more cohesive actor on the social dimension, we anticipated detecting a certain fluidity within network structures due to the unique characteristics of online environments mentioned earlier.

We used two different measures to derive users' social positions as a second subdimension of the social patterns of style, with a particular emphasis on leaders. First, we computed a "leadership score" l_{qt}. This score captures the concept of emergent leadership based on user interaction. It reflects the collective perception of prominence, measured by the proportion of retweeting users who retweeted a certain user. The leadership score l_{qt} for each identified Querdenken user q and timeframe t is expressed by the following formula:

$$l_{qt} = \frac{\sum_{r \in R^t} b_{qr}^t}{|R^t|}$$

The l_{qt} score quantifies the proportion of retweeting users within a specific timeframe ($\sum_{r \in R^t} b_{qr}^t$) who retweeted a distinct Querdenken user q, divided by the total number of unique retweeting users in that timeframe (R^t). This score ranges from 0 (indicating that no retweeting user in that timeframe retweeted a Querdenken user) to 1 (indicating that all retweeting users in that timeframe retweeted a Querdenken user). We chose this measure over simply counting the number of retweets per Querdenken user because it quantifies the engagement of the *entire* group of retweeting users within a specific timeframe. This approach reflects the collective consensus regarding who was considered prominent based on the engagement of the retweeting users. Moreover, we decided on the l_{qt} score over the widely used degree centrality because it focuses on interactions between retweeting users and Querdenken users rather than assessing the entire network, irrespective of the user's status. The emerging distribution of this score indicates which users and their tweets received broad support, thus playing a central role in defining the movement. If only a few users were retweeted by a broad base of retweeting users, while the majority received only scattered attention, it would indicate agreement among retweeting users regarding who was disseminating resonating content that defined the movement. Thus, we considered these selected users to be acknowledged leaders of the collective.

As a second measure, to assess the *stability* of users' social positions over time, we calculated the average positional difference based on the l_{qt} score for the subset of Querdenken users q who appeared in all three timeframes. To achieve this, we first ranked these users based on their l_{qt} score within each period. To assess whether users maintained their positions across all three timeframes and thus solidified their social position based on the retweeting behavior of the retweeting user, we calculated the absolute difference between their rankings in t_1 and t_2, t_1 and t_3, and t_2 and t_3. Based on these values, we computed the average positional difference for each user. A value closer to 0 indicates that a Querdenken user maintained their position over time, while higher values indicate greater average fluctuations in their position. The results of these analyses are detailed in Section 5.2.1.

4.2.3 Analysis of Cultural Patterns

For analyzing the solidification of style's cultural patterns, we constructed monopartite hashtag-by-hashtag networks for each timeframe by linking hashtags that co-occurred in at least one tweet published by any Querdenken user who remained in the Querdenken user network after the BiCM filtering (see Figure 2). Hence, these hashtag-by-hashtag networks are based solely on content from the most significant Querdenken users. In these networks, hashtags represent the nodes, the unweighted edges represent their co-occurrence in at least one tweet. These networks provided the basis for selecting tweets to extract the cultural patterns. Note that we also used this sample to evaluate the domain's characterizing topics and level of politicization (see Section 4.2.1). The final sample consisted of 960 tweets posted by 62 users (see Appendix B6 for details).

Based on this sample, we extracted the cultural patterns regarding collective agency, sameness, and distinctiveness by means of a qualitative content analysis of the tweets and respective hashtags, following Schreier (2012, 2014). Schreier recommends combining a deductive and data-driven strategy when deriving a coding frame (Schreier, 2012, p. 89). Given that our application context required both theoretically derived categories and exploratory analysis, this combined approach was ideal for this analysis.

As the first step in the qualitative content analysis, two coders identified all tweets featuring utterances pointing to collective agency, sameness, and distinctiveness—the "main categories" (Schreier, 2012, p. 59) of our coding frame. This step represents the deductive part of the qualitative analysis. The rules for this round of coding are provided in Appendix B7.[8] In the next step, through multiple rounds of open coding, three coders developed "subcategories" (Schreier, 2012, p. 60) to specify the content of the tweets within the framework of the three main categories. This refinement provided the foundation for examining the emergence of recurring cultural patterns over time. The codebook for the final round of coding is provided in Appendix B8.

Regarding collective agency, the subcategories examined types of actions users referred to as past or anticipated collective actions, underscoring Querdenken's capacity to influence the societal center. Hence, the development of this part of the coding scheme was purely data-driven. Regarding tweets that pointed to Querdenken's sameness and distinctiveness, we derived two types of subcategories: social groups and actors described as part of the collective or distinct from it, and characteristics attributed to Querdenken and its opponents. While the former was data-driven, we relied on deductive coding for the latter, as the attributed characteristics were unlikely to be random. We drew on the set of binary codes outlined by Alexander (2006, pp. 57-59), which define "the characteristics of actors, social relationships and institutions that are considered appropriate in a democratically functioning society" (Alexander & Smith, 1993, p. 161) as well as their corresponding antitheses (e.g., Alexander, 2001, 2006; Alexander & Smith, 1993; for a detailed description of these codes, see Appendix B8.3 and B8.5). To generate resonance within a democratic context, social movements—as well as other societal actors—typically label themselves and their adherents as proponents of the morally desirable characteristics, while their opponents are cast as embodying the antithesis, thereby preventing the outright realization of the morally desirable (Alexander, 2001, p. 163, 2006, p. 231; Alexander & Smith, 1993, p. 197).

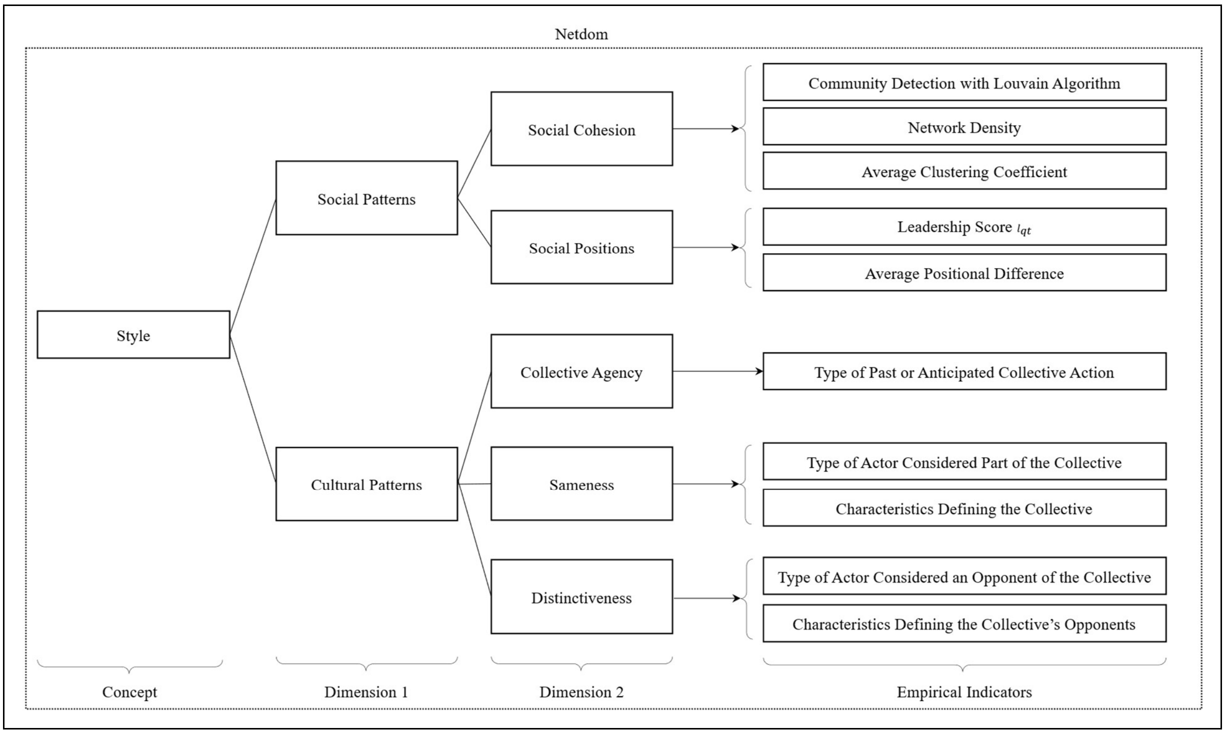


Figure 3. Empirical Indicators Operationalizing the Conceptual Dimensions of Style

[8] To ensure consistency, we assessed intercoder reliability for each dimension separately. For collective agency, we obtained a Fleiss's kappa of 0.864. For sameness of the collective's representatives, we obtained a Fleiss's kappa of 0.846. For the collective's distinctiveness, we obtained a Fleiss's kappa of 0.832.

To determine whether cultural patterns emerged in the portrayal of collective agency, sameness, and distinctiveness within the corpus—and thus identify an emerging style—we analyzed the frequency with which subcategories were assigned across different time periods. The results of this descriptive analysis are detailed in Section 5.2.2. Figure 3 presents an overview of all empirical indicators used to evaluate the conceptual dimensions of style.

5 Results

In this section, we present the findings of our study concerning the Querdenken-specific netdom and the social and cultural patterns that emerged within that form the movement's distinct style.

5.1 Analyzing the Netdom of Querdenken

To analyze the network side of the Querdenken-specific netdom on Twitter, we examined the network structures of the bipartite user-by-hashtag network with a focus on its actor layer. In t_1, we detected 61 users referencing 1409 hashtags, connected by 2233 edges. In t_2, we identified 84 users referencing 2030 hashtags, connected by 3712 edges. By t_3, this had increased to 99 users referencing 2577 hashtags, connected by 4702 edges. This progression highlights a steady increase in the number of users creating and maintaining the network of the Querdenken-specific netdom. A detailed analysis of the social patterns of these networks related to the movement's style is offered in Section 5.2.1.

The analysis of the netdoms' domain side shows that the domain was shaped most decisively by two topics in particular (see Table 2): *Criticism of political authorities* and *shared information on COVID-19* were consistently ranked among the decisive topics within the domain across all periods, confirming earlier findings that the political response to the COVID-19 pandemic was the movement's core concern (e.g., Holzer et al., 2021; Kern et al., 2023; Nachtwey et al., 2020). The topics *underestimation of COVID-19 threat* and *economic crisis* were prominent early on but declined in later periods, indicating a significant shift in the movement's concerns as the pandemic evolved—from initially critiquing political authorities for insufficient action in addressing the COVID-19 threat to later condemning them for perceived overreach and the imposition of excessive regulations. Accordingly, in t_2 and t_3, we found a noticeable increase in the prominence of topics such as *trivialization of COVID-19*, *speculation about conspiracy*, and *criticism of hygiene measures*, solidifying their importance alongside the ongoing criticism of political authorities and COVID-19 in general. Moreover, the growing significance of the topics *support of alternative science and media* and *criticism of science* underlines Querdenken's strong scientific skepticism (e.g., Kern et al., 2023; Kumkar, 2021; Pantenburg & Sepp, 2021). Overall, this trend highlights an increasing distrust directed at the societal center—particularly political authorities, the scientific community, and the public health system—that is typical for social conflicts in the public sphere (Rucht, 2023b, p. 75).

Table 2. Topic Persistence Over Time

Topic	t_1	t_2	t_3
Criticism of political authorities	●	●	●
Information on COVID-19	●	●	●
Underestimation of COVID-19 threat	●	○	○
Economic crisis	●	●	○
Criticism of hygiene measures	○	●	●
Criticism of science	○	●	●
Speculation about conspiracy	○	●	●
Support of alternative science and media	○	●	●
Trivialization of COVID-19	○	●	●
Criticism of COVID-19 statistics	○	●	○
Collective action	○	●	○
Criticism of health administration	○	○	●

Note: ● (filled circle) indicates topics with above-average frequency, based on tweets featuring these topics during a timeframe; ○ (empty circle) indicates topics without above-average frequency during a timeframe, measured in the same way. Topics consistently ranking below average in frequency are excluded from this table. Cells in dark gray indicate topics with above-average frequency across all three timeframes; cells in light gray indicate topics with above-average frequency in two timeframes.

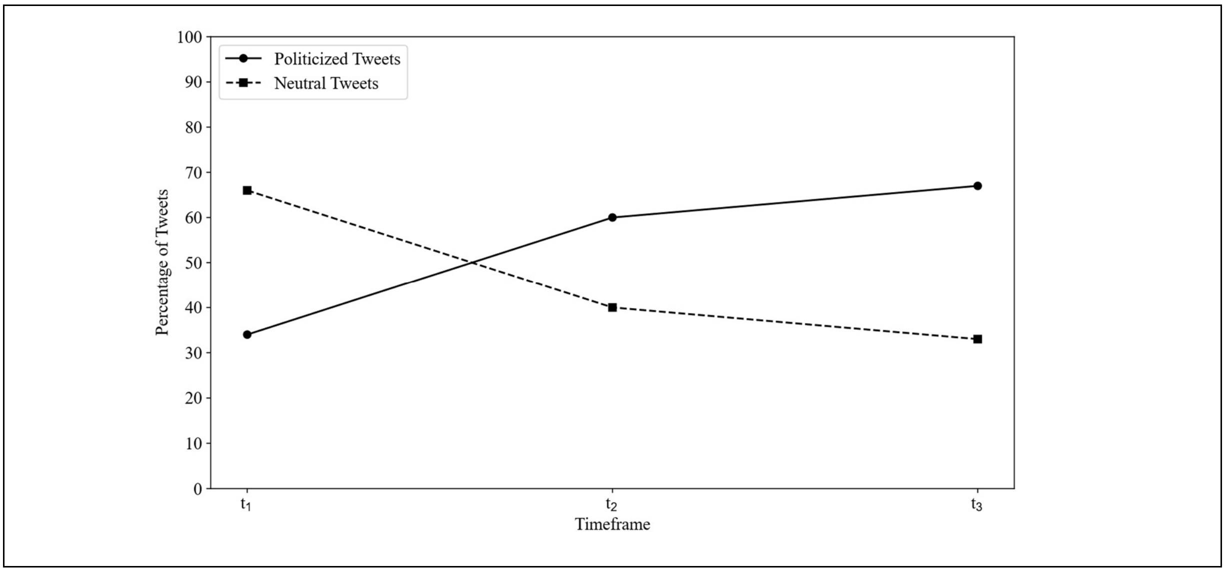


Figure 4. Politicization of the Domain Over Time

Table 2 summarizes the persistence and prominence of these topics across the three timeframes. Figure 4 complements this analysis by illustrating the percentage of tweets across the three time periods containing politicized content—that is, content addressing political conflict—as opposed to neutral content. Over time, there was a significant shift from neutral to politicized content, with politicized content surpassing neutral content by t_2 and reaching nearly 70% in t_3. This trend indicates the growing politicization of the domain.

Together, Table 2 and Figure 4 illustrate the domain's solidification and increasing politicization, which serve as the foundation for Querdenken's distinctive style. Building on this, we now examine the movement's style in detail.

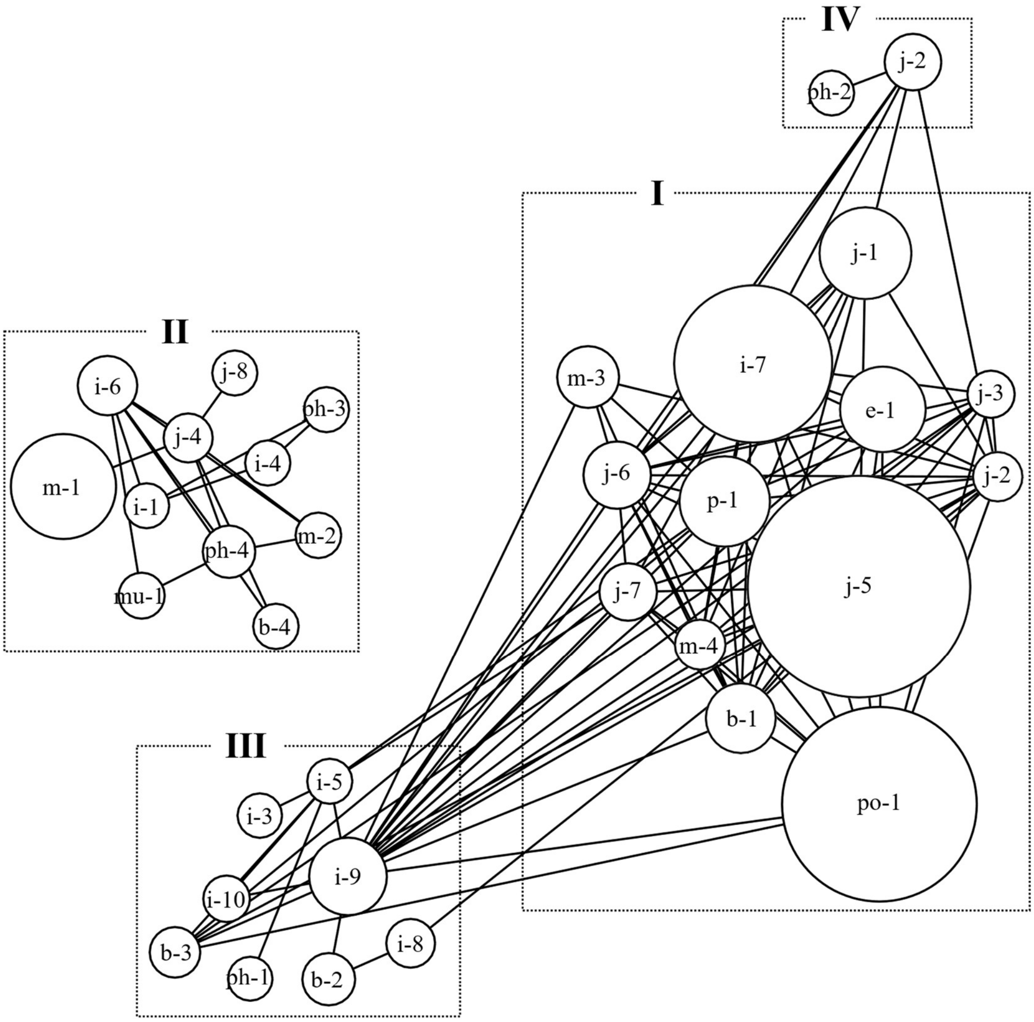
5.2 Detecting the Style of Querdenken

5.2.1 Social Patterns

Regarding Querdenken's *social cohesion*, the results of our community detection reveal that Querdenken was not a monolithic entity but dispersed into a decentralized "network of networks" (Bennett & Segerberg, 2013, p. 152)—a structure typical for social movements (e.g., Rucht, 2023b, p. 19). In t_1, the community detection divided the Querdenken user network into four communities; in t_2, five communities were detected; and in t_3, six communities were detected. This trend indicates increasing diversification over time, as visualized in Figures 5-7. To avoid violating any privacy restrictions, we anonymized individual Querdenken users with labels based on a qualitative coding of users' Twitter Bio—an expression of their fourth sense of identity in White's understanding.

To further substantiate the understanding of the detected communities, we conducted additional analyses assessing the stability of relationships between individuals over time within their respective communities (see Appendix C1 for details). The findings reveal constant changes in community composition and user networks change over time, consistent with prior research suggesting that social media foster fluid actor constellations (e.g., Johnson et al., 2015).

The impression of a "network of networks" is further supported by the inspection of the networks' density and transitive triplets measured through the average clustering coefficient. Table 3 presents these metrics for each timeframe, alongside the number of nodes and edges in the Querdenken-user networks for contextual information. Notably, t_1 exhibits the lowest density and transitivity, consistent with limited user activity—only 34 of 101 Querdenken users were active during this period. This reduced activity explains the lower network connectivity. An intriguing development emerges when comparing the networks from t_2 and t_3. Despite an increase in active Querdenken users from 65 in t_2 to 83 in t_3, the network density remains nearly unchanged (7.3% and 7.4%, respectively), because the number of edges shows only a slight increase from 367 edges in t_2 to 371 edges t_3. This finding indicates that Querdenken users in t_2 were more interconnected than in t_3, a notion supported by the higher average clustering coefficient in t_2 (35.1% vs. 33.1% in t_3). These findings imply that up until t_2, retweeting users retweeted fewer Querdenken users and reached a higher consensus regarding whose shared content to endorse, leading to a concentrated retweeting of posts and, consequently, relatively connected network structures among Querdenken users in t_2. By t_3, retweeting users had begun engaging with a more diverse range of Querdenken users, leading to a more dispersed and less cohesive Querdenken user network.

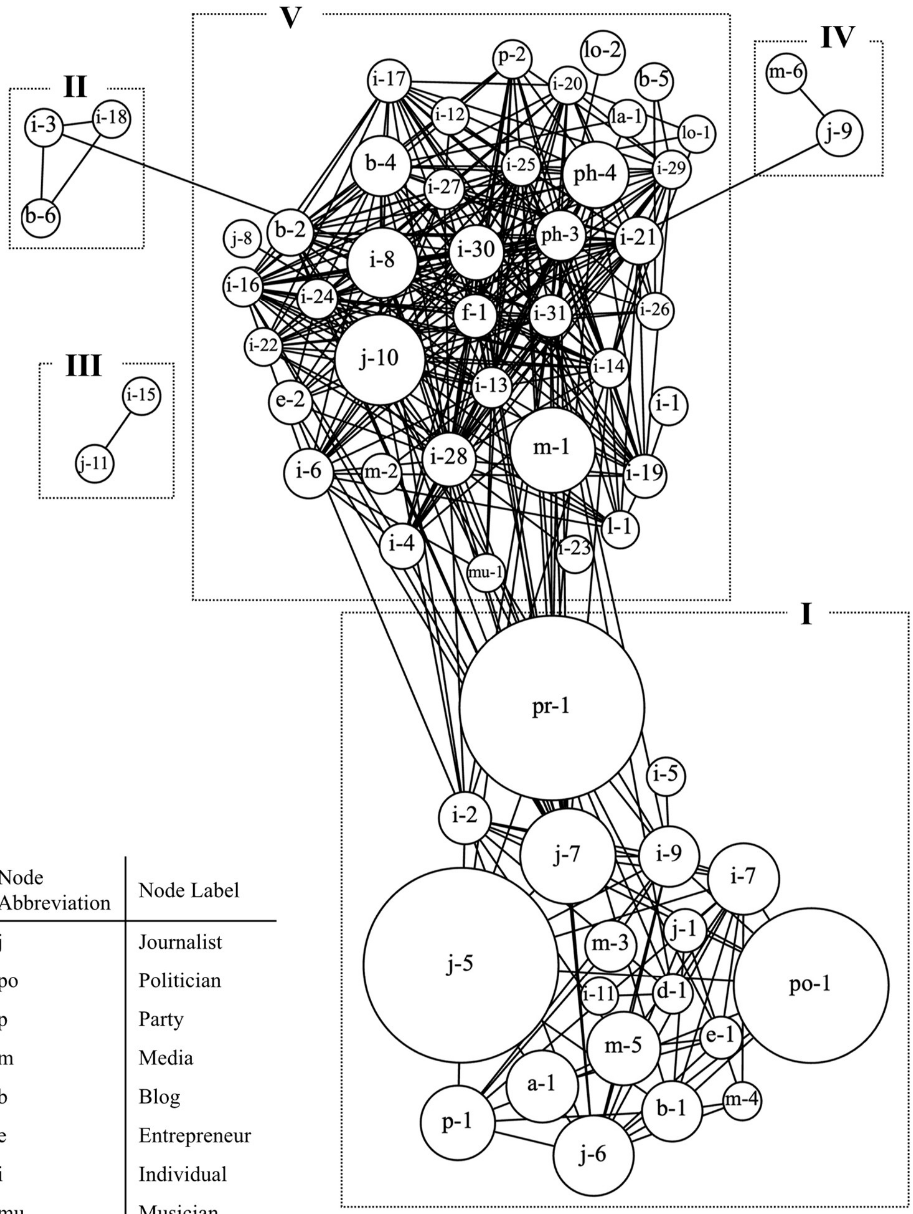


Node Abbreviation	Node Label
j	Journalist
po	Politician
p	Party
m	Media
b	Blog
e	Entrepreneur
i	Individual
mu	Musician
ph	Physician

Community Abbreviation	Community Role	Main Topic within Community
I	Emergent leader	Information on COVID-19
II	Challenger	Information on COVID-19
III	Supporter	Criticism of political authorities
IV	Supporter	Underestimation of COVID-19 threat

Note: Node size according to l_{qt} score; usernames anonymized; community role based on positional analysis; main topics as identified within netdom analysis based on tweet frequency.

Figure 5. Querdenken User Network in t_1

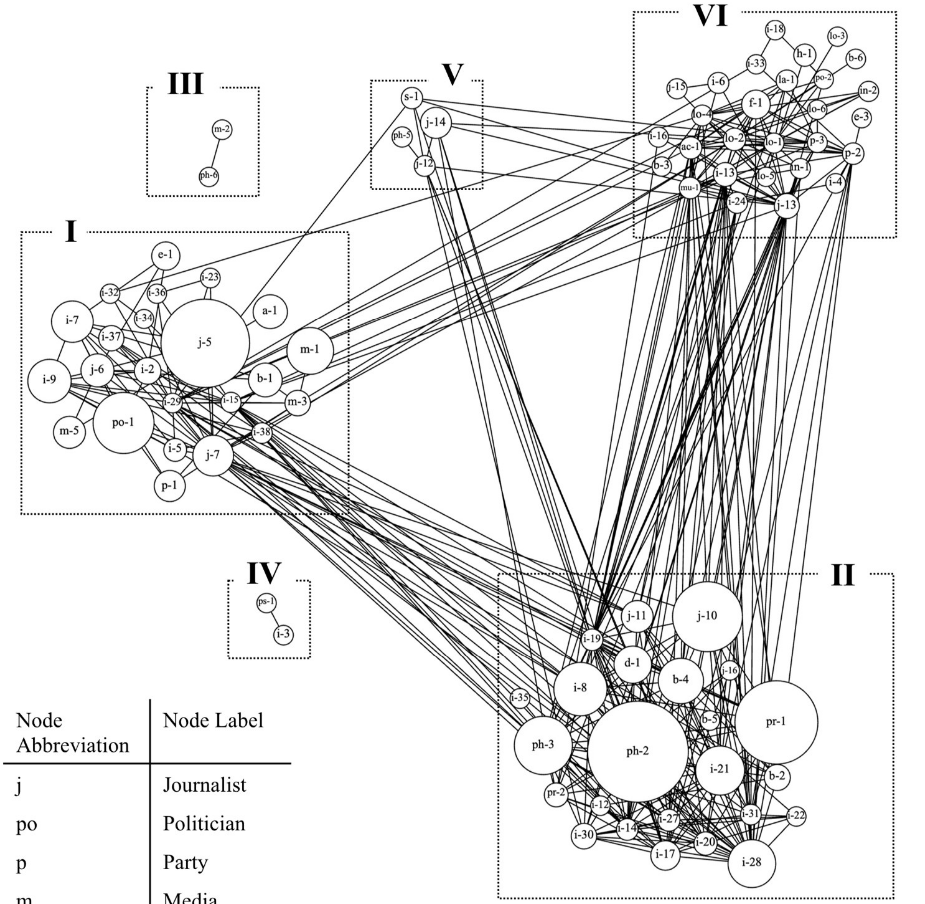


Node Abbreviation	Node Label
j	Journalist
po	Politician
p	Party
m	Media
b	Blog
e	Entrepreneur
i	Individual
mu	Musician
ph	Physician
pr	Professor
d	Director
f	Foundation
l	Lawyer
lo	Local chapter
la	Law firm
a	Author

Community Abbreviation	Community Role	Main Topic within Community
I	Emergent leader	Information on COVID-19
II	Challenger	Support of alternative science and media
III	Challenger	Criticism of hygiene measures
IV	Challenger	Speculation about conspiracy
V	Supporter	Information on COVID-19

Note: Node size according to l_{qt} score; usernames anonymized; community role based on positional analysis; main topics as identified within netdom analysis based on tweet frequency.

Figure 6. Querdenken User Network in t_2



Node Abbreviation	Node Label
j	Journalist
po	Politician
p	Party
m	Media
b	Blog
e	Entrepreneur
i	Individual
mu	Musician
ph	Physician
pr	Professor
d	Director
f	Foundation
lo	Local chapter
la	Law firm
a	Author
ac	Activist
in	Initiative
s	Scientist
h	Historian
ps	Psychologist

Community Abbreviation	Community Role	Main Topic within Community
I	Emergent leader	Information on COVID-19
II	Emergent leader	Criticism of hygiene measures
III	Challenger	Support of alternative science and media
IV	Challenger	Criticism of vaccination
V	Supporter	Criticism of health administration
VI	Supporter	Support of alternative science and media

Note: Node size according to l_{qt} score; usernames anonymized; community role based on positional analysis; main topics as identified within netdom analysis based on tweet frequency.

Figure 7. Querdenken User Network in t_3

Table 3. Overview of Density and Transitive Triplets Over Time

	t_1	t_2	t_3
Number of overall distinct Querdenken users (absolute)	101	101	101
Number of active Querdenken users (absolute)	34	65	83
Number of edges (absolute)	110	367	371
Density (in %)	2.2	7.3	7.4
Transitive triplets (average clustering coefficient) (in %)	20.0	35.1	33.1

Overall, the significantly higher values of density and transitivity in t_2 and t_3 compared to t_1 indicate increasing group engagement and cohesion within Querdenken over time. In the following, we turn to the results regarding social patterns based on *social positions*.

Figure 8 depicts the l_{qt} score distributions across timeframes. We plotted the values for each Querdenken user and timeframe as a scatterplot, with Querdenken users on the *y*-axis and their l_{qt} scores in percentages on the *x*-axis. For the three timeframes, it becomes evident that the majority of all Querdenken users were sparsely retweeted, whereas a few Querdenken users received significant attention from retweeting users. These Querdenken users who were able to unite a high share of retweeting users behind them may be considered—at least temporarily—de facto leaders of the movement.[9]

The pattern, evident across timeframes, follows a power-law distribution—a pattern found in “networks of very different types” (Bennett & Segerberg, 2013, p. 152), which is also a typical characteristic of social media networks and online communities (Johnson et al., 2014): A small number of users receive substantial recognition, while a large number of users receive only minimal recognition.[10]

These distributions reveal that, across timeframes, the retweeting users sustaining Querdenken online primarily redistributed content from only a few leading users. However, the prominence of individual users shifted over time. When comparing the top 10 users based on the l_{qt} score across the three timeframes, only three Querdenken users remained among the most retweeted users in all timeframes (see Figure 9), whereas 20 unique users held top positions at different times. Moreover, the top positions based on the l_{qt} score shifted over time. Notably, Twitter accounts of the local Querdenken chapters were at no point in time the leading voices in the Querdenken user networks. These findings suggest that subsuming the pandemic-skeptic protests exclusively under the Querdenken label may only partially capture its essence. They also demonstrate that particular users are not essential for upholding the idea of a greater collective; rather, we found evidence that online leadership is an emergent phenomenon that may shift over time, depending on the collective’s preferences (Faraj et al., 2011; Harysi et al., 2019; Johnson et al., 2015). In the context of social movements, this aligns with the claim that “social media reduces and even eliminates the need for central movement leadership” (Leong et al., 2015, p. 2). In sum, Querdenken’s leaders appear to be interchangeable.

This fluctuation in social positions is further illustrated by the average rank difference (see Figure 10). When analyzing the positions of the 22 unique Querdenken users who appeared in all three timeframes, it becomes clear that users were more likely to shift their positions than remain in similar ones. Only three users maintained a relatively stable rank, while the remaining 19 users experienced at least one positional change, on average. Notably, among the four Querdenken users with the least change in positions, three were among the top 10 leaders, suggesting that leaders tended to have more stable positions compared to other users, potentially based on the “cumulative advantage” (Merton, 1988) in prominence they conserve over time. Yet, overall, the average rank difference highlights the fluidity of social positions as users continually redefined who best represented the movement and its style.

An intriguing pattern emerged when interpreting the results of the community detection in light of the findings from the positional analysis. For each time period, this combined analysis allowed for a formal characterization of the communities, suggesting that these communities served different roles within the broader movement (see Figures 5-7)—even though communities with a specific role may not have been consistently sustained by the same users over time, as already evidenced by shifts in leadership described above.

In each timeframe, there was at least one community in which emergent leaders came together: Community I in t_1, Community I in t_2, and both Communities I and II in t_3. This differentiation in t_3 may be attributed to the leaders’ statuses: Community I consisted of leaders who were active in previous periods, whereas Community II was composed of newly emerging leaders.

[9] To validate the l_{qt} score findings, we also calculated the degree centrality of the user-by-user networks. The results support our findings, highlighting that the l_{qt} score effectively identified prominent users based on the retweet activity within the networks (see Appendix C2 for details).

[10] Since a power-law distribution should not be inferred only by visual analysis (Clauset et al., 2009), we also performed a statistical goodness-of-fit analysis of the data distribution by calculating the Kolmogorov-Smirnov statistic. The results confirm that all three distributions may indeed be described as a power-law (see Appendix C3 for details).

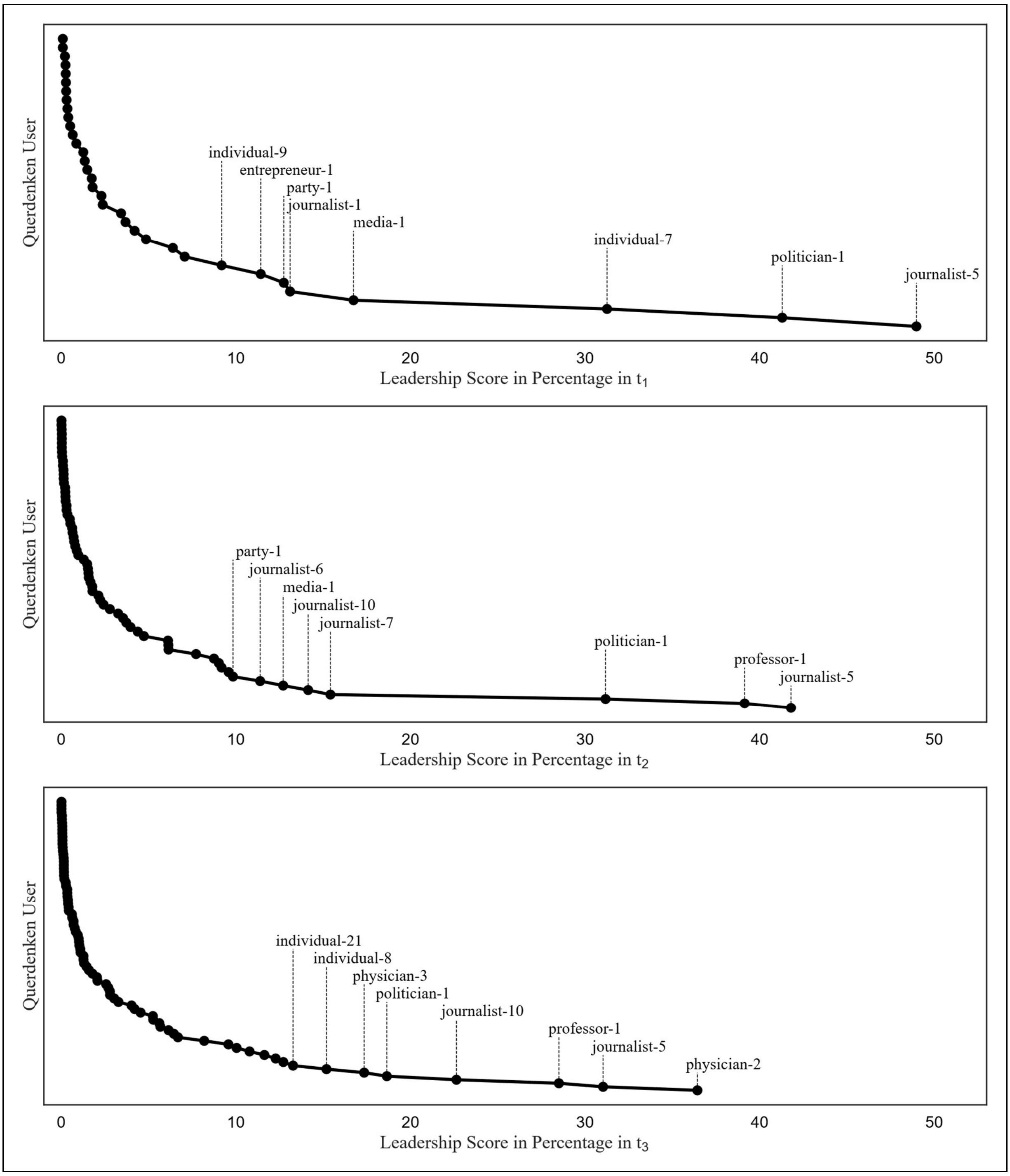


Figure 8. Leadership Score Distribution Over Time

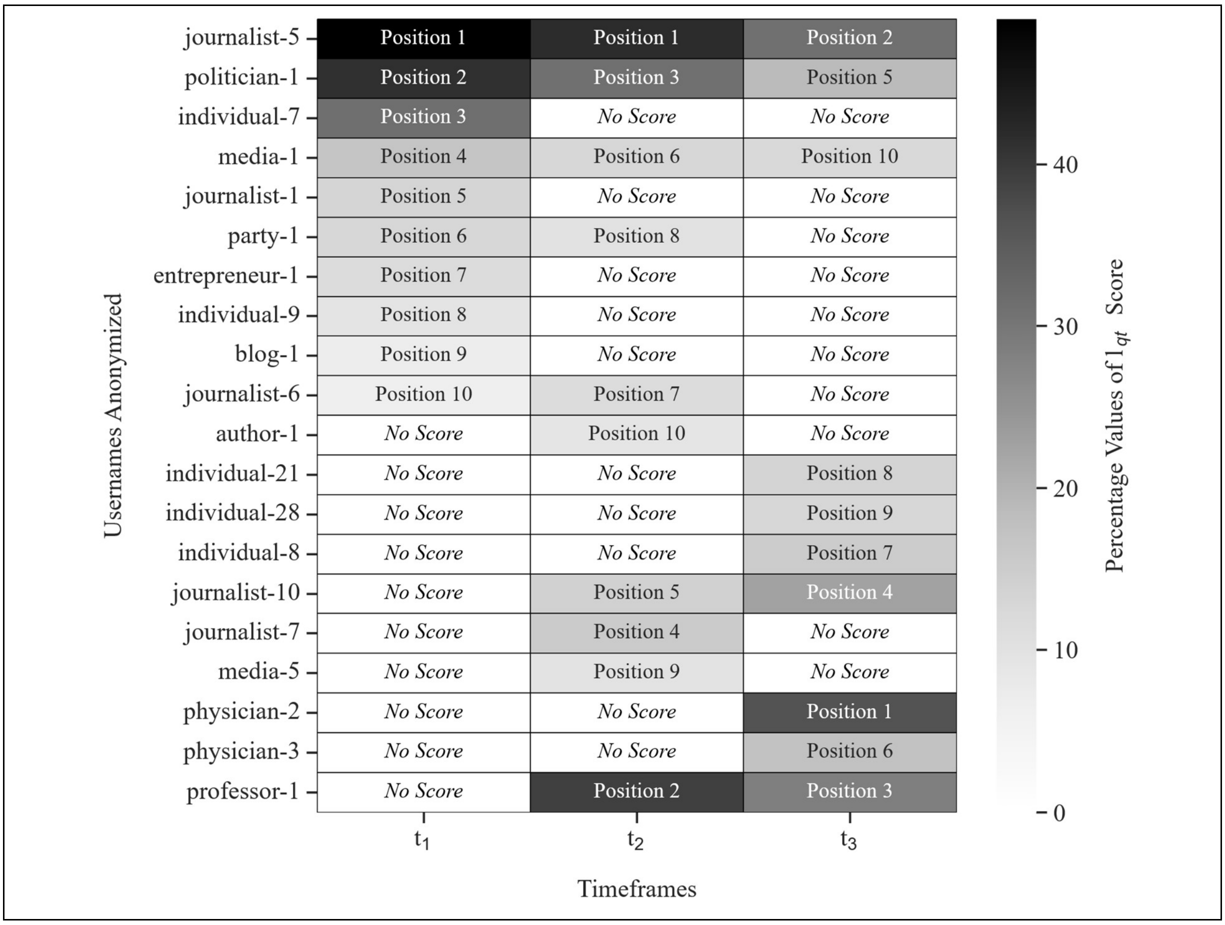


Figure 9. Shift of Top User Positions Based on $_{qt}$ Scores Over Time

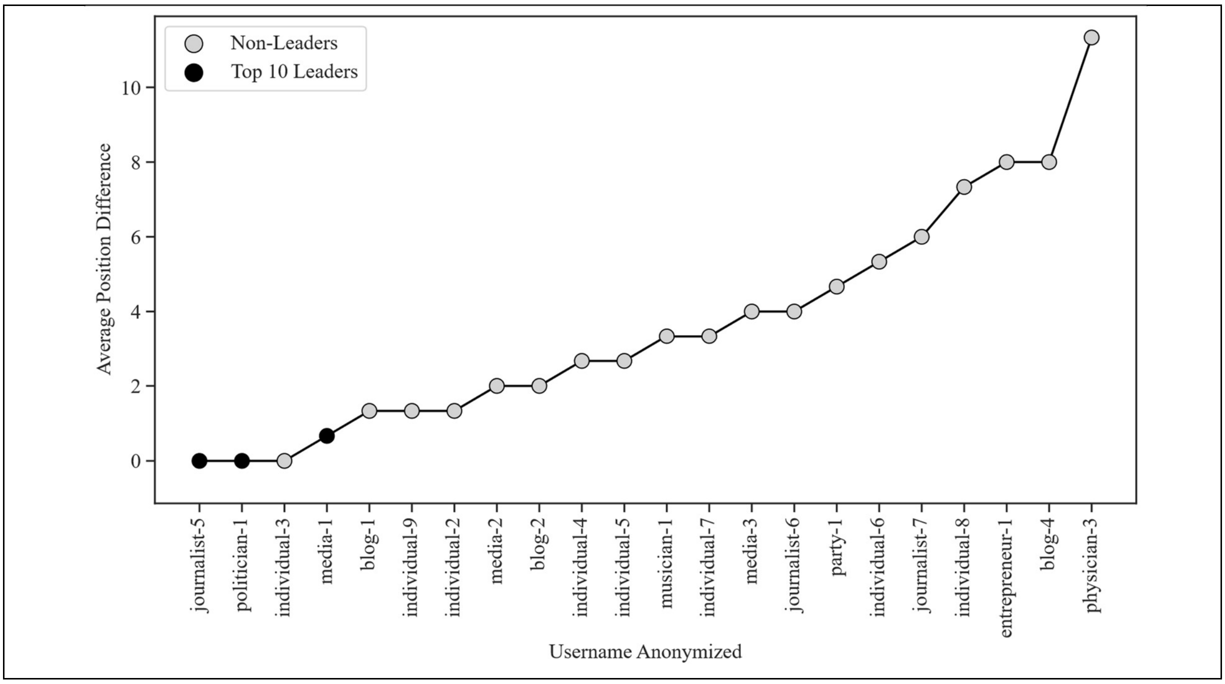


Figure 10. Average Position Difference of Users Present in All Three Timeframes According to $_{qt}$ Scores

Each period also featured one or more communities that were either weakly connected to the rest of the network or entirely isolated: Community II in t_1, Communities II, III, and IV in t_2, and Communities III and IV in t_3. Building on the work of Fligstein and McAdam (2012), these communities, which attract a different or more niche audience, can be characterized as "challenger" communities within the movement. They experiment with alternative interpretations within the overarching framework of Querdenken, potentially offering radical or divergent perspectives that challenge the mainstream narrative—thus offering creative divergence. Challenger communities can be seen as a vital source of creativity within the movement: New interpretations are tested in these semi-isolated environments (e.g., Kern & Nam, 2009), and if new interpretations resonate, they can be adopted later by other parts of the network. Conversely, if they fail to gain traction, any negative impact is confined to these relatively isolated communities and does not undermine the broader movement. These communities may also maintain specific niches, such as addressing local issues or targeting the more radical fringes of Querdenken, adapting the movement's message to different contexts. By maintaining a certain degree of separation, these communities can preserve ideological diversity and prevent the movement from becoming too homogeneous or rigid.

In each time period, there were also one or more communities connected to the communities of emergent leaders but consisting of less influential actors: Communities III and IV during t_1, Community V in t_2, and Communities V and VI in t_3. Drawing on the work of Vaast et al. (2017, pp. 1192-1193), these communities can be described as communities of "supporters." Their role extends beyond simply amplifying the leaders' messages. By sharing their own content related to the movement while still resonating with Twitter users who also retweet emergent leaders, supporters help refine the movement's framing, adding further perspectives. They act as a bridge between the emergent leaders and the broader public on Twitter, making the movement's worldview more accessible and relatable to a larger audience (e.g., Syed & Silva, 2020, p. 11; Vaast et al., 2017).

These three types of communities, however, do not operate in isolation; rather, they are interdependent. Emergent leaders generate core narratives, challengers introduce new ideas and challenge the status quo, and supporters ensure that these narratives reach and resonate with a broader audience (for a similar argument with respect to user roles, see e.g., Vaast et al., 2017, p. 1198).

In sum, these insights from the actors' network perspective present a complex image. Over time, the network relations became denser, and certain users gained influence. Yet, all Querdenken user networks fragmented into various communities, which continued to diversify as the number of communities increased. Further, neither the communities nor the most influential users remained consistent over time. Instead, different users sustained the communities and assumed leadership roles.

This conclusion must, however, be further qualified by accounting for the emergent systemic properties of the network. Despite considerable shifts at the individual user level, the networks exhibited considerable stability across all timeframes, as evidenced by the consistent power-law distribution—that is, the consistent development of emergent leadership—and the enduring functional differentiation among communities. This suggests that while individual positions shifted, the network's overall structure and functional differentiation exhibited strong stability and continuity.

To complement this analysis, we now turn to examining patterns at the cultural level.

5.2.2 Cultural Patterns

Before delving into the specific content shared by adherents of Querdenken on Twitter regarding recurring patterns of collective agency, sameness, and distinctiveness, we first examine the absolute frequencies with which statements pertaining to these three dimensions appeared over time (see Figure 11).

While the frequency of all three dimensions increased over time, beginning from t_1, statements emphasizing the demarcation of Querdenken's opponents appeared much more frequently than statements addressing the other two dimensions. This finding is in line with social movement research suggesting that the formation of collective identities usually begins with drawing boundaries between a collective actor and its adversaries (della Porta & Diani, 2006, pp. 93-94; Tilly, 2003, p. 608). It is also consistent with protest against the societal center as a core cultural feature of social movements (see Section 3.2). Translating these insights into our conceptual perspective, we posit that Querdenken's style on the cultural level is mainly characterized by differentiation from its environment.

A recognizable style, however, also requires consistent patterns in how the three dimensions—collective agency, sameness, and distinctiveness—are expressed. To explore whether such patterns emerged, we now examine the subcategories identified through the second step of our qualitative content analysis. As the development of a distinctive style relies on the solidification of cultural patterns, we focus on the patterns recurring in all three timeframes or solidifying in t_2 and t_3.

We begin by examining how collective agency was expressed in the tweets shared by Querdenken. This analysis reveals how the movement's adherents perceived and enacted their ability to influence and mobilize for their cause over time, based on the mentioned types of past and anticipated collective action. Figure 12 provides an overview of the forms of collective action, along with their respective frequencies expressed as percentages.

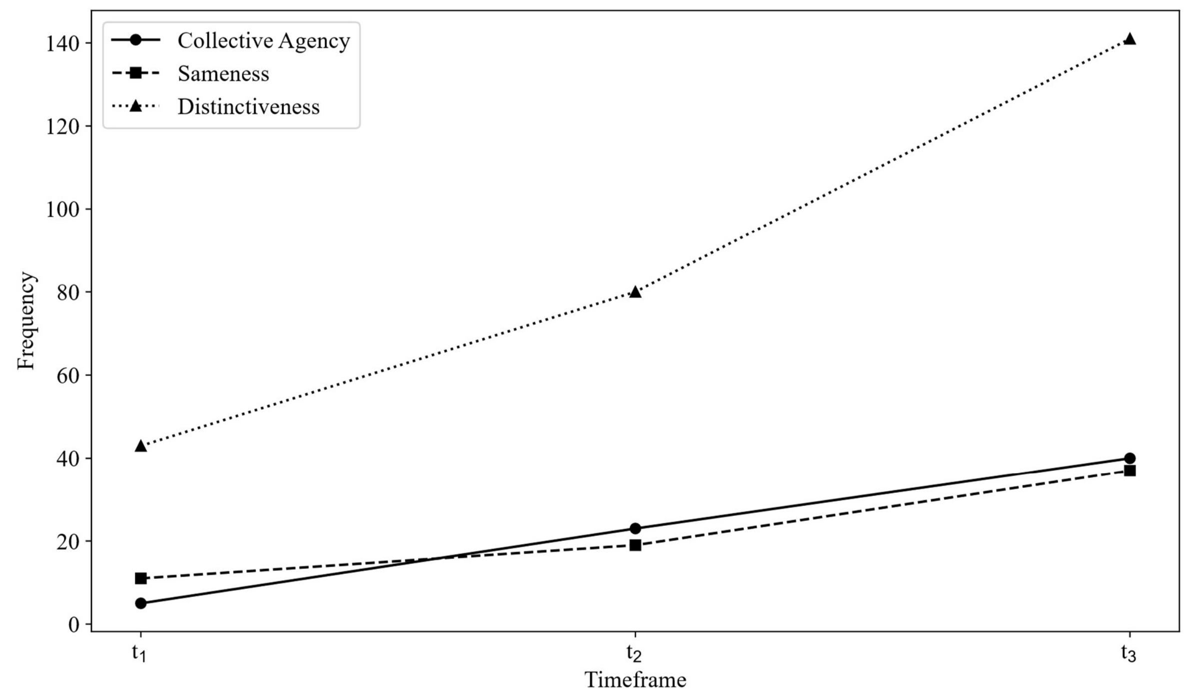
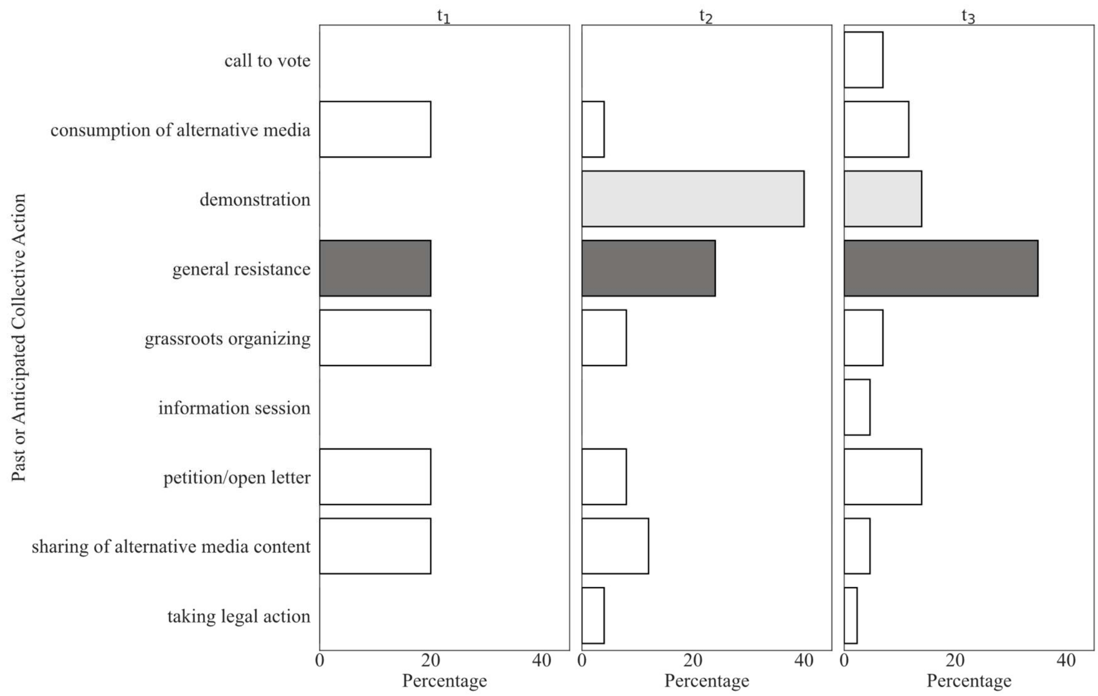
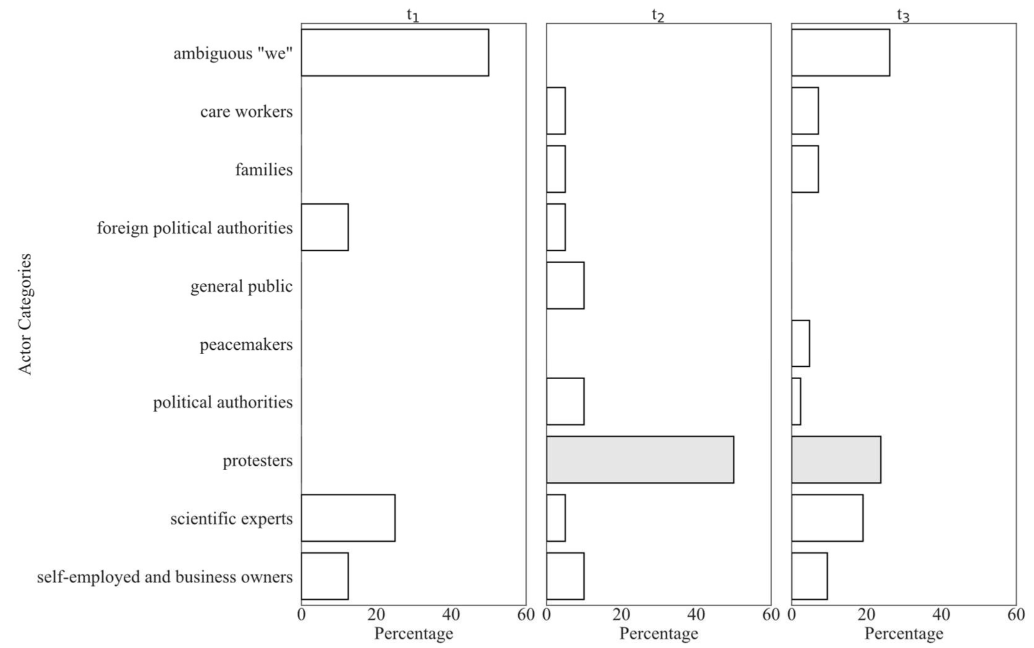


Figure 11. Occurrences of Collective Agency, Sameness, and Distinctiveness Over Time



Note: Categories represented above average in all three timeframes are highlighted in dark gray; categories represented above average in t_2 and t_3 are highlighted in light gray.

Figure 12. Types of Past or Anticipated Collective Action Expressing Querdenken's Collective Agency Over Time



Note: Categories represented above average in t_2 and t_3 are highlighted in light gray.

Figure 13. Type of Actors Considered as Part of Querdenken Contributing to its Sameness over Time

In t_1, there was a mix of *general resistance* and various forms of information exchange that illustrated a foundational stage during which the movement's adherents focused on internal communication. This phase was marked by the creation and dissemination of alternative narratives that challenged official accounts of COVID-19 and provided orientation and connection within the in-group during the outbreak of the pandemic in Germany. In t_2, the movement expanded its action repertoire, including more organized and direct action such as *demonstrations* and *taking legal action*, which became even more diversified in t_3.

Overall, the data reveal a consistent pattern: References to *general resistance* were dominant across all three timeframes—that is, broad calls for change without specifying particular actions or targets. This allowed a diverse range of groups or individuals to engage with the movement without committing to specific strategies. By avoiding the alienation of supporters who might have disagreed with extreme strategies (della Porta & Diani, 2006, pp. 179-180; Rucht, 2023b, p. 106), referencing such a flexible collective action repertoire maintained unity within the coalition, supported tactical flexibility, and helped to gain public approval—a useful strategy for social movements to mobilize supporters.

Alongside this somewhat vague form of collective action, another notable pattern emerged: while *demonstrations* were absent in t_1, they emerged as a significant form of collective action in t_2 and t_3, becoming a central element of the movement's repertoire during these later timeframes. In movement research, demonstrations are considered part of a conventional movement repertoire (e.g., Tarrow, 2011, p. 56) and possess a high level of societal legitimacy and popularity (Norris et al., 2005).

The increasing focus on *demonstrations* and the continual reference to *general resistance* suggests that the movement became increasingly aware of its own capacity for large-scale, publicly visible mobilization. Hence, the movement's style in this respect is characterized by the coexistence of specific, organized action in t_2 and t_3, alongside a more general yet vague call for resistance present throughout all three timeframes.

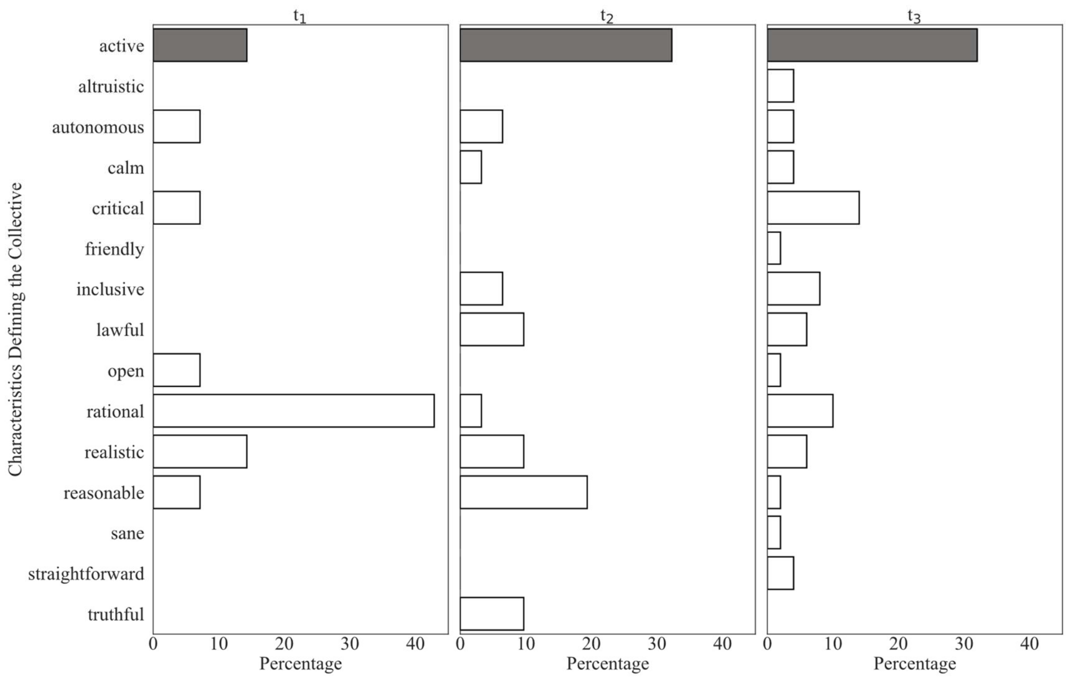
To assess Querdenken's sameness, we first evaluate how the movement's supporters categorized various actors as part of the movement across the three timeframes. Figure 13 presents an overview of the type of actors considered as part of Querdenken over time.

Again, we see a shift over time. In t_1, the movement most prominently identified with *scientific experts* and an *ambiguous "we"* that remained unspecified as representative of the group. In t_2, there was both a concretization and diversification of actors considered part of the movement. Most prominently, however, were adherents referred to as *protesters*—indicating that active dissent was a fundamental element of their sameness, more so than other group affiliations. In addition, the movement included a range of social groups that were particularly negatively affected by the preventive measures taken during the pandemic, such as *self-employed* and *business owners*. In t_3, the predominant self-perception as protesters solidified, along with the claim to represent a broad coalition of social groups. These groups were either affected by the regulations or possessed expertise in science or healthcare, enabling them to critically evaluate political decision-making.

Although the movement did not consistently highlight a single predominant actor group across all three timeframes, one pattern is evident: the movement's self-perception as protesters increasingly solidified in t_2 and t_3. Again, by embracing this broad, somewhat ambiguous category, Querdenken avoided alienating potential supporters. This openness allowed for the formation of coalitions with a diverse range of actors and points to a "tolerant identity" (della Porta, 2005)—a crucial trait for Querdenken, as it "helped the movement in dealing with its heterogeneous bases" (della Porta, 2005, p. 168; see also Section 4.1). Thus, Querdenken's style in this subdimension is characterized by a growing self-perception as a protest movement, particularly in t_2 and t_3, enabling coalition-building with various other groups. Figure 14 compiles the characteristics that define Querdenken adherents, the second indicator we assessed to understand the movement's perception of internal sameness.

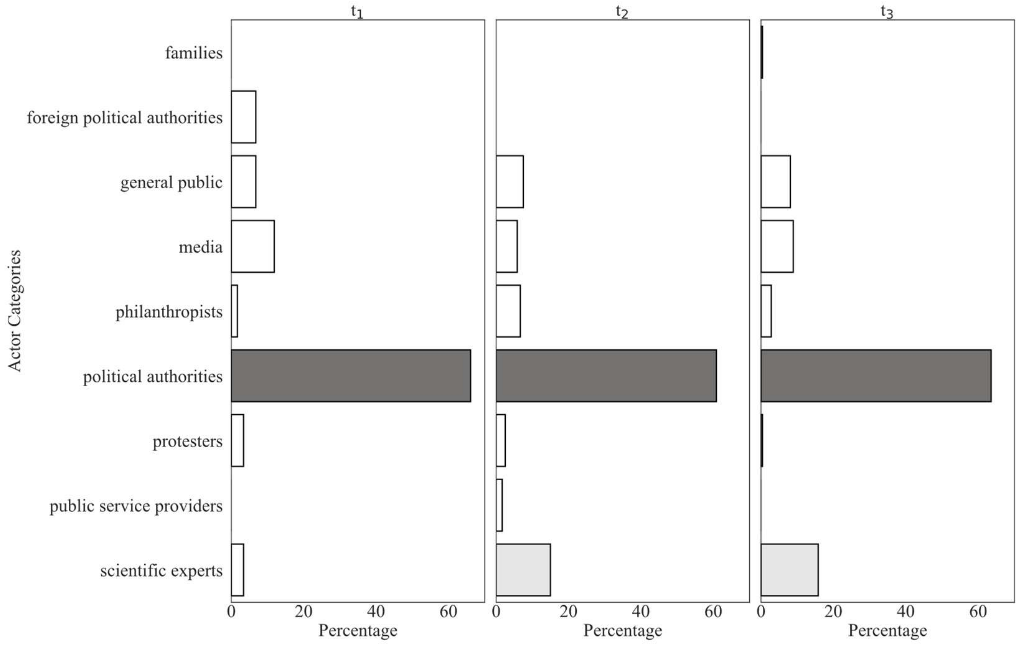
In the initial period t_1, Querdenken supporters primarily defined themselves through a *rational*, evidence-based worldview, positioning themselves as fundamentally skeptical of official accounts of COVID-19. As the movement evolved, this foundation of rationality was supplemented by attributes such as *reasonable* in t_2 and *critical* in t_3. By using these categories consistently across all three timeframes, Querdenken supporters portrayed themselves not merely as contrarians or "outside the box" thinkers—as the name Querdenken suggests—but as individuals whose skepticism was grounded in reason, logic, and common sense. Thus, our analysis confirms Alexander's (2006) assumption described in Section 4.2.3, which posits that actors in a democratic conflict present themselves with "civil"—that is, morally desirable—qualities.

While this recurring self-characterization can be seen as a stabilizing pattern, one unifying trait stands out across all three timeframes, shaping Querdenken's overall self-perception: the movement consistently depicting itself as *active*, highlighting a classical social movement trait of actively seeking to change the status quo against opposing forces from the societal center (Rucht, 2023b, pp. 18-22; see also Section 3.2). This emphasis on collective action parallels the shifts described above, that is, movement adherents increasingly describing themselves as *protesters* using *demonstrations* as a primary means to express agency. Thus, the movement's style in this subdimension is characterized by a critical stance and a strong emphasis on activism and engagement for political change, rather than passively enduring perceived political failures.



Note: Categories represented above average in all three timeframes are highlighted in dark gray.

Figure 14. Characteristics Defining Querdenken's Sameness Over Time



Note: Categories represented above average in all three timeframes are highlighted in dark gray; categories represented above average in t_2 and t_3 are highlighted in light gray.

Figure 15. Type of Actors Considered as Opponents of Querdenken Contributing to its Distinctiveness over Time

Next, as Figure 15 illustrates, we analyzed the distribution of groups from which the movement differentiated itself across the three timeframes—the first subdimension we assessed regarding the dimension of distinctiveness.

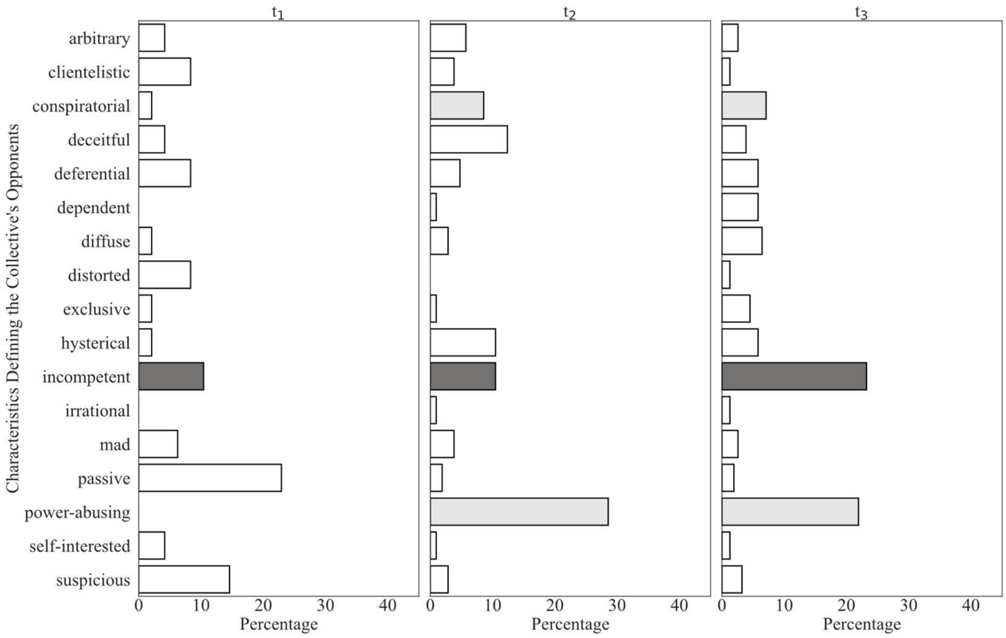
In all three timeframes, a clear pattern emerged: *political authorities* consistently dominated as the primary opponents, with approximately 60% representation across all three timeframes. The minimal variation across timeframes suggests a stable perception of opposition. While targeting *political authorities* as opponents is a common trait among social movements (della Porta & Diani, 2006, p. 21), it is more revealing to examine the unique opponents specific to the Querdenken movement. Notably, in t_2 and t_3, *scientific experts* advising and supporting the government emerged as significant opponents, while other categories, such as *philanthropists* like Bill Gates and *public service providers*, were present but addressed less frequently.

In examining the dimension of distinctiveness, we analyzed not only the actors from whom Querdenken differentiated itself but also the characteristics attributed to those actors. Figure 16 compiles the characteristics Querdenken attributed to its opponents across the three timeframes.

During t_1, Querdenken's opponents were primarily depicted as *passive*, reflecting the narrative that political decision-makers in Germany failed to protect citizens from COVID-19—as evidenced, for instance, by their reluctance to close national borders. In addition, political authorities and mainstream scientific experts were portrayed as *incompetent,*[11] reflecting the belief that they lacked the ability to manage the crisis. This attribution of incompetence persisted in t_2 and t_3. However, in t_2 and t_3, the enforcement of government decisions regarding protective measures was described as overly intrusive and, therefore, unlawful. Accordingly, political authorities—Querdenken's main adversaries—were depicted as *power-abusing* and *conspiratorial*, as reflected in the use of terms such as "surveillance state" and "dictatorship" in tweets, indicating that their actions were driven by ulterior, malevolent motives rather than genuine public health concerns. Across all timeframes, incompetence was the dominant attribute used to describe opponents.

[11] Please note that we added the code *incompetent* to the coding scheme derived from Alexander (2006). On the one hand, this addition was prompted by the significant number of tweets indicating a perceived deficiency in the abilities, knowledge, or qualifications necessary to effectively fulfill a particular role or perform a particular duty. The prevalence of such tweets justified the inclusion of this new category in Alexander's codes. On the other hand, this code aligns with existing literature, which underlines that the conflict over COVID-19 public health measures was often framed as an epistemic conflict (Bogner, 2021, pp. 21-25; Bogner & Menz, 2021; Pantenburg & Sepp, 2021).

This finding demonstrates that the movement's supporters increasingly converged online around a common set of derogatory attributes to define and demarcate their opponents. This trend indicates the solidification of the movement's style in this subdimension of distinctiveness from t_2 onwards. Overall, the movement sought to undermine the legitimacy of the pandemic measures and foster skepticism. This approach reflects a broader pattern of framing the pandemic as a conflict between the truth of the movement and the deception of the societal center. This aligns with existing literature, which indicates that the conflict over COVID-19 public health measures was often framed as an epistemic conflict, i.e., focusing on questions of expertise and competence (Bogner, 2021; Bogner & Menz, 2021; Pantenburg & Sepp, 2021). This framing is characteristic of many contemporary conflicts (e.g., Bogner, 2021; Büttner & Laux, 2021, p. 16). To conclude, Table 4 summarizes the solidifying social and cultural patterns that define Querdenken's style on Twitter, as identified in the results above.



Note: Categories represented above average in all three timeframes are highlighted in dark gray; categories represented above average in t_2 and t_3 are highlighted in light gray.

Figure 16. Characteristics Defining Querdenken's Distinctiveness Towards Opponents

Table 4. The Style of Querdenken on Twitter

Dimension	**Main findings**
Social patterns	**Social cohesion:** Querdenken user networks formed decentralized "networks of networks" (Bennett & Segerberg, 2013, p. 152) with only limited social cohesion, indicated by low density and average clustering. Communities within these networks can be distinguished based on their role for the overall movement—emergent leaders, supporters, and challengers. **Social positions:** Prominence within the Querdenken user networks follows a power-law distribution. Highly popular Querdenken users may be considered—at least temporarily—emergent leaders. However, individual prominence shifts over time.
Cultural patterns	**Collective agency:** Movement adherents blended organized collective action—especially demonstrations—with a more general call for resistance. **Sameness:** Movement adherents increasingly portrayed themselves as a collective of active protesters who were rational and critical, i.e., emphasizing only "civil" (Alexander, 2006) qualities in its self-description, legitimating their engagement for political change. **Distinctiveness:** Movement adherents increasingly demarcated themselves from political authorities and scientific experts using a set of "anti-civil" (Alexander, 2006) qualities to describe them, particularly incompetence, abuse of power, and conspiratorial behavior. The dimension of distinctiveness was most strongly invoked to form the style on the cultural level.

6 Discussion

Having conceptualized style online and applied our conceptualization to the case of Querdenken, thereby addressing RQ1, we now shift our focus to RQ2. To address this research question, we derive a set of propositions explaining the formation of style online based on our empirical results. The propositions are given in alignment with the structure of the results presentation.

Social movements' collective identity is generally understood as being "forged … in the course of conflict" (della Porta & Diani, 2006, p. 93). Not all conflicts, however, become politicized; some remain unaddressed and latent (e.g., Dahrendorf, 1972, pp. 35-36). The recognition of shared grievances and the attribution of adversarial roles regarding an issue that compels society to choose sides are necessary to politicize a latent conflict (Simon & Klandermans, 2001, pp. 324-326). In examining the Querdenken-specific netdom, we observed that the movement's adherents created an increasingly politicized domain. This heightened politicization became evident in the growing awareness of its adherents' role as participants in the collective contestation of the political response to the COVID-19 pandemic, thereby creating fertile ground for the formation of style.

Proposition 1: For the formation of a social movement's style online, the stabilization of a netdom comprising a politicized domain over time is a necessary condition.

The interactions among Querdenken adherents led to the formation of social networks composed of interconnected communities—a social pattern commonly observed in online interaction (e.g., Bennett & Segerberg, 2013, p. 152). Given the low entry and exit barriers inherent to online environments (e.g., Johnson et al., 2015, p. 166), the composition of these networked communities was in constant flux, and so the networks' social cohesion remained limited. Despite this, and in line with the literature on emergent leadership (e.g., Faraj et al., 2015; Harysi et al., 2019; Johnson et al., 2015), a formal structure characterized by a power-law distribution solidified over time. As noted earlier, emerging structures like these are particularly prevalent in social media environments (e.g., Johnson et al., 2014), where digital platforms with algorithmically curated timelines foster mechanisms such as preferential attachment, thereby contributing to these emerging structures.

A stable social pattern emerged and persisted across different timeframes as a result of these characteristic structures. In this respect, the platform may be seen as an "organizing agent" (Bennett & Segerberg, 2013, p. 34), assuming a role traditionally fulfilled by SMOs that aided Querdenken in appointing temporary de facto "opinion leaders" (Rogers, 1995). "Thus, organizing that occurs through social media is inherently socio-material" (Leonardi & Vaast, 2017, p. 180) because the platform's architecture introduces a degree of orderliness within a context of purely informal, spontaneous self-organization, devoid of any form of formal governance (e.g., Massa & O'Mahony, 2021, p. 1041; Vaast et al., 2017, p. 1180). The platform thereby alleviated the movement of some of the decision-making burdens that accompany the coordination of collective action (on the prerequisites and challenges of participatory decision-making in social movements, see, e.g., Haug, 2013; Haug & Rucht, 2013).

However, no single user's leadership consolidated over time. Instead, shifting users consistently adopted resonant patterns observed in the content of current opinion leaders they later replaced, thereby creating and perpetuating the collective's overarching style online. Thus, we found evidence of emergent leadership (e.g., Faraj et al., 2015; Harysi et al., 2019; Johnson et al., 2015; Majchrzak et al., 2013) in online social movements, based on recognition from other users.

In addition to the emergent leaders, we identified communities of supporters and challengers—suggesting a system of interdependent roles that collectively enhance the appeal and creativity of the movement as a whole.

Proposition 2: For the formation of a social movement's style online, emergent leadership patterns stabilize, resulting in a power-law distribution regarding users' influence that is supported by communities of other users fulfilling complementary roles.

Distinctiveness appeared as the key conceptual dimension in which cultural patterns emerged. This prominence may be attributed to the dynamics of social media, where emotionally charged and negatively framed content tends to attract particular user attention (Henn & Posegga, 2023; Kohout et al., 2023; Soroka et al., 2019) and diffuses with remarkable ease (Stieglitz & Dang-Xuan, 2013; Tsugawa & Ohsaki, 2015). These dynamics reinforce the visibility of tweets featuring occurrences of distinctiveness. Moreover, defining "others" or adversaries is inherently less complex and fraught than articulating the attributes and forms of agency of one's own group. Making definitive statements about the characteristics of a movement involves navigating internal diversity and the potential for disagreement, which may incite internal conflict stemming from ideological differences to competing interests of a movement's factions. For online collectives lacking agreed-upon procedures for decision-making, navigating, and collectively establishing an agreement on "who we are," this appears to be even more demanding. As a result, we infer that there is a tendency to avoid definitive selfcharacterizations online. Instead, gravitating towards establishing a style in opposition to a clearly defined "other" appears to be more effective for fostering cohesion through a shared sense of distinctiveness.

Proposition 3: When a social movement's style is formed online, the dimension of distinctiveness is the most pronounced.

The analysis of Querdenken's sense of collective agency revealed that its adherents consistently expressed this perception through references to broad, generalized calls for resistance, with demonstrations gaining importance over time. This development underlines the movement's growing capability to act as a resourceful conflict party that challenges the opponents from the societal center. Hence, the movement increasingly perceived itself to be the architect of its own social conditions, which reflects the "ideas of empowerment" (Gamson, 1991, p. 37) crucial for collective action.

Proposition 4: For the formation of a social movement's style online, the stabilization of cultural patterns pertaining to collective agency directed against the societal center is a necessary condition.

As previously noted, Querdenken's adherents' sameness remained notably ambiguous throughout the observation period. We have already established that defining the sameness in a movement as heterogeneous as Querdenken generally poses a challenge. Social movement scholars, however, highlight a broader trend among contemporary movements towards establishing what is described as "weak identities" (Rucht, 2011) or "tolerant identities" (della Porta, 2005). These movements typically coalesce around vague slogans, allowing supporters considerable flexibility in selecting their ideological stance, specific tactics, and goals. Weak identities, though, are not viewed as a deficit; rather, they prove to be advantageous for broad mobilization (Rucht, 2011, p. 82). Sustaining such a broadly defined sense of sameness, however, requires balancing two extremes: on the one hand, an overly diffuse self-concept may result in a movement's lack of profile and coherence, leading to the dissolution of sameness; on the other hand, movements always risk becoming too insular, leading to social closure and eventually culminating in sect-like behavior (Rucht, 1995, p. 19, 2011, p. 83). This tendency toward a relatively weakly defined self-conception is thus a phenomenon observable both online (e.g., Bennett & Segerberg, 2013) and offline (e.g., Rucht, 2011), suggesting it is a prevalent strategy in contemporary collective action. Bennett and Segerberg's findings on recent forms of connective action that avoid imposing strong ideological bonds (see Section 1) may, in fact, reflect a more general phenomenon witnessed across various protest settings.

Proposition 5: For the formation of a social movement's style online, it is beneficial to foster a tolerant identity towards similar actors.

Although we observed signs of a weakly defined self-concept within Querdenken, the movement nonetheless succeeded in establishing several defining traits of its adherents' sameness—an aspect potentially underestimated by Bennett and Segerberg. These traits align with Alexander's codes of binary opposition (Alexander, 2001, p. 163, 2006, p. 231; Alexander & Smith, 1993, p. 197; see also Section 4.2.3), enabling the movement to distinguish itself from its opponents with morally desirable traits typical of democratic societies.

Proposition 6: For the formation of a social movement's style online, the stabilization of cultural patterns pertaining to sameness necessarily relies on consistently attributing morally desirable traits to the movement's adherents.

The evaluation of the cultural pattern regarding distinctiveness revealed two solidifying patterns: Across all three timeframes, political authorities were consistently identified as the main opponents, while in later timeframes scientific experts advising the government emerged as a secondary opponent category. In addition, the attribution of incompetence remained a consistent characteristic applied to Querdenken's opponents, with narratives labeling them as conspiratorial and power-abusing emerging later. These ascribed characteristics align again with Alexander's (2006, pp. 57-59) scheme of binary codes, which guide Querdenken's boundary work.

Evaluating the overall pattern within this subdimension reveals that political authorities served as the primary external adversary for Querdenken. Since this category encompasses a wide range of political figures, it serves as an umbrella term that facilitates the creation of a unifying narrative without requiring nuanced internal consensus among the broad coalition of participants holding varying grievances and ideological perspectives (see Section 4.1) because it can encompass a wide range of critiques, from government passivity to government overreach and notions of conspiracy. Thus, the movement was able to cultivate a sense of sameness against a broadly appealing "other," which was adapted to address the movement's heterogeneous and evolving concerns across time (see Section 5.1). This clearly identifiable external adversary gave the movement a unified sense of direction.

By establishing this clear distinction from the onset, Querdenken established a well-defined concept of its opponents, reflecting a deliberate process of boundary work against them. Defining a movement in opposition to authorities, such as political leaders, in such a way is a widely recognized process for identity formation in social movement studies (della Porta & Diani, 2006, p. 112). These boundary processes are known to strengthen identity formation by enhancing movement mobilization and intensifying "hostility toward powerholders and their representatives" (della Porta & Diani, 2006, p. 112). This hostility is especially pronounced when combined with the attribution of conspiratorial intentions because it accelerates the radicalization of communities and movements, deepening their opposition to perceived enemies (Abdalla Mikhaeil & Baskerville, 2024).

Proposition 7: For the formation of a social movement's style online, the stabilization of cultural patterns pertaining to distinctiveness against clearly identifiable adversaries necessarily relies on consistently attributing morally undesirable traits to the movement's adversaries.

Upon reviewing our findings, a clear pattern emerges: Querdenken adherents found themselves in communities of alternating compositions and shifting leadership. However, at the cultural level, we observe a consolidation of patterns related to collective agency, sameness, and distinctiveness, while on the social level, the power law distribution also indicates a stable pattern of asymmetric influence among Querdenken adherents, and a functional differentiation among communities is also stabilizing. This is paralleled by the solidification of key topics and their increasing politicization at the domain level. Linking these two overarching dynamics—the fluidity of actors at the individual level and the consolidation of social and cultural patterns at the collective level—implies that these social structures, the set of cultural references defining Querdenken's style, as well as its domain, take on a structurally autonomous, emergent quality because they are sustained by shifting actor constellations (see Langer et al., 2020, who also demonstrate how shifting actor constellations can sustain a movement over time). These patterns consolidate without depending on the original authors of specific tweets or hashtags to further disseminate them or consistently sustaining a particular position within the network. Hence, as a recognizable social entity, a movement like Querdenken can prevail as long as *any* constellation of actors continues to uphold the core cultural and social patterns defining its style.

This dynamic may fundamentally arise from the fact that—unlike face-to-face interactions—social interactions online are "self-documenting" (Schwarz, 2021, p. 25). They serve as both "social interactions and the collective production of objects" (Schwarz, 2021, p. 29). As a result, their content becomes objectified, persists, and can potentially be revisited and redistributed by any user days, months, or even years later by means of asynchronous interactions (e.g., boyd, 2010, pp. 46-47; Milan, 2015, p. 7; Schwarz, 2021, pp. 15-16; Syed & Silva, 2023, p. 251; Treem & Leonardi, 2012, p. 155). We presume that self-coordination is enabled in such a context via the process of "mutual adjustment" (Scharpf, 1997), wherein an actor's intention for future action is shaped by the perceived status quo, "which itself is a result of the interplay of all actors involved up to now" (Schimank, 2005, p. 20; see also, e.g., Scharpf, 1997, p. 109). In the online context, this translates to the fact that the persistence of past interactions allows users to observe and adapt to each other's behavior, aligning their future postings with the observed patterns of conduct that represent a notion of "how things are done" (Berger & Luckmann, 1972, p. 77). Ultimately, mutual adjustment results in a recognizable, overarching style.

As already indicated above, mutual adjustment is influenced predominantly by emergent leaders, who can serve as "proxies" (Vaast et al., 2013, p. 1082) and guide alignment with the established cultural patterns, facilitating "mimetic isomorphism" (DiMaggio & Powell, 1983, pp. 150-152). This dynamic of mutual observation and mutual adjustment, guided by opinion leaders, enables the interactive solidification of recurring cultural patterns in an otherwise disordered, chaotic social environment—even amid shifting actor constellations.

Based on these considerations, we conclude that despite the fluidity of the online sphere, social movements can develop a structurally autonomous style online that constitutes and sustains the movement as a discernible collective actor, relatively independent of particular users on the individual level. This process is guided by temporary emerging leaders. This conclusion is possible only because we placed the duality of style at the center of our analysis.

Proposition 8: When a social movement's style is formed online, the associated social and cultural patterns become structurally autonomous from individual actors, meaning they persist in shaping and influencing style independently of specific users.

When placing the process of style formation observed in the case of Querdenken within a broader context, the outbreak of the COVID-19 pandemic can be understood as creating a so-called "situation" (Mische & White, 1998), characterized as "stochastic and uncertain" (Mische & White, 1998, p. 700). This uncertainty made it challenging to predict how interactions within movements would unfold, prompting a search for new social norms and expectations because actors perceived the future as "problematic and unresolved" (Mische & White, 1998, p. 700). This led to reinterpretations of the social context and realignments of social positions and status orders (Mische & White, 1998, p. 700).

We see this realignment within our Querdenken case study. At the onset of the COVID-19 pandemic, the movement displayed an ambiguous style; its cultural patterns, topics, and level of politicization within its netdom were still somewhat volatile. However, as described above, these patterns solidified over time as Querdenken became more concrete about how to position itself in relation to this new situation. This development process can be addressed by the notions of "weak" and "strong culture," which we elaborate on in the following.

A distinctive feature of weak culture is its ability to facilitate navigation through contexts that lack clear structures (Schultz & Breiger, 2010, p. 618). It helps to bridge connections and bonds with "strangers" (Schultz & Breiger, 2010, p. 613), that is, building netdoms beyond closed groups already aware of their pre-established shared interpretations. To build these connections, a weak culture creates a "sociable substrate" (Schultz & Breiger, 2010, p. 613) based on cultural references that are addressable with low risk, fostering high levels of relatability, even among previously unknown individuals, and thereby securing the continuity of interaction. The value of interactions based on weak culture lies in their focus on establishing basic common understandings and expectations of "how things are done" rather than conveying specific content (Schultz & Breiger, 2010, p. 613).

Although such interactions may seem trivial, they still build connections—while simultaneously creating boundaries and thus leading to a certain degree of social closure (Schultz & Breiger, 2010, p. 615). Over time, continued interaction, mutual observation, and adjustment may foster increased certainty of expectations, encouraging actors to express more high-risk opinions. This can lead to the development of more narrowly defined niches shaped by individual "idiosyncrasies" (Homans, 1961) and shared by fewer users (e.g., particular conspiracy theories or critiques based on professional knowledge) that nevertheless contribute to the emergence of a strong culture within those niches. We observed this process of specialization within our case study of Querdenken when comparing the development of topics and their polarization within the netdom, particularly when focusing on the smaller, more niche-oriented communities predominantly sharing these polarized topics (see Figures 5-7).

Thus, an online collective may transition from a weak culture to a strong culture as it develops increasingly specific interpretations of a situation, raising the bar for sharing such content online due to the necessity of a specific ideological stance concerning, for example, political elites, science, and vaccines. In the case of Querdenken and its expressed style, we see initial indications of this transition from a weak to a strong culture, resulting in a more distinctive style of the movement. It can be assumed, however, that the formation of a strong culture is more likely to occur within more closed and cohesive communication spaces, where shared values and norms can be deeply cultivated and reinforced among a small set of participants. In contrast, platforms with high fluidity and openness, such as Twitter, facilitate the persistence of a weak culture, which allows for broad engagement and mobilization.

Proposition 9: In an evolving social context online, a weak culture facilitates initial open interactions, which may gradually transition into patterns of a strong culture as actors develop more specific, high-risk interpretations and ideological stances, leading to increased specialization and social closure and resulting in a more distinctive style of the social movement.

7 Conclusion

We posed two research questions that guide this study. To address RQ1, we developed a novel framework for conceptualizing collective identity online, tailored to the unique characteristics of digital trace data. Rooted in Harrison White's notion of style, this conceptualization integrates the duality of culture and structure with established dimensions of collective identity. With that, we addressed a critical gap in previous research and offer a more nuanced understanding of how to conceptualize the collective identity of social movements and other collectives—online as well as offline. By combining this theoretical conceptualization with socio-semantic network analysis—a method that reflects our concept's duality—we provide an approach to empirically determine a movement's collective identity online.

To address RQ2, we examined the style formation of the Querdenken movement on Twitter over time. Our findings demonstrate that collective identity online becomes tangible through the consolidation of recurring social and cultural patterns, demarcating and describing Querdenken as a distinct collective actor. Thus, the social and cultural patterns of a movement's style emerge as a structurally autonomous phenomenon, as they are continuously reproduced independently of specific users. Instead, these patterns are perpetuated over time by a *multitude* of shifting users, with different individuals temporarily assuming emergent leadership roles. Based on these findings, we developed nine propositions that outline conditions for social movements to establish a distinctive style online, making them recognizable as unique actors. These propositions also theorize how style forms and manifests over time.

With our study, we contribute to the debate over what it means for social movements to be organized online, particularly in the absence of formal movement organizations structuring this process. Building on social movement theory (e.g., Gamson, 1991; Melucci, 1996), we highlight the pivotal role of collective identity, which serves as the foundation for social collectives to emerge as coherent, recognizable entities endowed with a sense of agency. In doing so, we challenge the perspective put forth by Bennett and Segerberg (2013), who argued that collective identity's relevance has diminished in contemporary digitally fueled social movements. Contrary to this view, our findings demonstrate that social movements can indeed cultivate a distinctive collective identity—understood as a collective's style—online, which marks them as unique social actors capable of deriving sameness, distinctiveness, and collective agency. Our findings underscore that the formation of collective identity is not bound to the offline or organizational context; it is equally relevant and achievable in online environments. This solidification of a collective identity allows online social movements to become recognizable entities that can be sustained over time, despite looser forms of organization and changing actor constellations in digital environments.

Overall, consistent with research proclaiming new forms of collectivity and collective engagement online (Bakardjieva, 2015; Chamakiotis et al., 2021; George & Leidner, 2019; Gerbaudo, 2014; Gerbaudo & Treré, 2015; Langer et al., 2020; Stewart & Schultze, 2019), we posit that the formation of a collective identity online is shaped by the unique characteristics of the online space, particularly the objectification of interaction. We contend that these characteristics enable the emergence of new forms of self-organizing—particularly via "mutual observation" (e.g., Schimank, 2005, p. 20) and "mutual adjustment" (Scharpf, 1997)—whereby a set of dispersed actors can coalesce and forge recurring social and cultural patterns defining their collective identity.

In the following, we derive theoretical and practical implications and highlight potential avenues for future research.

7.1 Theoretical Contribution

This study offers four contributions to the study of online collective identity, digital activism, and online organization.

First, by integrating the concept of style with well-defined subdimensions of collective identity on a theoretical level and employing socio-semantic network analysis on a methodological level, we developed a comprehensive conceptual and empirical framework for detecting, describing, and quantifying collective identity online. This framework enables analyzing the empirically detectable social signature of both social movements and other collective actors online, addressing both the cultural and social dimensions while accounting for the peculiarities of digital trace data. In doing so, we respond to the call for further studies combining network and content features to fully exploit the potential of digital trace data (Borgatti & Foster, 2003; Kane et al., 2014; Syed & Silva, 2020, 2023) and better understand "how social media networks sustain movements" (Leong et al., 2019, p. 192). Our work advances the understanding of digital activism (George & Leidner, 2019), helping this "raucous teenager" on its journey to adulthood.

Second, we provide deeper insights into how "de-centralized individuals come together to form a collective with a distinct identity" (Tarafdar & Ray, 2021, p. 1086), particularly in fluid contexts where hierarchical structures or organizational boundaries are absent (Young et al., 2019). Our results indicate that formal organizations, local chapters, and the continuity of leadership play only a minor role in forming and maintaining the collective identity of a movement. Instead, temporary leaders emerge, gaining their status through recognition from the collective. These findings contribute to the literature on the manifestation of online community leadership (e.g., Faraj et al., 2011, 2015; Harysi et al., 2019; Johnson et al., 2015; van Haperen & Uitermark, 2023) and respond to calls to move beyond the focus on formal organization that has historically dominated IS research (Cardoso et al., 2019; Ghobadi & Clegg, 2015; Kane et al., 2014; Vaast et al., 2017).

Third, our conceptualization of collective identity formation online draws on approaches from different disciplines. By integrating traditional, pre-digital theories from social movement research with insights from IS studies and applying them to the online context, we contribute to the field of research that advocates the interdisciplinary integration of social science theories with IS phenomena (George & Leidner, 2019, p. 17; Ren et al., 2012; Selander & Jarvenpaa, 2016; Tsai & Bagozzi, 2014) and enhance our understanding of "the interactions among technology, social movement members, and their environments" (McKenna, 2020, p. 187). Moreover, integrating theories from these two distinct disciplines enabled us to analyze phenomena such as online collective identity—situated at the intersection of these fields—more deeply and comprehensively. Thus, this theoretical foundation allowed us to contextualize our empirical findings within a robust theoretical framework, leading to the formulation of nine propositions concerning online identity formation. We therefore position our framework within Type II theories (explaining), as defined by Gregor (2006, p. 620).

Fourth, we provide profound insights into the case of Querdenken, a heterogeneous movement that emerged during the COVID-19 pandemic. Researchers interested in further analyzing science- and pandemic-skeptic movements might use our findings on Querdenken's collective agency, sameness, and distinctiveness as a starting point.

7.2 Practical Contribution

Our findings offer three practical insights for social movement adherents, activists, policymakers, and platform owners.

First, social movement adherents and SMOs can benefit from understanding that the development of a collective identity that manifests in the solidification of cultural and social patterns can be helpful in organizing and sustaining social movements online as discernible composite actors. By fostering such patterns, adherents and SMOs can extend a movement's momentum and ensure its longevity (see also Syed & Silva, 2023, for the importance of message framing to movement mobilization and survival).

Second, leadership within social movements online is emergent, depending on the collective's temporary consensus regarding who is considered an adequate opinion leader. Thus, the existence of the leadership role remains but becomes autonomous of individual actors. This finding, linked with the functional differentiation among movement communities, provides valuable insights into the composition of movements online, allowing movement leaders and SMOs to better comprehend the interdependencies among movement factions. Movement leaders and SMOs should tailor their strategies to engage with these highly influential yet temporary opinion leaders when seeking to mobilize or steer their movement. Being sensitive to shifting opinion leadership can also aid policymakers seeking a better understanding of public discourse (e.g., Stieglitz & Dang-Xuan, 2013) and platform providers interested in shaping online discussions through their support or regulation of these leaders (e.g., Faraj et al., 2015).

Third, in the Querdenken movement, we find first indications for the development from a weak to a strong culture (Schultz & Breiger, 2010), which policymakers and platform providers can monitor to prevent radicalization and the spread of conspiracy theories. Recognizing cultural solidification can be helpful in determining when countermeasures are needed to avoid fragmentation into more extreme subgroups.

7.3 Limitations and Implications for Future Research

Although we provide valuable insights into the conceptualization and formation of collective identity online, our study should be seen in the light of five limitations.

First, in our study, online interaction patterns with external actors, such as Querdenken's antagonists and audience, remain unobserved. Future research should focus on examining how social movements and collective actors are embedded in their online environments, offering deeper insights into the boundary processes that define and distinguish these movements as unique actors.

Second, we opted for a case study to empirically illustrate our conceptualization and derive propositions about the characteristic patterns shaping the style of social movements online—a research design often confronted with concerns of generalizability (e.g., Eisenhardt & Graebner, 2007, p. 27; Goertz & Mahoney, 2009, p. 307). In light of the different elements proposed in this study, it is necessary to engage in a nuanced discussion of their generalizability, differentiating between the generalizability of the concept of "style," its operationalization based on Twitter, and the propositions derived from the empirical case study.

We suggest that the conceptualization of style is generalizable to any type of social movement, whether observed online or offline, although the specific operationalization outlined in Figure 3 may need adaptation when applied beyond the context of Twitter. Consequently, our framework offers analytical categories that are not restricted to any particular movement or platform, making it widely applicable beyond our specific case. To stimulate future research, we followed Gerring's (2004) advice to "overreport" (p. 346) all potentially generalizable propositions derivable from our empirical case. We suggest that Propositions 1 through 7, as well as Proposition 9, apply to networked social movements online in general. Proposition 8, however, pertains to platform affordances. The mechanism identified here applies to any platform that allows asynchronous interaction and objectifies these interactions, which is crucial for mutual observation and adjustment among shifting users. Platforms that include features such as confirmation metrics and algorithmic curation based on such metrics can further facilitate these processes—whether within closed groups or public spaces like Twitter. Since theory-building through case studies is inherently a bottom-up process (Eisenhardt, 1989, p. 547), we encourage future research to undertake theory-testing case studies. Theory testing is crucial, as it helps distinguish features that are typical of a larger set of units from those specific to the case under study (Gerring, 2004, p. 346). Moreover, future research should investigate the potential for other variants of style and explore equifinal causal paths leading to the emergence of a social movement's style online (on equifinality, see, e.g., Goertz & Mahoney, 2009, pp. 314-315).

Third, building on the generalizability and platform-specific considerations discussed above, we encourage more systematic investigation and theorizing regarding social media's distinct "materiality" (e.g., Leonardi, 2012, p. 29), which would deepen our understanding of how online environments shape collective identity formation and continuity compared to the offline world. While we provided initial insights (see Sections 3.3 and 6), the concept of affordances (e.g., Strong et al., 2014; Treem & Leonardi, 2012) provides a fitting lens to assess the effect of this "imbrication between human and material agencies" (Leonardi, 2011, p. 149) in the process of collective identity formation online. Thus, we invite researchers to explore how social media platforms' affordances and design features influence the formation of style.

Fourth, in our study, direct retweets served as the basis for network construction. While this is a well-justified approach that allowed us to incorporate the notion of emergent leadership based on the collective’s retweeting behavior, it may still introduce certain biases. For instance, it may be assumed that less radical content is retweeted more often than more radical content. Consequently, such content might be overrepresented. This bias could, in turn, lead to an underrepresentation of the more radical groupings within the collectives under scrutiny. Thus, an avenue for future research would entail incorporating other types of ties as building blocks for constructing the networks that represent netdoms and the foundation for detecting cultural and social patterns.

Fifth, while we incorporated the notion of offline events within our data collection, it would be very interesting to evaluate how real-life events, such as new policy regulations and demonstrations important to the movement, influence style formation online by, for instance, shaping the emergence of certain opinion leaders. Hence, a promising avenue for future research would involve expanding the analysis of leadership dynamics and boundary processes in style formation in relation to real-life events.

References

Abdalla Mikhaeil, C., & Baskerville, R. L. (2024). Explaining online conspiracy theory radicalization: A second-order affordance for identity-driven escalation. *Information Systems Journal*, *34*(3), 711-735.

Ackland, R., & O'Neil, M. (2011). Online collective identity: The case of the environmental movement. *Social Networks*, *33*(3), 177-190.

Aladwani, A. M. (2015). Facilitators, characteristics, and impacts of Twitter use: Theoretical analysis and empirical illustration. *International Journal of Information Management*, *35*(1), 15-25.

Albert, S., & Whetten, D. (1985). Organizational Identity. *Research in Organizational Behavior*, *7*, 263-295.

Alexander, J. C. (2001). Toward a sociology of evil: Getting beyond modernist common sense about the alternative to "the good." In M. P. Lara (Ed.), *Rethinking evil: Contemporary perspectives* (pp. 153-172). University of California Press.

Alexander, J. C. (2006). *The civil sphere*. Oxford University Press.

Alexander, J. C., & Smith, P. (1993). The discourse of American civil society: A new proposal for cultural studies. *Theory and Society*, *22*(2), 151-207.

Azarian, G. R. (2005). *The general sociology of Harrison C. White: Chaos and order in networks*. Palgrave Macmillan.

Bakardjieva, M. (2015). Do clouds have politics? Collective actors in social media land. *Information, Communication & Society*, *18*(8), 983-990.

Barabási, A.-L. (2012). Luck or reason. *Nature*, *489*(7417), 507-508.

Barandiaran, X. E., Calleja-López, A., & Cozzo, E. (2020). Defining collective identities in technopolitical interaction networks. *Frontiers in Psychology*, *11*, Article 1549.

Barberá, P., Wang, N., Bonneau, R., Jost, J. T., Nagler, J., Tucker, J., & González-Bailón, S. (2015). The critical periphery in the growth of social protests. *PloS One*, *10*(11), Article e0143611.

Barkun, M. (2013). *A culture of conspiracy: Apocalyptic visions in contemporary America* (2nd ed.). University of California Press.

Basov, N., Breiger, R., & Hellsten, I. (2020). Socio-semantic and other dualities. *Poetics*, *78*, Article 101433.

Becatti, C., Caldarelli, G., Lambiotte, R., & Saracco, F. (2019). Extracting significant signal of news consumption from social networks: The case of Twitter in Italian political elections. *Palgrave Communications, 5*, Article 91.

Bennett, W. L., & Segerberg, A. (2012). The logic of connective action: Digital media and the personalization of contentious politics. *Information, Communication & Society*, *15*(5), 739-768.

Bennett, W. L., & Segerberg, A. (2013). *The logic of connective action: Digital media and the personalization of contentious politics*. Cambridge University Press.

Berger, P. L., & Luckmann, T. (1972). *The social construction of reality: A treatise in the sociology of knowledge*. Penguin.

Bhattacharya, S., Spann, B., & Agarwal, N. (2024). Solidarity to storming: Assessing the socio-technical factors behind modern social movements. *Proceedings of the 32nd European Conference on Information Systems*.

Blee, K. M. (2013). How options disappear: Causality and emergence in grassroots activist groups. *American Journal of Sociology*, *119*(3), 655-681.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, Article P10008.

Bogner, A. (2021). *Die Epistemisierung des Politischen: Wie die Macht des Wissens die Demokratie gefährdet*. Reclam.

Bogner, A., & Menz, W. (2021). Wissen und Werte im Widerstreit: Zum Verhältnis von Expertise und Politik in der Corona-Pandemie. *Leviathan*, *49*(1), 111-132.

Borgatti, S. P., & Foster, P. C. (2003). The network paradigm in organizational research: A review and typology. *Journal of Management*, *29*(6), 991-1013.

boyd, d. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *A networked self: Identity, community, and culture on social network sites* (pp. 39-58). Routledge.

Braccini, A. M., Sæbø, Ø., & Federici, T. (2019). From the blogosphere into the parliament: The role of digital technologies in organizing social movements. *Information and Organization*, *29*(3), Article 100250.

Büttner, S., & Laux, T. (2021). Umstrittene Expertise und die Wissensproblematik der Politik: Eine Einführung. *Leviathan*, *49*(Suppl. 38), 13-40.

Byrne, D. (2009). Case-based methods: Why we need them; What they are; How to do them. In D. Byrne & C. C. Ragin (Eds.), *The SAGE Handbook of case-based methods* (pp. 1-10). SAGE.

Cardoso, A., Boudreau, M.-C., & Carvalho, J. Á. (2019). Organizing collective action: Does information and communication technology matter? *Information and Organization*, *29*(3), Article 100256.

Caren, N., Andrews, K. T., & Lu, T. (2020). Contemporary social movements in a hybrid media environment. *Annual Review of Sociology*, *46*, 443-465.

Castells, M. (2015). *Networks of outrage and hope: Social movements in the internet age* (2nd ed.). Polity Press.

Chamakiotis, P., Petrakaki, D., & Pentali, N. (2021). Social value creation through digital activism in an online health community. *Information Systems Journal*, *31*(1), 94-119.

Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*, *8*(1), Article 15951.

Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A., & Caldarelli, G. (2019). The statistical physics of real-world networks. *Nature Review Physics, 1*, 58-71.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661-703.

Coretti, L., & Pica, D. (2015). The rise and fall of collective identity in networked movements: Communication protocols, Facebook, and the anti-Berlusconi protest. *Information, Communication & Society*, *18*(8), 951-967.

Dahrendorf, R. (1972). *Konflikt und Freiheit: Auf dem Weg zur Dienstklassengesellschaft*. Piper.

della Porta, D. (2005). Multiple belongings, tolerant identities, and the construction of “another politics”: Between the European Social Forum and the local social fora. In D. della Porta & S. Tarrow (Eds.), *Transnational protest and global activism* (pp. 175-202). Rowman & Littlefield.

della Porta, D., & Diani, M. (2006). *Social movements: An introduction* (2nd ed.). Blackwell.

Diani, M. (1992). The concept of social movement. *Sociological Review*, *40*(1), 1-25.

Diani, M. (2002). Network analysis. In B. Klandermans & S. Staggenborg (Eds.), *Methods of social movement research* (pp. 173-200). University of Minnesota Press.

Diani, M. (2003). “Leaders” or brokers? Positions and influence in social movement networks. In M. Diani & D. McAdam (Eds.), *Social movements and networks: Relational approaches to collective action* (pp. 105-122). Oxford University Press.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, *48*(2), 147-160.

Earl, J. (2015). The future of social movement organizations: The waning dominance of SMOs online. *American Behavioral Scientist*, *59*(1), 35-52.

Earl, J. (2018). Technology and social media. In D. A. Snow, S. A. Soule, H. Kriesi, & H. J. McCammon (Eds.), *The Wiley Blackwell companion to social movements* (2nd ed., pp. 289-305). John Wiley & Sons.

Eder, K. (2009). A theory of collective identity making sense of the debate on a “European identity.” *European Journal of Social Theory*, *12*(4), 427-447.

Edwards, B., & McCarthy, J. D. (2004). Resources and social movement mobilization. In D. A. Snow, S. A. Soule, & H. Kriesi (Eds.), *The Blackwell companion to social movements* (pp. 116-152). Blackwell.

Eisenhardt, K. M. (1989). Building theories from case study research. *The Academy of Management Review*, *14*(4), 532-550.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *The Academy of Management Journal*, *50*(1), 25-32.

Emirbayer, M., & Goodwin, J. (1994). Network analysis, culture, and the problem of agency. *American Journal of Sociology*, *99*(6), 1411-1454.

Falkowski, T. (2009). *Community analysis in dynamic social networks* (Doctoral Dissertation). Otto-von-Guericke-Universität Magdeburg.

Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization Science*, *22*(5), 1224-1239.

Faraj, S., Kudaravalli, S., & Wasko, M. (2015). Leading collaboration in online communities. *MIS Quarterly*, *39*(2), 393-412.

Fayard, A.-L., & DeSanctis, G. (2008). Kiosks, clubs and neighborhoods: The language games of online forums. Journal of the Association for Information Systems, 9(10), 677-705.

Fayard, A.-L., & DeSanctis, G. (2010). Enacting language games: The development of a sense of "we-ness" in online forums. *Information Systems Journal*, *20*(4), 383-416.

Flesher Fominaya, C. (2010). Collective identity in social movements: Central concepts and debates. *Sociological Compass*, *4*(6), 393-404.

Flesher Fominaya, C. (2019). Collective identity in social movements: Assessing the limits of a theoretical framework. In D. A. Snow, S. A. Soule, H. Kriesi, & H. J. McCammon (Eds.), *The Wiley Blackwell companion to social movements* (2nd ed., pp. 429-445). John Wiley & Sons.

Fligstein, N., & McAdam, D. (2012). *A theory of fields*. Oxford University Press.

Fontdevila, J. (2018). Switchings among netdoms: The relational sociology of Harrison C. White. In F. Dépelteau (Ed.), *The Palgrave Handbook of relational sociology* (pp. 231-269). Palgrave Macmillan.

Frei, N., Schäfer, R., & Nachtwey, O. (2021). Die Proteste gegen die Corona-Maßnahmen: Eine soziologische Annäherung. *Forschungsjournal Soziale Bewegungen*, *34*(2), 249-258.

Gamson, W. A. (1991). Commitment and agency in social movements. *Sociological Forum*, *6*(1), 27-50.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. MIT Press.

George, J. J., & Leidner, D. E. (2019). From clicktivism to hacktivism: Understanding digital activism. *Information and Organization*, *29*(3), Article 100249.

Gerbaudo, P. (2014). The persistence of collectivity in digital protest. *Information, Communication & Society*, *17*(2), 264-268.

Gerbaudo, P. (2015). Protest avatars as memetic signifiers: Political profile pictures and the construction of collective identity on social media in the 2011 protest wave. *Information, Communication & Society*, *18*(8), 916-929.

Gerbaudo, P. (2024). From individual affectedness to collective identity: Personal testimony campaigns on social media and the logic of collection. N*ew Media & Society, 26*(8), 4904-4921.

Gerbaudo, P., & Treré, E. (2015). In search of the 'we' of social media activism: Introduction to the special issue on social media and protest identities. *Information, Communication & Society*, *18*(8), 865-871.

Gerring, J. (2004). What is a case study and what is it good for? *American Political Science Review*, *98*(2), 341-354.

Ghobadi, S., & Clegg, S. (2015). "These days will never be forgotten…": A critical mass approach to online activism. *Information and Organization*, *25*(1), 52-71.

Giddens, A. (1984). *The constitution of society.* Polity Press.

Giddens, A. (1991). *Modernity and self-identity*. Stanford University Press.

Gieryn, T. F. (1983). Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review*, *48*(9), 781-795.

Gioia, D. A., Patvardhan, S. D., Hamilton, A. I., & Corley, K. G. (2013). Organizational identity formation and change. *The Academy of Management Annals*, *7*(1), 123-193.

Gioia, D. A., Schultz, M., & Corley, K. G. (2000). Organizational identity, image and adaptive instability. *The Academy of Management Review*, *25*(1), 63-81.

Gladwell, M. (2010). Small change: Why the revolution will not be tweeted. *The New Yorker*. https://www.newyorker.com/magazine/2010/10/04/small-change-malcolm-gladwell

Godart, F. C., & White, H. C. (2010). Switchings under uncertainty: The coming and becoming of meanings. *Poetics*, *38*(6), 567-586.

Goertz, G., & Mahoney, J. (2009). Scope in case study research. In D. Byrne & C. C. Ragin (Eds.), *The SAGE Handbook of case-based methods* (pp. 307-317). SAGE.

Goffman, E. (1969). *The presentation of self in everyday life*. Penguin Press.

Goffman, E. (1979). Footing. *Semiotica*, *25*(1-2), 1-29.

Grande, E., Hutter, S., Hunger, S., & Kanol, E. (2021). *Alles Covidioten? Politische Potenziale des Corona-Protests in Deutschland* (ZZ 2021-601; WZB Discussion Paper). Wissenschaftszentrum Berlin für Sozialforschung.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611-642.

Gruber, H. (2017). Quoting and retweeting as communicative practices in computer mediated discourse. *Discourse, Context & Media*, *20*(6), 1-9.

Hardy, C., Lawrence, T. B., & Grant, D. (2005). Discourse and collaboration: The role of conversations and collective identity. *The Academy of Management Review, 30*(1), 58-77.

Harysi, M. A., Negoita, B., & Nandhakumar, J. (2019). The evolution of leadership structures in online communities: A social network perspective. *Proceedings of the 27th European Conference on Information Systems*.

Haug, C. (2013). Organizing spaces: Meeting arenas as a social movement infrastructure between organization, network, and institution. *Organization Studies*, *34*(5-6), 705-732.

Haug, C., & Rucht, D. (2013). Structurelessness: An evil or an asset? In D. della Porta & D. Rucht (Eds.), *Meeting democracy: Power and deliberation in Global Justice Movements* (pp. 179-213). Cambridge University Press.

Henn, T., & Posegga, O. (2023). Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit. *PLoS One*, *18*(8), e0288962.

Hippert, J., & Saul, P. (2021). Von Stuttgart bis auf die Treppe des Reichstagsgebäudes. *Süddeutsche Zeitung*. https://www.sueddeutsche.de/politik/querdenken-chronologie-bundestag-1.5279496

Holzer, B., Koos, S., Meyer, C., Otto, I., Pannreck, I., & Reichardt, S. (2021). Einleitung: Protest in der Pandemie. In S. Reichardt (Ed.), *Die Misstrauensgemeinschaft der Querdenker: Die Corona-Proteste aus kultur- und sozialwissenschaftlicher Perspektive* (pp. 7-26). Campus.

Homans, G. C. (1961). *Social behavior: Its elementary forms*. Harcourt, Brace & World.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 767-797.

Hunt, S. A., & Benford, R. D. (2004). Collective identity, solidarity, and commitment. In D. A. Snow, S. A. Soule, & H. Kriesi (Eds.), *The Blackwell companion to social movements* (pp. 433-457). Blackwell.

Jackson, M. O. (2010). *Social and economic networks*. Princeton University Press.

Jarvenpaa, S. L., & Selander, L. (2023). Between scale and impact: Member prototype ambiguity in digital transformation. *European Journal of Information Systems*, *32*(3), 390-408.

Jarynowskyi, A., Semenov, A., & Belik, V. (2020). Protest perspective against COVID-19 risk mitigation strategies on the German internet. *Proceedings of the International Conference on Computational Data and Social Networks*.

Johnson, S. L., Faraj, S., & Kudaravalli, S. (2014). Emergence of power laws in online communities: The role of social mechanisms and preferential attachment. *MIS Quarterly*, *38*(3), 795-808.

Johnson, S. L., Safadi, H., & Faraj, S. (2015). The emergence of online community leadership. *Information Systems Research*, *26*(1), 165-187.

Jost, P., & Dogruel, L. (2023). Radical mobilization in times of crisis: Use and effects of appeals and populist communication features in Telegram channels. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231186372

Jungherr, A. (2015). *Analyzing political communication with digital trace data: The role of Twitter messages in social science research*. Springer.

Kane, G. C., Alavi, M., Labianca, G., & Borgatti, S. P. (2014). What's different about social media networks? A framework and research agenda. *MIS Quarterly*, *38*(1), 275-304.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research: Past, present and future. *Information Systems Frontiers*, *20*(3), 513-558.

Karahanna, E., Xin Xu, S., Xu, Y., & Zhang, N. (Andy). (2018). The needs-affordances-features perspective for the use of social media. *MIS Quarterly*, *42*(3), 737-756.

Kavada, A. (2009). Email lists and the construction of an open and multifaceted identity: The case of the London 2004 European Social Forum. *Information, Communication & Society, 12*(6), 817-839.

Kavada, A. (2015). Creating the collective: Social media, the Occupy movement and its constitution as a collective actor. *Information, Communication & Society, 18*(8), 872-886.

Kern, T., & Nam, S. (2009). The making of a social movement: Citizen journalism in South Korea. *Current Sociology*, *57*(5), 637-660.

Kern, T., Opitz, D., Polenz, J., Pruisken, I., & Tell, S. (2023). A peripheral movement? Querdenken in the critical public discourse about the COVID-19 pandemic. *German Politics and Society*, *41*(2), 16-34.

Khazraee, E., & Novak, A. N. (2018). Digitally mediated protest: Social media affordances for collective identity construction. *Social Media + Society*, *4*(1). https://doi.org/10.1177/2056305118765740

Klandermans, B. (2014). Identity politics and politicized identities: Identity processes and the dynamics of protest. *Political Psychology*, *35*(1), 1-22.

Kohout, S., Kruikemeier, S., & Bakker, B. N. (2023). May I have your attention, please? An eye tracking study on emotional social media comments. *Computers in Human Behavior*, *139*, Article 107495.

Koos, S. (2021a). *Die "Querdenker:" Wer nimmt an Corona-Protesten teil und warum? Ergebnisse einer Befragung während der "Corona-Proteste" am 4.10.2020 in Konstanz.* (Working paper). Universität Konstanz.

Koos, S. (2021b). Konturen einer heterogenen "Misstrauensgemeinschaft": Die soziale Zusammensetzung der Corona-Proteste und die Motive ihrer Teilnehmer:innen. In S. Reichardt (Ed.), *Die Misstrauensgemeinschaft der "Querdenker": Die Corona-Proteste aus kultur- und sozialwissenschaftlicher Perspektive* (pp. 67-89). Campus.

Kriesi, H., Wüest, B., Lorenzini, J., Makarov, P., Enggist, M., Rothenhäusler, K., Kurer, T., Häusermann, S., Wangen, P., Altiparmakis, A., Borbáth, E., Bremer, B., Gessler, T., Hunger, S., Hutter, S., Schulte-Cloos, J., & Wang, C. (2020). *PolDem-protest dataset 30 European countries: Version 1* (Working paper). European University Institute; Wissenschaftszentrum Berlin für Sozialforschung.

Kumkar, N. C. (2021). Unbestimmte Negation: Gegenexpertise als alternative Fakten. *Leviathan*, *49*(Suppl. 38), 90-110.

Lamont, M., & Molnár, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology*, *28*, 167-195.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, *80*(5), Article 056117.

Langer, N., Khurana, S., & Vaast, E. (2020). Activating the sisterhood: A structural and temporal analysis of sustained connective action in #MeTooIndia. *Proceedings of the International Conference on Information Systems*.

Lee, J., Yang, C.-S., Hsu, C., & Wang, J.-H. (2018). Emerging leaderships in an online community: A longitudinal network analysis. *Proceedings of the 51st Hawaii International Conference on System Science*.

Lee, M., & Martin, J. L. (2018). Doorway to the dharma of duality. *Poetics*, *68*, 18-30.

Leidner, D. E., & Gregory, R. W. (2024). About theory and theorizing. *Journal of the Association for Information Systems*, *25*(3), 501-521.

Leonardi, P. M. (2011). When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS Quarterly*, *35*(1), 147-167.

Leonardi, P. M. (2012). Materiality, sociomateriality, and socio-technical systems: What do these terms mean? How are they different? Do we need them? In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), *Materiality and organizing: Social interaction in a technological world* (pp. 24-48). Oxford University Press.

Leonardi, P. M., & Vaast, E. (2017). Social media and their affordances for organizing: A review and agenda for research. *Academy of Management Annals*, *11*(1), 150-188.

Leong, C., Faik, I., Tan, F. T. C., Tan, B., & Khoo, Y. H. (2020). Digital organizing of a global social movement: From connective to collective action. *Information and Organization*, *30*(4), Article 100324.

Leong, C., Pan, S. L., Bahri, S., & Fauzi, A. (2015). Digital enablement: Social media in empowering the grassroot environmental movement of Malaysia. *Proceedings of the International Conference on Information Systems*.

Leong, C., Pan, S. L., Bahri, S., & Fauzi, A. (2019). Social media empowerment in social movements: Power activation and power accrual in digital activism. *European Journal of Information Systems*, *28*(2), 173-204.

Levina, N., & Arriaga, M. (2014). Distinction and status production on user-generated content platforms: Using Bourdieu's theory of cultural production to understand social dynamics in online fields. *Information Systems Research*, *25*(3), 468-488.

Luhmann, N. (2013). *Theory of social systems: Volume 2* (Trans. R. Barrett). Stanford University Press. (Original work published 1997)

Lundgaard, D., Razmerita, L., & Tan, C.-W. (2018). Emergent collaboration on Twitter: A case study of the #BlackLivesMatter movement. *Proceedings of the 51st Hawaii International Conference on System Science.*

Maghrabi, R., & Salam, A. F. (2011). Social media, social movement and political change: The case of 2011 Cairo revolt. *Proceedings of the International Conference on Information Systems*.

Majchrzak, A., Faraj, S., Kane, G. C., & Azad, B. (2013). The contradictory influence of social media affordances on online communal knowledge sharing. *Journal of Computer-Mediated Communication*, *19*(1), 38-55.

Marsili, N. (2021). Retweeting: Its linguistic and epistemic value. *Synthese*, *198*(11), 10457-10483.

Marwell, G., Oliver, P. E., & Prahl, R. (1988). Social Networks and Collective Action: A Theory of the Critical Mass III. *The American Journal of Sociology*, *94*(3), 502-534.

Massa, F. G., & O'Mahony, S. (2021). Order from chaos: How networked activists self-organize by creating a participation architecture. *Administrative Science Quarterly*, *66*(4), 1037-1083.

McCall, G. J., & Simmons, J. L. (1978). *Identities and interactions: An examination of human associations in everyday life*. Free Press.

McKenna, B. (2020). Creating convivial affordances: A study of virtual world social movements. *Information Systems Journal*, *30*(1), 185-214.

Melucci, A. (1995). The process of collective identity. In H. Johnston & B. Klandermans (Eds.), *Social movements and culture* (pp. 41-63). University of Minnesota Press.

Melucci, A. (1996). *Challenging codes: Collective action in the information age*. Cambridge University Press.

Merton, R. K. (1988). The Matthew effect in science, II. Cumulative advantage and the symbolism of intellectual property. *Isis*, *79*(4), 606-623.

Milan, S. (2015). When algorithms shape collective action: Social media and the dynamics of cloud protesting. *Social Media + Society*, *1*(2). https://doi.org/10.1177/2056305115622481

Miranda, S. M., Young, A., & Yetgin, E. (2016). Are social media emancipatory or hegemonic? Societal effects of mass media digitization in the case of the SOPA discourse. *MIS Quarterly*, *40*(2), 303-330.

Mische, A. (2008). *Partisan publics: Communication and contention across Brazilian youth activist networks*. Princeton University Press.

Mische, A., & White, H. C. (1998). Between conversation and situation: Public switching dynamics across network domains. *Social Research*, *65*(3), 695-724.

Mohr, J. W., & White, H. C. (2008). How to model an institution. *Theory and Society*, *37*(5), 485-512.

Monterde, A., Calleja-López, A., Aguilera, M., Barandiaran, X. E., & Postill, J. (2015). Multitudinous identities: A qualitative and network analysis of the 15M collective identity. *Information, Communication & Society*, *18*(8), 930-950.

Morozov, E. (2011). *The net delusion: The dark side of internet freedom*. Public Affairs.

Moutidis, I., & Williams, H. T. P. (2021). Community evolution on Stack Overflow. *PLOS One, 16(*6), e0253010.

Mundt, M., Ross, K., & Burnett, C. M. (2018). Scaling social movements through social media: The case of Black Lives Matter. *Social Media + Society*, *4*(4). https://doi.org/10.1177/2056305118807911

Nachtwey, O., Schäfer, R., & Frei, N. (2020). *Politische Soziologie der Corona-Proteste* (Working paper). Universität Basel.

Norris, P., Walgrave, S., & Van Aelst, P. (2005). Who demonstrates? Antistate rebels, conventional participants, or everyone? *Comparative Politics*, *37*(2), 189-205.

Oh, O., Eom, C., & Rao, H. R. (2015). Research note: Role of social media in social change: An analysis of collective sense making during the 2011 Egypt revolution. *Information Systems Research*, *26*(1), 210-223.

Pachucki, M. A., & Breiger, R. L. (2010). Cultural holes: Beyond relationality in social networks and culture. *Annual Review of Sociology*, *36*, 205-224.

Pantenburg, J., & Sepp, B. (2021). Wissen, hausgemacht: Selbstverständnis, Expertisen und Hausverstand der Querdenker. *Leviathan*, *49*(Suppl. 38), 468-482.

Priante, A. (2021). Temporal evolution of communication networks in social movements: Evidence from the November health campaign on Twitter. *Proceedings of the International Conference on Information Systems*.

Priante, A., Ehrenhard, M. L., van den Broek, T., & Need, A. (2018). Identity and collective action via computer-mediated communication: A review and agenda for future research. *New Media & Society*, *20*(7), 2647-2669.

Radicioni, T., Saracco, F., Pavan, E., & Squartini, T. (2021a). Analysing Twitter semantic networks: The case of 2018 Italian elections. *Scientific Reports*, *11*(1), Article 13207.

Radicioni, T., Squartini, T., Pavan, E., & Saracco, F. (2021b). Networked partisanship and framing: A socio-semantic network analysis of the Italian debate on migration. *PloS One*, *16*(8), Article e0256705.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. University of California Press.

Rainie, L., & Wellman, B. (2012). *Networked: The new social operating system*. MIT Press.

Ray, D., & Tarafdar, M. (2017). How does Twitter influence a social movement? *Proceedings of the 25th European Conference on Information Systems*.

Ren, Y., Harper, F. M., Drenner, S., Terveen, L., Kiesler, S., Riedl, J., & Kraut, R. E. (2012). Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *MIS Quarterly*, *36*(3), 841-864.

Reyes-Menendez, A., Saura, J. R., & Thomas, S. B. (2020). Exploring key indicators of social identity in the #MeToo era: Using discourse analysis in UGC. *International Journal of Information Management*, *54*, Article 102129.

Rogers, E. M. (1995). *Diffusion of innovations*. Free Press.

Rucht, D. (1995). Kollektive Identität: Konzeptionelle Überlegungen zu einem Desiderat der Bewegungsforschung. *Forschungsjournal Neue Soziale Bewegungen*, *8*(1), 9-23.

Rucht, D. (2011). The strength of weak identities. *Forschungsjournal Soziale Bewegungen*, *24*(4), 73-84.

Rucht, D. (2023a). *Kollektiver Protest und soziale Bewegungen*. Beltz Juventa.

Rucht, D. (2023b). *Social movements: A theoretical approach*. Oxford University Press.

Rucht, D., & Neidhardt, F. (1998). Methodological issues in collecting protest event data: Units of analysis, sources and sampling, coding problems. In D. Rucht, R. Koopmans, & F. Neidhardt (Eds.), *Acts of dissent: New developments in the study of protest* (pp. 65-89). Edition Sigma.

Sæbø, Ø., Federici, T., & Braccini, A. M. (2020). Combining social media affordances for organising collective action. *Information Systems Journal*, *30*(4), 699-732.

Saracco, F., Di Clemente, R., Gabrielli, A., & Squartini, T. (2015). Randomizing bipartite networks: The case of the world trade web. *Scientific Reports*, *5*, Article 10595.

Saracco, F., Straka, M. J., Di Clemente, R., Gabrielli, A., Caldarelli, G., & Squartini, T. (2017). Inferring monopartite projections of bipartite networks: An entropy-based approach. *New Journal of Physics*, *19*(5), Article 053022.

Scharpf, F. W. (1997). *Games real actors play: Actor-centered institutionalism in policy research*. Westview.

Schimank, U. (2005). From "clean" mechanisms to "dirty" models: Methodological perspectives of an up-scaling of actor constellations. In K. Fischer, M. Florian, & T. Malsch (Eds.), *Socionics: Scalability of complex social systems* (pp. 15-35). Springer.

Schreier, M. (2012). *Qualitative content analysis in practice*. SAGE.

Schreier, M. (2014). Varianten qualitativer Inhaltsanalyse: Ein Wegweiser im Dickicht der Begrifflichkeiten. *Forum Qualitative Sozialforschung*, *15*(1), Article 18.

Schultz, J., & Breiger, R. L. (2010). The strength of weak culture. *Poetics*, *38*, 610-624.

Schwarz, O. (2021). *Sociological theory for digital society: The codes that bind us together*. John Wiley & Sons.

Selander, L., & Jarvenpaa, S. L. (2016). Digital action repertoires and transforming a social movement organization. *MIS Quarterly*, *40*(2), 331-352.

Shannon, C. E. (1948). A mathematical theory of communication. T*he Bell System Technical Journal, 27*(3), 379-423.

Simmel, G. (1955). The web of group affiliations. In G. Simmel, *Conflict and the web of group affiliations* (pp. 125-195). Free Press.

Simon, B., & Klandermans, B. (2001). Politicized collective identity: A social psychological analysis. *American Psychologist*, *56*(4), 319-331.

Snow, D. A. (2001). *Collective identity and expressive forms* (Working Paper). Center for the Study of Democracy, UC Irvine.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *PNAS*, *116*(38), 18888-18892.

Soule, S. A. (2004). Diffusion processes within and across movements. In D. A. Snow, S. A. Soule, & H. Kriesi (Eds.), *The Blackwell companion to social movements* (pp. 294-309). Blackwell.

Stewart, M., & Schultze, U. (2017). A performative identity perspective of cyberactivism: The case of My Stealthy Freedom. *Proceedings of the International Conference on Information Systems*.

Stewart, M., & Schultze, U. (2019). Producing solidarity in social media activism: The case of My Stealthy Freedom. *Information and Organization*, *29*(3), Article 100251.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. *Journal of*

Management Information Systems, *29*(4), 217-248.

Straka, M. J., Caldarelli, G., & Saracco, F. (2017). Grand canonical validation of the bipartite international trade network. *Physical Review E, 96*(2), Article 022306.

Strang, D., & Soule, S. A. (1998). Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, *24*, 265-290.

Strong, D. M., Volkoff, O., Johnson, S. A., Pelletier, L. R., & Tulu, B. (2014). A theory of organization: EHR affordance actualization. *Journal of the Association for Information Systems*, *15*(2), 53-85.

Sunstein, C. R., & Vermeule, A. (2008). C*onspiracy theories* (University of Chicago Public Law & Legal Theory Working Paper). The University of Chicago.

Syed, R., & Silva, L. (2020). Core and peripheral users and social movement mobilization. *Proceedings of the International Conference on Information Systems*.

Syed, R., & Silva, L. (2023). Social movement sustainability on social media: An analysis of the Women's March movement on Twitter. *Journal of the Association for Information Systems*, *24*(1), 249-293.

Tarafdar, M., & Ray, D. (2021). Role of social media in social protest cycles: A sociomaterial examination. *Information Systems Research*, *32*(3), 1066-1090.

Tarrow, S. (2011). *Power in movement: Social movements and contentious politics* (3rd ed.). Cambridge University Press.

Taylor, V., & Whittier, N. E. (1992). Collective identity in social movement communities: Lesbian feminist mobilization. In A. D. Morris & C. McClurg Mueller (Eds.), *Frontiers in social movement theory* (pp. 104-129). Yale University Press.

Theocharis, Y., Lowe, W., van Deth, J. W., & Garcá-Albacete, G. (2015). Using Twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society*, *18*(2), 202-220.

Tilly, C. (1978). *From mobilization to revolution*. Random House.

Tilly, C. (2003). Political identities in changing polities. *Social Research*, *70*(2), 605-619.

Touraine, A. (1985). An introduction to the study of social movements. *Social Research*, *52*(4), 749-787.

Treem, J. W., & Leonardi, P. M. (2012). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association*, *36*(1), 143-189.

Treré, E. (2015). Reclaiming, proclaiming, and maintaining collective identity in the #YoSoy132 movement in Mexico: An examination of digital frontstage and backstage activism through social media and instant messaging platforms. *Information, Communication & Society*, *18*(8), 901-915.

Tsai, H.-T., & Bagozzi, R. P. (2014). Contribution behavior in virtual communities: Cognitive, emotional, and social influences. *MIS Quarterly*, *38*(1), 143-164.

Tsugawa, S., & Ohsaki, H. (2015). Negative messages spread rapidly and widely on social media. *Proceedings of the ACM on Conference on Online Social Networks*.

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Vaast, E., Davidson, E. J., & Mattson, T. (2013). Talking about technology: The emergence of a new actor category through new media. *MIS Quarterly*, *37*(4), 1069-1092.

Vaast, E., Safadi, H., Lapointe, L., & Negoita, B. (2017). Social media affordances for connective action. *MIS Quarterly*, *41*(4), 1179-1206.

Vajapeyam, S. (2014). *Understanding Shannon's entropy metric for information*. arXiv. https://arxiv.org/abs/1405.2061

van Haperen, S., & Uitermark, J. (2023). Dynamics of digitally networked leadership in #blacklivesmatter. *First Monday*, *28*(8). https://doi.org/10.5210/fm.v28i8.12936

Van Laer, J., & Van Aelst, P. (2010). Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication & Society*, *13*(8), 1146-1171.

Venkatesan, S., Valecha, R., Yaraghi, N., Oh, O., & Rao, H. R. (2021). Influence in social media: An investigation of tweets spanning the 2011 Egyptian revolution. *MIS Quarterly*, *45*(3), 1679-1714.

Virchow, F., Häusler, A., & Döring, M. (2020). *Pandemie-Leugnung und extreme Rechte in Nordrhein-*

Westfalen (CoRE-NRW Kurzgutachten 3). Bonn International Center for Conversion.

Weisskircher, M. (2024). Far-right parties and divisions over movement-party strategy: The AfD and the anti-Corona protests of Querdenken. In M. Weisskircher (Ed.), *Contemporary Germany and the fourth wave of far-right politics: From the streets to parliament* (pp. 159-173). Routledge.

White, H. C. (1997). Can mathematics be social? Flexible representations for interaction process and its sociocultural constructions. *Sociological Forum*, *12*(1), 53-71.

White, H. C. (2008). *Identity and control: How social formations emerge* (2nd ed.). Princeton University Press.

White, H. C., Godart, F. C., & Corona, V. P. (2007). Mobilizing identities: Uncertainty and control in strategy. *Theory, Culture & Society*, *24*(7-8), 181-202.

Winter, H., Gerster, L., Helmer, J., & Baaken, T. (2021). *Disinformation overdose: A study of the crisis of trust among vaccine sceptics and anti-vaxxers*. Institute for Strategic Dialogue.

Wry, T., Lounsbury, M., & Glynn, M. A. (2011). Legitimating nascent collective identities: Coordinating cultural entrepreneurship. *Organization Science*, *22*(2), 449-463.

Young, A., Selander, L., & Vaast, E. (2019). Digital organizing for social impact: Current insights and future research avenues on collective action, social movements, and digital technologies. *Information and Organization*, *29*(3), Article 100257.

Zhang, K., Bhattacharyya, S., & Ram, S. (2016). Large-scale network analysis for online social brand advertising. *MIS Quarterly*, *40*(4), 849-868.

Appendix A: Literature Review

Table A1: Literature Summary: Empirical Investigation of Collective Identity Formation of Social Movements Online

Study	Approach	Central definition of collective identity (CI)	Data and empirical analysis	Key empirical findings on collective identity (CI) formation online
Ackland & O'Neil (2011)	Focus on network dimension	"Collective identity is a mutually agreed upon definition of membership, boundaries, activities and norms of behavior used to characterize a grouping of actors. … Melucci (1995) has argued that collective identity is the interactive and shared process by which social actors come together to form a collectivity: this process involves shared cognitive definitions such as language and rituals; networks of active relationships between actors; and a degree of emotional investment which enables people to feel part of a common unity" (p. 178).	Network analysis based on two separate types of networks: hyperlink networks among social movement organizations' (SMO) website and frame networks based on SMO websites' content.	Supports existence of CI online: "our model emphasizes the importance for online SMOs of participation in informal networks and direct control over the means of communication, both of which favor the preeminence of expressive behavior leading to the formation of collective identity" (p. 187).
Barandiaran et al. (2020)	Focus on network dimension	"a collective identity is the interaction network that is both the result and the source of recurrent, cohesive, and coordinated communicative interactions between different agents across different communication spaces, distinguishing itself from the environment and other identities within a communication scope" (p. 6).	Network analyses based on public interaction on Facebook and Twitter.	Supports existence of CI online: "We believe that digital platforms both mediate and simplify the ways in which social identities emerge" (p. 19).
Bhattacharya et al. (2024)	Focus on network and cultural dimension but no dual approach	"The notion of collective identity was introduced by Alberto Melucci in the realm of social movement theory. Collective identity, also referred to as group identity, entails a shared sense of affiliation to a particular group" (p. 3).	Quantitative examination of Twitter data on the Brazil presidential election and the US presidential election to analyze the Brazil anti-government protest campaign and the U.S. Capitol Riots by evaluating the "inter-topic distance and network modularity across a temporally evolving network" (p. 5).	The findings of the paper show that collective identity plays a pivotal role in sustaining social movements with "communities that share a common interest demonstrat[ing] significantly higher coherence in both similarity of content and network cohesiveness" (p. 11).
Coretti & Pica (2015)	Focus on network and cultural dimension but no dual approach	"According to Melucci (1995, p. 45), 'collective identity as a process refers … to a network of active relationships between the actors, who interact, communicate, influence each other, negotiate, and make decisions'" (p. 952).	Case study combining "a quantitative overview of patterns of membership and interaction" (p. 954) and content analysis based on a social movement's Facebook page.	While the paper generally recognizes that collective identity formation might take place online, it "suggests that the communication protocols of commercial social networking media lead to organizational centralization and fragmentation in social movements by eroding one of the preconditions of collective identity, solidarity" (p. 951).

Fayard & DeSanctis (2008)	Focus on cultural dimension	*No definition provided*	Qualitative discourse analysis based on three online forums unrelated to protest.	Supports existence of CI online: "Our discourse analysis shows that each forum created … a more or less developed sense of collective identity" (p. 686).
Fayard & DeSanctis (2010)	Focus on cultural dimension	"a label associated with a set of shared attributes meaningful to the group" (p. 384).	Qualitative discourse analysis based on two online forums unrelated to protest.	Supports existence of CI online: "Our discourse analysis illustrates how, as they engage in the language game, participants produce a set of five discursive practices that define and enact their sense of we-ness" (p. 385).
Gerbaudo (2015)	Focus on cultural dimension	"collective 'sense of self', what binds together different individuals in a group, allowing them to be externally recognized as an 'actor', and to conceive of themselves internally as a collective subject with a unified intentionality (Touraine, 1988)" (p. 920-921).	Qualitative case study combining interviews, online observations, and social media content based on Facebook and Twitter posts focused on avatars/ pictures.	Supports existence of CI online: Protest avatars used across the analyzed movements proved to be a significant "forms of identification of protest movements" (p. 916) online by "temporarily revers[ing] the experience of individualization of digital connectivity" (p. 927).
Gerbaudo (2024)	Focus on cultural dimension	"Taylor and Whittier (1992: 172) have described collective identity as 'the shared definition of a group that derive from members' common interests, and solidarity'" (p. 4907).	Qualitative discourse analysis based on Twitter posts from "personal testimony campaigns."	Supports existence of CI online: CI emerges via personal testimony campaigns online because "personal *experiences* become the basis for a 'collective story-telling' … in a collective process in which 'the collective' is forged by 'collecting the personal'" (p. 4915, emphasis in original).
Hardy et al. (2005)	*No empirical approach*	CI "'addresses the 'we-ness' of a group, stressing the similarities or shared attributes around which group members coalesce' (Cerulo, 1997: 386)" (p. 61).	*No empirical analysis provided*	*No empirical results provided*
Kavada (2009)	Focus on cultural dimension	CI "is 'the (often implicitly) agreed upon definition of membership, boundaries, and activities for the group. … It is built through shared definitions of the situation by its members, and it is a result of a process of negotiation and 'laborious adjustment' of different elements relating to the ends and means of collective action and its relation to the environment' (Johnston et al. 1994, p. 15)" (p. 820).	Qualitative case study combining interviews and content analysis of three email lists.	Supports existence of CI online: "each [email] list constituted a different 'network of active relationships' (Melucci 1996, p. 71) or locus of collective identity, giving rise to different imaginings of the social context and its boundaries or the 'we' that was referred to" (p. 835).
Kavada (2015)	Focus on cultural dimension	"Melucci (1996) views collective identity as 'an interactive and shared definition produced by a number of individuals (or groups at a more complex level) concerning the orientations of their action and the field of opportunities and constraints in which such action is to take place' (p. 70)" (p. 874).	Qualitative case study based on interviews, document analysis, and social media content.	Supports existence of CI online: the author draws attention to the interplay of online and face-to-face identization. "Social media were used to broadcast and amplify the process of 'identization' taking place face-to-face" (p. 884) and, thus, aided in the process of identity formation.

Khazraee & Novak (2018)	Focus on cultural dimension	"Melucci (1985) states, 'Collective identity is … a shared definition of the field of opportunities and constraints offered to collective action …' (p. 793)" (p. 3-4).	Qualitative case study combining textual and visual analysis based on a social movement's Facebook posts.	Supports existence of CI online: The authors identify two types of affordances that contribute to the "formation of group identities" (p. 1) online.
Lundgaard et al. (2018)	*No empirical approach*	*No definition provided*	*Other empirical focus*	Supports existence of CI online: CI can manifest online, fostering "emergent collaboration" of a "united 'we'" (p. 2198).
Maghrabi & Salam (2011)	Focus on cultural dimension	*No definition provided*	Qualitative evaluation of tweets from the 2011 Cairo revolts.	Supports the existence of CI online: Finds the use of words such as "our," "we," and "us" in tweets, indicating that protesters "have perceived themselves as a one unit that thinks and feels in a same way" (p. 8).
Monterde et al. (2015)	Focus on network dimension	"He [Melucci] gave a system- and network-friendly definition of collective identity by considering it 'a network of active relationships between the actors, who interact, communicate, influence each other, negotiate and make decisions. Forms of organization and models of leadership, communicative channels and technologies of communication are constitutive parts of this network of relationships' (Melucci, 1995, pp. 44-45)" (p. 931).	Case study combining network analysis public interaction on Facebook, participant observation, and an online survey.	Supports existence of CI online: "we introduce a novel method for synchrony detection in Facebook activity to identify the distributed, yet integrated, coordinated activity behind collective identities. Applying this analytical strategy to the 15M movement, we show how it displays a specific form of systemic collective identity" (p. 930).
Priante (2021)	*No empirical approach*	"Social movement theory defines collective identity as 'a network of active relationships between the actors, who interact, communicate, influence each other, negotiate and make decisions' (Melucci 1995, p. 45)" (p. 4).	*Other empirical focus*	*Other thematic focus*
Priante et al. (2018)	*No empirical approach*	"Collective identity is a process that involves cognitive definitions of such goals and means of action, networks of relations between individuals who interact with each other, and a certain emotional investment that contributes to a sense of common unity (Melucci, 1995)" (p. 2650).	*No empirical analysis provided*	*No empirical results provided*
Ray & Tarafdar (2017)	Focus on cultural dimension	*No definition provided*	Qualitative content analysis of tweets in the context of the Delhi gang rape.	Supports existence of CI online: CI manifestation through the common usage of hashtags as well as shared concerns and beliefs ("identity framing") in tweets.
Stewart & Schultze (2017)	Focus on cultural dimension	"collective identity, that is, the movement members' understanding of what binds them together and motivates them to act (Polletta and Jasper 2001)" (p. 2).	Qualitative case study combining interviews and visual analysis based on a social movement's Facebook posts.	Supports existence of CI online: "Our findings indicate that the specific configuration of cyberactivism's affective sociomaterial practices produce … an individual's activist identity as well as the movement's collective identity" (p. 1).

Stewart & Schultze (2019)	*No empirical approach*	"Collective identity … is generally defined as 'a shared sense of 'oneness' or 'we-ness' anchored in either real or imagined shared attributes and experiences among those who comprise the collectivity and in relation or contrast to one or more actual or imagined sets of others' (Snow, 2001, p. 2). Melucci (1995), theorizes collective identity as a process by which the movement as a unified actor is formed" (p. 3).	*Other empirical focus*	*Other thematic focus*
Syed & Silva (2023)	*No empirical approach*	"Collective identity is a perception among individuals that they are members of a larger community by virtue of the grievances they share" (p. 254).	*Other empirical focus*	*Other thematic focus*
Treré (2015)	Focus on cultural dimension	"At its most basic level, collective identity is a shared sense of 'we-ness' and 'collective agency' (Snow, 2001). Italian sociologist Alberto Melucci describes it as 'an interactive and shared definition produced by several interacting individuals who are concerned with the orientation of their action as well as the field of opportunities and constraints in which their action takes place' (1995, p. 44)" (p. 903).	Ethnographic case study based on participant observation, interviews and content analysis based on a movement's website, Facebook and Twitter posts, internal communication via messenger, and official documents such as posters, leaflets, and manifestos.	Supports existence of CI online: "The findings point out how that internal cohesion and collective identity are fundamentally shaped and reinforced in the social media backstage by practices of 'ludic activism', which indicates that social media represent not only the organizational backbone of contemporary social movements, but also multifaceted ecologies where a new, expressive and humorous 'communicative resistance grammar' emerges" (p. 901).

Appendix B: Methodology

B1. Classifying Querdenken Content and Creating Final Querdenken User Sample

Based on previous findings (e.g., Nachtwey et al., 2020), we classified content as Querdenken-related if tweets contained either content that was (1) critical of the government or its imposed restrictions, (2) critical toward vaccination, (3) critical toward approved scientific findings, (4) posted by a central figure of the Querdenken milieu, (5) calls for joining a Querdenken demonstration, or (6) related to previous Querdenken events.

To obtain our final Querdenken sample, we analyzed the 25 most recent tweets from each of the Querdenken followee accounts on October 11, 2021, and categorized them according to the coding schema outlined earlier. Each tweet was marked with a "1" if it included content from one of the six categories above and a "0" if it did not. If at least one of the 25 tweets received a "1" classification, we labeled the profile as a Querdenken-related user. After classifying the Querdenken followees' profiles accordingly, we obtained 245 additional users associated with Querdenken. Combined with the 39 accounts maintained by the Querdenken local chapters, this yielded a list of 284 Twitter users; the banning or deplatforming of Twitter accounts during the period of analysis reduced this to 272 unique user accounts for our final analysis.

B2. Codebook for Topics of Domains

- **General coding rules:**
  - The entire tweet text (including hashtags) serves as a unit of analysis.
  - If possible, only one label should be assigned; the code that best summarizes the general topic of the tweet should be picked; multiple codes may be labeled if necessary.
  - If no code fits the tweet text, the tweet should be labeled "0." In this case, the level of politicization will also not be coded (= "99").
  - Codes are applied to tweets according to the following inclusion criteria and exclusion criteria:

Code	Inclusion criteria	Exclusion criteria
Attribution of blame	A tweet claims that a certain actor, country, or other social entity is guilty of causing the COVID-19 pandemic.	N/A
Collective action	A tweet reports on planned, current, or past protest events organized by Querdenken or related actors in Germany. Protest events are defined as "a collective, public action by a non-governmental actor who expresses criticism or dissent and articulates a societal or political demand" (Rucht & Neidhardt, 1998, p. 68). Collective action includes any form of protest, e.g., petitions, boycotts, and any other form it may take.	N/A
COVID-19 consequences for schools and pupils	A tweet explicitly addresses the negative consequences for schools and pupils, e.g., learning deficits and difficulties of home schooling.	Criticism of masks for children/pupils are coded as part of *criticism of hygiene measures*. General feelings of isolation and loneliness among children are coded as *criticism of hygiene measures*.
Criticism of climate activism	A tweet delegitimizes climate activism during the COVID-19 pandemic, criticize priorities of climate activists.	N/A
Criticism of conspiracy myths	A tweet delegitimizes or criticizes conspiracy myths in general or certain conspiracy myths explicitly.	N/A
Criticism of COVID-19 statistics	A tweet questions the informative value of infection rates or death rates as a basis for political measures. Reasons may be, e.g., apparent lack of validity of tests, criticism of tests in general, and criticism of guidelines for counting deaths due to COVID-19.	Criticism that goes beyond COVID-19 statistics and includes criticism about hygiene measures or political actions taken against COIVD-19 is coded as *criticism of hygiene measures* or *criticism of political authorities*.

Criticism of health administration	A tweet criticizes the situation in hospitals and nursing homes, e.g., lack of nursing staff and difficulties in the health system given the workload during the COVID-19 pandemic.	Criticism that addresses concrete hygiene measures applied in hospitals and nursing homes is coded *criticism of hygiene measures*.
Criticism of hygiene measures	A tweet criticizes COVID-19 measures (e.g., criticism of mandatory masks, lockdowns, and quarantine as unnecessary) as well as negative consequences of these protection measures (e.g., psychological stress, depression, and feelings of isolation). Measures/policies need to be addressed explicitly.	Downplaying the threat posed by COVID-19 is coded as *trivialization of COVID-19*. Calls for more rigid measures are coded as part of *underestimation of COVID-19 threat*. Criticism related to politicians or political institutions is coded as *criticism of political authorities*.
Criticism of political authorities	A tweet criticizes politicians and/or political institutions for their management of the pandemic and/or its consequences. Politicians/institutions and measures need to be named explicitly. Politicians and/or institutions may also be addressed colloquially or framed ideologically, such as in "#eu-komiker" ("EU comedian").	General dissatisfaction with hygiene measures without addressing a politician/an institution explicitly is coded as *criticism of hygiene measures*.
Criticism of science	A tweet criticizes individual scientists or scientific institutions, challenge scientific studies/results or lack of such studies. We consider the Robert-Koch-Institute (RKI) to be a scientific institution.	N/A
Criticism of vaccination	A tweet criticizes vaccines or delegitimizes vaccines against SARS-CoV-2.	Concerns/rumors regarding a possible mandatory vaccination are coded as part of *speculation about conspiracy*.
Economic crisis	A tweet criticizes the negative economic consequences of COVID-19 pandemic (and hygiene measures, especially lockdowns), e.g., rise in unemployment. A tweet addresses the need for financial aid for certain groups or economic actors.	N/A
Encroachment on civil rights	A tweet criticizes (disproportionate) encroachment of civil rights as a consequence of lockdowns or other COVID-19-related policies. Tweets are included only if they explicitly mention certain civil rights encroachments.	N/A
Future after COVID-19	A tweet addresses (uncertainty about) the post-COVID-19 future.	N/A
Globalization criticism	A tweet criticizes the opening of markets and international trade as well as international cooperation among states. A tweet expresses skepticism regarding immigration.	N/A
Information on COVID-19	A tweet features news-type reports that cover developments caused by COVID-19, e.g., COVID-19 infection rates (nationally, internationally), newly introduced or suspended measures against COVID-19 (nationally, internationally), developments of vaccines, and COVID-19 infections of prominent people such as politicians and others.	Expressions of an emotional stance, such as by combining infection rates in nursing homes with the hashtag "#falscherfokus" (wrong focus) are coded as part of *criticism of health administration.*
International collective action	This code applies the same inclusion criteria as the *collective action* code but focuses exclusively on protest events against COVID-19-related politics outside of Germany.	Reports of collective action within Germany are coded as *collective action*.

Lack of infrastructure	A tweet addresses infrastructure deficits, especially in terms of transportation.	N/A
Media bias	A tweet claims that there are systematic biases or misinformation about COVID-19 in Germany's public broadcast media and "mainstream" newspapers.	N/A
Self-awareness	A tweet features the feeling of self-awareness and "we-ness" of protesters against COVID-19 policies.	N/A
Social inequality	A tweet addresses growing inequality (national, international) or supply shortages for disadvantaged groups due to the pandemic.	N/A
Social interaction	A tweet addresses changes in social interaction resulting from the COVID-19 pandemic (e.g., difficulties in interactions due to masks and changing social climate).	Direct criticism of masks (or other forms of hygiene measures) is coded as *criticism of hygiene measures*.
Speculation about conspiracy	A tweet addresses presumed conspirators or uncover a presumed conspiracy. Apparent facts must be questioned in terms of appearances being deceptive, true causes or conspirators possibly being hidden and kept secret (Barkun, 2013): "a conspiracy theory can generally be counted as such if it is *an effort to explain some event or practice by reference to the machinations of powerful people, who have also managed to conceal their role*" (Sunstein & Vermeule, 2008, p. 4, emphasis in original). Hashtags that hint at common conspiracy myths among Querdenken, as listed by Virchow et al. (2020, pp. 25-30), i.e., conspiracies regarding QAnon, Bill Gates, etc.	N/A
Support alternative science and media	A tweet endorses alternative sources of information/"truth" (e.g., by enhancing the contributions of scientists who contradict "mainstream" scientists).	N/A
Support of other countries	A tweet supports COVID-19-related policies by countries other than Germany. Countries need to be addressed explicitly.	N/A
Trivialization of COVID-19	A tweet questions the extraordinary severity of the health risks posed by SARS-CoV-2, e.g., by equating COVID-19 with other, less-severe illnesses or by downplaying its health risk, framing the general social/political reaction regarding the pandemic as unnecessary, exaggerated panicking.	N/A
Underestimation of COVID-19 threat	A tweet claims that society/political actors underestimate the health risk posed by SARS-CoV-2, e.g., because they did not introduce measures necessary for pandemic prevention or they introduced measures too late.	N/A
Victimhood	A tweet claims that critics of the government or COVID-19 related policies are treated unfairly (in the media, in general); claim victimhood for their own position/group.	N/A
Other (= "0")	The object of reference or context is not clear. Sharing media links that cannot be assessed because the content was deleted.	Content is related to the COVID-19 pandemic and can be classified as one of the codes mentioned before.

B3. Codebook for Politicization of Domains

- **General coding rules:**
  - The entire tweet text (including hashtags) serves as a unit of analysis.
  - "1" indicates that a tweet is politicized, meaning that its content is described as a subject of political conflict; if a tweet does not exhibit any signs of politicization, it is assigned a value of "0."

Code	Description
1	A tweet is considered politicized if it contains any of the three elements for politicization defined by Simon and Klandermans (2001): (1) Raises awareness of shared grievances based on perceived injustices or inequalities arising "when social comparisons reveal that one's in-group is worse off than relevant out-groups" (p. 324); (2) Identifies an external opponent or enemy who is blamed for the group's predicament (p. 325); and (3) Involves society at large, compelling it to take sides in the power struggle (pp. 325-326).
0	A tweet features neutral coverage of COVID-19 related topics, as would typically be found in a news context.

B4. Entropy-Based Null Models

"The entropy of a variable is the 'amount of information' contained in the variable. … Similarly, the information in a variable is tied to the amount of surprise that value of the variable causes when revealed" (Vajapeyam, 2014, p. 1). The idea behind such filtering models, which originate from statistical physics (Cimini et al., 2019), is to compare the observation of a link between nodes i and j against properly defined benchmark models aimed at maximizing an entropy function. In our case, we maximized the Shannon (1948) entropy, which is "standard in information theory" (Straka et al., 2017, p. 3) and can be mathematically translated into the formula:

$$S = -\sum_{G_B \in \mathcal{G}_B} P(G_B) lnP(G_B)$$

where $\mathcal{G}_B$ stands for "an ensemble of bipartite graphs in which the number of links can vary, and the number of nodes per layer is fixed to the value of the network under analysis" (Straka et al., 2017, p. 3). $P(G_B)$ reflects the likelihood of the graph $G_B \in \mathcal{G}_B$ keeping in mind that the network is composed of a distinct node set $\top$ on the top layer and another distinct node set $\bot$ on the bottom layer. Based on this estimation, two nodes of a set $\top$ or set $\bot$ are linked if their connection through their neighbors (so-called V-motifs) is statistically significant. The significance level is estimated by calculating a p-value for each pair of nodes based on the comparison between observed V-motifs and V-motifs obtained within the benchmark model (for more information, see Straka et al., 2017). Thus, a new $m \times n$ matrix filled with the respective p-values for each node pair can be created. We linked nodes i and j if their calculated p-value satisfies the statistically common and significant threshold value of $p < 0.05$. For calculating the p-values of the BiCM null model, we used the code provided in the GitHub repository https://github.com/tsakim/bicm.

B5. Polarization Score

The validity of the detected communities can also be validated by testing whether retweeting users can unambiguously be assigned to a respective Querdenken community. For an empirically sound retweeting user assignment, we computed a polarization score (Radicioni et al., 2021a) p_r^t per retweeting user r for each timeframe t which is defined as:

$$p_r^t = max_c\{I_{rc}^t\}$$

and

$$I_{rc}^t = \frac{|Q_c^t \cap N_r^t|}{|N_r^t|}$$

where Q_c^t stands for the set of Querdenken users belonging to community c in timeframe t and N_r^t for the set of neighbors, in our case Querdenken users, a retweeting user r has retweeted in that timeframe (Becatti et al., 2019, p. 4; Radicioni et al., 2021b, p. 3). p_r^t uses the maximum value of the obtained I_{rc}^t values and is the polarization score of a retweeting user r in a specific timeframe t. Thus, if a retweeting user only retweeted Querdenken users belonging to one community, their polarization score would be 1. In our case, if a retweeting user reached a polarization score of >0.5, we assumed that they belong to a specific community and assigned them to the respective community. If a retweeting user could not be assigned to a community (polarization score <0.5), we labeled them as -1. As proven elsewhere (Becatti et al., 2019), our assignment outcomes show that retweeting users tended to retweet tweets of Querdenken users belonging to the same community. In t_1 242 of overall 8263 (2.93%) retweeting users could not be assigned to a community, in t_2 425 of overall 9640 (4.41%) retweeting users could not be assigned to a community, and t_3 859 of overall 10,238 (8.39%) retweeting users could not be assigned to a community. With these results, we validated the detected communities in which Querdenken and retweeting users tended to assemble around similar discussion topics and shared cultural patterns.

B6. Receiving Tweet Sample for Qualitative Content Analysis

To receive our tweet sample for qualitative content analysis, we used the hashtag-by-hashtag networks within each timeframe and mapped each hashtag to the Querdenken user communities, identified using the Louvain algorithm. It is important to note that this step is crucial for preventing the blending of discourses from smaller Querdenken user communities with the broader discourse occurring within the Querdenken user networks throughout the entire timeframe. Thus, we can tag hashtags with a community label based on the community affiliation of the Querdenken user who posted the tweet containing the associated hashtag. We then aimed to find the "conversational drivers" (Radicioni, Saracco, et al., 2021, p. 8) within each community—in our case, the most central hashtags per community in the hashtag-by-hashtag networks—as these strongly affect identity formation on the cultural dimension. We derived the threshold for hashtags considered "conversational drivers" with a k-core decomposition of each Querdenken community's hashtag-by-hashtag network. Note that this technique is commonly applied for identifying important nodes and structuring a network into subgraphs (Priante, 2021; Syed & Silva, 2020).

We identified all hashtags within the two innermost k-core shells—thus, with the two highest possible core values in each hashtag-by-hashtag network—as the most important "conversational drivers." We then detected all tweets for each Querdenken user community, including at least one of these hashtags. From this data set, we randomly sampled 10 percent of the tweets per community for qualitative analysis. If 10 percent resulted in fewer than 30 tweets, we included all available tweets from that community. The sampling procedure resulted in a final sample of 960 tweets posted by 62 users.

B7. Codebook: Identifying Collective Agency, Sameness, and Distinctiveness

- **General coding rules:**
  - The entire tweet text (including hashtags) serves as a unit of analysis.
  - For each dimension, only one code, either "0" or "1," can be attributed to each tweet.
  - Code "0" is applied if the post contains no claim(s) related to the respective dimension, "1" is applied if such claim(s) are present.

Dimension	**Description**
Collective agency	Collective agency is understood as capacity for action (Giddens, 1984), i.e., it "suggests the possibility of collective action in pursuit of common interests" (Snow, 2001, p. 3). Thus, posts are interpreted as containing perceptions of collective agency if they address a group's capacity to act together toward shared goals or interests. This is evidenced by references to past or anticipated successes in influencing, e.g., society in general or political elites in particular through, e.g., sustained collective protest, sustained collective criticism, or the sharing of information.
Sameness	Sameness addresses the perception that a group possesses shared values, experiences, characteristics, interests, etc. that unify them (Snow, 2001, p. 3; Stewart & Schultze, 2019, p. 20). This is evidenced when posts contain references to a collective's values, characteristics, interests, etc. (when a post's author is part of the collective). This is also evidenced when specific social categories are described as part of the collective. Generally, sameness addresses the question: "Who are we?".
Distinctiveness	Distinctiveness addresses a collective's boundary of its environment represented by "actual or imagined sets of "others"" (Snow, 2001, p. 3). This is evidenced by demarcating out-groups, particularly opponents, by means of, e.g., criticism of their values, characteristics, interests, etc. Generally, distinctiveness addresses the question: "Who are we not?".

B8. Codebook: Subcategories of Collective Agency, Sameness, and Distinctiveness

- **General coding rules applying to Sections 8.1-8.5:**
  - The entire tweet text (including hashtags) serves as a unit of analysis.
  - If possible, only one label should be assigned; multiple codes may be labeled if necessary. If no code fits the tweet text, the tweet should be labeled “0.”
  - Codes are applied to a tweet if it contains references to the following descriptions:

B8.1Collective Action: Types of Past or Anticipated Collective Action

Code	Description
Call to vote	A tweet exhibits an appeal or endorsement urging individuals to support and vote for a specific political party in an election. The goal is to mobilize voters and influence election outcomes by aligning public support with the favored party.
Consumption of alternative media	A tweet encourages consuming alternative media content, i.e., engaging with media sources that provide perspectives or information differing from conventional media outlets. This may include accessing and relying on alternative news websites, independent journalism, social media channels, or publications that offer, e.g., critiques of dominant narratives, or focus on issues often overlooked by media outlets perceived as “mainstream.”
Demonstration	A tweet encourages participation in demonstrations, or reports on the success of previous demonstrations. Following Kriesi et al. (2020), demonstrations are defined as “politically motivated public gatherings” (p. 35).
General resistance	A tweet encourages participation in general resistance, or reports on the success of previous general resistance. General resistance refers to broad, somewhat vague calls to oppose, resist or criticize political authorities’ decisions without specifying particular forms of action or strategies.
Grassroots organizing	A tweet encourages participation in grassroots organizing, or reports on the success of previous grassroots organizing.
Information session	A tweet encourages participation in a specific information session, or reports on the success of a previous information session.
Petition/open letter	A tweet encourages support for a specific petition or open letter campaign, or reports on the success of a previous petition or campaign.
Sharing of alternative media	A tweet endorses the sharing of particular content of alternative media. For our understanding of alternative media, see the code *consumption of alternative media*.
Taking legal action	A tweet encourages taking legal action on behalf of the movement, or reports on previous successful attempts of taking legal action. Taking legal action is understood as strategic use of legal actions and processes to advance a cause or challenge a policy, decision, or law.

B8.2. Sameness: Type of Actors Considered as Part of Querdenken

Codes in this section are applied to a tweet if the mentioned actors are described as being aligned with or part of Querdenken.

Code	Description
Ambiguous "we"	A tweet uses an ambiguous or generalized collective pronoun, such as "we," to imply a shared experience, responsibility, or perspective without clearly defining the specific group or individuals being referenced. This often creates a sense of collective involvement or consensus on a particular issue or action, such as questioning measures or policies, but lacks precise identification of who the "we" encompasses.
Care workers	A tweet describes individuals or groups employed in the healthcare sector, including both nursing homes and hospitals. This category includes nurses, medical assistants, caregivers, and other professionals involved in providing daily living assistance or medical care for patients or the elderly.
Families	A tweet describes individuals or groups who fulfill functions within a family context. Examples include parents and children.
Foreign political authorities	A tweet expresses alignment between Querdenken and foreign political authorities—especially foreign political decision-makers—and highlights perceived similarities in values, ideologies, and actions. This code applies when the post suggests that foreign political authorities embody principles or stances that resonate with or support Querdenken's goals or beliefs.
General public	A tweet portrays the general public as being aligned with or involved in the movement. A tweet encompasses descriptions that suggest widespread support, participation, or identification with the movement's goals, values, or actions among ordinary citizens. This may also include the depiction of the movement as representing the general population.
Peacemakers	A tweet describes individuals or groups identified as promoting peace.
Political authorities	A tweet contains references to individuals, political parties, and governmental structures within the political sphere. This category may include elected officials, government leaders, and other figures with political decision-making power, as well as specific institutions, such as ministries and governmental bodies. It also covers political parties.
Protesters	A tweet describes individuals or groups actively participating in demonstrations.
Scientific experts	A tweet contains references to scientific experts. This category encompasses researchers as well as trained medical professionals. They are recognized for their advanced knowledge and expertise in specific fields and provide informed insights and analyses based on that expertise. Typically, this includes scientific experts who critique COVID-19 measures and policies.
Self-employed and business owners	A tweet contains references to self-employed persons and/or business owners. This category includes individuals who operate their own businesses or work independently. Examples are, e.g., artists, hairdressers, and restaurant owners.

B8.3. Sameness: Characteristics Defining Querdenken’s Sameness

The codes and their opposites are derived from Alexander (2006, pp. 57-59). Codes in this section are applied to a tweet if it characterizes individuals, groups, or entities perceived as part of or speaking on behalf of Querdenken as:

Code	Description	Alexander’s paired opposite
Active	Engaging actively or taking initiative in response to a situation or issue. In the context of this study, it reflects actions such as resistance, protest, or proactive behavior, suggesting that the subject is not passive but rather is involved in direct, deliberate efforts to influence or address the matter at hand.	Passive
Altruistic	Caring about the needs of or improving conditions for others, typically highlighting social or ethical considerations over personal gain and reflecting a concern for collective well-being and fairness rather than self-interest or profit.	Greedy
Autonomous	Acting independently, i.e., free from external control or influence. It highlights the ability to self-direct and take initiative, to think autonomously.	Dependent
Calm	Being composed, relaxed, or untroubled in their response to a situation.	Excitable
Critical	Having a questioning stance, rather than showing unquestioned approval or support. “Critical” indicates a scrutinizing attitude and often involves assessing actions, policies, and statements from a perspective of healthy skepticism.	Deferential
Friendly	Expressing support, solidarity, or friendship towards others.	Antagonistic
Inclusive	Advocates for the involvement or consideration of all relevant individuals, groups, and/or perspectives. It indicates a focus on broad participation, collective benefit, and/or the integration of diverse viewpoints.	Exclusive
Lawful	Acting within or advocating for actions aligned with legal norms or principles, e.g., advocating for fundamental rights. *Note:* To use only adjectives as characteristics, we substituted “lawful” for Alexander’s code “law.”	Power
Open	Communicating transparently and openly. Hence, they are depicted as candid and forthright, in contrast to secretive.	Secretive
Rational	Making decisions based on scientific evidence.	Irrational
Realistic	Making decisions based on practical reasons and aligned with the actual circumstances, emphasizing an understanding of what is achievable or likely based on the current conditions and factual information.	Distorted
Reasonable	Making decisions based on common sense. Hence, their decisions are based on sound judgment, rather than emotion, ideology, or extreme positions.	Hysterical
Sane	Being mentally stable or exhibiting clear thinking. It denotes a sense of normalcy.	Mad
Straightforward	Acting clear, direct, and honest. It praises a style of communication that is perceived as open and transparent, lacking manipulation or strategic intent.	Calculating
Truthful	Exhibiting a commitment to presenting truthful and reliable information, often in contrast to misinformation or deceptive narratives.	Deceitful

B8.4. Distinctiveness: Type of Actors Considered as Opponents of Querdenken

Codes in this section are applied to a tweet if the mentioned actors are described as opposed to Querdenken.

Code	**Description**
Families	A tweet describes individuals or groups fulfilling functions within a family context. This category includes individuals, such as parents.
Foreign political authorities	A tweet expresses rejection towards foreign political authorities, highlighting perceived differences or negative aspects of their actions or policies.
General public	A tweet contains references to the general public. It encompasses the general populace as a collective entity and often contrasts its attitudes, actions, or responses with those of Querdenken. The portrayal typically highlights the perceived support of the general public for government measures and policies, positioning the general public as antagonistic or indifferent to the movement's views.
Media	A tweet contains references to media outlets and journalists. The portrayal often highlights a perceived conflict or divergence between the media's reporting and the movement's stance.
Philanthropists	A tweet contains references to philanthropists. This category includes individuals known for their charitable activities and philanthropic efforts, often involving significant financial contributions to various causes.
Political authorities	A tweet contains references to individuals, political parties, and governmental structures within the political sphere. This category includes elected officials, government leaders, and other figures with political decision-making power, as well as specific institutions, such as ministries and governmental bodies. It also covers political parties.
Protesters	A tweet contains references to individuals or groups actively participating in demonstrations. This category entails individuals and groups active in counter-protests, which are organized by e.g., antifa activists.
Public service providers	A tweet contains references to individuals or groups who deliver services to the public. This category entails, e.g., teachers.
Scientific experts	A tweet contains references to scientific experts. This category encompasses researchers, trained medical professionals, and research institutions. They are recognized for their advanced knowledge and expertise in specific fields and provide informed insights and analyses based on their expertise. Typically, these individuals or institutions were serving on advisory boards and committees and offered advice to governments, health organizations, and the public on best practices for preventing infection, managing healthcare resources, and implementing public health measures. Examples are, e.g., the Robert Koch Institute, but also scientists such as Christian Drosten and Lothar Wieler who are public figures.

B8.5. Distinctiveness: Characteristics Defining Querdenken's Distinctiveness Towards Opponents

All codes and their opposites are derived from Alexander (2006, pp. 57-59). Codes in this section are applied to a tweet if it characterizes individuals, groups, or entities perceived as distinct from Querdenken as:

Code	Description	Alexander's paired opposite
Arbitrary	Making decisions or actions based on random, capricious, or inconsistent criteria rather than on established principles and the rule of law.	Rule regulated
Clientelist	Engaging in clientelism, which involves providing favors, resources, or benefits in exchange for political support or loyalty. This often includes actions that prioritize the interests of a specific group (e.g., close associates, refugees) or constituency to maintain or gain support, rather than acting based on impartial principles. *Note:* To use only adjectives as characteristics, we substituted "clientelist" for Alexander's code "bonds of loyalty."	Contracts
Conspiratorial	Having a deliberate, hidden agenda or malicious motives. It often implies that the actions of individuals or organizations are part of a larger, secretive plan intended to deceive or manipulate the public. Therefore, this code reflects a belief in or promotion of conspiracy theories.	Deliberative
Deceitful	Acting dishonest or deliberately misleading. This can include presenting false or misleading information, fabricating facts, omitting critical details, or intentionally misrepresenting a situation to deceive or manipulate public opinion.	Truthful
Deferential	Showing excessive respect or yielding to authority without critical examination. This may involve accepting statements or actions at face value, avoiding challenging or scrutinizing them, and generally adopting a compliant or supportive stance.	Critical
Dependent	Lacking the capacity for independent thought or critical analysis, suggesting that they are overly reliant on external authorities, narratives, or influences. It implies that they are seen as incapable of forming their own opinions or making decisions without external guidance.	Autonomous
Diffuse	Characterization of actors is too vague to fit into more precise codes derived from Alexander (2006).	N/A
Distorted	Having a worldview that is out of touch with reality, implying that their focus or perspective is misguided or misplaced.	Realistic
Exclusive	Deliberately excluding others from discussions, considerations, or benefits. It reflects a view that certain entities are being marginalized or disregarded, suggesting that the actions or attitudes being described are creating or reinforcing boundaries.	Inclusive
Hysterical	Acting excessively emotional, panicky, or overdramatic in their reaction to events or issues.	Reasonable
Incompetent	Lacking the necessary abilities, knowledge, or qualifications to effectively perform their roles or duties. We added this code to those proposed by Alexander during our coding process, prompted by the significant number of tweets indicating a perceived deficiency in the abilities, knowledge, or qualifications necessary to effectively perform a particular role or duty, without specifying particular traits that would fit an existing code. This code aligns with the existing literature that indicates that the conflict over COVID-19 public health measures was often framed as an epistemic problem, i.e., an issue of knowledge, expertise, and competence (Bogner, 2021, pp. 21-25; Bogner & Menz, 2021; Pantenburg & Sepp, 2021).	N/A
Irrational	Making decisions or statements that lack logical coherence or evidence-based reasoning. "Irrational" indicates that the actions or beliefs in question are not supported by facts, data, or sound argumentation. This code focuses on the lack of evidence or logical reasoning.	Rational

Mad	Acting in a manner perceived as emotionally extreme or erratic. It suggests a lack of coherence in the actions or statements of the person or group being described. It conveys a sense of disordered or frantic behavior that may seem out of touch with calm or measured responses.	Sane
Passive	Lacking active engagement, involvement, or resistance. It denotes a stance where individuals, groups, or institutions are depicted as merely observing or accepting a situation without taking proactive measures or showing strong opinions.	Active
Power-abusing	Misusing power. This includes actions where officials are seen as exploiting their authority to impose restrictions or make decisions that excessively limit individual freedoms or rights. The term "power-abusing" suggests that the authorities are acting in a way that is perceived as unjust, overreaching, or oppressive in the context of their role and responsibilities. *Note:* To use only adjectives as characteristics, we substituted "power-abusing" for Alexander's code "power."	Law
Self-interested	Acting based on their own personal benefits or advantages rather than considering the broader public interest or collective well-being.	Honorable
Suspicious	Having flaws or inconsistencies. This may include questioning their transparency, or accuracy, and typically implies a lack of trustworthiness. In contrast to the code "conspiratorial," "suspicious" does not imply a broader, hidden agenda or malicious intent.	Trusting

Appendix C: Supporting Results

C1. User-Community Stability Score

To empirically substantiate the understanding of the detected communities by the Louvain algorithm, we also assessed the level of stable user networks within communities over time via a user-community stability score per Querdenken user. By leaning on features of Falkowski's (2009) and Moutidis and Williams' (2021) calculations of community persistence, we defined the user-community stability score ucs_q^t for each Querdenken user q for each timeframe t as follows:

$$ucs_q^t = \frac{|N_c^t \cap N_c^{t+1}|}{|N_c^t \cup N_c^{t+1}|}$$

where N_c^t stands for the direct neighbors of a Querdenken user in one community in one of the three timeframes t_1, t_2 or t_3. N_c^{t+1} constitutes the same Querdenken user's direct neighbors within one community in $t+1$. $N_c^t \cup N_c^{t+1}$ adds all direct neighbors of one Querdenken user in t and $t+1$. Thus, each Querdenken user received three community stability scores that compared the consistency of each Querdenken user's ego networks within communities between $t_1$and t_2, t_2 and t_3 as well as t_1 and t_3. Querdenken users who did not appear in both of the compared timeframes and users who were connected to entirely new neighbors within a community received a value of 0. Querdenken users connected to the same users within the same communities in both timeframes under scrutiny received a value of 1. In all three timeframe comparisons, 212 out of 303 cases received the value of 0, indicating that Querdenken users changed their entire network structure from one timeframe to another. Only 12 users received the highest value of 0.579 (in the comparison between t_1 and t_2), which indicates that 57.9% of a user's network remained the same. When calculating the average of all scores within the three timeframe comparisons, we received a value of 0.093 for $t_1 t_2$, a value of 0.144 for $t_2 t_3$, and a value of 0.039 for $t_1 t_3$. These results support the finding that actor constellations changed over time.

C2. Degree Centrality User-by-User Networks

To validate the findings of our leadership score l_{qt}, we created monopartite user-by-user networks for each timeframe t with the one node set $U^t = \{1, \dots, m\}$. The matrices related to these networks can be expressed by C^t with the dimensions $m \times m$, where m corresponds to the total number of both Querdenken users and retweeting users within t. Users were linked with each other if they retweeted each other at least once during the respective timeframe t. In this case, the element $c_{uu'}^t$ equals 1, otherwise 0. We chose to create these networks because they capture the retweeting behavior of both Querdenken users and retweeting users, reflecting their collective decisions regarding who was considered prominent within the group. This approach aligns with the concept behind our leadership score l_{qt}. However, unlike our leadership score, these networks did not distinguish between different user types, allowing us to validate whether the leaders we previously identified were still considered prominent without any special status within the network.

In t_1 the network features 8306 unique users connected by 87,158 edges; in t_2 the network features 9713 users linked by 103,139 edges; and in t_3 the network consists of 10,329 users linked by 96,764 edges. We calculated the normalized degree centrality for all users to identify those who were retweeted the most, which indicates who was perceived as prominent by the collective and, therefore, as a leader.

When evaluating the users with the highest normalized degree centrality across all three timeframes, we again found that our previously identified leaders, based on the leadership score l_{qt}, consistently appeared as the top retweeted users (see Figure C1). This finding validates our previous leadership analysis.

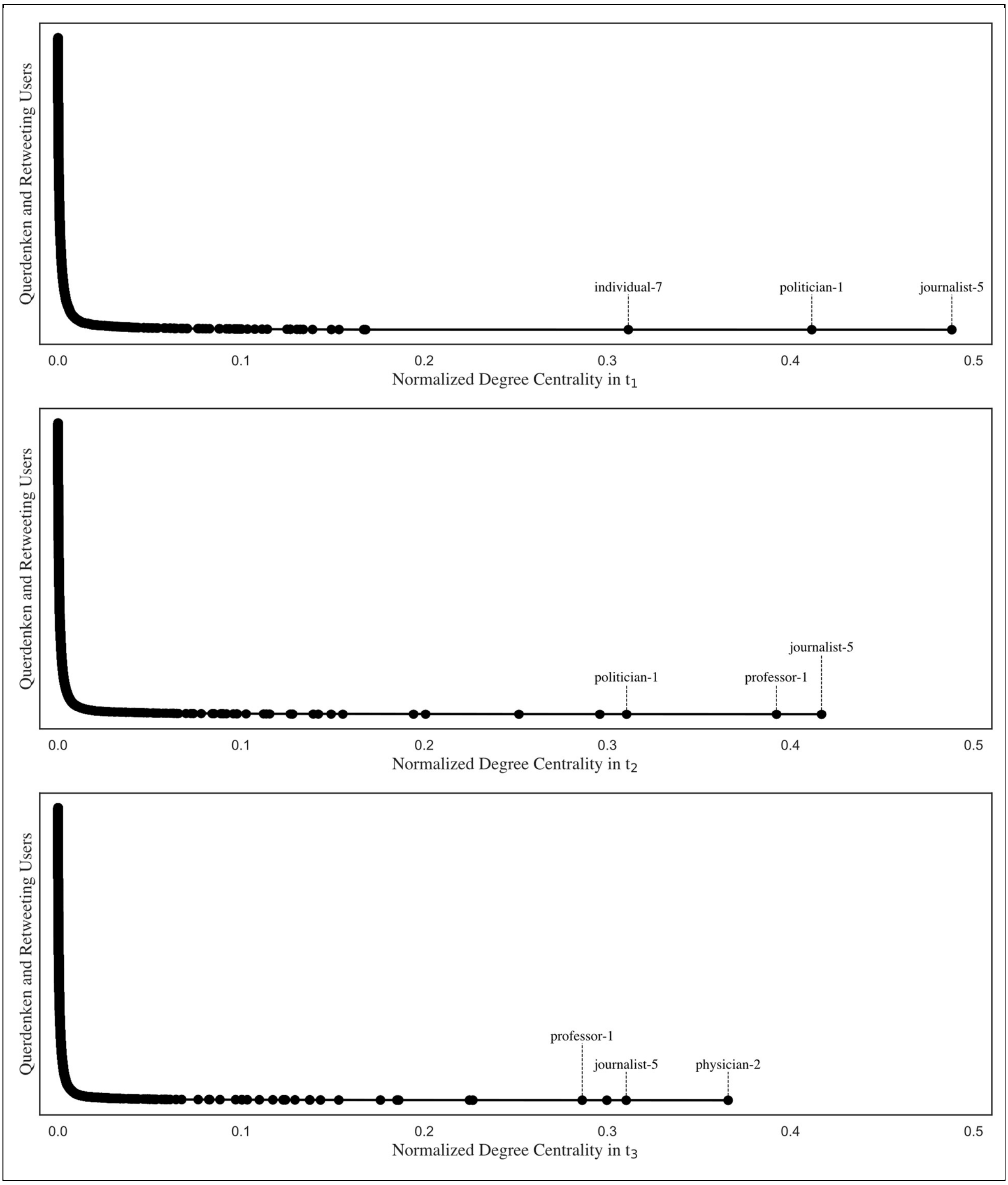


Figure C1. Normalized Degree Centrality Distribution Over Time

C3. Power-Law Distribution Goodness-of-Fit Analysis

Since a power-law distribution should not be inferred by a visual analysis alone (Clauset et al., 2009), we also performed a statistical goodness-of-fit analysis of the data distribution of the l_{qt} scores. These analyses, presented in the following, generated p-values as empirical evidence for the plausibility that the data distribution followed a power-law distribution. We applied the approach introduced by Clauset et al. (2009) and fit a power-law model to the l_{qt} scores distribution, and compared this distribution with a perfect power-law distribution. Thus, we randomly generated 1000 new distribution samples based on our fitted data. We calculated the distance between a perfect power-law distribution and these synthetically generated distributions by computing the Kolmogorov-Smirnov statistic for each pair. Subsequently, we calculated the respective p-values by counting "the fraction of the time that the resulting [Kolmogorov-Smirnov] statistic is larger than the value for the [initial] data" (Clauset et al., 2009, p. 677).

When applying this procedure to the l_{qt} score distributions in all three timeframes, we received values that surpassed the threshold of $p > 0.1$ by far (see Table C1; note that the p-values change when reconducting this procedure since this test is a heuristic estimation based on the comparison with 1000 newly generated distribution samples each time, yet even after conducting multiple runs, we always received values surpassing the threshold of 0.1 for each timeframe). In contrast to usual hypothesis testing, a power-law distribution was present if the H_0 could not be rejected (Clauset et al., 2009), indicating that the initial data resembled the distribution of a power-law (note that larger p-values, however, did not indicate a better power-law fit). Based on the resulting p-values, we could not reject H_0, showing that a power-law distribution is a plausible fit for the data structure.

Table C1. Significance Testing for a Power-Law Distribution

	t_1	t_2	t_3
p-values	0.458	0.319	0.895

About the Authors

Theresa Henn-Latus is a PhD student and research associate in the Department of Information Systems and Social Networks at the University of Bamberg. Her research interests include online communities, social network analysis, protest movements on social media, and image analysis.

Sarah Tell is a PhD student and research associate at the SOCIUM Research Center on Inequality and Social Policy at the University of Bremen. Her research interests include political sociology, digitalization, sociological theory, and computational social sciences.

Julian Polenz is a PhD student and research associate in the Department of Sociology at the University of Bamberg. His research interests include political sociology, sociological theory, and computational social sciences.

Thomas Kern is a professor of sociological theory and the director of the Institute of Sociology at the University of Bremen, Germany. He has held professorships in sociology at the Universities of Heidelberg, Chemnitz, and Bamberg. His research interests include social theory, protest movements, digitalization, democracy, and religion.

Oliver Posegga is an affiliate and former professor of the Department of Information Systems and Social Networks at the University of Bamberg, Germany. His main research interests are social networks, digital collaboration, and the sociotechnical dynamics of hybrid media systems. He is also interested in the ethical management of AI, AI literacy, and digital responsibility.

Appendix A.2: Paper II

Bibliographic data	Henn, T., Posegga, O. 2023. What Do They Meme? Exploring the Role of Memes as Cultural Symbols of Online Communities. International Conference on Information Systems (ICIS). Hyderabad, India.
Abstract	Analyzing symbols shared within online communities (OCs) is essential to better understand communities' expressed cultures. To evaluate how OCs differ in their expressed culture and analyze the effects of community rules (CR) and moderation policies (MP), we examined meme sharing of subreddit and interaction communities on Reddit. To detect memes shared within subreddits automatically, we trained a convolutional neural network and applied a feature-matching algorithm to create meme networks with components consisting of visually similar memes. Based on each community's component composition, we created community-specific meme languages that we compared across subreddit and interaction communities. Our results show that memes can be aggregated to characteristic meme languages linked to individual OCs; yet MP and CR do not impact the homogeneity of shared memes. Based on these findings, we plan to analyze dynamically the relationship between memes and OCs, examining memes' textual content and diving deeper into users' individual meme languages.
URL	https://aisel.aisnet.org/icis2023/socmedia_digcollab/socmedia_digcollab/8

What Do They Meme? Exploring the Role of Memes as Cultural Symbols of Online Communities

Short Paper

Theresa Henn
University of Bamberg
An der Weberei 5, 96047 Bamberg, Germany
theresa.henn@uni-bamberg.de

Oliver Posegga
University of Bamberg
An der Weberei 5, 96047 Bamberg, Germany
oliver.posegga@uni-bamberg.de

Abstract

Analyzing symbols shared within online communities (OCs) is essential to better understand communities' expressed cultures. To evaluate how OCs differ in their expressed culture and analyze the effects of community rules (CR) and moderation policies (MP), we examined meme sharing of subreddit and interaction communities on Reddit. To detect memes shared within subreddits automatically, we trained a convolutional neural network and applied a feature-matching algorithm to create meme networks with components consisting of visually similar memes. Based on each community's component composition, we created community-specific meme languages that we compared across subreddit and interaction communities. Our results show that memes can be aggregated to characteristic meme languages linked to individual OCs; yet MP and CR do not impact the homogeneity of shared memes. Based on these findings, we plan to analyze dynamically the relationship between memes and OCs, examining memes' textual content and diving deeper into users' individual meme languages.

Keywords: Online Communities, Memes, Cultural Identity, Social Media, Network Analysis, Image Classification

Introduction

We all use memes daily, but know little about their symbolic value and cultural meaning for online communities (OCs). Analyzing memes as cultural symbols representing an OC's shared values, identity, and ideology has, however, gained attention in recent years as a promising avenue for research (see, e.g., DeCook 2018; Nissenbaum and Shifman 2017; Nowotny and Reidy 2022). This is a welcomed development, as furthering the understanding of OCs' culture is essential for (1) gaining deeper insights into online user interaction that can be beneficial (e.g., support in mental health forums) but that may also turn toxic (e.g., sexist or xenophobic discussions) (Massanari 2017); (2) expanding the knowledge on OCs' shared norms and values, which have a significant impact on, for example, an OC's sustainability (Seraj 2012)); and (3) improving our understanding of reciprocal spillover effects between online behavior and offline consequences, such as effects on electoral turnouts (Nowotny and Reidy 2022). Given these far-reaching consequences of OCs' online culture, it is somewhat surprising that IS research has paid little attention to this field of research compared to other disciplines (see, e.g., Kozinets 2010). Only a few works have addressed, for instance, the effects of organizational culture on IS adoption in organizations (Jackson 2011), the impact of nationality and culture on knowledge sharing (Moser and Deichmann 2021), and the influence of language on an OC's culture and identity creation (Fayard and DeSanctis 2010).

This research gap should not mistakenly be attributed to a lack of relevance; rather, it speaks to the complex challenge of conceptualizing, operationalizing, and tracking OCs' culture in the online space's fluidity. While online culture has been studied through ethnographic (Kozinets 2010) and linguistic (Fayard and

DeSanctis 2010) lenses, for example, analyzing it through memes as cultural symbols of OCs' common values and identities has proven to be a promising complementary approach (e.g., DeCook 2018; Gal et al. 2016) that accounts for the increasing importance of user-generated visual content online (Marwick 2015).

This paper adds to this line of research by evaluating whether individual OCs share a characteristic "meme language"–by which we mean the composition of distinct memes and their variants shared within an OC– that symbolizes their shared culture, norms, and identity, demarcating them from other OCs. In doing so, we enter a debate. One line of the literature confirms memes' symbolic and identity-shaping effect for different OCs by highlighting their characteristic cultural value for distinct OCs (e.g., DeCook 2018). These findings suggest that we should be able to detect characteristic meme languages for distinct OCs that provide valuable insights into their shared culture. Another line of research emphasizes the rapid diffusion of memes across communities (e.g., Spitzberg 2014), thus assigning them a universal role that makes them rather inadequate for offering profound insights into distinguishable OC cultures.

Research findings have also shown that community rules (CR) and moderation policies (MP) significantly impact the type of content shared within an OC and how users interact with it (Seering et al. 2019). From this, we follow that the CRs and MPs in place can also affect the sharing of memes within an OC. Therefore, it is plausible to assume that OCs with defined CRs and MPs allow the sharing only of those memes that closely comply with their established norms and identity. Hence, the existence of CRs and MPs could lead to less diverse memes shared within an OC and, thus, a more uniformly expressed culture through a homogenous meme language.

So, to evaluate the characteristic value of meme languages for distinct OCs and analyze the effects of CRs and MPs thereupon, we pose two research questions using Reddit:

RQ1: *To what extent do OCs on Reddit share a characteristic meme language?*

RQ2: How do Reddit CRs and MPs affect the diversity of an OC's meme language?

To answer the two research questions, we analyze memes posted on Reddit's 100 largest subreddits (measured by the number of members). To test whether specific memes can be aggregated to a characteristic meme language of individual OCs, we examine the extent to which distinct communities share particular memes and their variants. We differentiate between two types of Reddit OCs, as both evolve through different logics of formation and vary in their levels of MPs and CRs that could affect the exchange of inherent symbols such as memes: subreddit communities (SCs) that develop through Reddit's internal platform structures and that create predetermined boundaries for communities of interest with defined CRs and MPs; and interaction communities (ICs) that go beyond platform-imposed community boundaries and evolve across individual subreddits and their MPs and CRs based on user interactions.

Memes have symbolic value, and we assume that users from individual OCs share specific memes and their variants as an expression of their culture, leading toward a characteristic meme language, that demarcates them from other communities. In addition, we assume that this behavior is moderated by an OC's CRs and MPs. To verify these assumptions empirically, we need to detect memes and their variants within our data. We use the method introduced by Courtois and Frissen (2022), who uncover "meme families" by linking similar memes by matching of visual features associated with the original meme. Based on the similarity detected, memes are mapped to a network graph that comprises components with the original memes and their variants. Knowing which meme was posted in which community enables us to detect all meme components of visually similar memes within each community (SCs and ICs). The composition of these components then constitutes an OC's individual meme language. To answer RQ1, we examine the detected meme language of all SCs and ICs and evaluate their resemblance. To answer RQ2, we calculate a similarity score s for each community that quantifies the homogeneity of a community's meme language and statistically compare these scores between SCs and ICs.

The remainder of this paper is structured as follows. We briefly define our conception of culture and memes and provide an overview of the current state of research on the identity-shaping power of memes for OCs, their diffusion across OCs, and our theoretical foundation and causal model grounded in Stuart Hall's two systems of representation; we then describe our methods, including our data collection on Reddit and our procedures for creating the interaction network and labeling the SCs and ICs, generating a network graph of memes and their variants within each community, comparing the meme languages across communities, and evaluating the homogeneity of shared meme languages between SCs and ICs. We conclude our paper with an overview of our preliminary results and a description of the next research steps.

Theoretical Background

Definition of Culture and Internet Memes

"Culture" is a multifaceted and complex concept that authors from various disciplines have tried to narrow down for decades. In this paper, we decided to follow Donal Carbaugh, a cultural communication researcher, who defines "culture" as "a system of symbols, premises, rules, forms, and the domains and dimensions of mutual meanings associated with these" that is "implicated, employed, or creatively invoked in conversation" (Carbaugh 1993, p. 182). Thus, for Carbaugh, culture is constituted through communication. This understanding of culture aligns perfectly with the online sphere, as the communicative traces left by users can be used as a proxy for their expressed culture. In our case, we focus on one specific type of trace: Internet memes that are often regarded as "a key fixture of digital culture" (DeCook 2018, p. 485) having symbolic and identity-shaping value for OCs (Nowotny and Reidy 2022).

Richard Dawkins introduced the meme concept, defining a meme as "the idea of a unit of cultural transmission, or a unit of *imitation*" (Dawkins 2016, p. 249, emphasis in original). In the online world, an Internet meme can be described as "(a) a group of digital items sharing common characteristics of content, form and/or stance, which (b) were created with awareness of each other, and (c) were circulated, imitated, and/or transformed via the Internet by many users" (Shifman 2014, p. 41). Like Nissenbaum and Shifman (2018), we focus herein on "image macros" as a specific type of meme that "consist of a repeating image overlaid with text and usually deal with a specific topic or situation." (Nissenbaum and Shifman 2018, p. 298). These memes are known for their "cultural biographies" (Nowotny and Reidy 2022, p. 34), as their initial composition of text and image is continuously altered, refined, and transformed, which supports their rapid dissemination through the Internet often related to their humoristic and sarcastic messages. This type of meme is, however, more than a simple compilation of varying humorous texts and images; instead, it depicts reproducible combinations of cultural elements that are continuously remixed and constitute, based on their aggregated interplay, specific values and meanings for OCs (Nowotny and Reidy 2022). Not all memes on the Internet are image macros, but they are one of the most common types and have been used by several other researchers as their analytical foundation (e.g., Nissenbaum and Shifman 2018). Thus, we are confident in considering them to be representational.

Analyzing OC's Culture: The Role of Memes as Cultural Symbols

There have been several attempts in research to analyze OCs' culture. One prominent example, widely applied, is Kozinets' (2010) "Netnography" approach, which uses methods from ethnography to study online cultures and communities. Other scholars have focused on the analysis of language as a characteristic feature of online culture that creates a common sense of "we-ness" (Fayard and DeSanctis 2010). However, as online social media platforms have increasingly become a "visual medium" (Marwick 2015), shared images have become an essential part of an OC's expressed culture, resulting in a "digital visual culture" that affects how groups communicate and are received by others (McSwiney et al. 2021).

Memes are a significant expression of this digital visual culture. One line of research has addressed the growing online importance of memes by focusing on them as characteristic cultural symbols of specific online groups that express an OC's values, identity, and shared ideologies. Posting memes is regarded as important for creating a feeling of belonging to an in-group (DeCook 2018; Nissenbaum and Shifman 2017) as it aids in developing a common identity and attachment to a community of like-minded people (DeCook 2018; Nissenbaum and Shifman 2017; Nowotny and Reidy 2022; Shifman 2014). As memes shared by OCs tend to be "typically unique to each individual community" (DeCook 2018, p. 488), they aid OCs in their boundary work toward other online groups (e.g., Gal et al. 2016). Thus, although memes are known to mutate, their origins can usually be associated with specific OCs, particularly grounded in a distinct type of humor shared by specific OCs (Nowotny and Reidy 2022). Hence, memes can be regarded as identity-shaping cultural symbols for individual communities that "function as cues of membership" and "constitute forms of cultural capital." (Nissenbaum and Shifman 2017, p. 485). This identity-shaping effect becomes quite evident when analyzing the use of memes in alt-right online groups. Several researchers have identified the prominent "Pepe the Frog" meme, for instance, as symbolic of the alt-right (see, e.g., Nowotny and Reidy 2022). Moreover, research has shown that particularly memes' humoristic nature is used by alt-right groups as "weaponized irony" (Nowotny and Reidy 2022, p. 139) to expand these group's reach toward mainstream audiences, as they try to win them over to their ideology.

Although memes can usually be traced back to "local" cultures related to specific OCs that coin them with an initial meaning, many memes eventually diffuse through the Internet and are picked up by the general public (Nowotny and Reidy 2022). Thus, there is another line of research that has addressed memes as a "phenomenon that transgresses social and cultural boundaries" (Bauckhage 2011, p. 42) and analyzed how memes from "peripheral communities … diffuse to the mainstream core" (Morina and Bernstein 2022). Spitzberg (2014) introduced a multilevel model to track the complex landscape of meme diffusion on the Internet; it emphasizes that memes, depending on internal and external factors, "will continue replicating throughout the systems to, between, and beyond which they are communicated." (Spitzberg 2014, p. 316). Nowotny and Reidy (2022, pp. 121–136) thematized how memes of Captain America were simultaneously used by supporters of the so-called "social justice" movement and adherents of the alt-right to promote their values and ideology. These findings stress that memes not only have symbolic value for distinct OCs but that the same memes have the potential to connect opposite communities.

This paper makes two main contributions. First, the literature review reveals a theoretical discrepancy regarding the role of memes for OCs; here, we address the two research strands by empirically testing the extent to which specific memes, because of their symbolic value, can be linked to distinct OCs. By introducing the concept of "meme languages" (composition of distinct memes and their variants), we provide a novel conceptual approach to empirically measure and differentiate between OC's cultures, aiming to initiate a fruitful discussion in IS research. The second is that we analyze how CRs and MPs impact the sharing of similar memes by differentiating between communities based on platform boundaries (SCs) and cross-cutting communities that build upon user interaction beyond subreddit affiliation (ICs). These insights will be valuable to enhance the understanding of how MPs and CRs can be applied to prevent negative user behavior (e.g., toxicity) and promote positive user interaction (e.g., social support). These contributions of our short paper are the first step toward a more profound analysis that will continue to evaluate (1) the role of memes that are strongly connected to individual OCs and memes that "bridge" the cultural boundaries between different OCs; (2) meme languages on a visual and textual level, which enables a more profound meme analysis; and (3) the detailed effects of different CRs and MPs on an individual user's meme language across communities.

Stuart Hall's Two Systems of Representation

We use cultural sociologist Stuart Hall's work on the two systems of representation to create a causal link between the sharing of memes and the representation of an OC's culture while also including the notion of CRs and MPs. Like Carbaugh's definition of culture introduced above, Hall (1997) emphasizes the role of language and meaning as the two essential components representing culture. Against this background, he developed his idea of two systems of representation in which the first system of representation comprises the idea that all of us possess a mental "conceptual system" that correlates "all sorts of objects, people and events … with a set of concepts or *mental representations* (1997, p. 17; emphasis in original). Thus, based on our mental representation of real-world objects, people, and events and their interconnected relationships, we create "conceptual maps" that aid us in interpreting the world in a meaningful way. To adequately communicate these conceptual maps with other people and construct meaning, we need Hall's second system of representation: a common language system. For Hall, "language" refers to spoken or written words and comprises sounds and images, which he summarizes as "signs" that carry meaning. Thus, the composition of these "signs" constitutes a culture's common "language" for communicating meaning.

When translating Hall's theoretical considerations to the online sphere, we can only make inferences about the second system of representation, as digital traces do not provide insights into a user's interpretation of signs' cultural meaning. We can, however, evaluate whether shared signs (memes), based on their composition, lead to a characteristic cultural language (meme language) that represents a specific culture (OC culture). Following these considerations, we assume that SCs' and IC's individual cultures are expressed through distinct memes that can be aggregated into characteristic meme languages for each community. Thus, our first hypothesis:

H1: Individual SCs and ICs share characteristic meme languages.

Sharing these signs in the online space frequently happens within the boundaries of MPs and CRs. These

MPs and CRs guide what signs should be shared within a community (Seering et al. 2019). Therefore, it is plausible to assume that they affect the diversity of shared signs within a community and lead to a more homogenous cultural language and representation. As a first analysis, we thus assume that SCs share more homogenous meme languages than ICs, as the former is constituted around at least a minimum level of concrete CRs and MPs. In contrast, the latter evolves around user interaction across subreddit boundaries that is free of centralized supervision. From this, we derive our second hypothesis:

H2: The meme language of SCs is more homogenous than that of ICs.

Methods

Collecting Submissions and Comments on Reddit

Reddit, our choice for analyzing the relationship between memes and their characteristic association with specific OCs, is regarded as "the front page of the Internet" (Gaudette et al. 2021), bringing together users who "participate in, contribute to, and even develop like-minded communities whose members share a common interest." (Gaudette et al. 2021, p. 3493). These OCs, called subreddits, have their own moderators, CRs, and topics of interest that facilitate the development of a collective identity among users (Gaudette et al. 2021). In addition to these pre-defined subreddit communities based on boundaries set by the platform, users are free to engage with multiple communities and interact across SC boundaries. The patterns that arise from the interactions resemble structures of social interaction and reveal communities of users that span the boundaries of SCs. Thus, these interaction communities (ICs) evolve naturally and serve as the ideal foundation for analyzing the diffusion of memes in heterogeneous communication spaces. The availability of data on both types of OCs makes Reddit a suitable platform for our analysis.

We downloaded six months' worth (August 2022–January 2023) of Reddit submission and comment dumps obtained from Pushshift. These files comprise the entire 245-GB Reddit feed during this period. Our goal was to analyze the 100 largest of all unique 709,747 subreddits as measured by number of members in January 2023. We used that month's submission dump as our reference and extracted the number of members per subreddit. We then extracted all submissions and comments posted within those 100 largest subreddits, yielding 7,109,414 submissions and 178,286,358 comments (28.6 GB). Thus, although representing only 0.014% of the number of subreddits, our sample covers 11.673% of the shared submissions and comments within that period, allowing us to analyze a substantial proportion of the shared Reddit content and thus, obtain insights regarding the role of memes within the most prominent subreddits. To control for potential bots that would bias our results, we tested whether users within our dataset were part of a publicly available list of 392 bots as of April 14, 2023, published by Reddit.[1]. After dropping all submissions and comments that were posted by bots or whose authors were deleted, we ended up with 4,925,264 submissions and 155,598,892 comments, posted by 8,710,321 unique authors.

Collecting Images and Detecting Memes

To analyze whether the composition of specific memes is characteristic of individual OCs, we needed to detect memes within our data. We began by downloading, where possible, all images attached to the submissions from the 100 largest subreddits posted in August 2022. This yielded 173,771 images (106.3 GB) in what we called the "image dataset." We then filtered these data to obtain those images matching our definition of "meme" above. We avoided laboriously going through all images manually by using a convolutional neural network (CNN) model–having demonstrated outstanding performances in similar classification contexts (e.g., Sherratt et al. 2023)–to automate meme detection.

To train the binary CNN (meme/non-meme), we used the following openly available datasets: (1) the Memotion Dataset 7k for obtaining memes, which contains 6,992 memes from different genres and thus complyies with our meme definition (images overlayed with text); (2) a subsample of four different datasets available from TensorFlow. These include PASCAL VOC, which covers twenty general image classes; SVHN, which comprises images of digits; LFW, which includes images of faces and persons; and wider_face, a face detection benchmark dataset to receive non-memes. As training a CNN with a balanced data sample is advised, we extracted 6,990 images from these four datasets (60% PASCAL VOC, 15% SVHN, 15% LFW,

[1] https://www.reddit.com/r/autowikibot/wiki/redditbots

and 10% wider_face) to match the meme sample obtained from the Memotion Dataset 7k. These approximately 14,000 images were the training and testing data for the CNN model.

We used the CNN architecture of Keras' pre-trained EfficientNetB4 as our baseline model. To adjust the model to our context, we used transfer learning, freezing all pre-trained layers and retraining the last layer by compiling the model with a sigmoid function. After performing a hyperparameter tuning, we decided on adam as an optimizer, with a learning rate of 0.01. We included an early stop callback function to monitor the testing accuracy, with a patience parameter of 5. As increasing the sample size leads to better-performing models, we also applied image augmentation throughout the training process. We used 80% of the data as training data and 20% as testing data, beginning the training process with 100 epochs and a batch size of 32. Due to the early stop callback function, the model stopped at epoch 18 and achieved an accuracy of 89.9%, a precision of 95.0%, a recall of 84.3%, and a loss of 24.0% on the testing dataset.

Applying the meme classifier to our image dataset from Reddit resulted in 29,778 images classified as memes and 143,809 as non-memes. To ensure the quality of this classification, we manually labeled 100 meme-classified and 100 non-meme classified images; we identified 64 of the former as memes and 82 of the latter as non-memes–receiving an overall accuracy rate of 73%. This value was within an acceptable range compared to the accuracy obtained with the testing dataset and was above 70%, so we continued working with the detected memes. We called this the "meme dataset."

Network Creation and OC Labeling

To receive ICs, we created the interaction network I. Our node set $U = \{1, \dots, n\}$ comprises all unique users of the 100 largest subreddits who either authored a submission or comment. Each user $i \in U$ is connected with another user $j \in U$ if they interacted at least once, with an interaction defined as a user commenting on another user's submission or comment. The network this created comprised 7,743,544 nodes and 93,107,759 edges. In its structure, we detected one giant component with 7,721,402 nodes and 92,843,239 edges; the remaining 5,599 components had at most 10 nodes. The giant component accounted for 99.7% of all the nodes depicting user interaction, so we continued working with it, dropping all nodes and edges from the other components.

To create SCs, we simply used Reddit's pre-defined subreddit boundaries, which in our case led to 100 SCs. To detect interaction communities, we applied to the interaction network I the commonly used Louvain algorithm, which is based on a modularity-maximizing approach and known for its scalable performance when applied to large networks. To control for the quality of the detected communities, we calculated the modularity score, which ranges from -1 (partitioning is worse than chance) to 0 (partitioning is as good as change) to 1 (perfect partitioning). After computing the communities with different resolution parameters, we decided on the resolution = 1.5, as this parameter leads to the modularity score of 0.44. By using this parameter, we yielded 1,846 different ICs.

Detection of Memes and their Variants and Analysis of Meme Languages

To evaluate an OC's meme language and answer our two research questions, we need to detect memes and their variants. To do so, we followed the approach introduced by Courtois and Frissen (2022), who created a mapping algorithm that detects similar meme features across images. Thus, even modified images that still contain essential elements of the original image are considered similar. The authors calculate the distance between image pairs based on visual similarity for their feature matching: the smaller the distance, the more likely that images share the same features and belong to the same original meme. If a meme cannot be matched with any other meme, it is dropped and not further processed. Once similar memes are detected through feature matching, a meme network X is created, consisting of components B that include memes detected as similar. For further information, please see Courtois and Frissen (2022).

The Courtois and Frissen (2022) model is finetuned for a context very similar to ours, so we used the model with their hyperparameters. We began with a subsample of 3,000 detected memes posted in August 2022 (approximately 10% of memes in that period). After manually screening these images, we continued to work with 1,892 memes that matched our definition (63%, aligning with the previous sanity check). The discarded images were most often screenshots of social media conversations or news articles that would have biased our result if not dropped. These 1,892 memes were posted in 30 SCs and 22 ICs; we dropped all communities with less than ten shared memes to guarantee a meaningful visual feature-matching process, resulting in 11 SCs and 9 ICs, each containing a certain number of memes defined as $y \in Y$.

We applied the feature-matching algorithm to all memes $y \in Y$ of each of the remaining SCs and ICs and created a meme network X for each community that consisted of components B comprised of memes and their variants as nodes defined as $x \in X$. We manually validated whether the matched memes belonged to the same original meme and adjusted–if necessary–the feature matching parameter m (see Courtois and Frissen 2022) and the individual components until all components represented one meme and its variants. To create and compare meme languages, we manually labeled each component B of each meme network X with a number representing the distinct type of meme it comprises (note that components of the same meme were labeled alike). Thus, an OC's meme language is constituted by its individual component composition and can be compared between SCs and ICs, respectively.

To analyze how the meme language of SCs and ICs differs in their homogeneity, we calculate a similarity score s for each community, with $s = \frac{x}{y}(1 - \frac{c}{y})$, where x represents the number of matched memes used as nodes in the network graph X, y the number of initial memes from each individual community and c the number of matched meme components B in each community's meme network X. Thus, scores closer to 1 denote that a community shared predominantly visually similar memes (nearly all memes could be matched with other memes and could be aggregated to a few meme components), representing a relatively homogenous meme language, whereas scores closer to 0 indicate a heterogenous meme language with many visually different memes (see Table 1 for an overview of all SC and IC values). Because of the small sample size, we performed a Wilcoxon Rank-Sum Test comparing the s values between SCs and ICs, assuming that SCs should display greater values (more homogenous meme languages). Yet, with a p-value of 0.65, we cannot reject the null hypothesis and, thus, cannot conclude that the s scores for SCs are significantly higher than for ICs. Therefore, SCs' meme languages are not more homogenous than the ones of ICs.

OC	OC name	y	x	c	m	Meme language composition	s
SC	AdviceAnimals	43	22	9	4	1;2;3;4;5;6;7;8;9	0.40
SC	dankmemes	283	29	13	20; B_1: 50	10;11;12;13;14;15;16;17;18;2;19;20;21	0.10
SC	funny	58	0	0	1	/	0
SC	gaming	37	0	0	1	/	0
SC	HistoryMemes	155	18	6	20; B_1: 90	10;22;23;24;25;26	0.11
SC	MadeMeSmile	11	0	0	1	/	0
SC	me_irl	61	0	0	15	/	0
SC	memes	1,020	170	49	20; B_1: 200	27;28;29;15;30;31;21;22;33;34;2;35;36;37;11;38;39;40;41;42;43;44;45;46;47;48;33;49;50;14;12;51;52;53;54;55;56;47;17;58;59;60;61;62;63;64;23;65;66;67	0.16
SC	pcmmasterrace	27	2	1	4	68	0.07
SC	wallstreetbets	85	12	4	4	69;70;71;22	0.13
SC	wholesomememes	37	0	0	4	/	0
IC	10	1,075	212	63	4; B_1:400	44;72;47;19;40;29;45;24;15;73;52;74;75;14;11;76;53;41;66;2;58;22;77;42;12;28;78;79;80;56;36;81;82;83;84;85;13;86;33;87;54;31;65;49;37;46;43;17;34;57;88;62;89;60;90;59;91;61;64;23;48;92;93	0.19
IC	24	123	8	3	20	38;94;10	0.06
IC	69	53	8	4	4	2;3;6;7	0.14
IC	81	92	12	4	4	71;95;70;22	0.12
IC	89	137	0	0	15	/	0
IC	121	16	0	0	4	/	0
IC	679	38	2	1	4	41	0,05
IC	1133	127	12	4	4; B_2: 25	25;96;26;22	0.09
IC	1187	25	2	1	4	2	0.08

Table 1. Overview of different values of SCs and ICs.

Preliminary Findings and Next Steps

To answer RQ1, we compare the meme language compositions between the individual SCs and ICs, respectively (see Table 1), making two interesting observations. First, for both types of OCs, meme languages vary highly between the individual OCs, with only a few memes and their variants appearing in several OCs (e.g., memes 2, 10, and 22). Second, for seven out of the 20 OCs evaluated, the memes could not be mapped to a network graph X, which indicates that these OCs shared visually diverse memes that do not constitute a characteristic meme language. Based on these findings, we can accept *H1* insofar as the following: if SCs or ICs feature a detectable meme language, it differs between the individual communities and is thus characteristic for distinct OCs. These preliminary findings show that certain memes can be seen as cultural symbols that are, in their composition, tied to particular OCs. To answer RQ2, we evaluated the meme languages' homogeneity between SCs and ICs by calculating a similarity score s. The Wilcoxon Rank-Sum Test's result showed, however, that SCs and ICs do not vary in their meme language homogeneity, meaning that an OC's CRs and MPs did not impact the sharing of homogenous memes within our data. Thus, we must reject *H2*.

In conclusion, we found initial evidence supporting research that considers memes as cultural symbols for individual OCs (SCs and ICs) that can be aggregated to characteristic meme languages. Yet, our results also show that some memes, and thus the culture expressed through them, overlap in some of the SCs and ICs. This suggests that specific memes have the potential to "bridge" the cultural boundaries of communities and are not tied exclusively to any individual OC. In addition, our preliminary results demonstrate that CRs and MPs do not affect the sharing of homogenous memes within an OC, as the meme language of SCs and ICs did not significantly differ in their level of homogeneity. This finding shows that communities based on platform-imposed boundaries and communities established through cross-cutting user interaction feature a complex and multifaceted culture that a single symbol cannot represent; instead, the compositions of these symbols provide more profound insights into an OC's expressed culture. These findings entail first important implications as they not only aid in providing a deeper understanding of an OC's shared values and norms being essential for, an OC's sustainability, for example, but also in expanding our knowledge on cultural diffusion online, which, for instance, can be important to track and prevent online radicalization.

These preliminary findings provided valuable initial insights into the analysis of memes as cultural symbols for OCs and pave the way for several promising future research avenues. However, while the analysis of memes shared within the 100 largest subreddits allowed us to cover a substantial portion of the shared mainstream content on Reddit, our findings are not representative of all subreddits on Reddit. It is reasonable to assume that smaller, niche-oriented subreddits with a more solidified user base share a more homogenous meme language with specific memes tightly linked to individual OCs. Yet, these niche-oriented subreddits are also known to influence each other reciprocally across subreddit boundaries, making a combined analysis of ICs and SCs an interesting subject for future research. Therefore, our findings should be seen as a first step toward analyzing the role of memes as cultural symbols for OCs, which we plan to complement with the evaluation of meme languages shared in different types of subreddits. Likewise, as we analyzed only a subsample of memes posted in August 2022, we would like to expand our analysis by examining the relationship between meme languages and OCs within the entire six-month period of our collected data. Doing so would allow us to dive deeper into analyzing newly emerging meme languages and explore, for instance, which memes are tightly connected to OCs, and which ones diffuse within the network, bridging the cultural border across (opposing) communities. After answering such questions for the platform Reddit, we would like to analyze whether similar findings could be reproduced on other platforms. These findings will be valuable in evaluating how different platform architectures, user bases, and affordances affect the sharing of memes, providing a more holistic picture of OCs' culture. We recognize that defining an OC's meme language based on memes' visual similarity is only the first step in uncovering the profound potential of memes as cultural symbols of different OCs. To fully understand memes' symbolic and cultural meaning for distinct OCs, we want to complement our quantitative visual analysis with a qualitative examination of the memes' content, which would enable us to define an OC's meme language more profoundly and compare them across OCs in more depth. One possible approach for an in-depth evaluation of a meme's visual and textual components could be a multimodal discourse analysis, as applied by, for instance, Yoon (2016). Finally, we would like to investigate how varying levels of CRs and MPs affect the meme language within an OC and an individual user's meme language across communities. Such insights could help to answer questions about the mechanisms behind the emergence of an OC's culture and whether users adapt their cultural communication according to changing environmental conditions. Thus, our work has great potential to be expanded, initiating novel ways to analyze, track, and compare OCs' online culture and offering profound insights into OC's reciprocal cultural influence.

References

Bauckhage, C. 2011. "Insights into Internet Memes," in *Conference on Web and Social Media*, pp. 42–49.

Carbaugh, D. 1993. "'Soul' and 'Self': Soviet and American Cultures in Conversation," *Quarterly Journal of Speech* (79:2), pp. 182–200.

Courtois, C., and Frissen, T. 2022. "Computer Vision and Internet Meme Genealogy: An Evaluation of Image Feature Matching as a Technique for Pattern Detection," *Communication Methods and Measures* (17:1), pp. 17–39.

Dawkins, R. 2016. *The Selfish Gene*, (40th ed.), Oxford: Oxford University Press.

DeCook, J. R. 2018. "Memes and Symbolic Violence: #proudboys and the Use of Memes for Propaganda and the Construction of Collective Identity," *Learning, Media and Technology* (43:4), pp. 485–504.

Fayard, A.-L., and DeSanctis, G. 2010. "Enacting Language Games: The Development of a Sense of 'We-ness' in Online Forums," *Information Systems Journal* (20:4), pp. 383–416.

Gal, N., Shifman, L., and Kampf, Z. 2016. "'It Gets Better': Internet Memes and the Construction of Collective Identity," *New Media & Society* (18:8), pp. 1698–1714.

Gaudette, T., Scrivens, R., Davies, G., and Frank, R. 2021. "Upvoting Extremism: Collective Identity Formation and the Extreme Right on Reddit," *New Media & Society* (23:12), pp. 3491–3508.

Hall, S. 1997. "The Work of Representation," in *Representation: Cultural Representations and Signifying Practices* (2nd ed.), London: SAGE, pp. 15–64.

Jackson, S. 2011. "Organizational Culture and Information Systems Adoption: A Three-Perspective Approach," *Information and Organization* (21:2), pp. 57–83.

Kozinets, R. V. 2010. *Netnography: Doing Ethnographic Research Online*, London: SAGE.

Marwick, A. E. 2015. "Instafame: Luxury Selfies in the Attention Economy," *Public Culture* (27:1), 137–160.

Massanari, A. 2017. "# Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures," *New Media & Society* (19:3), pp. 329–346.

McSwiney, J., Vaughan, M., Heft, A., and Hoffmann, M. 2021. "Sharing the Hate? Memes and Transnationality in the Far Right's Digital Visual Culture," *Information Communication and Society* (24:16), pp. 2502–2521.

Morina, D., and Bernstein, M. S. 2022. "A Web-Scale Analysis of the Community Origins of Image Memes," in *Conference on Human-Computer Interaction*, pp. 74:1-74:25.

Moser, C., and Deichmann, D. 2021. "Knowledge Sharing in Two Cultures: The Moderating Effect of National Culture on Perceived Knowledge Quality in Online Communities," *European Journal of Information Systems* (30:6), pp. 623–641.

Nissenbaum, A., and Shifman, L. 2017. "Internet Memes as Contested Cultural Capital: The Case of 4chan's /b/ Board," *New Media & Society* (19:4), pp. 483–501.

Nissenbaum, A., and Shifman, L. 2018. "Meme Templates as Expressive Repertoires in a Globalizing World: A Cross-Linguistic Study," *Journal of Computer-Mediated Communication* (23:5), pp. 294–310.

Nowotny, J., and Reidy, J. 2022. *Memes-Formen Und Folgen Eines Internetphänomens*, Bielefeld: transcript Verlag.

Seering, J., Wang, T., Yoon, J., and Kaufman, G. 2019. "Moderator Engagement and Community Development in the Age of Algorithms," *New Media & Society* (21:7), pp. 1417–1443.

Seraj, M. 2012. "We Create, We Connect, We Respect, Therefore We Are: Intellectual, Social, and Cultural Value in Online Communities," *Journal of Interactive Marketing* (26:4), pp. 209–222.

Sherratt, V., Pimbblet, K., and Dethlefs, N. 2023. "Multi-Channel Convolutional Neural Network for Precise Meme Classification," in *Conference on Multimedia Retrieval*, pp. 190–198.

Shifman, L. 2014. *Memes in Digital Culture*, Boston: MIT Press.

Spitzberg, B. H. 2014. "Toward A Model of Meme Diffusion (M ^{3} D): Toward A Model of Meme Diffusion," *Communication Theory* (24:3), pp. 311–339.

Yoon, I. 2016. "Why Is It Not Just a Joke? Analysis of Internet Memes Associated with Racism and Hidden Ideology of Colorblindness," *Journal of Cultural Research in Art Education* (33:1), pp. 92–123.

Appendix A.3: Paper III

Bibliographic data	Henn, T. 2024. Follow the Memeing: Analyzing the Cultural Diffusion between Mainstream and Alt-Right Communities based on Shared Memes. European Conference on Information Systems (ECIS). Paphos, Cyprus.
Abstract	In this paper, I evaluate the role of memes as cultural symbols of online communities and their diffusion across community boundaries. Using the concept of meme languages, I compare examples of cultural symbols shared by alt-right and mainstream communities and analyze their similarities and differences based on a visual feature-matching algorithm introduced by Courtois and Frissen (2022). My findings show that alt-right and mainstream communities possess a greater cultural resemblance between each other than within each community type. As a community's purpose and cultural boundaries significantly impact meme-sharing, I conclude there is no "one-size-fits-all" interpretation of cultural symbols. Moreover, I use the theoretical lens of memetics to evaluate the cultural diffusion process between both communities. I find evidence that several memes cross cultural boundaries after surviving a competitive selection process. My insights expand current research findings on online community boundaries and cultural diffusion, initiating interesting avenues for future research.
URL	https://aisel.aisnet.org/ecis2024/track24_socialmedia/track24_socialmedia/11

FOLLOW THE MEMEING: ANALYZING THE CULTURAL DIFFUSION BETWEEN MAINSTREAM AND ALT-RIGHT COMMUNITIES ON REDDIT BASED ON SHARED MEMES

Completed Research Paper

Theresa Henn, University of Bamberg, Bamberg, Germany, theresa.henn@uni-bamberg.de

Abstract

In this paper, I evaluate the role of memes as cultural symbols of online communities and their diffusion across community boundaries. Using the concept of meme languages, I compare examples of cultural symbols shared by alt-right and mainstream communities and analyze their similarities and differences based on a visual feature-matching algorithm introduced by Courtois and Frissen (2022). My findings show that alt-right and mainstream communities possess a greater cultural resemblance between each other than within each community type. As a community's purpose and cultural boundaries significantly impact meme-sharing, I conclude there is no "one-size-fits-all" interpretation of cultural symbols. Moreover, I use the theoretical lens of memetics to evaluate the cultural diffusion process between both communities. I find evidence that several memes cross cultural boundaries after surviving a competitive selection process. My insights expand current research findings on online community boundaries and cultural diffusion, initiating interesting avenues for future research.

Keywords: Online Communities, Memes, Cultural Diffusion, Social Media.

1 Introduction

Sharing memes has become ubiquitous in our daily online interactions; most often, though, we are unaware of a meme's origin, cultural connotation, and, thus, its potential meaning. This lack of knowledge is unfortunate, as research has proven memes' profound symbolic value for online communities (OCs), representing their culture, ideology, and identity (e.g., DeCook, 2018; Gal et al., 2016; Henn & Posegga, 2023; Nowotny & Reidy, 2022). Yet, while known as a rich "source of cultural production" (Heikkilä, 2017, p. 11), OCs and their shared symbols are still considered "a site of contestations of specific meanings" (Heikkilä, 2017, p. 11), which highlights the importance of enhancing our understanding of memes and their cultural meaning for OCs. By doing so, we can gain profound insights into the cultural expression of OCs, which are of particular interest to information system (IS) scholars. Improving our knowledge of OCs and their expressed culture aids in understanding the formation and sustainability of online collectives (Seraj, 2012), enhancing the understanding of spillover effects between shared online symbols and offline events (Rieger et al., 2021), and expanding our knowledge about how specific cultural artifacts become popular within OCs by "going viral" beyond initial community boundaries (Ling et al., 2021). Hence, it is not surprising that research on online culture in the context of OCs has become more prominent in the IS community over the years (e.g., Fayard & DeSanctis, 2010).

When analyzing memes as cultural symbols of OCs, previous research has stressed two unique characteristics that highlight memes' significant and symbolic role for OCs. First, memes are considered identity-shaping symbols that "function as cues for membership" (Nissenbaum & Shifman, 2017, p. 485), being an integral part of in-group identification and representation online (DeCook, 2018; Rogers & Giorgi, 2024). Thus, posting memes is regarded as "an important facet of identity and community building" (DeCook, 2018, pp. 485–486) that represents an OC's cultural self-conception while establishing a cultural boundary with other OCs (Gal et al., 2016). Yet, while "any specific meme might carry profound meaning within a specific community at a specific moment" (Nooney & Portwood-Stacer, 2014,

p. 249), their second significant characteristic is their evolving and diffusing nature across OC boundaries (e.g., Zannettou et al., 2018). On the one hand, this diffusing nature is related to social media as an environment that fosters online collectives' fluid boundaries; on the other hand, it is rooted in memes' nature of being transformed when shared within the online space (Nowotny & Reidy, 2022). Hence, memes are often described as a "phenomenon that transgresses social and cultural boundaries" (Bauckhage, 2011, p. 42) – also referred to as "memetic diffusion" (Nissenbaum & Shifman, 2018, p. 295).

Memes thus play a dual – and at first glance a potentially conflicting – role in the cultural representation and expression of OCs: while they are cultural and identity-shaping symbols that represent an in-group and serve as a sign of identification, memes also bridge the cultural boundaries of OCs by diffusing across the web. Expanding our knowledge on both characteristics is of great value to IS research, as it allows us to gain insights into several important phenomena, such as online radicalization processes (Gaudette et al., 2021), the reciprocal cultural influence of OCs and the establishment of boundaries between them (Massachs et al., 2020), and the identification of cultural artifacts that play a pivotal role in shaping popular online culture within and across OC boundaries (Nowotny & Reidy, 2022).

Since they are creative recombinations of culturally charged elements that go beyond the simple compilation of textual and visual features (Nowotny & Reidy, 2022), previous research has stressed the need for "new literacies" (Knobel & Lankshear, 2007) and conceptualizations to evaluate memes. These advancements are necessary to better understand memes – once referred to as "conceptual troublemaker[s]" (Shifman, 2013a) – as symbols that represent a cultural facet of OCs. Thus, to evaluate and understand their characteristic cultural value for individual OCs and their cultural bridging function between OCs, we need conceptual approaches that account for memes' unique characteristics.

With this paper, I add to this line of research by using the concept of "meme language" as introduced by Henn and Posegga (2023), that is, "the composition of distinct memes and their variants shared within an OC" (2023, p. 2), which is distinct for each OC and can be understood as an expression of its "shared culture, norms, and identity, demarcating [it] from other OCs" (2023, p. 2). The authors' concept builds upon the theoretical work of Stuart Hall (1997), whose two systems of representation theory allows for the connection between memes as symbols that, in their distinct combination, allow for the creation of an OC's individual cultural meme language. To operationalize meme languages for different OCs, I apply a method introduced by Courtois and Frissen (2022), who detect similar meme templates constituted of memes that visually resemble each other (for a similar application, see Henn & Posegga, 2023) – so-called "meme families" (e.g., Lugea, 2019). The composition of these meme families then defines an OC's distinct meme language, which is an expression of its culture based on shared symbols.

I use the social media platform Reddit to apply the concept of meme language to different OCs and evaluate the cultural boundaries of and diffusion between them. Reddit brings together users who "participate in, contribute to, and even develop like-minded communities whose members share a common interest" (Gaudette et al., 2021, p. 3493) and is, therefore, a commonly used platform for studying OCs (e.g., Massachs et al., 2020; Morina & Bernstein, 2022). Its so-called "subreddits" feature particular community rules and moderation policies that naturally structure Reddit's communication space into distinct OCs, making it a suitable platform for this analysis.

On Reddit, I apply the concept of meme language to subreddits associated with two different types of OCs to analyze how they resemble each other in their cultural expression based on their shared memes. The first type are subreddits whose sole purpose is the creation and sharing of memes without being linked to any particular political orientation or ideology. I subsume these subreddits under the term "mainstream OCs" since they are highly popular for meme-sharing on Reddit (based on the number of submissions and members) and without an additional topical focus besides memes, attracting a large number of users who contribute, share, and engage with memes on the platform. I compare these mainstream OCs, representing the general meme-sharing discourse on Reddit, with OCs that feature explicit cultural boundaries and a clear understanding of in and out-groups. The second type are an example of ideologically coined OCs with explicit cultural boundaries – subreddits associated with the alt-right movement (i.e., the "alternative right"), "a right-wing, primarily online political movement or grouping based in the U.S. whose members reject mainstream conservative politics and espouse extremist beliefs and policies typically centered on ideas of white nationalism" (Merriam-Webster, 2023). The movement gained momentum during the 2016 U.S. presidential election campaign as supporters of Donald Trump and is known for its explicit cultural boundaries, a clear delineation between in- and out-groups, and a "distinct verbal and visual language that is characterized by the use of memes, subcultural terms, and references to the wider web culture" (Rieger et al., 2021, p. 3). I call these subreddits "alt-right OCs."

I chose to analyze the cultural expression of and diffusion between mainstream and alt-right OCs based on shared memes for three reasons. First, the two types of communities differ significantly in their ideological orientation and purpose. Whereas mainstream OCs are devoted to creating and sharing memes fostering rather loose community boundaries, OCs associated with the alt-right are communities that represent a concrete normative and ideological worldview, creating explicit community boundaries that clearly distinguish between in- and out-groups. Thus, it is highly interesting to evaluate how different levels of concrete community boundaries and disparities in communities' main purpose affect the cultural expression of and reciprocal influence and diffusion between OCs. Second, "culture [for the alt-right] is seen as the primary value and a site of political action, often expressed in racial terms as preservation of white culture and identity" (Heikkilä, 2017, p. 3). Hence, culture and identity play a pivotal role in the alt-right's self-conception, making the movement a particularly interesting case for analyzing how they use symbols to express their culture and identity online. Finally, research has shown that even though the alt-right describes itself as an "alternative" to mainstream politics and curates the narrative of being anti-establishment (Heikkilä, 2017), groups related to the alt-right use memes as "weaponized irony" (Nowotny & Reidy, 2022, p. 139) to expand their reach to mainstream audiences, younger people in particular (DeCook, 2018), and win them over (Nowotny & Reidy, 2022). Since memes shared by the alt-right commonly feature hateful, racist, and misogynistic content (Rieger et al., 2021) and memes' playful character makes it more likely to spread these previously too controversial topics into the mainstream (Heikkilä, 2017), it is important to evaluate the extent to which memes bridge the cultural boundaries between alt-right and mainstream OCs and reach a broader audience.

So, to analyze how mainstream and alt-right OCs on Reddit differ in their expressed cultures based on their shared memes and to evaluate which memes might bridge the cultural boundaries between the two types of OCs, I pose two research questions:

RQ1: Does mainstream community culture – expressed through the sharing of memes – resemble that of alt-right communities on Reddit?

RQ2: Which memes bridge the cultural boundaries between the two types of communities on Reddit?

To answer my research questions, I apply the concept of meme languages to both types of OCs. By evaluating and comparing the meme languages of mainstream and alt-right OCs on Reddit, I can determine the cultural resemblance of each OC type based on common memes while also analyzing memes that bridge the cultural boundaries of the two types of OCs, indicating a cultural diffusion between them.

This paper is structured as follows. In the next section, I give an overview of the theoretical background by briefly defining the main concepts "culture" and "meme" and summarizing the current state of literature regarding memes as cultural symbols of OCs that can also bridge OC boundaries by diffusing through the Internet. In this context, I highlight memes' connection to the alt-right movement by illustrating current research findings on how groups on the alt-right spectrum use memes to expand their reach toward mainstream audiences. I then briefly introduce Stuart Hall's (1997) idea of two systems of representation as my theoretical basis for understanding memes as an expression of an OC's culture. Here, I pay particular attention to Richard Dawkins' (2016) idea of memetics as a way to explain cultural diffusion online. I next describe my data collection on Reddit, the extraction of images and memes, and the visual matching between memes based on the method of Courtois and Frissen (2022). After applying this method to each OC, I obtain an OC's characteristic meme language. I compare all meme languages with respect to their similarity and analyze memes that bridge the cultural boundaries between the two types of OCs. To answer my two research questions, I discuss the findings of my analysis by focusing on community boundaries and memes that appear in both mainstream and alt-right OCs. I end the paper by summarizing my main findings and highlighting limitations and potential avenues for future research.

2 Theoretical Background

2.1 Memes as diffusing cultural symbols between OCs

This paper's understanding of "culture" is guided by the definition of Donal Carbaugh, a researcher who approaches culture from a communication perspective. He describes culture as "a system of symbols, premises, rules, forms, and the domains and dimensions of mutual meanings associated with these" that is "implicated, employed, or creatively invoked in conversation" (Carbaugh, 1993, p. 182). This definition of culture stresses the importance of conveying certain symbols and norms with a specific meaning through conversation understood by people who share a common cultural understanding.

I decided on this culture definition because it stresses the importance of communicating specific signs, which we can easily translate to the online sphere. Thus, we can follow users' "digital trace data," that is, their communication traces on social media. For this analysis, I evaluate memes as a specific type of symbolic trace, as images on social media platforms have become increasingly popular over the past years (e.g., Nowotny & Reidy, 2022), and "memes and digital culture seem like a match made in heaven" (Shifman, 2013b, p. 7), with memes being an essential part of digital culture (Rogers & Giorgi, 2024), often described as identity-shaping symbols for OCs (e.g., DeCook, 2018; Shifman, 2013a).

The term "meme," in today's understanding, was coined by Richard Dawkins (2016), who described it as "the idea of a unit of cultural transmission, or a unit of *imitation*" (2016, p. 249, emphasis in original). For Dawkins (2016), memes have an impact on a society's cultural and social evolution that is similar to that of genes in biology: they are (cultural) artifacts that spread and duplicate through communication and can, therefore, be inherited by different people (Dawkins, 2016; Nowotny & Reidy, 2022). Although the analogy between memes and genes has been criticized for falling too short by comparing two artifacts with different transmissions logics (Lugea, 2019), the idea of understanding memes as "contagious ideas, styles and modes, and ways of doing things that spread from person to person within a culture" (Lankshear & Knobel, 2019, p. 43) has become established. Transferring Dawkins' understanding to the online world, Internet memes can be understood as "digital variants" of the initial definition (Merrill & Lindgren, 2021, p. 2404) that emphasize "'catchy' and widely propagated ideas and phenomena" (Knobel & Lankshear, 2007, p. 201) rather than Dawkins initial and more extensive definition of memes (Lugea, 2019). Therefore, Internet memes are considered "units of popular culture that are circulated, imitated, and transformed by individual Internet users, creating a shared cultural experience in the process" (Shifman, 2013a, p. 367) that is interpreted through "socially constructed rules and conventions" (Rogers & Giorgi, 2024, p. 75).

In this paper, I focus on a specific type of meme, so-called "image macros," – which "have become an established sub-genre" (Lugea, 2019, p. 81) of memes known as "a repeating image overlayed with text and usually deal[ing] with a specific topic or situation" (Nissenbaum & Shifman, 2018, p. 298). A key characteristic of image macros is that they are transformed when shared online by exchanging visual and textual features of the original image (Shifman, 2013b). This continuous recombination of culturally charged elements carries profound cultural meaning for different OCs and fundamentally differentiates a meme from other types of images shared online (Nowotny & Reidy, 2022). Although image macros do not express the rich diversity of all possible online memes, such as dance video remakes, gifs, and text-only memes (Lankshear & Knobel, 2019, p. 49), nor do they fully account for Dawkins' initial extensive definition of memes (Lugea, 2019), I decided on this type of meme because it is one of the most popular types shared online and has been used as the analytical basis for several works (e.g., Nissenbaum & Shifman, 2018). As short video platforms such as TikTok rise in prominence, it will become more important to consider in particular memes in the form of videos for future analysis.

The notion of memes from the early 2000s as "seemingly superficial and trivial elements of popular culture" (Johnson, 2007, p. 27), has been increasingly challenged over the past decade with recent research findings proving memes' symbolic and identity-shaping character for specific OCs (DeCook, 2018; Gal et al., 2016). Posting memes is considered a sign of identity work (Lankshear & Knobel, 2019) and community affiliation that aids OCs to "strengthen their own sense of an 'in-group'"

(DeCook, 2018, p. 489). Hence, memes can be regarded as identity-coining symbols that "shape the mindsets, forms of behavior, and actions of social groups [online]" (Shifman, 2013a, p. 365) and are an expression of an OC's culture. For instance, memes' symbolic and identity-shaping effect has been evaluated in the context of the online LGBTQ community by analyzing the "It Gets Better" suicide-prevention campaign on YouTube as an example of performative meme culture. The analysis showed that memes aid online groups in constructing their collective identity and norms (Gal et al., 2016).

Even though memes typically come from an identifiable origin (Nowotny & Reidy, 2022) and are often associated with specific OCs (DeCook, 2018), they are also known to diffuse across the web by being altered and modified when traversing OC boundaries (Bauckhage, 2011; Nowotny & Reidy, 2022; Zannettou et al., 2018). Research has shown that memes posted on different platforms such as Twitter, 4chan, and Reddit, and within different types of OCs diffuse to different platforms and communities, reciprocally impacting the posting of further memes (Zannettou et al., 2018). Similarly, when analyzing memes' diffusion process between core and periphery communities across "all indexable online communities that principally share meme images with English text overlays" (Morina & Bernstein, 2022, p. 74:1), Morina and Bernstein showed that – contrary to their assumptions – memes posted within periphery communities are less likely to diffuse. The analysis of meme evolution within pro-vaccination and anti-vaccination communities showed that while both communities "comprehend memes from totally different perspectives and deliver opposing opinions" (He et al., 2016, p. 61), they used similar memes when discussing vaccination. In a similar vein, memes related to Captain America have been shared by supporters of the social justice movement and adherents of the alt-right, while memes featuring Pepe the Frog – currently associated with the alt-right – have also been used by the Hong Kong protestors as a sign for their fight for democracy (Nowotny & Reidy, 2022). Thus, "memetic logic" works in both directions: initial social justice memes can be captured by movements associated with the political right and vice versa (Heikkilä, 2017). These findings stress memes' bridging function between OCs, even if the communities are at opposite ends of the (political) spectrum.

Several researchers have addressed the use of memes by the alt-right movement, which is well-known for its frequent meme-sharing, particularly in support of Donald Trump (Heikkilä, 2017) –the aforementioned Pepe the Frog being one of the most prominent symbols (e.g., Nowotny & Reidy, 2022). For instance, DeCook (2018) analyzed the symbolic effects of memes on the identity construction of the alt-right group "Proud Boys." Social media "are a socializing institution for youth," she wrote (2018, p. 502), concluding that sharing ideologically coined memes related to the alt-right not only activates a shared identity but is also used as propaganda to attract recruits. These novel mobilization tactics have been the subject of several recent research projects that have shown alt-right groups purposefully taking advantage of memes' humoristic nature (Heikkilä, 2017; Nowotny & Reidy, 2022), also referred to as "weaponized irony" (Nowotny & Reidy, 2022, p. 139) to reach a broader audience with their ideology. As memes are often perceived as "harmless jokes," implying that the content is generally acceptable, younger users in particular may be unaware of their true ideological meaning and continue sharing memes across the Internet (DeCook, 2018). This diffusion of alt-right content toward the mainstream should, however, be treated with caution, as research has shown reciprocal effects between online meme-sharing and offline events, with, for instance, supporters of the alt-right wearing Pepe the Frog masks during the January 6, 2021, storming of the U.S. Capitol (Nowotny & Reidy, 2022, pp. 164–165).

2.2 Two systems of representation and cultural diffusion through the lens of memetics

I rely on Stuart Hall's (1997) idea of two systems of representation to establish a causal link between memes and their cultural representation of OCs. According to Hall (1997), language and meaning are essential for cultural representation and constitute the two elements of his two systems of (cultural) representation. His first system of representation addresses the "meaning" dimension of his theory by including the notion that every person correlates "all sorts of objects, people and events … with a set of concepts or *mental representations*" (1997, p. 17; emphasis in original). The resulting "conceptual maps" aid in understanding and interpreting the world around us. The second system of representation involves the notion of "a common language system" to communicate our individual conceptual maps with other people. "Language," in Hall's understanding, goes beyond spoken words and comprises all signs that carry meaning, including even facial expressions and images.

Henn and Posegga (2023) translated Hall's second system of cultural representation to the online world by understanding memes as shared signs that, in their composition, create a characteristic cultural language that the authors term "meme language." Based on this understanding, each OC carries an individual meme language comprising a combination of specific memes to represent its culture. However, since the online space fosters fluid boundaries for collectives, and memes are known to continuously transgress social and cultural boundaries (Bauckhage, 2011), online culture – memes in particular – diffuse through the Internet (e.g., Zannettou et al., 2018). Thus, even though some memes might (originally) be tied to a specific OC, they can still be captured by OCs from the (political) opposite spectrum, resulting in "online culture wars" of OCs fighting over a memes' group affiliation (Nowotny & Reidy, 2022, p. 150), such as the aforementioned Pepe or Captain America. Thus, it is likely that parts of individual meme languages blend with the meme languages of other OCs.

When evaluating cultural diffusion within the online sphere and in the context of memes, it is often connected to the "memetics" approach, which explains how online culture is replicated and refined based on memes. Rooted in Dawkins' (2016) introduction of memes as cultural units that are reproduced, altered, and transmitted between people, this approach can be broadly defined as "a field dedicated to understand the spread of information and culture through memes" (Lugea, 2019, p. 82) and is often linked with the analogy between memes and genes. Although the analogy has been criticized, it is still helpful in understanding the evolution of memes, with some memes spreading throughout the web while others do not. Based on the understanding of memetics, memes – like genes – behave in a Darwinian manner: the human consciousness hosts memes just as an organism may host genes (Ermakov & Ermakov, 2021, p. 4) and can, like a contagious process, pass them on, "infecting" other consciousness with the same cultural idea (Ermakov & Ermakov, 2021). In a similar analogy, memes are placed within what is a competitive environment in which an abundance of social media content competes for user attention; thus, "memes suited to their sociocultural environment spread successfully, while others become extinct" (Shifman, 2013a, p. 363). Consequently, memes that resonate with a large audience are more likely to diffuse across the web and are, therefore, likely to be part of multiple meme languages.

Based on these theoretical considerations, I assume that OCs with a more identical thematic focus feature a more similar meme language that expresses their culture than OCs with different themes and understanding of in and out-groups. When applied to this paper's subject, my first hypothesis is:

H1: Mainstream and alt-right OCs share more similar meme languages within each community type than the other community type.

As memes are known to diffuse and evolve through the web, some parts of the meme languages of OCs will resemble each other, as some memes bridge OCs' cultural boundaries even if they feature different topics and cultural boundaries. Moreover, research has shown that the alt-right tries to expand its reach toward mainstream audiences through the playfulness of memes (Heikkilä, 2017). Thus, I expect to detect fragments of mainstream and alt-right meme languages that resemble each other. Based on these considerations, I derive my second hypothesis:

H2: Fragments of alt-right and mainstream OCs' meme language will resemble each other, as some memes bridge the cultural boundaries between the two types of OCs.

By analyzing the result linked to *H2*, I also exploratory evaluate which memes bridge the cultural boundaries between both types of OCs, thereby answering *RQ2*.

3 Methods

3.1 Data collection on Reddit

Reddit, a platform often described as "the front page of the Internet" (Gaudette et al., 2021), is my source for collecting data on mainstream and alt-right OCs. It is particularly popular for political discussions (e.g., Massachs et al., 2020), for example, and for sharing memes (e.g., Zannettou et al., 2018). Since Reddit's internal subreddit structure simplifies the identification of OCs related to the alt-right movement and mainstream meme OCs, it is a suitable platform for my analysis.

One of the most famous subreddits related to the alt-right movement is r/The_Donald, which Reddit banned in June 2020 for violating platform policies regarding hateful content (Gaudette et al., 2021). r/altright, another popular subreddit among the alt-right movement, had been banned in February 2017 for similar violations (Hern, 2017). Another subreddit, r/conspiracy, is also popular among alt-right users and has "significant correlations" with r/The_Donald. (Massachs et al., 2020, p. 50). As all three communities have ranked among the most popular alt-right subreddits, are often used in research on this movement (e.g., Gaudette et al., 2021; Massachs et al., 2020), and have a clear topical focus and cultural boundaries (see subreddits' self-description), I set my data collection timeframe accordingly and downloaded Reddit's submission dump from December 2016, obtained through Pushshift (Baumgartner et al., 2020); December 2016 was the earliest month prior the banning of r/alt-right that included its data. I called this 2.53 GB file, which contains the entire Reddit feed within this period, the "Reddit Dataset."

To compare these three subreddits related to the alt-right movement with the same number of mainstream subreddits devoted to creating and sharing memes as their main purpose, I filtered the December 2016 submission dump toward subreddits containing the word "meme" within their name. To identify subreddits within this reduced sample that attract a large audience interested in sharing memes, thus depicting the "general" meme-sharing discourse on Reddit, I used the number of submissions and identified the three largest subreddits within this period: r/dankmemes, r/wholesomememes, and r/memes. These three subreddits are often used for analyzing meme-sharing on Reddit (e.g., Chabot & Chen, 2020; Morina & Bernstein, 2022) and define themselves as communities dedicated to creating and sharing memes (see subreddits' self-description). I combined these submissions with the submissions detected for r/The_Donald, r/conspiracy, and r/altright and named this the "Submission Dataset."

As my goal is to analyze the memes shared within these six OCs, I downloaded, in the cases possible, all images attached to each submission in the Submission Dataset. This totaled 21,542 images (5.89 GB); I call this the "Image Dataset." To detect memes within these images, I manually screened all images according to the "image macro" definition introduced above. Note that I also included memes featuring Pepe the Frog within this dataset that did not strictly follow the image macro definition by showing, in some cases, only Pepe or a respective mutation without a textual element. I decided on this expansion of the dataset since Pepe depicts one of the most popular memes currently associated with the alt-right and serves as an essential cultural symbol for the movement (Heikkilä, 2017). I opted for a manual screening for two reasons. First, manual screening guarantees high accuracy in correctly identifying memes that follow my definition; the same degree of accuracy cannot be achieved with automatized classification approaches such as convolutional neural networks. Likewise, as the period of analysis takes place right after the November 2016 U.S. presidential election, the Image Dataset likely features several memes related to the election that are very context-specific and could deviate from commonly used memes, which, in turn, might jeopardize automatized methods correctly classifying these images as memes. Second, by screening the images posted within each community, I obtained an initial understanding of the content shared, which was helpful for further analyzing their individual meme language and, in particular, the memes bridging the cultural boundaries between the OCs.

Table 1 provides a concise overview of each subreddit's number of submissions, unique users, images, and detected memes. Note that 318 users appear in both types of OCs compared to 35,739 unique users within all six subreddits, highlighting that both communities attract a different type of user base.

	r/dankmemes	r/wholesomememes	r/memes	r/The_Donald	r/conspiracy	r/alt-right
n of submissions	13,839	4,642	4,338	150,986	14,847	5,536
n of unique users	4,361	2,461	2,457	23,392	3,569	1,046

n of images	3,481	1,715	832	15,158	208	140
n of memes	2,633	1,163	712	4,899	35	45

Table 1. Overview of different subreddits' submissions, users, images, and memes.

3.2 Creating an OC's meme language

After identifying all memes within the six subreddits, I continued to detect specific memes and their visual variants, essential for creating individual meme languages and operationalizing the culture expressed by alt-right and mainstream OCs. To detect memes and their visual variants, I applied the visual feature-mapping algorithm introduced by Courtois and Frissen (2022) (see, Henn & Posegga, 2023 for a similar application). In their paper, Courtois and Frissen (2022) used the algorithm to detect visually similar memes, so-called meme families (Lugea, 2019), which resembles this paper's context and thus makes it a suitable fit for operationalizing meme languages.

The algorithm compares a manually set number m of visual features across images and evaluates whether images contain similar features based on a manually set distance threshold d. Note that the algorithm blurs the textual features of each image to avoid a similarity matching based on the same words. If the compared visual image features do not meet the set threshold d, they are considered unequal and are discarded; if they meet the threshold criterion, they are considered similar and part of one meme family. Hence, images that are modified but contain essential features of the original images are grouped together, which is very helpful when working with memes known to evolve over time. Similar memes are mapped toward a meme network graph G, with memes as nodes connected through edges representing memes that met the visual distance threshold d (and are, therefore, considered similar). The network graph G comprises individual meme families as components B featuring the memes that were detected as similar. The composition of these meme families constitutes each OC's meme language.

In their model, Courtois and Frissen (2022) set the hyperparameters d=27 and m=4. As my application context resembles theirs, I applied the algorithm to the memes of each of the six communities with the same hyperparameter configuration. After manually checking the results, I increased the minimal number of macheted features to m=10 for all components B in each OC, as unequal memes were considered similar multiple times before. In most cases, each OC featured one component B that was much larger than the remaining components, resembling the finding of Courtois and Frissen (2022). For those components, I increased the minimal number of matched features to m=25, as those components featured a compilation of multiple subcomponents with visually similar memes. By increasing m accordingly, I split these large components into smaller subcomponents featuring visually similar memes. Thus, when applying the feature-matching algorithm to the memes $y \in Y$ of each of the six communities, I received a meme network graph G for each community that comprises an individual number of components B (meme families) with memes and their visual variations as nodes $g \in G$. I manually validated each component B regarding its composition of visual similar memes and split the component into subcomponents in the cases necessary (e.g., the component still comprised two or more memes and their visual variants). Figure 1 shows examples of components B representing two meme families as a fragment of r/The_Donald's meme language.

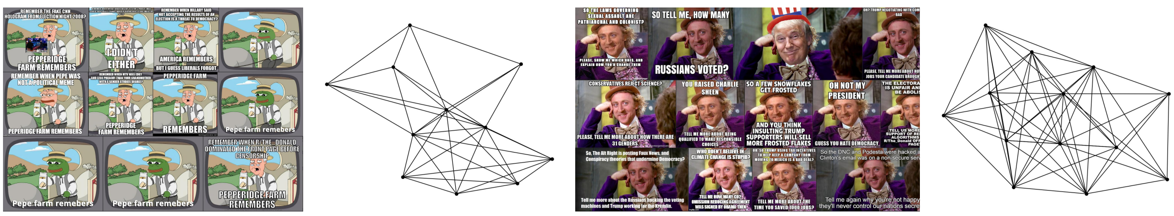


Figure 1. Two examples of meme families of r/The_Donald's meme language and their respective network components B (read from left to right).

4 Results

4.1 Comparing alt-right and mainstream OC's meme languages

Overall, I detected 892 components (sum of components in the second column of Table 2) across all six communities (848 visually distinct components); Table 2 provides an overview of each OC's component structure and matched memes.

Subreddit (OC)	n of components B	n of initial memes $y \epsilon Y$	n of memes as nodes $g \epsilon G$	Percentage of memes used in graph G
r/dankmemes	137	2,633	506	19.22%
r/wholesomememes	144	1,163	422	36.29%
r/memes	32	712	79	11.10%
r/The_Donald	576	4,899	2,127	43.42%
r/conspiracy	2	35	4	11.43%
r/altright	1	45	2	4.44%

Table 2. Overview of different subreddits' component structures and matched memes.

The meme language characteristics of the three mainstream OCs offer interesting insights into their individual culture based on memes. The subreddit r/wholesomememes appears to feature the most similar memes, as more than 36 percent of initial memes were mapped to the network graph G: more than one-third of the memes from r/wholesomememes could be matched with at least one visually similar meme. r/memes features the fewest initial memes of the three mainstream OCs and also exhibits the smallest proportion of memes that could be matched with similar memes (11.10%). This finding indicates that memes shared within r/memes are visually more heterogeneous, leading toward a meme language comprising only approximately one in 10 memes shared within that community. r/dankmemes is situated between r/memes and r/wholesomememes, with about one in five memes matched with at least one visually similar meme (19.22%). Yet, when comparing the number of components between r/dankmemes and r/wholesomememes, it becomes evident that r/dankmemes features more memes originating from the same meme family, as the number of components is slightly smaller than that for r/wholesomememes (137 and 144, respectively).

When evaluating the meme language composition of alt-right OCs, it becomes evident that the subreddit r/The_Donald not only features most initial memes compared to all other OCs (being, again, proof of the alt-right's affinity for sharing memes) but also shares the most similar memes, as more than 43 percent of the memes were matched and turned into nodes of the meme graph G. Thus, nearly half of the memes could be matched with at least one visually similar meme. The OCs r/altright and r/conspiracy, in contrast, contain the fewest number of initial memes compared to all other OCs and share a set of visually diverse memes, as only 4.44 and 11.43 percent, respectively, of the memes turned into nodes of the meme graph G. Note that the meme matching frequency within r/The_Donald is 10 times higher than in r/altright, meaning that memes shared in r/The_Donald are 10 times more likely to feature similar visual characteristics. This finding indicates that r/The_Donald features a meme language comprised of more visually similar memes than either of the other two alt-right-related OCs (and all mainstream OCs).

To answer *RQ1*, I compare the resemblance of each OC's meme language and, therefore, their cultural similarity based on shared symbols. I do so by detecting all components B that were shared within at least two of the six OCs. Table 3 is an overview of the number of the same components that appear in multiple OCs.

	r/dankmemes	r/memes	r/wholesomememes	r/altright	r/conspiracy	r/The_Donald

r/dankmemes		6	11	None	None	11
r/memes	6		5	None	None	14
r/whole-somememes	11	5		None	None	12
r/altright	None	None	None		None	1
r/conspiracy	None	None	None	None		None
r/The_Donald	11	14	12	1	None	

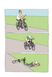
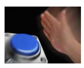
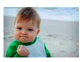
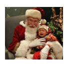
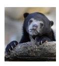
Table 3. Overview of the number of components appearing in different OCs.

When evaluating the components shared by at least two of the three mainstream OCs, I find that r/dankmemes and r/wholesome memes feature 11 of the same components, which is more than twice as many as r/memes shares with r/wholesomememes (five components) and nearly twice as many shared by r/memes and r/dankmemes (six components). Thus, the two former communities share several of the same memes and feature a more similar culture based on common symbols than r/memes and the two communities. The three communities related to the alt-right show a very different picture. The subreddits r/The_Donald and r/altright have only one component in common, and r/conspiracy has none of the same components as the two alt-right communities. Although I can detect only one component match across all three subreddits related to the alt-right, it seems plausible that the subreddits r/The_Donald and r/altright exhibit a greater resemblance in their meme sharing than r/conspiracy, as the latter is known for featuring alt-right content but not for being entirely devoted to the movement. These findings show that OCs related to the alt-right seem to feature more visually different memes compared to mainstream OCs. However, the fact that r/altright and r/conspiracy shared few memes to begin with affects the composition of their meme languages, since it is difficult to match any memes if few are shared.

The comparison *between* mainstream and alt-right OCs reveals several interesting findings. The meme languages of r/The_Donald and mainstream OCs resemble each other more than the meme language of mainstream OCs among each other. r/The_Donald shares 11 common components with r/dankmemes, 12 common components with r/wholesomememes and 14 common components with r/memes (compared to 11 common components between r/dankmemes and r/wholesomememes as the most similar mainstream OC meme languages). When comparing the components of r/altright and r/conspiracy with the ones from the three mainstream OCs, I detect no matches. These findings indicate that based on the components featured in r/The_Donald, alt-right and mainstream OCs share more common memes across the boundaries of the two types of communities as within them. Hence, the cultural symbols shared by the two types of communities diffuse across community boundaries.

4.2 Memes bridging mainstream and alt-right community boundaries

To sustain the previous cultural diffusion findings and answer *RQ2*, I evaluate which memes and their variants bridge the cultural boundaries of both types of OCs. I detected all meme families shared by at least one mainstream OC and r/The_Donald, as it is the only alt-right OC that shared memes appearing in mainstream communities. Figure 2 shows 26 meme families that are part of the meme language of a mainstream community and r/The_Donald, along with their year of origin as obtained from *knowyourmeme.com*, a website widely used in research to obtain memes' metadata (e.g., Bauckhage, 2011).

B_{5}  2011	B_{8}  2011	B_{42}  2015	B_{50}  2015	B_{59}  2012	B_{64}  2015	B_{96}  2012	B_{103}  2015	B_{108}  2014
B_{109}  2016	B_{112}  2014	B_{138}  2007	B_{140}  2011	B_{145}  2007	B_{146}  2012	B_{147}  /	B_{152}  2012	B_{154}  2012

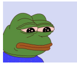
B_{155}  2016	B_{157}  2016	B_{158}  2008	B_{164}  2010	B_{192}  2007	B_{211}  2016	B_{255}  2014	B_{278}  2011

Figure 2. Overview of memes appearing in mainstream OCs and r/The_Donald.

The evaluation of meme families that appear in at least one mainstream OC and r/The_Donald results in the detection of 26 memes that bridge the cultural boundaries between the two types of OCs. Some 16 of these memes feature a human or a part of a human body; nine memes comprise fictional characters (human and non-human) and comics; and one meme depicts an animal. Noteworthy is that 22 of the 26 memes were shared before 2016 (remember, the data are from December 2016), indicating that it is established memes with a historical track record that bridge the cultural boundaries between OCs. These established memes are often shared within r/The_Donald (especially components B_8, B_{152}, B_{164}, and B_{255}). At the same time, this subreddit also features the most recent memes from 2016 (especially components B_{155}, B_{157}, and B_{211}) compared to mainstream OCs. Overall, six users posted memes that were part of r/The_Donald's meme language and of a mainstream OC. Of these six users, three users posted the exact same memes in r/The_Donald and one of the mainstream OCs (meme related to templates B_{147} and B_{211}). Finally, it must be noted that in absolute numbers, Pepe the Frog is the most prominent meme across all communities, with r/The_Donald's meme language featuring 452 Pepe the Frog memes and its variations and r/memes sharing two Pepe the Frog memes (454 Pepe the Frog memes in total).

5 Discussion

When evaluating the expressed culture of mainstream and alt-right OCs and comparing their differences (*RQ1*), two unexpected findings can be observed. First, it is striking that memes shared in r/The_Donald are, by far, visually more similar than the memes shared within any of the other mainstream or alt-right OCs; more than 43 percent of the memes were matched with at least one similar visual variant (compared to, e.g., 11% in r/memes and r/conspiracy). This finding is conceivable when evaluated in relation to the purpose, self-conception, and cultural boundaries of the individual communities. r/The_Donald is known to be the most popular subreddit for users related to the alt-right (Gaudette et al., 2021) and was, therefore, an important, identity-shaping online meeting place for alt-right adherents (before its banning in June 2020). Based on this explicit purpose and self-conception, users are likely to (re)share certain memes that resonate with the groups' norms, values, and ideology. Thus, once memes cross the cultural boundaries of r/The_Donald, they are likely to be reshared, remixed, and reproduced within these boundaries, Pepe the Frog being the most prominent example. This resharing of common symbols – that is, visually similar memes – is essential for the creation of a group's collective identity and demarcation from other communities (e.g., DeCook, 2018), highlighting memes' important role as identity-shaping symbols for the alt-right and an expression of their community culture (DeCook, 2018).

Once these memes are shared continuously within the community's boundaries, they become identifiable as a cultural symbol that is associated with the group and its identity. When continually confronted with memes such as Pepe the Frog, users become "meme literate" (Lugea, 2019, p. 97) and understand which memes represent the group's identity. By sharing these popular memes, users can officially signal their affiliation with the in-group, reinforcing the prominence of these memes while demarcating themselves from other out-groups that do not belong to the community, as suggested by Tajfel and Turner's social identity theory (2004). When not complying with the community's norms and values, memes are considered "rule-breaking" (Lugea, 2019, p. 90) and either seldomly cross the cultural boundaries of r/The_Donald in the first place, or, if they do, are not reshared by the majority of users since they are not considered a symbol of the in-group (for a similar finding on 4chan's /b/ board, please see Nissenbaum and Shifman, 2017). These findings also resonate with Nancy Fraser's (1990) concept of counterpublics: the alt-right considers itself as an "alternative" toward mainstream (conservative) politics (Heikkilä, 2017). This demarcation toward mainstream politics fosters the creation of boundaries toward other groups by "formulat[ing] oppositional interpretations of … identities …" (Fraser, 1990, p. 67), making communication spaces such as r/The_Donald more likely to share symbols that distinguish between in-groups and out-groups. Thus, the disproportionally high number of visually similar memes emphasizes that r/The_Donald depicts an OC with better-defined cultural boundaries and a more concrete understanding of cultural artifacts representing the in-group compared to any other OC examined.

In contrast, the two mainstream communities, r/memes and r/dankmemes, are known for sharing innovative and constantly novel memes as their main purpose. Therefore, the identity-shaping effect of individual memes is less pronounced within these communities; it is the diversity and plurality of memes that constitute their cultural identity, explaining the visual heterogeneity of shared memes. Based on these community characteristics, it seems plausible that only 11.10 percent (r/memes) and 19.22 percent (r/dankmemes) of memes were matched with a visual variant. Unlike these two communities, r/wholesomememes features a higher ratio of visually similar memes (36.29%), yet it is still below the level of r/The_Donald's meme similarity. Although this subreddit is also devoted to the creation and sharing of memes, it slightly follows different meme-sharing norms compared to r/memes and r/dankmemes; r/wholesomememes is known for "revolv[ing] around themes such as meaningful family relationships, friendships, or romantic partnerships," including memes that "feature feel-good stories or messages of thanksgiving and frequently include images of cute animals" (Chabot & Chen, 2020, p. 2). Thus, it is often "dogs and cats dominating the field" (Chabot & Chen, 2020, p. 2), which helps explain the more homogenous meme-sharing compared to the other mainstream OCs. Memes shared in the two alt-right communities, r/altright and r/conspiracy, feature the greatest heterogeneity, as only 4.44 and 11.43 percent of the memes, respectively, were matched with a visually similar meme. This finding can be explained by the overall low importance of meme-sharing within these OCs (only 45 and 35 shared memes, respectively) and the fact that r/conspiracy is not only dedicated to sharing alt-right content but also "devoted to conspiracy theories" (Massachs et al., 2020, p. 55). Thus, memes that are considered a cultural symbol of r/conspiracy representing the norms and values of the OC are visually more diverse since the OC comprises a greater topical spectrum than r/The_Donald. Against this background (less prominence of memes and greater topical focus), it seems plausible that both OCs feature a set of visually diverse memes as an expression of a more heterogeneous OC culture based on shared memes.

The second interesting finding concerns comparing meme languages and expressed culture *between* mainstream and alt-right OCs. When contrasting their meme languages' component compositions, mainstream OCs do not feature more of the same memes among each other. Instead, the meme language of r/The_Donald exhibits more commonality with the meme languages of the mainstream OCs. This finding may seem surprising at first glance, as it contradicts my assumption regarding the similarity of OC culture (communities related to a similar topic were expected to feature a more similar meme language). Yet, when evaluated in light of the previous arguments regarding community purpose, boundaries, and identity, an explanation emerges: memes shared within mainstream OCs that are focused on novel and innovative memes commonly diffuse through the web and are picked up and reused by other communities (see Morina & Bernstein, 2022 for a similar finding). Thus, as the continuous sharing of novel memes constitutes the identity of subreddits such as r/memes and r/dankmemes, the disparity between their shared memes and meme language becomes plausible. Users from r/The_Donald, in contrast, are likely to pick up some of the diverse memes shared within these communities and ultimately coin some of them as symbols of their own in-group. Based on these findings, I find indications to reject *H1*.

The results regarding memes bridging the cultural boundaries between the two types of OCs (*RQ2*) provide two interesting insights regarding the diffusion of cultural symbols online. First, when evaluating the memes that appear in both types of OCs, it becomes evident that a meme's age is decisive for its cultural diffusion across communities. It is memes that have been established for a long or intermediate period, such as the "Success Kid" (B_{138}) and "Xzibit Yo Dawg" (B_{192}) or "Bad Luck Brian" (B_{96}) and "Brace Yourselves" (B_{140}), that predominately crossed the boundaries of mainstream and alt-right OCs. In contrast, new memes (remember, the dataset is from December 2016) also related to specific events, such as Donald Trump's dinner with Mitt Romney (B_{155}) in late November 2016, are less likely to be shared within the two types of communities.

These findings resonate with the theoretical idea of memetics and cultural diffusion online: memes "follow the rules of competitive selection" (Shifman, 2013a, p. 372), meaning they are placed within a competitive environment in which a multitude of cultural artifacts compete for users' attention. Memes that survived this competitive selection process have proven to feature a high level of adaptiveness (Aunger, 2000), allowing them to resonate with different sociocultural environments, that is, their"[s]ocial norms, perceptions, and preferences" (Shifman, 2013a, p. 366). Thus, the "established" memes, measured in years of existence, have survived this competitive selection process and have shown that they resonate with a broader audience. It is, therefore, plausible that memes originating between 2007 and 2015 are more likely to "travel on the digital highway" (Shifman, 2013a, p. 366) between alt-right and mainstream OCs as compared to more recent memes since 2016, which have not yet undergone this selection process. Thus, by becoming detached from their community of origin, these memes diffuse through the web as part of the broader digital culture (Shifman, 2013b).

Different OCs can then use these "universal" meme templates and alter them according to their shared identity and values, making them "community-specific" within OC boundaries and using them as an identity-shaping symbol representing their in-group. Meme researchers are well aware of such "hijacking" of memes and using them as a cultural symbol for specific communities or movements, leading, in extreme cases, to "online culture wars" (Nowotny & Reidy, 2022). Pepe the Frog is a prime example; initially posted within mainstream OCs, it was later adopted by the Trump campaign as an alt-right symbol to "capitalise upon the growing culture of trolling deployed by the alt-right" (Glitsos & Hall, 2019, p. 393). Thus, Pepe is no longer considered a "harmless cartoon character" (Heikkilä, 2017, p. 12) but has become "a political weapon" (Glitsos & Hall, 2019), with mainstream and alt-right communities fighting over its hegemony (Ling et al., 2021) (see also the example of the Captain America meme mentioned in 2.1). Likewise, it is interesting that memes featuring Donald Trump (one with Mitt Romney and one with Ye, formerly Kanye West) appear in alt-right and mainstream OCs. At first glance, one could assume that these types of memes should be associated only with alt-right OCs, as Trump is one of the leading figures for the alt-right. Yet, as Trump has become a part of the overall digital meme culture, Trump memes have diffused toward mainstream OCs (Glitsos & Hall, 2019), indicating the reciprocal cultural influence between alt-right and mainstream communities.

Second, memes featuring persons and figures from movies and comics (e.g., B_8, B_{50}, B_{59}, and B_{152}) are prominent for bridging the cultural boundaries between alt-right and mainstream OCs. It seems plausible that memes featuring more "universal" content that is not necessarily linked to a single event or situation are shared more frequently within different OCs (for a similar finding, see Bauckhage, 2011), as they are quite customizable for different contexts, which is in line with argument of adaptiveness mentioned before (Aunger, 2000). Likewise, the virality and diffusion of memes are highly dependent on the memes' content, with memes featuring characters – that is, "people … or anthropomorphized creatures/objects, such as cartoon characters" (Ling et al., 2021, p. 81:10) – more likely to go viral and be shared by multiple OCs than memes featuring objects or scenes. My results support these previous research findings, as 24 of 26 memes that appeared in alt-right and mainstream OCs include characters. Employing these "easy-to-use" and commonly known memes can also be regarded as a strategic choice, particularly by supporters of the alt-right, to reach broader audiences and win them over, as discussed before. Overall, I find indications to accept *H2*.

6 Conclusion

In this paper, I evaluate memes' role as cultural symbols of alt-right and mainstream OCs and their cultural bridging function across OC boundaries. My findings show that more similar memes are shared between mainstream and alt-right OCs than within each community type (*RQ1*). In line with Henry Jenkins' statement, "If It Doesn't Spread, Its Dead" (2009), my findings also indicate that memes are more likely to bridge the cultural boundaries between the two types of OCs if they survive a competitive selection process (*RQ2*).

These results allow for two significant theoretical contributions, expanding our knowledge of memes as identity-shaping, cultural symbols for OCs and their diffusion across community boundaries. First, memes can serve as cultural symbols of OCs that facilitate the comparison of cultural facets across communities. By using the concept of meme languages, I find evidence that alt-right and mainstream OCs have a greater cultural resemblance based on memes between each other than among each type of community (*RQ1*). For some OCs, a limited number of individual memes serve as an expression of their culture and identity, with, for instance, Pepe the Frog, a popular symbol of r/The_Donald and memes featuring "feel-good" cats and dogs being predominantly shared in r/wholesomememes. If a community is built upon a clear (ideological) understanding of in- and out-groups, such as in the case of r/The_Donald, more homogenous memes will be shared as an expression of the group identity, whereas memes violating the community norms are considered "rule-breaking" and will not be propagated. Communities built upon the idea of sharing a plurality of diverse memes and featuring less strict community boundaries, such as r/memes and r/dankmemes, are, in contrast, characterized by a heterogeneous meme language as an expression of their culture. Thus, we cannot interpret an OC's meme language in a "one-size-fits-all" way; instead, the reciprocal effects between an OC's shared norms, purpose, and identity must be considered when evaluating an OC's meme language and expressed culture based on memes.

These insights allow for an important extension of current research on memes as cultural symbols for OCs as "a key manifestation of participatory [online] culture" (Rogers & Giorgi, 2024, p. 73): It is essential to consider a community's purpose and self-conception when evaluating its expressed culture based on memes. Moreover, concrete cultural boundaries and a strong understanding of in- and out-groups affect the homogeneity of similar memes shared within OCs. These insights aid in better understanding the identity creation of OCs and their demarcation from other groups online, which are essential for an OC's sustainability and longevity (Seraj, 2012). Further, the analysis of memes expands our knowledge of OC boundaries and cultural diffusion processes online, as we see that several memes cross the cultural boundaries between alt-right and mainstream OCs (*RQ2*). Memes with a more extensive historical record appear more often in the two OCs than recently created memes. In line with the theoretical approach of memetics, these findings indicate that memes need to survive a competitive selection process over time to become more likely to diffuse through the web (Shifman, 2013a). Moreover, a meme's adaptiveness and generalizability are essential factors that influence their likelihood of bridging the cultural boundaries between OCs and becoming "famous drops in [the] memetic ocean" (Shifman, 2013a, p. 364). With my particular attention to memes that bridge the cultural boundaries between OCs, this paper's findings aid in expanding our understanding of OCs' reciprocal cultural influence and the kind of memes that connect opposing OCs. These insights are of particular interest to the IS community as enhancing our knowledge of memes' diffusion processes as cultural symbols aids in better understanding phenomena such as online radicalization.

Since memes "have carved out a significant role in public discourse for their ability to spread ideas and influence debates during elections, protests, and social movements" (Rogers & Giorgi, 2024, p. 73), the findings also entail important practical contributions. Previous findings have shown that sharing memes online can have reciprocal spillover effects with real-life events (Nowotny & Reidy, 2022), particularly in the context of the alt-right (Heikkilä, 2017). For instance, as Richard Spencer, one of the leading figures of the alt-right movement and a white supremacist, said regarding Donald Trump's election, "though we may use [memes] half-jokingly, they represent something truly important, the victory of will … We willed Donald Trump into office, we made this dream into reality" (Heikkilä, 2017, p. 13). Thus, increasing the knowledge of memes as cultural symbols of OCs and memetic diffusion is important for academics and policymakers alike, as memes can significantly impact the public discourse, affecting real-life opinions and group behavior (Nowotny & Reidy, 2022).

While these findings provide important insights into memes as cultural symbols for OCs and their role in cultural diffusion processes online, they should be seen in light of their limitations, which offer interesting avenues for future research. First, it would be interesting to dive deeper into mechanisms that steer cultural diffusion processes online by considering the network structures of different OCs, influential users whose level of popularity might impact the virality of certain memes, the organizational structure of OCs, and the role of OC leadership. These additional insights would also allow us to analyze the direction of cultural diffusion and expand our knowledge of influential communities affecting the likelihood of memes going viral (for a first approach, see Morina & Bernstein, 2022). Moreover, these insights would provide the basis for answering broader questions, such as whether the organizational structure of communities affects the cultural diffusion of their shared symbols and whether the cultural diffusion between OCs depends on memes' features or user characteristics. In this paper, I studied meme-sharing based on image macros between two types of OCs on Reddit. Although it is a valid first step, as Reddit's natural community structures provide a solid basis for comparing different types of OCs and image macros depicting one of the most popular types of memes, it would be very interesting to expand this analysis by evaluating different types of memes shared on multiple platforms. Since memes' "composition varies depending on the site where they are sourced" (Rogers & Giorgi, 2024, p. 85), analyzing, for instance, video-based memes on platforms such as TikTok would be a valuable extension of current research on online culture. Doing so would allow us to expand our knowledge of different types of memes, how different platforms foster different types of OC structures, and their effects on meme-sharing and diffusion. Moreover, in this study, I analyzed the meme-sharing between two types of OCs to better understand their cultural similarity based on shared symbols and their reciprocal cultural influence. To expand this analysis by also emphasizing the importance of cultural boundaries, it would be very interesting to analyze memes shared exclusively by individual OCs. This would allow for expanding the analysis toward "core" memes that could provide profound insights into an OC's identity and differentiation between in- and out-groups. Finally, while evaluating an OC's culture based on the visual similarity of memes is a valid first approach, an important next step is to combine the visual analysis with the analysis of a meme's content and featured emotions. This combined evaluation would allow us to gain deeper insights into individual memes and expand our knowledge of how visually similar memes become "hijacked" and are shared by opposing communities.

References

Aunger, R. (2000). *Darwinizing culture: The status of memetics as a science*, Oxford University Press.

Bauckhage, C. (2011). "Insights into internet memes," in: *International Conference on Web and Social Media*, Barcelona, Spain.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). "The pushshift reddit dataset," in: *International Conference on Web and Social Media*, Atlanta, USA.

Carbaugh, D. (1993). ""Soul" and "self": Soviet and American cultures in conversation," *Quarterly Journal of Speech* 79 (2), 182–200.

Chabot, R. and Chen, Y. (2020). "Living your best life and radiating positivity: Exploratory conceptions of wholesome memes as the new sincerity," in: *Annual Conference of CAIS.*

Courtois, C. and Frissen, T. (2022). "Computer Vision and Internet Meme Genealogy: An Evaluation of Image Feature Matching as a Technique for Pattern Detection," *Communication Methods and Measures* 17 (1), 17–39.

Dawkins, R. (2016). *The selfish gene*, 40th Edition. Oxford University Press.

DeCook, J. R. (2018). "Memes and symbolic violence: #proudboys and the use of memes for propaganda and the construction of collective identity," *Learning, Media and Technology* 43 (4), 485–504.

Ermakov, D. and Ermakov, A. (2021). "Memetic approach to cultural evolution," *Biosystems* 204, 1–9.

Fayard, A.-L. and DeSanctis, G. (2010). "Enacting language games: The development of a sense of 'we-ness' in online forums," *Information Systems Journal* 20 (4), 383–416.

Fraser, N. (1990). "Rethinking the public sphere: A contribution to the critique of actually existing democracy," *Social Text* 25/26, 56–80.

Gal, N., Shifman, L., and Kampf, Z. (2016). ""It Gets Better": Internet memes and the construction of collective identity," *New Media & Society* 18 (8), 1698–1714.

Gaudette, T., Scrivens, R., Davies, G., and Frank, R. (2021). "Upvoting extremism: Collective identity formation and the extreme right on Reddit," *New Media & Society* 23 (12), 3491–3508.

Glitsos, L., Hall, J. (2019). "The Pepe the Frog meme: An examination of social, political, and cultural implications through the tradition of the Darwinian Absurd," *Journal for Cultural Research* 23 (4).

Hall, S. (1997). "The Work of Representation" in: Hall, S. (ed.) *Representation: Cultural Representations and Signifying Practices*, 2nd Edition, SAGE, pp. 15–64.

He, S., Xiaolong, Z., Wang, J., Chang, Z., Yin, L., and Zeng, D. D. (2016). "Meme extraction and tracing in crisis events," in: *Conference on Intelligence and Security Informatics*, Tucson, USA.

Heikkilä, N. (2017). "Online antagonism of the alt-right in the 2016 election," *European Journal of American Studies* 12 (12–2), 1–23.

Henn, T., Posegga, O. (2023)."What Do They Meme? Exploring the Role of Memes as Cultural Symbols for Online Communities," in: *International Conference on Information Systems*, Hyderabad, India.

Hern, A. (2017). *Reddit bans far-right groups altright and alternativeright.* URL: https://www.theguardian.com/technology/2017/feb/02/reddit-bans-far-right-groups-altright-alternativeright (visited on November 13, 2023).

Jenkins, H. (2009). *If It Doesn't Spread, It's Dead: Media Viruses and Memes.* URL: http://henryjenkins.org/blog/2009/02/if_it_doesnt_spread_its_dead_p.html (visited on November 14, 2023).

Johnson, D. (2007). "Mapping the Meme: A Geographical Approach to Materialist Rhetorical Criticism," *Communication and Critical/Cultural Studies* 4 (1), 27–50.

Knobel, M. and Lankshear, C. (2007). "Online memes, affinities, and cultural production," *A New Literacies Sampler* 29, 199–227.

Lankshear, C. and Knobel, M. (2019). "Memes, macros, meaning, and menace: Some trends in internet memes," *The Journal of Communication and Media Studies* 4 (4), 43–56.

Ling, C., AbuHilal, I., Blackburn, J., De Cristofaro, E., Zannettou, S., and Stringhini, G. (2021). "Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes," in: *Conference on Human-Computer Interaction*.

Lugea, J. (2019). "The pragma-stylistics of 'image macro'internet memes" in: Ringrow, H. and Pihlaja, S. (eds.) *Contemporary media stylistics*, 1st Edition, Bloomsbury Publishing, pp. 81–106.

Massachs, J., Monti, C., Morales, G. D. F., and Bonchi, F. (2020). "Roots of Trumpism: Homophily and Social Feedback in Donald Trump Support on Reddit," *Conference Web Science*, Southampton, UK.

Merriam-Webster. (2023). *Alt-right.* URL: https://www.merriam-webster.com/dictionary/alt-right (visited on November 09, 2023).

Merrill, S. and Lindgren, S. (2021). "Memes, brands and the politics of post-terror togetherness: Following the Manchester bee after the 2017 Manchester Arena bombing," *Information, Communication & Society* 24 (16), 2403–2421.

Morina, D. and Bernstein, M. S. (2022). "A Web-Scale Analysis of the Community Origins of Image Memes," in: *Conference on Human-Computer Interactio*n.

Nissenbaum, A. and Shifman, L. (2017). "Internet memes as contested cultural capital: The case of 4chan's /b/ board," *New Media & Society* 19 (4), 483–501.

Nissenbaum, A. and Shifman, L. (2018). "Meme Templates as Expressive Repertoires in a Globalizing World: A Cross-Linguistic Study," *Journal of Computer-Mediated Communication* 23 (5), 294–310.

Nooney, L. and Portwood-Stacer, L. (2014). "One does not simply: An introduction to the special issue on internet memes," *Journal of Visual Culture* 13 (3), 248–252.

Nowotny, J., Reidy, J. (2022). *Memes-Formen und Folgen eines Internetphänomens*. transcript Verlag.

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., and Groh, G. (2021). "Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit," *Social Media+ Society* 7 (4), 1–14.

Rogers, R. and Giorgi, G. (2024). "What is a meme, technically speaking?," *Information, Communication & Society* 27 (1), 73–91.

Seraj, M. (2012). "We create, we connect, we respect, therefore we are: Intellectual, social, and cultural value in online communities," *Journal of Interactive Marketing* 26 (4), 209–222.

Shifman, L. (2013a). "Memes in a digital world: Reconciling with a conceptual troublemaker," *Journal of Computer-Mediated Communication* 18 (3), 362–377.

Shifman, L. (2013b). *Memes in digital culture*. MIT press.

Tajfel, H. and Turner, J. C. (2004). "The social identity theory of intergroup behavior" in: Jost, J.T. and Sidanius, J. (eds.) *Political psychology*, 1st Edition, Psychology Press, pp. 276–293.

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Suarez-Tangil, G. (2018). "On the origins of memes by means of fringe web communities," in: *Internet Measurement Conference*, Boston, USA.

Appendix A.4: Paper IV

Bibliographic data	Henn, T., Posegga, O. 2024. Peeking behind the Memes: Evaluating the Boundary Work of Online Communities through Shared Memes. International Conference on Information Systems (ICIS). Bangkok, Thailand.
Abstract	This paper evaluates how online communities (OC) use memes as part of their boundary creation vis-à-vis other groups online. To examine how an OC's boundaries are depicted within memes, we build on social identity theory and derive three different types of boundaries essential for demarcating an OC as a distinct actor. To ensure a systematic evaluation of memes' featured boundaries and sustain our findings, we evaluate each meme along with its featured themes and the type of role attributed to its main character(s). This framework guides our qualitative content analysis of memes shared by the subreddit r/The_Donald on the platform Reddit. Moreover, we demonstrate that significant offline events, such as Trump's presidential election, affect the boundary work of this subreddit. Our results underscore memes' merit for establishing boundaries between OCs, highlighting the complex interplay between boundary types, themes, and roles contributing to an OC's boundaries and identity.
URL	https://aisel.aisnet.org/icis2024/socmedia_digcollab/socmedia_digcollab/8

Peeking behind the Memes: Evaluating the Boundary Work of Online Communities through Shared Memes

Completed Research Paper

Theresa Henn
University of Bamberg
An der Weberei 5, 96047 Bamberg, Germany
theresa.henn@uni-bamberg.de

Oliver Posegga
University of Bamberg
An der Weberei 5, 96047 Bamberg, Germany
oliver.posegga@uni-bamberg.de

Abstract

This paper evaluates how online communities (OC) use memes as part of their boundary creation vis-à-vis other groups online. To examine how an OC's boundaries are depicted within memes, we build on social identity theory and derive three different types of boundaries essential for demarcating an OC as a distinct actor. To ensure a systematic evaluation of memes' featured boundaries and sustain our findings, we evaluate each meme along with its featured themes and the type of role attributed to its main character(s). This framework guides our qualitative content analysis of memes shared by the subreddit r/The_Donald on the platform Reddit. Moreover, we demonstrate that significant offline events, such as Trump's presidential election, affect the boundary work of this subreddit. Our results underscore memes' merit for establishing boundaries between OCs, highlighting the complex interplay between boundary types, themes, and roles contributing to an OC's boundaries and identity.

Keywords: Online Communities, Memes, Collective Identity, Social Media, Social Identity Theory

Introduction

We know that memes are essential cultural and identity-shaping symbols for online communities (OC) (e.g., DeCook, 2018; Henn & Posegga, 2023). Their ability to encapsulate profound cultural messages in a highly condensed and often humorous format makes them an important communication medium within OCs, helping shape the discourse on pivotal topics such as protests (Milner, 2013). We also know that sharing memes in accordance with an OC's norms and values is regarded as an indication of membership and "communal belonging" (Nissenbaum & Shifman, 2017, p. 485). What remains unclear, however, is the role memes play in creating and sustaining boundaries between different OCs, thereby demarcating communities as distinct actors vis-à-vis other groups online and contributing to an OC's identity formation.

Increasing our knowledge of the role of memes as identity-shaping symbols that create and sustain boundaries between an OC's members and outsiders is of great value to the information systems (IS) community. It allows us to better understand how OCs develop collective identities and a sense of "we-ness" – a task particularly challenging in fluid environments such as social media (Fayard & DeSanctis, 2010) but essential for an OC's sustainability (e.g., Syed & Silva, 2023). Likewise, given the difficulties in creating and maintaining boundaries in such dynamic environments, this perspective sheds light on how the meme-sharing of users allows OCs to establish themselves as distinct collective actors, thereby shedding light on their users' shared values and ideology, which can be particularly insightful for community-building and moderation (Seering et al., 2019). It enables us to evaluate how meme-mediated identity formation and boundary creation affect user engagement and attachment to an OC, essential factors for an OC's success and longevity (Ren et al., 2012). Further, it provides the opportunity to analyze how memes can be used as "discursive weapons" (Nissenbaum & Shifman, 2017, p. 483) that can foster the creation and maintenance of OCs known for spreading hateful messages, thereby – in the most extreme case – supporting the dissemination of "online radicalisation [and] far-right extremism" (Askanius & Keller, 2021, p. 2523).

To expand this line of research and examine how an OC's boundaries are depicted and reinforced within memes, we build upon Tajfel and Turner's (2004) social identity theory. This theory offers valuable insights into group dynamics, social conflict, and social change; our primary focus, though, is on the framework's concept of differentiating between the positively characterized in-groups and negatively depicted out-groups, effectively creating boundaries between members and outsiders and thereby differentiating OCs and shaping their identities. These processes are often referred to as "boundary work" in related literature (e.g., Flesher Fominaya, 2019, pp. 435–436). Signals of this boundary work can potentially be observed within the digital traces of users' online communication. Given that memes depict a very popular form of these traces (Shifman, 2013a) known to hold considerable importance in shaping and expressing the identity of an OC (e.g., DeCook, 2018; Henn, 2024), we evaluate whether memes feature in-group and out-group references and how these references contribute to the boundary work of OCs.

For this analysis, we have chosen the subreddit r/The_Donald on the platform Reddit as a strong OC example. Before it was banned in June 2020, this subreddit was widely recognized as the primary OC for adherents of the alt-right (i.e., "alternative right") movement as well as for supporters of Donald Trump on Reddit (Trujillo & Cresci, 2022). We opted for this specific OC to evaluate boundary work based on shared memes for three reasons. First, the alt-right is known as a movement inherently positioned as an "alternative" to mainstream politics and embodies an explicit ideological worldview that differentiates between in-groups and out-groups, coupled with a clear comprehension of its own culture and collective identity (Heikkilä, 2017). We anticipate encountering these clearly articulated boundaries and expression of identity within the online meme-sharing culture of r/The_Donald. Second, r/The_Donald serves as a prime example of a highly polarizing OC with explicit ideological views and a strong, distinct identity (Heikkilä, 2017), making it a representative case for studying both its own boundary work and that of other strongly ideologically coined OCs. It is this clear articulation of ideology and identity that makes r/The_Donald particularly well-suited for examining how users' meme-sharing contributes to and reinforces the boundary work of OCs. Therefore, to gain initial insights into how memes aid in an OC's demarcation from others and to understand their role within an OC's boundary formation, we chose this subreddit over more moderate OCs that are likely to exhibit less pronounced signs of boundary work. Third, r/The_Donald and the alt-right are known for their frequent use of memes featuring references to pop culture (Heikkilä, 2017). This particularly playful nature of memes aids these types of OCs in disseminating hateful, racist, and misogynistic content (e.g., Askanius & Keller, 2021) that demarcates the OC from outsiders. Similarly, these OCs also share memes that portray the in-group in a favorable light, such as depicting Trump and his supporters in a heroic manner (e.g., Heikkilä, 2017, p. 12), thereby fostering in-group cohesion. Thus, we anticipate identifying memes featuring these clear boundary references within this subreddit.

Studies demonstrate that certain significant offline events can profoundly influence online behavior within OCs (e.g., Chung et al., 2021). These events have proven to strongly affect a group's self-image, cohesion, and boundary creation, including far-right online groups (Bliuc et al., 2019). Therefore, as "memes can be quickly produced and shared, and therefore can agilely respond to diverse public events" (Milner, 2013, p. 2359), we expect users from r/The_Donald to adjust their meme-sharing practices on Reddit in response to significant offline events pivotal to the subreddit. To understand what constitutes a "significant" offline event that potentially influences meme-sharing and featured boundaries related to in-groups and out-groups within a subreddit such as r/The_Donald, we turn to realistic group conflict (RGC) theory, as introduced by Sherif (1966). RGC theory posits that intergroup hostility arises from competition over scarce resources such as power and influence. Consequently, we assume that offline events related to resource scarcity are likely the enhance the hostility between groups, thereby reinforcing the boundaries between in-groups and out-groups as outlined by social identity theory (Tajfel & Turner, 2004).

For r/The_Donald, we consider the 2016 U.S. presidential election a significant event due to its substantial impact on the reallocation of power and influence (for a similar interpretation, see Oc et al., 2018). The reallocation of these resources could significantly impact Trump supporters, such as the users of r/The_Donald, by shaping how effectively they can pursue their political agenda, which includes, for instance, opposition to multiculturalism and feminism (e.g., Heikkilä, 2017, p. 2). We assume that this significant offline event likely influenced the subreddit's meme-sharing and boundary dynamics, as the alt-right gained prominence as supporters of Donald Trump during the 2016 U.S. presidential campaign, and r/The_Donald was notably active in supporting Trump (Heikkilä, 2017).

To analyze how users from r/The_Donald share memes that feature references toward in- and out-groups, thereby answering the call to expand the knowledge on memes' role as identity-shaping symbols that demarcate different OCs (e.g., Henn, 2024; Nissenbaum & Shifman, 2017, p. 499) while also evaluating the impact of significant offline events on the subreddit's boundary work, we pose two research questions:

RQ1: How do users from r/The_Donald use memes to distinguish between in-groups and out-groups, thereby demarcating the OC from other groups online?

RQ2: Did the election of Donald Trump as the 45th president of the United States affect the boundaries featured in memes shared by users from r/The_Donald?

To answer our research questions, we evaluate memes shared within r/The_Donald in November 2016, which is the month in which the presidential election took place, thus allowing us to evaluate the boundaries featured in memes shared by users from r/The_Donald (*RQ1*) both before and after the election (*RQ2*). For our analysis, we consider only memes with both a visual and textual component, commonly referred to as "image macros" (Nissenbaum & Shifman, 2018). To analyze memes' references to in- and out-groups, we conduct a qualitative content analysis. Based on Tajfel and Turner's (2004) social identity theory, we derive three deductive coding categories that capture the essential function of boundaries as described within this theory. After categorizing each meme into one of these categories or determining whether it does not fit any of them, we continue our qualitative content analysis by systematically assessing each categorized meme along two dimensions: its featured *theme*, representing a more concrete manifestation of a meme's featured boundaries; and the type of *role* attributed to the main character(s) within the meme, since research has shown that roles are particularly relevant for a meme's featured narrative (De Saint Laurent et al., 2021), thereby highly affecting the boundaries it features.

This paper is structured as follows. The next section provides a brief overview of the definition of memes, coined by Richard Dawkins (2016), and summarizes the current status of research regarding memes as identity-shaping symbols for OCs and their use as boundaries with other OCs, particularly in the context of the alt-right movement. We then describe our theoretical foundation, rooted in social identity theory (Tajfel & Turner, 2004), the concept of boundary work (Flesher Fominaya, 2019), and RGC theory (Sherif, 1966). The methods section provides an overview of our data collection on Reddit, identification of memes, and coding procedure. After categorizing each meme and identifying its themes and roles, we present our results and discuss them in light of the mechanisms driving the identified boundary work. We close with a summary of our findings, thus answering our research questions, highlight our contributions, discuss the limitations of our work, and offer avenues for future research.

Theoretical Background

Memes as Identity-Shaping Symbols for OCs

Richard Dawkins, who first used the term "meme," considered it "the idea of a unit of cultural transmission, or a unit of *imitation*" (Dawkins, 2016, p. 249, emphasis in original). In today's understanding of memes, his definition has been subject to transformation (Lugea, 2019) and, little by little, has mutated into memes being known as "units of popular cultural experience that are circulated, imitated, and transformed by individual Internet users, creating a shared cultural experience in the process" (Shifman, 2013a, p. 367).

Memes are known to take on various forms and shapes – videos, photo remixes, images, and so on (Shifman, 2013a). In today's digital culture, though, people who use the term "meme" are most commonly referring to so-called "image macros" (Lugea, 2019), which are "repeating image[s] overlayed with text [that] usually deal with a specific topic or situation" (Nissenbaum & Shifman, 2018, p. 298). These are considered among the most popular memes online since they can easily be altered and remixed through the exchange of visual and textual elements, which contributes to their virality (Shifman, 2013b). We deem this type of meme particularly suitable for our analysis because we can make use of its multimodal nature, as pointed out in several works (e.g., Henn, 2024; Henn & Posegga, 2023), and evaluate how the combination of textual and visual elements contributes to the demarcation of an OC toward other groups online.

There is a long strand of research addressing how meme-sharing can serve as a signal of group affiliation and thus foster a feeling of communal belonging (e.g., Henn & Posegga, 2023; Nissenbaum & Shifman, 2017). By sharing memes that adhere to community-specific conventions, IS research has shown that users can signal their belonging to an OC (e.g., Henn, 2024; Henn & Posegga, 2023), demonstrating sufficient "meme literacy" (Nissenbaum & Shifman, 2017) and "in-group knowledge" (Lugea, 2019, p. 97). Defying these conventions by sharing memes that are considered "rule-breaking" (Lugea, 2019) is, on the other side, "judged as failing to be part of the group" (Nissenbaum & Shifman, 2017, p. 485). Thus, sharing memes aids OCs to "strengthen their own sense of an 'in-group'" (DeCook, 2018, p. 489) by creating a mutual understanding of a common "we" and allowing users to signal their affiliation and "distinguish[] members from outsiders" (Nissenbaum & Shifman, 2017, p. 488).

Another line of research explores the role of memes as identity-shaping symbols representing an OC's culture and values while demarcating it from other groups (e.g., DeCook, 2018; Henn, 2024). For instance, Miltner (2014) thematizes the topic of identity creation based on her analysis of the meme-sharing of "LOLCat enthusiasts." She categorizes users into three distinct groups, showing that they highly differ in the types of memes shared – thereby demarcating themselves from the other groups detected. This boundary-drawing characteristic of memes is also addressed by Zeng and Abidin (2021), who analyzed the use of memes on the platform TikTok with respect to the differentiation and intergenerational conflict between Gen Z and the "Boomer" generation. The authors concluded that memes are highly important for expressing a group's self-image and self-presentation, thereby separating the group from other groups.

Sharing memes as identity-shaping symbols has been particularly prominent in the context of the alt-right (e.g., DeCook, 2018), especially when examining r/The_Donald (e.g., Heikkilä, 2017) and other OCs related to the far-right spectrum (e.g., Askanius & Keller, 2021). DeCook (2018), for instance, evaluated the meme-sharing of the alt-right group "Proud Boys" on Instagram and found that memes serve not only as a recruiting and propaganda tool but also as important symbols for the group's identity creation, which is particularly appealing to boys and young men. In a similar vein, Heikkilä (2017) analyzed the online activity of the alt-right movement during the 2016 presidential election, also focusing on the alt-right's meme-sharing that featured an anti-establishment sentiment. Notably, Pepe the Frog emerged as one of the movement's most prominent memes, as shown in several other works (e.g., Glitsos & Hall, 2019). He concluded that "the alt-right was not so much a political force, but a cultural force online" (Heikkilä, 2017, p. 14), and identified memes and offensive language as pivotal features of the alt-right's identity.

While some memes are temporarily associated with particular OCs as a representation of their identity, such as Pepe the Frog with the alt-right (e.g., Glitsos & Hall, 2019), memes are known to diffuse through the web, becoming part of the broader digital culture (e.g., Shifman, 2013b) while also shifting group affiliations. In extreme cases, these shifting group affiliations can lead to "online culture wars" (Glitsos & Hall, 2019) between different groups fighting over who has hegemony with particular memes. In these "battles," memes are known to "switch" group affiliation, exemplified by Pepe the Frog's evolution from a cartoon character shared by mainstream audiences (Glitsos & Hall, 2019) to a political symbol "hijacked" by many Trump supporters.

After reviewing the literature, it becomes evident that memes play a crucial role in signaling group affiliation through their sharing; are important symbols that shape the identity of OCs, especially within the context of OCs related to the alt-right; and are fiercely contested between different groups, also leading to meme diffusion across the web. What remains unclear, however, is memes' role in establishing concrete boundaries between OCs based on their references to in- and out-groups. IS research has shown, however, that drawing boundaries and creating identities are essential for an OC as they significantly influence its longevity and long-term success (Ren et al., 2012). Hence, with this paper, we fill this research gap and use social identity theory and the concept of boundary work to evaluate the meme-sharing of users from r/The_Donald with respect to in- and out-group references and the OC's demarcation toward other groups.

Social Identity Theory, Boundary Work, and Realistic Group Conflict Theory

Social identity theory, as introduced by Tajfel and Turner (2004), underscores the pivotal role of group categorization in shaping both behavior and attitudes within and across groups. The theory suggests that individuals create their social identity based on group membership through the categorization of themselves and members into "in-groups" and outsiders into "out-groups." Accordingly, individuals strive to establish a positive social identity for themselves by creating "positive distinctiveness through direct competition with the out-group" (Tajfel & Turner, 2004, p. 382). These intergroup comparisons result in "in-group bias—that is, the tendency to favor the in-group over the out-group in evaluations and behavior" (Tajfel & Turner, 2004, p. 373). Thus, to "maintain or achieve superiority over an out-group on some dimensions" (Tajfel & Turner, 2004, p. 378), individuals strive to elevate the status of the in-group by devaluing the out-group through direct comparison. As a result, the in-group is portrayed positively, especially when directly compared to the out-group; in contrast, the out-group is depicted in a notably derogatory manner. Based on these considerations, we identify three types of boundaries for differentiating between in-groups and out-groups, thereby establishing the in-group as the superior actor and contributing to an individual's positive social identity: *in-group favoritism* that portrays the in-group positively stemming from a positive self-conception; *in-group enhancement* achieved through direct comparisons with out-groups, thereby emphasizing the superiority of the in-group compared to the inferior out-group; and *out-group degradation* based on the negative depiction of the out-group.

Differentiating between in- and out-groups is a well-documented strategy in research for fostering and sustaining collectives (e.g., Flesher Fominaya, 2019; Snow, 2001). This process, typically called *boundary work* in the related literature (e.g., Flesher Fominaya, 2019, pp. 435–436), is instrumental in establishing the group as a distinct actor, influencing members' commitment to, support of, and identification with the group (Flesher Fominaya, 2019, pp. 435–436). According to Snow, this differentiation between in- and out-groups is often facilitated by so-called "boundary markers" (2001, p. 7), which are symbolic resources such as songs, slogans, and shared symbols. These markers aid in "distinguishing insiders from outsiders … in a fashion that heightens awareness of in-group commonalities and connections and out-group differences" (2001, p. 7). Hence, these symbolic resources embody the core values of the community, giving "symbolic substance to the claimed distinctive 'we'" (2001, p. 7), and enable members of the in-group to officially pledge allegiance to the group by reproducing these symbolic resources. Simultaneously, these symbolic resources render the in-group detectable by outsiders and demarcate it from other out-groups.

In the online world, users leave traces when communicating that potentially feature signals of boundary work. Memes are a very popular type of trace in today's online interactions (Nissenbaum & Shifman, 2017). Since memes have been proven to be identity-shaping symbols for different OCs that aid users in signaling group membership, as discussed before (see, e.g., DeCook, 2018; Henn, 2024; Nissenbaum & Shifman, 2017), memes can be seen as digital variants of these "symbolic resources," as Snow (2001) states. While all memes shared within an OC contribute to a group's distinct representation of identity (see, e.g., Henn, 2024; Henn & Posegga, 2023), this paper focuses on memes featuring references to in- and out-groups as an expression of an OC's explicit boundary work toward other groups. We chose to prioritize these types of memes because they are particularly significant in shaping a group's identity by clearly distinguishing between members and outsiders (Tajfel & Turner, 2004), thereby embodying the "symbolic substance" crucial for establishing a shared sense of a distinctive "we" as defined by Snow (2001, p. 7).

To better understand the boundaries depicted in memes, it is crucial to explore the narratives they convey. Research highlights that the characters featured in memes and the way they are framed play a pivotal role in influencing the "partial stories" conveyed within memes and their narrative construction (De Saint Laurent et al., 2021). This focus on specific characters is particularly pronounced within the alt-right, where figures such as Donald Trump are most often depicted in a heroic light, profoundly influencing the overall narrative (e.g., Lamerichs et al., 2018). Building upon the extensive literature on narrative construction, De Saint Laurent et al. (2021) identified four prominent roles commonly associated with characters in memes: *hero, victim, persecutor*, and *fool*. These categories offer valuable insights into classifying the main characters in memes and enabling a deeper understanding of the broader narrative featured within a meme (De Saint Laurent et al., 2021, p. 5). Given that these narratives strongly influence how boundaries are presented within memes—and considering the prominent role of Donald Trump and similar figures in our context—evaluating these roles is particularly useful for analyzing the boundaries memes feature.

Applying these considerations to our concrete research context, we expect users from r/The_Donald to share memes featuring explicit references to in-groups, out-groups, and roles. Sharing such memes serves a dual purpose: it acts as an assertion of their positive in-group identity, demarcating them as a distinct and superior actor, while simultaneously functioning as a means of boundary reinforcement vis-à-vis other groups, depicting the out-groups in a negative light. Given r/The_Donald's association with the alt-right, a movement characterized by its explicit ideological worldview and its clear differentiation between in- and out-groups, (e.g., Heikkilä, 2017), we anticipate encountering these clearly defined boundaries reflected in the meme-sharing within r/The_Donald as an important facet of the subreddit's identity representation.

Moreover, research has shown that offline events can impact online interaction within OCs, such as "fostering interactions between actors and communities of diverse backgrounds" (Chung et al., 2021, p. 1). Given their influence on user interaction, it seems likely that offline events also impact how OCs create and express their identity and boundaries toward other groups. In their work on identity formation in far-right online communities, Bliuc et al. (2019), for instance, showed that offline events such as riots can profoundly influence a group's self-image and perception of in- and out-groups online. Given the adaptable nature of memes, which can be rapidly modified in response to contemporary events (Milner, 2013, p. 2359), we also expect to detect a change in the meme-sharing of users from r/The_Donald after Trump's election.

To expand the research on the interplay between offline events and online user behavior, we, therefore, assess whether this offline event impacted the sharing of memes by r/The_Donald users. To underpin this analysis theoretically, we turn to *RGC theory* (Sherif 1966). RGC theory suggests that in situations with conflicting goals and a scarcity of resources such as power and influence, we can expect a heightened hostility toward the out-group. This shift in sentiment toward the out-group is driven by the perceived threat of the out-group's success, leading to the in-group's failure and diminished access to valuable resources. The competition over these resources strengthens the identification with one's own group and enhances the differentiation between in-groups and out-groups, as outlined by social identity theory and the concept of boundary work. Thus, offline events that affect the distribution of resources and amplify the competition between groups are likely to impact the boundaries depicted within memes, reflecting in-group and out-group dynamics. Since an election involves a profound competition for resources (such as power) and in a U.S. presidential election only one candidate (and party) can win, we assume that this offline event affects the portrayal of featured boundaries within memes pertaining to *in-group favoritism*, *in-group enhancement*, and *out-group degradation* (for a similar analysis, please see Oc et al., 2018).

Figure 1 is an overview of our theoretical considerations regarding meme-sharing and boundary work.

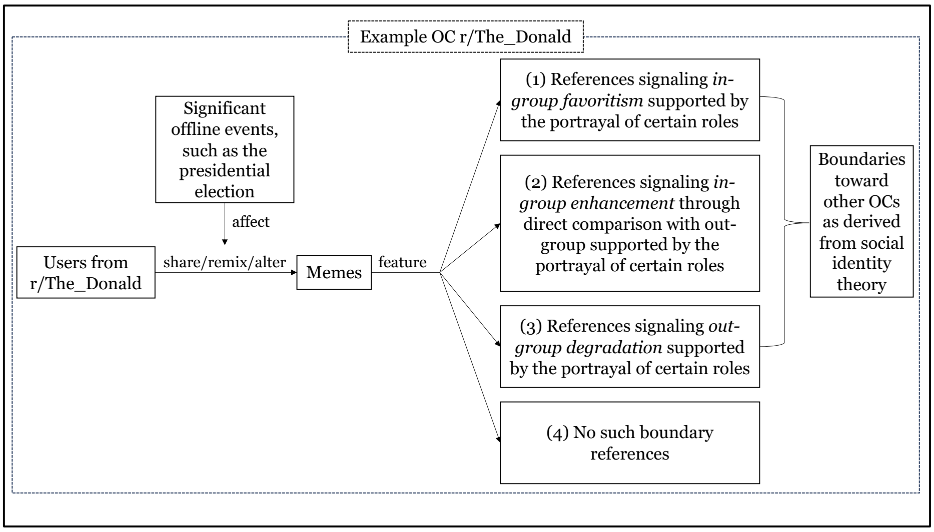


Figure 1. Visualization of Theoretical Considerations Regarding Meme-Sharing and Boundary Creation toward other OCs Based on Social Identity Theory

Methods

Data Collection on Reddit

For our data collection, we use the social news aggregator platform Reddit, commonly referred to as "the front page of the Internet." Reddit ranks among the most popular social platforms on the web and is a frequent data source in academic studies (e.g., Henn & Posegga, 2023). Reddit is known for its distinctive structure of subreddits, which are "like-minded communities whose members share a common interest"

(Gaudette et al., 2021, p. 3493). This structure is particularly useful for our analysis, as it facilitates the detection of OCs and enables us to assess their boundary work, as exemplified in the case of r/The_Donald as the most prominent OC for alt-right adherents and Trump supporters before its banning in June 2020 (Trujillo & Cresci, 2022). Moreover, Reddit is known for its relatively weak moderation policies, "stop[ping] short of preventing users from posting controversial ideas" (Gaudette et al., 2021, p. 3494). While this weak moderation could potentially have negative consequences, such as allowing the spread of hateful content (e.g., Gaudette et al., 2021), it also allows platform users to clearly express boundaries toward other groups. This aspect is crucial for our analysis, rendering Reddit a suitable platform for our work.

Prior to its ban, r/The_Donald was known as the most prominent community space for alt-right and Trump supporters, particularly discussing topics related to Trump's campaign and then presidency with weak moderation policies that permitted, for instance, the expression of hateful sentiments toward out-groups, such as anti-Muslim and anti-immigrant views (Gaudette et al., 2021, pp. 3494–3495). We anticipate observing these boundaries in a pronounced manner within the shared memes in r/The_Donald as we analyze data from November 2016, a period that includes the presidential election itself. Based on the assumptions of RGC theory, we expect this timeframe to capture a peak of r/The_Donald's boundary work, aimed at increasing in-group cohesion, thereby promoting Donald Trump and demarcating the OC from other out-groups, as the election depicts a significant competition over resources.

To obtain data shared in r/The_Donald, we used the Pushshift submission dumps from November 2016 and extracted all submissions shared within r/The_Donald (292,326 unique submissions posted by 43,427 unique users). Whenever possible, we downloaded the attached images, resulting in 24,943 images (7.02GB). To detect memes within these images, we went through all of them manually and labeled an image as "meme" if it followed our definition of an image macro. We opted for a manual coding procedure to familiarize ourselves with the data while also assuring high labeling accuracy that would not have been possible with an automatized approach. After screening all images, we detected 7,239 memes (1.86GB).

Each submission on Reddit receives a score based on the difference between its up- and downvotes, indicating how well the shared content aligns with the community's norms and identity (Reddit, 2023). Memes that deviate from a group's self-image and identity are likely to be disapproved of by its members, resulting in a lower score. We used this score metric to detect the most popular memes among the OC before and after the election. We separated the detected memes into those posted on or prior to November 8 (2,455 memes) and those posted from November 9 to November 30, 2016 (4,784 memes). We sorted memes within both these groups according to their scores, following a method similar to Gaudette et al. (2021), who analyzed prominent comments on r/The_Donald according to their scores. We then identified the 500 most popular memes from each period, resulting in a total of 1,000 memes. These memes, being the highest-rated within the observation periods, are expected to convey significant signals about the identity and boundaries of r/The_Donald, as they depict those memes that receive broad support among OC members.

Operationalizing Boundary Work Featured in Shared Memes

To evaluate the boundaries featured in memes shared within r/The_Donald, we develop a manual coding scheme based on the boundary types as derived from social identity theory. We conduct a qualitative content analysis and scrutinize each meme with respect to whether it includes references signaling *in-group favoritism, in-group enhancement, out-group degradation*, or *no such boundaries*. To obtain an initial understanding of what content relates to the in-group and out-group in our specific context, we rely on the insights of previous research on this OC, such as Gaudette et al. (2021), who examined comments shared within r/The_Donald concerning references to perceived enemies and supporters of the group. The authors detected a particularly anti-Muslim and anti-Left sentiment within the OC.

Table 1 is an overview of the different types of boundaries, along with their corresponding definitions and example memes illustrating each respective boundary.

Type of Boundary	**Definition**	**Example Meme**
in-group favoritism	The in-group is portrayed in a favorable and positive light without being compared directly with the out-group.	

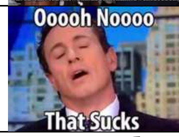
in-group enhancement	The superiority of the in-group is illustrated through comparison with the out-group, which is portrayed as inferior to the in-group.	
out-group degradation	The out-group is depicted in a derogatory and unfavorable manner without being compared directly with the in-group.	
no boundary reference	None of these references are present within the meme.	

Table 1. Overview of Boundary Types Including a Definition and Meme Example

After labeling each meme accordingly, we continue with the second part of our qualitative content analysis to explore reoccurring patterns in in- and out-group references by systematically evaluating each meme according to its featured *theme* (inductive coding) and depicted *roles* (deductive coding). For the former, we assess the concrete manifestation of the type of boundary featured within the meme, thereby receiving a deeper understanding of the OC's boundary work; for the latter, we rely on the role typology outlined by De Saint Laurent et al. (2021), who classified characters in COVID-19-related memes into four derived roles: *hero, victim, persecutor*, and *fool*. According to their typology, heroes are characterized as benevolent and strong figures; victims are portrayed as benevolent, but are "harmed by a situation or the actions of others" (De Saint Laurent et al., 2021, p. 5); persecutors embody the role of villains: and a fool is a character "who just does not 'get it' or acts in an embarrassing manner" (De Saint Laurent et al., 2021, p. 4).

For the deductive coding process, two researchers engaged in several rounds of open coding to familiarize themselves with the data. After numerous discussions about the application of each boundary type and role, they reached a consensus for each individual code. In our inductive coding process, we opted for a grounded theory approach, following the methodology outlined by Thornberg and Charmaz (2014). Thus, we conducted both initial and focused coding to derive our final theme codes. The two researchers used the insights gained from the deductive coding to discover "the most significant or frequent initial codes that make the most analytical sense" (Thornberg & Charmaz, 2014, p. 158). Subsequently, the researchers defined the main themes present in the data by aggregating these initial codes (initial coding) to the final six theme codes (focused coding), which were then mapped to the different types of boundaries from our deductive coding and served as subcategories. Thus, the types of boundaries illustrate the primary function of a boundary featured in memes, while the themes represent a more concrete manifestation of this function. After agreeing on the final boundary type, theme, and role codes, one researcher systematically applied them to each meme in the sample.

Results

After applying the coding scheme to our data, for each type of boundary, we discovered two related overarching themes. For *in-group favoritism*, our coding unveiled the themes *caring*, where the meme portrays in-group members supporting each other, and *strong leader*, with memes featuring leading figures of the in-group in, for instance, a heroic position. With respect to *in-group enhancement*, we identified the themes *dominance*, where the meme illustrates the in-group asserting superiority over the out-group, and *victimization*, with memes portraying the in-group as unjustly targeted by the out-group. For *out-group degradation*, we observed the emergence of the themes *exclusion*, where memes signal that the in-group does not consider the out-group as a legitimate actor, and *incivility*, wherein memes portray the out-group in a derogatory manner using vulgar language or exaggerated visual representations. Table 2 is an overview of our coding scheme, including the type of boundary, themes, their definition, and theme and meme examples.

Type of Boundary (Deductive)	**Themes (Inductive)**	**Definition of Theme**	**Concrete Theme Examples**	**Example Meme**

in-group favoritism	caring	Meme contains signals of caring directed toward the in-group. Caring can e.g., manifest in providing security for people from the in-group in need.	Welcoming everyone to the in-group, such as Pepe the Frog signaling that he supports everyone; signals to "we as a group can do it," such as winning the meme war together or aiding allies such as Marie Le Pen in France.	
	strong leader	Meme portrays figures from the in-group in a heroic light, suggesting they will lead the in-group to success.	Leaders such as Donald Trump are depicted in a heroic light; one central figure of the in-group is depicted with typical American signs such as the flag or a bald eagle.	
in-group enhancement	dominance	Meme contains signals showing that the in-group dominates the out-group, which can be on an individual or group level.	The in-group defeats the out-group with, e.g., Trump supporters being depicted as strong soldiers compared to weak Democrats; the in-group is more popular than the out-group with, e.g., Trump shown as supported by a broad mass compared to Clinton with no supporters.	
	victimization	Meme features signals that the in-group is the victim of the out-group.	The in-group is bullied by the out-group with, e.g., the media depicting Trump and his supporters in a negative light, such as with an accusation (perceived by the in-group to be false) that Trump and his supporters are racist.	
out-group degradation	exclusion	Meme contains signals indicating that the in-group does not consider the out-group to be a legitimate actor.	The out-group is highlighted as different by, e.g., depicting the media as an untrustworthy institution; or negative sentiments toward the out-group are articulated, such as depicting Hillary Clinton as untrustworthy based on her leaked e-mails.	
	incivility	Meme portrays out-groups in a highly derogatory manner, using vulgar language and/or an exaggerated visual depiction.	The out-group is depicted in an inhuman way, e.g.: as an evil character and/or animal; Hillary Clinton is depicted as a satanist or a pig; the negative traits of the out-group are exaggerated, such as a caricature of Barack Obama highlighting his allegedly bad policy decisions.	

Table 2. Overview Boundary Types (Deductive) and Themes (Inductive) with Examples

To answer our research questions, we counted the detected instances of boundary types, themes, and roles within our meme sample. To assess the impact of Donald Trump's election as a significant offline event on the boundary work of r/The_Donald, we divided these distributions into the pre- and post-election periods. Table 3 is a detailed overview of these distributions. Note that the instances of boundary types add up to 1,000, as each meme was assigned one of the four codes. Since memes with the code *no such boundary reference* were not coded any further, the number of detected themes add up to 1,000 minus those memes with no further coding. In each meme, we identified at least one of the roles as introduced by De Saint

Laurent et al. (2021). However, since multiple roles can be present in a single meme (see, De Saint Laurent et al., 2021) the total count of roles does not add up to 1,000.

Period	***n* Boundary Type**	***n* Themes**	***n* Roles**
Before election	in-group favoritism: 91 in-group enhancement: 130 out-group degradation: 256 no such boundary reference: 23	caring: 34 strong leader: 63 dominance: 82 victimization: 42 exclusion: 113 incivility: 143	hero: 196 victim: 39 persecutor: 202 fool: 146
After election	in-group favoritism: 88 in-group enhancement: 195 out-group degradation: 198 no such boundary reference: 19	caring: 47 strong leader: 44 dominance: 113 victimization: 77 exclusion: 121 incivility: 79	hero: 202 victim: 83 persecutor: 150 fool: 183

Table 3. Distribution of Boundary Types, Themes, and Roles Before and After Election

Three interesting findings emerge when comparing the distribution and types of codes before and after the election. First, the evaluation of types of boundaries reveals that *out-group degradation* significantly dominates as the most prevalent type of boundary reference before the election, with 256 occurrences, compared to 130 instances of *in-group enhancement* and 91 instances of *in-group favoritism*. The balance shifts, though, after the election, with *in-group enhancement* and *out-group degradation* nearly equal in frequency, at 195 and 198 occurrences, respectively. *In-group favoritism* is nearly the same in both periods, with 91 instances before the election and 88 instances after the election. In addition, the code *no such boundary reference* is distributed almost equally in both periods, with slightly more instances observed before the election (23) compared to after the election (19).

Second, the distribution of themes varies highly between both periods. Before the election, the top two themes detected are *incivility* (143) and *exclusion* (113), highlighting the importance of memes featuring *out-group degradation* boundaries during this period. Examples of *incivility* are, among others, conspiracy narratives such as depicting Hillary Clinton as a satanist and pedophile person who allegedly supports child trafficking; the theme *exclusion* most often revolves around the untrustworthiness of the media and the Democrats, who are accused of working together. The themes *dominance* and *strong leader* are moderately important, with 82 and 63 occurrences, respectively. The former often involves comparisons between Trump and Clinton, typically portraying Trump as the superior presidential candidate. The latter theme depicts Trump as a strong and heroic leader who will save America and make it great again. The themes of *victimization* (42) and *caring* (34) are deemed least significant. *Victimization* predominantly involves the media supposedly portraying Trump supporters in a negative light compared to the Democrats. Meanwhile, *caring* refers to users from r/The_Donald sharing common signs of the group, such as Pepe the Frog or centipedes, to strengthen the feeling of the in-group and mutual support.

Conversely, after the election, we observe a notable increase in the occurrences of the themes *dominance* and *victimization* (113 and 77, respectively), compared to 82 and 42 instances before the election. Instances of *dominance* include Trump's triumph over Clinton, while *victimization* is exemplified by incidents where Trump supporters allegedly faced attacks from Democrats after the election or were supposedly unjustly labeled as racist or misogynistic by the Left. The themes *exclusion* and *incivility* are still important during that time but less salient than before (121 and 79, respectively). Here, *exclusion* mostly refers to the degradation of the Left, "woke" people, and the media, occasionally including anti-immigrant, anti-Muslim, and anti-Black sentiments, albeit less frequently than anticipated based on the findings of Gaudette et al. (2021). The theme *incivility* continues to portray Clinton in an inhuman and derogatory light; yet the exaggerated depiction of "woke" people becomes increasingly prevalent within this theme. *Caring* also gains in prominence, with 47 occurrences after the election compared to 34 before the election. Following the election, *caring* predominantly involves Trump supporters demonstrating solidarity, as they seek to support international allies such as Marie Le Pen during her election campaign. The theme *strong leader* declines, with 44 instances after the election compared to 63 before the election. Despite this development, Trump and his vice presidential running mate Mike Pence continue to be portrayed as heroic figures.

Third, the distribution of roles between both periods features interesting insights in the boundary work of r/The_Donald. In both periods, the role of *hero* seems equally important, with 196 occurrences before and 202 instances after the election. Hence, characters that protect the in-group and lead them to success are essential for r/The_Donald's self-image, with Trump depicted as the main *hero* in most cases. Similar to the results of De Saint Laurent et al. (2021), these heroes typically confront and combat persecutors, thereby emphasizing their heroic nature. Consequently, it is unsurprising that we observe more instances of *persecutors* before the election (202) compared to after the election (150), as the narrative of Trump fighting Clinton was particularly prominent before the election. Predictably, the persecutors during both periods are predominantly Hillary, the Obamas, the Left, and the Democrats. The role of the *fool* maintains significance during both periods, with 146 instances before the election and 183 occurrences after the election. This finding underscores the specificities of memes, known for their humor and the exaggerated depictions of characters "not getting it" as their punchline (De Saint Laurent et al., 2021). Here, the primary foolish characters are Hillary and Bill Clinton and their daughter Chelsea, as well as people associated with "wokeness." Finally, the role of *victim* reveals an interesting contrast between both periods, as the instances of *victims* more than doubled after the election (83) compared to before the election (39). These results show that users from r/The_Donald increasingly emphasize their supposedly unjust treatment by out-groups, depicting themselves as victims of, for instance, the "liberals" or "woke" culture.

Discussion

Boundary Work Featured in Shared Memes

As outlined by social identity theory (Tajfel & Turner, 2004), individuals portray their in-group in a positive light to cultivate a positive social identity, thus fostering *in-group favoritism*. This positive self-image of the in-group is sustained by two themes identified in our data: *caring* and *strong leader*. *Caring* and reciprocal support within OCs are well-known themes in the IS literature on group cohesion within OCs (e.g., Sahharon et al., 2023) and play a pivotal role in fostering member attachment, thereby influencing an OC's longevity and success (Ren et al., 2012). Particularly in online support and self-help communities, we see that caring and support can foster a sense of community and belonging, thereby increasing the attachment toward the group (e.g., Weis et al., 2003). In our data, we observe similar tendencies with users from r/The_Donald portraying the OC as an inclusive community that welcomes everyone, exemplified by efforts to attract Bernie Sanders' supporters after his defeat against Clinton. Moreover, we detect a strong emphasis on the group's commitment to supporting its members, illustrated by references like Pepe fighting for the group's goals and members. Therefore, *caring* emerges as a pivotal theme in fostering a sense of belonging within the community, an essential mechanism for establishing a distinct identity and a sense of "we-ness" (Snow, 2001), thereby strengthening in-group cohesion and creating boundaries toward other OCs.

The theme *strong leader* connects to the literature on leadership in OCs. Despite a substantial body of IS research emphasizing the emergent and shifting nature of leadership in online environments (Johnson et al., 2015), scholars agree that leaders play a pivotal in establishing and maintaining successful and functioning OCs, while also affecting member identification with the OC (e.g., Panteli & Sivunen, 2019). This significance is particularly pronounced in OCs centered around ideological and political beliefs, such as r/The_Donald. The literature indicates that such OCs often idealize a leading figure as the enforcer of their demands and interests (Krämer, 2017). Given r/The_Donald's focus on Donald Trump, as already indicated by its name, it is unsurprising that Trump is frequently depicted as a heroic leader. Therefore, the theme of *strong leader* not only plays a vital role in sustaining an OC but also influences member identification with and attachment to the OC, portraying it as a favorable in-group led by a powerful leader.

To underscore the superiority of the in-group and cultivate a positive social identity, it is essential to compare the in-group with relevant out-groups, as suggested by social identity theory (Tajfel & Turner, 2004), resulting in *in-group enhancement*. In our data, we found that this type of boundary is sustained by the themes *dominance* and *victimization*. The theme of *dominance* portrays the in-group as superior by highlighting its strength, success, and overall superiority compared to different out-groups. By drawing these concrete boundaries between groups, a social hierarchy is established, underscoring the dominance of the in-group while marginalizing out-groups as inferior actors. IS research on OCs shows that this kind of intergroup competition can enhance members' identity-based attachment toward the in-group (e.g., Ren et al., 2012), as it emphasizes the boundaries between members and outsiders. This dynamic is reflected within our data, with direct comparisons between Trump and his supporters and various out-groups, such as the Left and the Clintons, consistently depicting the former as superior to the latter.

Conversely, the theme *victimization* does not underscore the strength or success of the in-group. Instead, the in-group's superiority manifests through allegedly unjust attacks by the out-group, thereby portraying the out-group as persecutors harming the supposedly innocent in-group. *Victimization* often emerges as a prevalent theme in the literature on intergroup conflict and boundary creation, separating groups into "'victims' and 'perpetrators' around binary distinctions of innocence and guilt, good and bad, legitimacy and illegitimacy" (Jankowitz, 2018, p. 260). Portraying the victims as "the good guys" unjustly targeted by "the bad guys" creates a "favorable sense of self, derived from assumptions about the innocence and morality of the victim" (Jankowitz, 2018, p. 263). Hence, boundary creation is inherent to the theme of *victimization*, as it legitimizes the in-group as the innocent and superior actor compared to the misbehaving out-group as the persecutor. This behavior reinforces the distinction between positively perceived in-groups and negatively viewed out-groups, as explained by social identity theory and the concept of boundary work. In our data, *victimization* becomes particularly prominent in the context of Trump supporters claiming to be falsely accused of racist and misogynistic behavior by the Left or alleging that they are discriminated against by the media. It is interesting to note that this victimization is often reversed by portraying the Left and Hillary Clinton as racist and antisemitic, resulting in a *dominance* theme that depicts Trump and his supporters as the (morally) superior group.

Out-group degradation emerges as the third type of boundary and is sustained by the themes *exclusion* and *incivility* as detected within our data. Existing research recognizes social *exclusion* as a pivotal mechanism for maintaining in-group cohesion and establishing the in-group as a legitimate actor while delegitimizing the out-group (e.g., Pickett & Brewer, 2004). Thus, the in-group discredits the out-group by depicting it in a negative light and establishes a clear boundary between "us" and "them." IS research shows that this delegitimization of out-groups is a common tactic for OCs, such as in the context of brand rivalry, with OCs often portraying out-groups in a negative light to strengthen loyalty and attachment to the in-group (e.g., Ewing et al., 2013). In our data, this theme is notably evident in the derogatory portrayal of the Left, the media, the Democrats, and people associated with "woke" culture. Conspiracy narratives are particularly prominent, emphasizing these groups' untrustworthiness, alleging collaboration with and manipulation by the "establishment" and global elite, and thereby delegitimizing these types of groups.

The theme *incivility* extends beyond mere exclusion, portraying the out-group in a highly derogatory manner using vulgar language and exaggerated visual depictions. Research has shown that *incivility* can be used as a means of an OC's identity performance: insulting the out-group can elevate the status of the in-group and strengthen its cohesion (e.g., Rains et al., 2017). Thus, *incivility* not only delegitimizes the out-group but severely discredits it, exaggerating the out-group's negative traits and leading, in the most extreme cases, to a dehumanization of the out-group. This significant differentiation between in- and out-groups underscores the pivotal role of *incivility* in consolidating in-group identity (Rains et al., 2017), thereby drawing boundaries toward the uncivil out-group. In our data, *incivility* manifests in highly conspiratorial narratives, such as allegations that the Clintons are involved in child trafficking and adhere to satanist cults.

Having discussed the various types of boundaries and their corresponding themes within the broader conceptual framework of boundary work, we now shift our focus to the different roles identified within the data as an important aspect of a meme's narrative creation and an OC's boundary work. In our data, the role of the *hero* stands out, particularly as embodied by Donald Trump. This role closely correlates with the theme of *strong leader*, which revolves around the heroic depiction of the leading figures within the OC. To further emphasize the strength and superiority of the in-group, the *hero* is often accompanied by a *persecutor* (De Saint Laurent et al., 2021) who must be defeated or from whom the leader's people need protection. In our specific case, Donald Trump or his supporters most often need to fight against Hillary Clinton, the media, the global elite, and people from the Left. This antagonistic narrative pervades all six themes, thereby contributing to all three types of boundaries: Donald Trump and his supporters are portrayed as taking *care* of the in-group and protecting their people by fighting the out-group; the strength of the OC's *leaders* is also underscored in this narrative; when successfully combating the persecutors, the heroes assert *dominance* over the out-group; by emphasizing the "good" and the "evil," a notion of morality is attached, associated with the theme *victimization*; and the out-group is depicted as an illegitimate actor that needs to be defeated, thereby *excluding* it from the in-group, often in an *uncivil* way. Another main antagonistic narrative centers around the roles of *victim* and *persecutor*. The role of the *victim* does not embody the strength of the *hero*, yet it still garners sympathy because victims are not accountable for their circumstances and are harmed by the *persecutors*. Predictably, this role is associated with the theme of *victimization*, as memes featuring this theme typically depict Trump or his supporters as victims targeted by the out-group persecutors. Finally, the role of the *fool* highlights a unique characteristic of memes, which frequently portray humorous figures in a foolish manner. In the context of boundary work, the role of the *fool* is associated with boundaries based on *out-group degradation* and, thus, to the themes *exclusion* and *incivility*.

Our findings align with those of De Saint Laurent et al. (2021), who highlight the significance of roles in shaping the narratives featured within memes. We identified that the four roles—*hero*, *persecutor*, *victim*, and *fool*—provide a robust framework for analyzing a character's portrayal in memes, extending beyond the specific context of their study on COVID-19. Evaluating these roles not only offers deeper insights into the narratives of memes but also enhances our understanding of the types of boundaries depicted within them, as roles such as the *hero* are integral to themes such as the *strong leader*. Thus, the use of specific roles within memes is a crucial element of an OC's boundary work.

After discussing the various types of boundaries, themes, and roles, it becomes evident that memes featuring signals of boundary work are an integral part of r/The_Donald. In contrast to the findings of Gaudette et al. (2021), we did not observe an abundance of anti-Muslim references. However, strong anti-Left, anti-media, anti-woke, and anti-establishment sentiments were prevalent within the memes. Moreover, we also find evidence that an OC's boundary work featured in memes is a complex interplay of various boundaries, themes, and roles that interconnect and reinforce each other. For example, the *hero* is often complemented by the *persecutor* role, which together support themes such as the *strong leader* as part of *in-group favoritism*, but also the theme *dominance*, which ties into *in-group enhancement*. Lastly, our findings confirm our assumption that the explicit ideological boundaries of r/The_Donald, as the most popular subreddit for alt-right supporters, are also prominently reflected in the memes shared within the community. Thus, this subreddit offers a valuable initial insight into the boundary work of OCs through shared memes, demonstrating that our study depicts a first important step toward better understanding this complex phenomenon and revealing how boundary work is expressed within memes.

Impact of Offline Events on Boundary Work in Memes

When analyzing the shift of boundary types, themes, and roles before and after the election (see Table 3), three interesting insights emerge. First, *out-group degradation* is by far the most prominent type of boundary reference before the election. The predominance of this degrading strategy shifts after the election, as r/The_Donald users share an equal number of memes featuring references to *out-group degradation* and *in-group enhancement*. Therefore, it appears that after winning the election, it becomes more important to compare the in-group with the out-group, particularly by depicting the in-group as a victim of supposedly unjust attacks from the out-group. Notably, *in-group favoritism* is the least frequently referenced boundary type in both periods, highlighting that r/The_Donald's boundary work during the election campaign is mostly focused on depicting out-groups in a negative light or as inferior to the in-group rather than solely portraying the in-group in a positive light. This strategy, known as "negative campaigning," (Lau & Rovner, 2009) is widely used in political campaigns, as it has often proven more effective to undermine the opponent than focusing solely on positive self-promotion.

Second, we see a shift in the distribution of themes before and after the election. Prior to the election, *exclusion* and *incivility* are the most prominent themes. After the election, *dominance* receives a notable increase, indicating the growing importance of directly depicting the in-group's superiority over the out-group. While demonstrating the superiority of Trump and his supporters, r/The_Donald users also share an increased number of memes featuring signals of *victimization* after the election. This *victimization* featured in memes also follows real-world events, such as Mike Pence being verbally attacked while attending a theater performance or the protests in the days after Trump's election.

Third, the distribution of different roles also aligns with the previous findings. Before the election, *hero* and *persecutor* are the most prominent roles, emphasizing the antagonistic narrative between Trump and his supporters and Clinton and the Democrats mentioned before. However, while *hero* and *persecutor* remain prominent roles after the election, the frequency of the *victim* role nearly doubles. This emphasizes the shift in r/The_Donald to an increased number of shared memes with depictions of the *victim* role with respect to the Left and "woke" culture. The role of the *fool* remains significant during both periods.

Considering RGC theory, these findings can be interpreted as follows. Trump's success in the presidential election—a significant event for the alt-right movement and r/The_Donald—was a major victory over competing out-groups such as Hillary Clinton and the Democrats. Based on RGC theory, it seems plausible to anticipate memes featuring more boundary references related to a hostile demarcation *before* the election compared to *after* the election, as the competition over scarce resources had been won. Our results indeed reveal this shift: Before the election, memes predominantly featured references to *out-group degradation*, with themes of *exclusion* and *incivility* most prevalent; the antagonistic roles of *hero* and *persecutor* were also more pronounced during this period. However, while some of these boundary references remained important after the election was won, there was a notable shift toward *in-group enhancement,* with the themes *dominance* and *victimization* becoming particularly prominent; also, the role of *victim* gained significant prominence. Thus, we observe a shift from merely degrading the out-group, often in an uncivil and hostile manner, to a more nuanced comparison between in-groups and out-groups by depicting the in-group as either dominating the out-group or as its victim. This pattern aligns with RGC theory, as boundaries related to a hostile demarcation predominated before the election, while post-election, these were largely replaced by boundaries emphasizing either the in-group's superiority or its victimization.

Overall, these findings emphasize that offline events, such as the presidential election, can influence users' online meme-sharing, thereby supporting previous research that suggests a reciprocal relationship between offline events and online behavior (e.g., Bliuc et al., 2019; Chung et al., 2021) while also underscoring the value of theories like RGC theory in enhancing our understanding of this dynamic.

Conclusion

To understand an OC's boundary work, we built upon social identity theory and derived *in-group favoritism, in-group enhancement,* and *out-group degradation* as the main types of boundaries that aid in establishing distinct OCs vis-à-vis other groups online. Through qualitative coding, of the memes shared in r/The_Donald, we inductively detected six emerging themes that clarify the concrete function of boundaries essential for an OC's boundary work. Moreover, we analyzed each meme with respect to the roles of *hero*, *victim*, *persecutor*, and *fool*, thereby gaining insights into how these characters contribute to creating boundaries for OCs. The results of this analysis answer our *RQ1* and show that signals of boundary work are prominently featured in memes shared within r/The_Donald. In these memes, different types of boundaries, themes, and roles interlink and reinforce each other, revealing the complex interplay of dimensions that shape the OC's boundary work. By dividing our meme sample into a pre- and post-election period, we were able to evaluate whether the depiction of boundaries featured in memes changed after this significant event. As anticipated based on RGC theory, our results reveal shifts in the distribution of boundary types, themes, and roles after the election; for instance, we found a decrease in the hostile demarcation of out-groups such as Hillary Clinton and the Democrats, coupled with an increased portrayal of Trump supporters as victims of these out-groups. Furthermore, we consistently observed references to real-life events within memes, such as the incident involving Mike Pence being confronted at a theater performance. Therefore, in response to our *RQ2*, we conclude that offline events such as elections are reflected in the meme-sharing of r/The_Donald, influencing the featured boundaries and illustrating the reciprocal relationship between the online and offline worlds.

These findings emphasize the importance of memes' multimodal nature for OCs' boundary work. Memes are more than simply images or texts or an aggregation of both. Rather, they are (digital) artifacts that, through the recombination of culturally charged elements, can convey profound cultural messages with complex meanings condensed within a single image macro (e.g., Nissenbaum & Shifman, 2018; Shifman, 2013b). This characteristic becomes particularly evident in instances where neither the image nor the text alone could effectively convey the intended message. The iconic meme template of "Condescending Wonka," featuring the text "Tell me again what the media says about Trump because they have been so right so far," is one example from our data. This meme is known for its sarcastic undertones that implies the opposite of the actual words (e.g., Aguilar et al., 2017). This added layer of meaning enables the meme to express its true message, which would be lost if only the image or text were used independently: it satirizes the media's dissemination of misinformation about Trump. Therefore, assessing an OC's boundary work based on shared memes is an important next step for gaining profound insights into the types of OC boundaries, which may not be fully captured through text or images alone.

Our study offers four significant theoretical and practical contributions to the field of meme research and OCs' boundary work. First, by using social identity theory, we derived three different types of boundaries suitable for analyzing the boundary work of OCs. By complementing these boundary types with concrete themes and roles, we developed an innovative framework for analyzing the boundary creation of OCs, contributing to the establishment of their unique identity and differentiation from other online groups. Given the framework's adaptability and allowing for potential extensions, such as additional types of boundaries, themes, and roles, we encourage future research to test and expand upon it by analyzing the boundary creation of diverse OCs across various platforms. We believe that our framework serves as a robust foundation for investigating how OCs establish boundaries in relation to other online groups, providing a valuable starting point for further exploration in this field of research. Second, by using memes as the unit of analysis, we expand current research on the boundary creation of OCs. Through our analysis, we underscore memes' unique ability to convey complex messages in a highly condensed format, going beyond the sole analysis of text and images and thus showing memes' merit in boundary creation. We consider our work an important next step in the analysis of multimodal data, particularly in the context of OCs' boundary creation. Third, our results expand research on the reciprocal relationship between offline events and online behavior. By applying RGC theory, we offer valuable insights into identifying significant offline events that impact user behavior in OCs and provide guidance on understanding how intergroup conflict influences the boundary work of these communities. Fourth, we deem our findings to have value beyond their theoretical merit because they can aid practitioners in several ways. Our findings have shown that memes can be used to demarcate OCs by distinguishing between in-groups and out-groups. These demarcation processes can contribute to online polarization, as out-groups are often portrayed in a derogatory manner while in-groups are depicted positively. Gaining deeper insights into these processes is crucial, as polarization can have negative consequences, such as incivility. Platform providers and designers can use our findings to reduce and mitigate these negative outcomes and create more inclusive and engaging community spaces. In addition, our insights are valuable for managing OCs, as they enhance the understanding of OC identity formation and reinforcement, which significantly impacts member attachment, community building (e.g., Ren et al., 2012), information flow, and collaboration (e.g., Faraj et al., 2011).

While our analysis provides important insights into the meme-sharing and boundary work of OCs, they need to be seen in light of their limitations. First, given that our analysis spans part of the presidential election period, we anticipate a peak in boundary work during this time. As our sample comprises memes with the highest Reddit score from this period, we likely assessed memes with more explicit boundaries compared to the entire corpus of memes shared within this OC. We, therefore, invite future research to extend the period of analysis, including additional offline events and those that do not center on the competition of resources, and to examine all memes shared within the OC, thereby receiving a more nuanced understanding of this OC's boundary work. Second, for a more comprehensive analysis of an OC's boundary work, we encourage future research to expand our framework by including more layers of analysis, such as the memes' metadata, the submission title, the context in which memes were shared, and information about the users who posted them. Doing so would allow the development of a unique "meme DNA," similar to a "fingerprint," that distinguishes one OC from another. Shifman's (2013a) framework for analyzing memes based on their content, form, and stance offers a promising first approach in this direction. Third, as this study represents the first attempt to evaluate boundary work in shared memes, we focused on an OC known for its explicit ideological connotations and identity as a representative example of communities known to feature similar explicit worldviews. It would be very valuable to investigate whether similar boundary dynamics are present in more moderate OCs. For instance, boundary themes such as *strong leader* may be less pronounced in communities with less political focus on specific figures. We invite future research to broaden our analysis by examining the boundary work of diverse OCs, also across different platforms, to receive a more nuanced understanding of online boundary dynamics.

References

Aguilar, G. K., Campbell, H. A., Stanley, M., & Taylor, E. (2017). Communicating mixed messages about religion through internet memes. *Information, Communication & Society*, *20*(10), 1498-1520. https://doi.org/10.1080/1369118X.2016.1229004

Askanius, T., & Keller, N. (2021). Murder fantasies in memes: Fascist aesthetics of death threats and the banalization of white supremacist violence. *Information, Communication & Society*, *24*(16), 2522-2539. https://doi.org/10.1080/1369118X.2021.1974517

Bliuc, A.-M., Betts, J., Vergani, M., Iqbal, M., & Dunn, K. (2019). Collective identity changes in far-right online communities: The role of offline intergroup conflict. *New Media & Society*, *21*(8), 1770-1786. https://doi.org/10.1177/1461444819831779

Chung, T. L. D., Johnson, O., Hall-Phillips, A., & Kim, K. (2021). The effects of offline events on online connective actions: An examination of #BoycottNFL using social network analysis. *Computers in Human Behavior*, 115, 106623. https://doi.org/10.1016/j.chb.2020.106623

Dawkins, R. (2016). *The selfish gene*. Oxford University Press.

De Saint Laurent, C., Glăveanu, V. P., & Literat, I. (2021). Internet memes as partial stories: Identifying political narratives in coronavirus memes. *Social Media + Society*, *7*(1), 205630512198893. https://doi.org/10.1177/2056305121988932

DeCook, J. R. (2018). Memes and symbolic violence: #proudboys and the use of memes for propaganda and the construction of collective identity. *Learning, Media and Technology*, *43*(4), 485-504. https://doi.org/10.1080/17439884.2018.1544149

Ewing, M. T., Wagstaff, P. E., & Powell, I. H. (2013). Brand rivalry and community conflict. *Journal of Business Research*, *66*(1), 4-12. https://doi.org/10.1016/j.jbusres.2011.07.017

Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization Science*, *22*(5), 1224-1239. https://doi.org/10.2307/41303115

Fayard, A.-L., & DeSanctis, G. (2010). Enacting language games: The development of a sense of 'we-ness' in online forums. *Information Systems Journal*, *20*(4), 383-416. https://doi.org/10.1111/j.1365-2575.2009.00335.x

Flesher Fominaya, C. (2019). Collective identity in social movements: Assessing the limits of a theoretical framework. In D. A. Snow, S.A. Soule, H. Kriesi, & H.J. McCammon (Eds.), *The Wiley Blackwell Companion to Social Movements* (pp.429-445). Wiley-Blackwell. https://doi.org/10.1002/9781119168577.ch24

Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2021). Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society*, *23*(12), 3491-3508. https://doi.org/10.1177/1461444820958123

Glitsos, L., & Hall, J. (2019). The Pepe the Frog meme: An examination of social, political, and cultural implications through the tradition of the Darwinian Absurd. *Journal for Cultural Research*, *23*(4), 381-395. https://doi.org/10.1080/14797585.2019.1713443

Heikkilä, N. (2017). Online antagonism of the alt-right in the 2016 election. *European Journal of American Studies*, *12*(12-2). https://doi.org/10.4000/ejas.12140

Henn, T. (2024). Follow the Memeing: Analyzing the Cultural Diffusion between Mainstream and Alt-Right Communities based on Shared Memes. *Proceedings of the 32nd European Conference on Information Systems*, 11. https://aisel.aisnet.org/ecis2024/track24_socialmedia/track24_socialmedia/11/

Henn, T., & Posegga, O. (2023). What Do they Meme? Exploring the Role of Memes as Cultural Symbols of Online Communities. *Proceedings of the 44th International Conference on Information Systems*, 8. https://aisel.aisnet.org/icis2023/socmedia_digcollab/socmedia_digcollab/8/

Jankowitz, S. (2018). Intergroup struggles over victimhood in violent conflict: The victim-perpetrator paradigm. *International Review of Victimology*, *24*(3), 259-271. https://doi.org/10.1177/0269758017745617

Johnson, S. L., Safadi, H., & Faraj, S. (2015). The emergence of online community leadership. *Information Systems Research*, 26(1), 165-187. https://doi.org/10.1287/isre.2014.0562

Krämer, B. (2017). Populist online practices: The function of the Internet in right-wing populism. *Information, Communication & Society*, *20*(9), 1293-1309. https://doi.org/10.1080/1369118X.2017.1328520

Lamerichs, N., Nguyen, D., Melguizo, M. C. P., Radojevic, R., & Lange-Böhmer, A. (2018). Elite male bodies: The circulation of alt-Right memes and the framing of politicians on Social Media. *Participations*, *15*(1), 180-206.

Lau, R. R., & Rovner, I. B. (2009). Negative campaigning. *Annual Review of Political Science*, *12*(1), 285-306. https://doi.org/10.1146/annurev.polisci.10.071905.101448

Lugea, J. (2019). The pragma-stylistics of 'image macro' internet memes. In H. Ringrow & S. Pihlaja (Eds.), *Contemporary Media Stylistics* (pp. 81-106). Bloomsbury Publishing. https://doi.org/10.5040/9781350064119.0009

Milner, R. M. (2013). Pop polyvocality: Internet memes, public participation, and the Occupy Wall Street movement. *International Journal of Communication*, *7*, 2357-2390.

Miltner, K. M. (2014). 'There's no place for lulz on LOLCats': The role of genre, gender, and group identity in the interpretation and enjoyment of an Internet meme. *First Monday*. https://doi.org/10.5210/fm.v19i8.5391
Nissenbaum, A., & Shifman, L. (2017). Internet memes as contested cultural capital: The case of 4chan's /b/board. *New Media & Society*, *19*(4), 483-501. https://doi.org/10.1177/1461444815609313
Nissenbaum, A., & Shifman, L. (2018). Meme templates as expressive repertoires in a globalizing world: A cross-linguistic study. *Journal of Computer-Mediated Communication*, *23*(5), 294-310. https://doi.org/10.1093/jcmc/zmy016
Oc, B., Moore, C., & Bashshur, M. R. (2018). When the tables are turned: The effects of the 2016 US Presidential election on in-group favoritism and out-group hostility. *Plos one*, *13*(5), e0197848. https://doi.org/10.1371/journal.pone.0197848
Panteli, N., & Sivunen, A. (2019). 'I Am Your Fan; Bookmarked!' Members' Identification Development in Founder-Led Online Communities. *Journal of the Association for Information Systems*, *20*(6), 824-841. https://doi.org/10.17705/1jais.00552
Pickett, C. L., & Brewer, M. B. (2004). The role of exclusion in maintaining ingroup inclusion. In D. Abrams, M. A. Hogg, & J. M. Marques (Eds.), in *Social Psychology of Inclusion and Exclusion* (pp. 107-130). Psychology Press.
Rains, S. A., Kenski, K., Coe, K., & Harwood, J. (2017). Incivility and political identity on the Internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*, *22*(4), 163-178. https://doi.org/10.1111/jcc4.12191
Reddit. (2023). *Reddiquette*. Reddit. https://reddit.zendesk.com/hc/en-us/articles/205926439-Reddiquette.
Ren, Y., Harper, F. M., Drenner, S., Terveen, L., Kiesler, S., Riedl, J., & Kraut, R. E. (2012). Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *MIS Quarterly*, *36*(3), 841-864. https://doi.org/10.2307/41703483
Sahharon, H., Zulkefli, M. Y., & Ibnu, I. N. (2023). Enhancing the sense of social cohesion through virtual community: a systematic review. *Journal of Information and Knowledge Management*, *2*, 378-401.
Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, *21*(7), 1417-1443. https://doi.org/10.1177/1461444818821316
Sherif, M. (1966). *Group conflict and co-operation: Their social psychology*. Routledge & Kegan Paul.
Shifman, L. (2013a). Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication*, *18*(3), 362-377. https://doi.org/10.1111/jcc4.12013
Shifman, L. (2013b). *Memes in digital culture*. MIT press.
Snow, D. A. (2001). *Collective identity and expressive forms* (Center for the Study of Democracy Working Paper). University of California, Irvine. https://escholarship.org/content/qt2zn1t7bj/qt2zn1t7bj.pdf
Syed, R., & Silva, L. (2023). Social movement sustainability on social media: An analysis of the women's March Movement on Twitter. *Journal of the Association for Information Systems*, *24*(1), 249-293. https://doi.org/10.17705/1jais.00776
Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Eds.), *Political Psychology* (pp. 276-293). Psychology Press. https://doi.org/10.4324/9780203505984-16
Thornberg, R., & Charmaz, K. (2014). Grounded theory and theoretical coding. In U. Flick (Ed.), *The SAGE Handbook of Qualitative Data Analysis (Vols. 1-5)* (pp. 153-169). SAGE Publications. https://doi.org/10.4135/9781446282243.n11
Trujillo, A., & Cresci, S. (2022). Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the Conference on Human-Computer Interaction*, 6, 1-28. https://doi.org/10.1145/3555639
Weis, R., Stamm, K., Smith, C., Nilan, M., Clark, F., Weis, J., & Kennedy, K. (2003). Communities of care and caring: The case of MSWatch.Com®. *Journal of Health Psychology*, *8*(1), 135-148. https://doi.org/10.1177/1359105303008001449
Zeng, J., & Abidin, C. (2021). '#OkBoomer, time to meet the Zoomers': Studying the memefication of intergenerational politics on TikTok. *Information, Communication & Society*, *24*(16), 2459-2481. https://doi.org/10.1080/1369118X.2021.1961007

Appendix A.5: Paper V

Bibliographic data	Henn-Latus, T., Posegga, O. Submitted. Classifying Protest Frames in Online News Reporting Using Multimodal Models with Zero-Shot and Few-Shot Learning. Journal of Computational Social Science.
Abstract	News framing of protests is important because it can shape public discourse and influence opinions about protest movements. Research has shown that media outlets often portray protests in a negative light, a pattern referred to as the protest paradigm. This study assesses the performance of multimodal models in classifying frames associated with the protest paradigm and in identifying whether protests are portrayed as violent and/or peaceful in news reporting on Instagram. Specifically, we evaluate different CLIP architectures and GPT4.1 by comparing their performance when using image-only inputs and using combined image and caption inputs. We also distinguish between zero-shot and few-shot learning approaches. The models' predictions are benchmarked against a human-labeled ground truth. Our results show that multimodal models perform considerably well in detecting protest frames, with performance improving when few-shot learning is applied. However, performance varies depending on the model architecture, prompt configuration, and the specific frames being classified.

Classifying Protest Frames in Online News Reporting Using Multimodal Models with Zero-Shot and Few-Shot Learning

Theresa Henn-Latus[1,2] and Oliver Posegga[1,2]

[1]Chair of Information Systems and Social Networks, University of Bamberg, An der Weberei 5, Bamberg, 96047, Germany.
[2]Google Scholar: Theresa Henn-Latus; Oliver Posegga.

Contributing authors: theresa.henn@uni-bamberg.de; oliver.posegga@uni-bamberg.de;

Abstract

News framing of protests is important because it can shape public discourse and influence opinions about protest movements. Research has shown that media outlets often portray protests in a negative light, a pattern referred to as the protest paradigm. This study assesses the performance of multimodal models in classifying frames associated with the protest paradigm and in identifying whether protests are portrayed as violent and/or peaceful in news reporting on Instagram. Specifically, we evaluate different CLIP architectures and GPT4.1 by comparing their performance when using image-only inputs and using combined image and caption inputs. We also distinguish between zero-shot and few-shot learning approaches. The models' predictions are benchmarked against a human-labeled ground truth. Our results show that multimodal models perform considerably well in detecting protest frames, with performance improving when few-shot learning is applied. However, performance varies depending on the model architecture, prompt configuration, and the specific frames being classified.

Keywords: multimodal models; protest paradigm; zero-shot learning; few-shot learning; vision-language models; large language models

1 Introduction

Media framing plays a crucial role in shaping public perception, attitudes, and discourse on protest movements, influencing whether they are seen as legitimate or delegitimized actors [1–4]. We know from research that mainstream media coverage often leans toward delegitimizing protests and protesters by emphasizing their disruptive and deviant behavior [3, 5, 6]. This pattern is known as the "protest paradigm" [7, 8].

While most research in this area has focused on traditional media coverage, recent years have seen a significant increase in studies that examine how protests are framed on social media [9–11]. This is a welcome development, as platforms such as Facebook, X, and Instagram facilitate the rapid spread of news [12], which amplifies the visibility of protest frames in media coverage, and serve as increasingly important news sources in the United States [13]. Moreover, the growing ubiquity of multimodal data on social media, combining text, images, and other forms of content, presents new opportunities for analyzing media framing of protest and the online manifestation of the protest paradigm.

Thus, detecting and classifying news frames on social media has become a rapidly evolving area of research. While early studies relied primarily on manual approaches to classify frames by using image and text data and a combination thereof [9, 14, 15], the field has increasingly embraced automated methods for news frame detection and classification. Recent advancements have focused largely on using text as input data for detecting and classifying news frames, primarily leveraging transformer architectures (e.g., [16]), such as implementations of Bidirectional Encoder Representations from Transformers (BERT) models [17]. Despite the inherent complexity of frame detection and classification—often challenging even for human coders [16]—research shows that BERT-based models perform considerably well in identifying and categorizing frames [16, 18].

In addition to using text data as input, few studies have explored the automated detection and classification of news frames based on images (e.g., [19–21]) and a combination of text and images [22]. Since the visual framing of protest can significantly influence public perception of a movement and protest [23], analyzing image framing in news coverage is especially crucial, particularly in visual-centric environments like social media (for the significance of multimodal and image data for the protest paradigm, see [9]). While these approaches represent an important first step toward a more profound analysis of frames and the integration of image and text data, they come with two key limitations.

First, some approaches (e.g., [21]) use object detection for frame classification by identifying elements such as people or police equipment, which are later used to determine the featured frame. While this modular approach is valid for analyzing individual components of an image, it fails to account for the interplay between detected objects. This oversight is crucial, as the overall composition of an image significantly shapes the frame it conveys. Second, for classifying frames based on image and/or text data, these approaches rely on large training datasets to fine-tune models for their task-specific application context. Annotating data, however, is a labor-intensive process, often limiting the availability of sufficient training data or requiring substantial human resources. Also, since these models are typically fine-tuned for specific contexts, their generalizability to different settings may be limited.

The advancements in pre-trained and prompt-based Large Language Models (LLMs) in recent years, such as various versions of OpenAI's Generative Pre-Trained Transformer (GPT), have partly mitigated the second challenge. Since these models are pre-trained on vast datasets for broad applicability across multiple tasks and domains, they enable frame classification through zero-shot learning [24, 25], reducing the need for task-specific labeled training data while still yielding solid results. Their performance can be enhanced by embedding a small set of labeled examples directly in their prompt (e.g., [26])—a technique known as few-shot learning—or by tuning prompts and a small set of parameters. Both learning approaches typically demand far less manual effort than earlier fully supervised methods.

While GPT models have shown promising results in news frame classification based on text input [24, 25], we are not aware of any approach that has leveraged advances in prompt-based multimodal models for analyzing images or a combination of images and text in the context of news frames. This paper addresses this research gap by evaluating the performance of two state-of-the-art multimodal model architectures in classifying protest-related news frames against a human-labeled benchmark. The models include different architectures of OpenAI's Contrastive Language-Image Pre-Training (CLIP) model and OpenAI's GPT4.1 model. Using Instagram data, we assess these models' performance in classifying news protest frames through zero-shot and few-shot learning approaches, utilizing both image and text inputs. Based on their training, these models can "interpret" the overall composition of images rather than merely detecting individual objects, thereby addressing the first challenge outlined above.

Our contributions are threefold. First, by leveraging the potential of image and text data, we provide a more comprehensive perspective on media framing, which is particularly relevant in multimedia environments such as social media. Second, advancing the understanding of framing in social media environments is crucial for two reasons: on the one hand, news content shared on platforms such as Instagram presents a unique case for frame analysis, as framing occurs in a highly condensed format (e.g., very short captions) compared to traditional media such as newspapers; on the other hand, Instagram is inherently image-centric, further emphasizing the importance of the frame analysis of visual content. Third, we evaluate whether multimodal models like CLIP, originally trained on comparably simple image and text pairs, can also handle more complex tasks, such as identifying frames in news reporting using combined text and image input. Moreover, by assessing the classification performance of zero-shot and few-shot learning approaches, we explore whether state-of-the-art multimodal models can help mitigate the labor-intensive process of labeling large amounts of training data for complex classification tasks.

Hence, we pose the following two research questions:

RQ1 How accurately do multimodal models classify protest paradigm frames in news reporting in Instagram posts, particularly when differentiating between zero-shot and few-shot approaches?

RQ2 How do model performances vary when using only the image of a post compared to combining the image with its caption?

This paper is structured as follows. In the next section, we provide a concise overview of media framing and the protest paradigm, highlighting typical associated frames while expanding the classical framework by emphasizing the central role of violence. We also review current research on protest frame classification based on text and images featured in news media. The methodology section details our data collection on Instagram and the coding of our ground truth. We then present the model architectures used for frame classification and compare the classification results against a human-labeled benchmark. We differentiate between zero-shot and few-shot learning approaches as well as classification outcomes based solely on images versus those combining images and text. We conclude with a discussion that situates our findings within the broader context of existing research, acknowledge the study's limitations, and outline directions for future research.

2 Theoretical Background

2.1 Framing and the Protest Paradigm

Framing is a well-established concept in research, and can be defined as a way to "*select some aspects of a perceived reality and make them more salient*" ([27, p. 52], emphasis in original). Media frames expand on this general definition and are understood as "*persistent patterns of cognition, interpretation, and presentation, of selection, emphasis, and exclusion, by which symbol-handlers routinely organize discourse, whether verbal or visual*" ([28, p. 7], emphasis in original).

In the context of protest coverage, research has shown that news media often "organize discourse" in a particular way, typically following a routinized pattern that delegitimizes and discredits protests (e.g., [2]). This pattern, referred to as the "protest paradigm" [7, 8], emphasizes the deviant behavior of protesters, encounters with the police, and protest violence rather than focusing on protesters' motives [8, 9]. Since the framing of protests can influence whether the public supports or opposes a protest's cause (e.g., [5]) and "affect a movement's ability to mobilize supporters and gain legitimacy in the eye of the public" [2, p. 334], evaluating the media coverage of protests is crucial.

The protest paradigm is typically associated in the literature with four distinct frames employed by the media: the *riot frame*, which portrays the protest as violent, causing physical harm to people and property; the *confrontation frame*, which focuses on clashes between protesters and the police; the *spectacle frame*, which highlights sensational aspects of a protest, such as celebrity involvement, emphasized emotions, or a protest's size; and the *debate frame*, which emphasizes a protest's demands, larger context, and grievances (e.g., [2, 29]). While the first three frames are used to delegitimize protests, the latter is typically employed to legitimize them.

Although these frames are well-established in the context of the protest paradigm, recent research has called for a reassessment, particularly regarding the role of violence [2, 5]. In the protest paradigm's traditional conceptualization, there is typically "no real distinction [made] between violence by protesters and violence against protesters" [2, p. 336]. As a result, protesters are most often blamed for any violence featured, thereby delegitimizing the protest. This framing, which portrays protesters as *perpetrators*, overlooks the possibility that violence may have been committed by the police and that the protesters are *victims* [2]. Thus, an increasing body of research now considers the evaluation of violence committed by the police as a crucial factor in this context [5, 29–34], as studies have shown that portraying police violence can potentially serve as a legitimizing device for protest (e.g., [5, 34]).

We follow this line of research and the call for a more nuanced evaluation of violence featured in news reporting. To that end, we introduce a fifth frame to complement the traditional four protest paradigm frames, which we refer to as the *victimization frame*. Conceptually, this frame serves as a counterpart to the riot and confrontation frames, focusing on physical violence committed by the police and/or other state authorities against protesters. It also includes protests with the central theme of police violence, such as those following the deaths of George Floyd and Mahsa Amini. Importantly, we do not assess whether the featured violence is justified; rather, we evaluate whether visual and textual cues are present that emphasize physical police violence against protesters.

Since the peaceful nature of protests is often underemphasized in the traditional conceptualization of the protest paradigm despite its potential as a legitimizing device (e.g., [9, 10, 35]), we also focus on whether a protest is explicitly depicted as particularly *peaceful* or *violent*. In addition to the five frames, we, therefore, incorporate this dimension into our analysis, aligning with established research in the field [9, 10].

2.2 Detection of Media Frames and Violence

Previous research on the protest paradigm has typically identified the four traditional frames through manual coding based on text input (e.g., [10, 11, 33, 36]). Less frequently, multimedia elements, such as images in combination with text, have also been used to identify these frames, although these approaches still depended on manual coding [9, 11, 37].

Several studies, though, have explored automated frame detection using both textual and visual inputs. For instance, when using text data, BERT-based models have been successfully applied to identify frames in news coverage [16, 24]. Mavridis et al. [38] analyzed YouTube videos from news outlets, using video transcriptions and metadata as input for frame classification. They applied models such as support vector machines and neural networks, comparing their classification performance against expert and crowd-annotated frame labels. Their findings indicated that frame detection is challenging for both humans and machines.

Other research used image data as input for frame detection. In their analysis of news coverage of the Black Lives Matter (BLM) movement, Jiang, Jin, and Deng [19] examined images from TikTok video snippets to identify featured frames. Using various Python libraries, they applied automated object detection for frame classification. Kim and Bas [21], too, used object detection tools, such as Google Cloud Vision, to analyze media portrayals of the BLM movement on Facebook. They used the identified tags to cluster images with a structural topic modeling algorithm, which revealed recurring topics that news outlets used to emphasize specific aspects of the movement.

Lastly, Tourni et al. [22] analyzed the framing of U.S. gun violence news coverage using both image and text data. They applied convolutional neural network (CNN) architectures combined with various BERT models and Google Web Entity tags for a multimodal analysis. Their findings indicate that integrating image and text features outperforms news frame prediction based on a single type of information.

Thus, most existing approaches for automated frame detection, particularly those using image data, rely either on object detection or CNN-based classification, which require task-specific training data. These models typically do not account for the overall image composition during classification. Moreover, none have been applied to the automatic detection of the more complex and nuanced frames associated with the protest paradigm. In this paper, we aim to address these shortcomings by leveraging multimodal models with zero-shot and few-shot learning approaches, classifying the five protest paradigm frames outlined before.

For the detection of violent and peaceful protests, most studies have relied on manual coding [11, 39]. Recent research, however, has increasingly focused on automating the detection of violence in protest media coverage, particularly when analyzing visual content [20, 40, 41]. With this study, we aim to expand this line of research by examining how effectively zero-shot and few-shot models perform in this context.

3 Methods

3.1 Data Collection and Reduction

Since the goal of this study is to evaluate the news framing of media outlets on social media, it was necessary to identify prominent news providers that share their content online. In this study, we focus on news featured in major U.S. outlets. Using the Reuters Institute's Digital News Report from 2021, 2022, and 2023 [42–44], we identified the most visited online news outlets. Based on this research, we identified the following 16 news outlets: Yahoo News, CNN, Fox News, The New York Times, MSN, USA Today, The Washington Post, MSNBC, BBC, Buzzfeed, ABC, CBS, The Wall Street Journal, The Huffington Post, NPR, and Newsmax.

Instagram, the platform of choice for this study, was selected for two key reasons. First, Instagram is known for its emphasis on images, making it well-suited for our focus on the automated visual detection of frames. Second, while previous studies have concentrated on platforms such as Facebook in this context (e.g., [11]), Instagram is a popular and important platform that has often been overlooked in frame classification research (for an exception, see, [37]).

We used the platform CrowdTangle to collect Instagram data from the 16 news outlets, downloading all posts from these outlets' profiles, starting from their profile inception up until the download date (the download dates ranged from February 8, 2024, to February 13, 2024). We obtained 221,614 posts. After filtering the data to include only posts featuring images, this number was reduced to 168,600 posts. To identify protest-related news posts, we further narrowed down the dataset through a filtering process using the string "protest OR protesters OR protesting OR demonstration OR demonstrators" (for a list of similar keywords, see, [31, 32]). This yielded 4,942 posts, from which we downloaded the first image in each case.

To establish a reliable ground truth for evaluating model performance, two researchers manually coded these posts. A preliminary coding round revealed, however, that the sample still contained posts unrelated to protests. For instance, the keyword "demonstration" also captured posts about aircraft flight demonstrations. In addition, following established research in the field of the protest paradigm [9, 29, 31], we decided to include only posts that referenced a protest event executed by a protest movement or an individual protester affiliated with a movement. After several rounds of open coding between two authors, one author filtered the dataset accordingly, reducing it to 3,444 posts (for more information, see Appendix). These posts serve as the basis for our frame analysis.

3.2 Labeling of Ground Truth

To operationalize the five protest paradigm frames, we drew on established definitions for the first four frames while developing a definition for the victimization frame ourselves (for more information, see Appendix). To establish a shared understanding of the five protest frames, as well as the distinction between peaceful and violent protests, both authors engaged in multiple rounds of open coding. We independently coded different subsamples of the dataset and discussed the results to align our interpretations. Once a common understanding of the labels was reached, one author proceeded to code the presence (1) or absence (0) of each frame within the caption, the image, and their combination for every post. The combination reflects the union of all frames identified in both the image and the caption. Note that frames are not mutually exclusive, meaning a post can feature multiple frames. In addition, one author assessed whether a post emphasizes a protest as violent (1) or not (0), and whether it emphasizes a protest as peaceful (1) or not (0). If no frame was present, the author assigned the value "*None.*"

3.3 Model Architecture

We employ different CLIP model architectures [45] for our analysis, as CLIP is widely regarded as one of the most advanced multimodal models and achieves strong performance across a range of vision-language tasks, outperforming earlier computer vision models based on, for example, CNN architectures [45]. CLIP follows a dual-encoder design consisting of a separate text encoder and image encoder, trained on extensive image-text pairs. This pretraining enables CLIP models to perform vision-language tasks beyond their original training context, resulting in strong performance in zero-shot scenarios (e.g., [46]). Through contrastive learning, CLIP aligns these visual and textual representations within a shared embedding space, using cosine similarity as the alignment measure. This enables CLIP not only to recognize individual objects within an image but also to capture the overall composition and semantic relationships between them. While CLIP was originally designed for image and text alignment, the shared embedding space also allows for meaningful comparisons between textual inputs, making it adaptable for tasks involving both image-to-text and text-to-text similarity.

Although CLIP models have traditionally been trained on comparably simpler image and text pairs, such as animals and their descriptions [45], this paper explores their potential in more complex tasks such as frame classification, where meanings are often latent rather than explicit. Early evidence suggests that CLIP can support more advanced vision-language tasks, such as evaluating the alignment between news thumbnails and the content of their corresponding articles [46].

For our analysis, we employed OpenCLIP, the open-source implementation of OpenAI's CLIP model [47], as the repository provides access to a range of pretrained models along with their benchmarked performances across various tasks. We opted to use vision transformer (ViT) architectures as our backbone, given their demonstrated superiority over traditional CNN-based models [45]. According to the OpenCLIP repository, the best-performing architecture for classification tasks is the vision transformer "ViT-H-14-378-quickgelu" (H14), pretrained on the "dfn5b" dataset. This model achieves an average accuracy of 0.71 across 35 benchmark datasets in image-text classification tasks. To ensure architectural comparability while allowing for variation in model size, we selected the top-performing ViT-based backbones besides the H-series models. Specifically, we employed "ViT-L-14-quickgelu" (L14; Acc: 0.67) pretrained on the "dfn2b" dataset and "ViT-bigG-14" (G14; Acc: 0.67) pretrained on the "laion2b_s39b_b160k" dataset as our base models. Model sizes are abbreviated, where "L" stands for "large," "H" for "huge," and "G" for "giant," reflecting progressively larger model architectures. The number "14" refers to the patch size used in each model.

For the zero-shot models, we applied a structured prompting approach, gradually increasing the level of detail in the prompts (for the full prompts, see Appendix). We began with basic definitions of each class, then extended these by incorporating additional explanations, and finally added concrete examples. This process can be summarized by the following three configurations:

$$
\begin{aligned}
\mathit{Configuration}_1 &= \mathit{basic\ definitions} \\
\mathit{Configuration}_2 &= \mathit{basic\ definitions} \ + \ \mathit{explanations} \\
\mathit{Configuration}_3 &= \mathit{basic\ definitions} \ + \ \mathit{explanations} \\
&\quad + \ \mathit{examples}
\end{aligned}
$$

During prompting, we paid particular attention to ensuring that the semantic embeddings of each label were as distinct as possible to maximize the models' ability to differentiate between classes. Since extending the prompt length reduced the distinctiveness of class embeddings, we ultimately decided against using models such as Long-CLIP [48], which are designed to handle longer text inputs beyond the 77-token limit of standard CLIP models. To ensure clearer separation between classes, we, therefore, compressed our codebook definitions, explanations, and examples into prompts capped at 77 tokens. This approach improved embedding distinctiveness and enhanced performance and comparability across classes. While this approach also required truncating captions to the 77-token limit, we still prioritized maximum class distinguishability. Given that Instagram captions are typically brief (average of 131.5 tokens in our data) compared to other forms of media, such as newspaper reporting, and that key information is often conveyed early in the text, we consider the use of CLIP models with a 77-token limit to be a reasonable choice.

In addition to CLIP, we selected a GPT-series model for benchmarking our results. While these models offer fewer possibilities for adjustment compared to CLIP and are considered black-box systems, they have demonstrated strong performance in classification tasks (e.g., [24]). We chose GPT4.1, which is described by OpenAI as a "highly intelligent model with largest content window" handling bot text and image input [49]. All CLIP models were run locally on a MacBook Pro featuring an M1 Max chip and 64 GB of RAM. All GPT4.1 model queries were executed via the OpenAI API.

3.4 Zero-Shot Classification Approach

All three prompt configurations were first tested in a zero-shot setting, using the three previously introduced CLIP model architectures. To address our *RQ2*, we distinguished between classification results based solely on images and those based on both images and captions.

We began by evaluating model performance using *only images*, comparing the predicted labels to our manually assigned labels. Our manual labeling was, however, influenced not only by visual elements but also by textual content embedded in the images, such as protest signs, as their combination plays a significant role in shaping how a protest image is framed. To account for this, we incorporated Optical Character Recognition (OCR) into our classification pipeline. After testing various OCR libraries, including the Python library EasyOCR and Microsoft's Transformer-based OCR (TrOCR), we selected the Python library PaddleOCR, as it delivered the most accurate results. To process only meaningful text, we filtered out text related to news outlet names and photo credits for individual photographers.

For each model architecture, we then encoded the images, OCR-extracted text, and prompts using the CLIP model's image and text encoders. CLIP models are trained so that semantically related image-text pairs have high cosine similarity between their embeddings. To compute these similarities accurately and make them comparable across inputs, we applied L2 normalization to all embeddings, scaling them to unit length. Next, we calculated cosine similarity scores separately for each class between the normalized image embeddings $S_{(\text{img},c)}$ and all normalized prompt embeddings and, when OCR text was available, between the normalized OCR-extracted text embeddings $S_{(\text{ocr},c)}$ and all normalized prompt embeddings. We then fused the two cosine similarity scores into a single image-OCR similarity score for each class $S_{(\text{io},c)}$:

$$S_{(\text{io},c)} = \alpha \times S_{(\text{img},c)} + (1 - \alpha) \times S_{(\text{ocr},c)}$$

where α is the weight assigned to image features, and $1 - \alpha$ to OCR-extracted text features. If no OCR text was detected for an image, we set $S_{(\text{io},c)} = S_{(\text{img},c)}$.

Because frames are not mutually exclusive, we assigned predicted labels ($\hat{y}_c$) on a per-class basis using the rule:

$$\hat{y}_c = \begin{cases} 1, & \text{if } S_{(\text{io},c)} \geq t \\ 0, & \text{otherwise} \end{cases}$$

where t is the classification threshold. For each class c, a label was assigned if the similarity score $S_{(\text{io},c)}$ met or exceeded t; otherwise, no label was assigned. The optimal values for α and t were determined jointly through a grid search for each model. This search was performed on a subsample of 700 images, with at least 100 samples per class. We optimized for the F_β score, which is the weighted harmonic mean of precision and recall, setting $\beta = 0.5$ to give precision twice the weight of recall.

For the image-only setting, the best results were obtained when either (1) image features were weighted at 100% and OCR features at 0%, or (2) image features at 90% and OCR features at 10%, underscoring the dominant role of image features in accurate classification. The optimal thresholds in this setting ranged from 0.21 to 0.32 across models.

We applied the same procedure for classifying images combined with captions, adding the similarity scores for the caption embeddings $S_{(\text{capt},c)}$ (obtained from the CLIP text encoder) as a third input. For each class c, the fused cosine similarity score from images, OCR-extracted text, and captions ($S_{(\text{ioc},c)}$) was computed as:

$$S_{(\text{ioc},c)} = \gamma \times S_{(\text{io},c)} + (1 - \gamma) \times S_{(\text{capt},c)}$$

where γ is the weight given to the fused image–OCR score $S_{(\text{io},c)}$ from the previous step, and $1 - \gamma$ is the weight assigned to the caption-based score. The value of $S_{(\text{io},c)}$ was computed in the same way as before, with the parameter α jointly re-optimized together with γ and t.

Label assignment again followed:

$$\hat{y}_c = \begin{cases} 1, & \text{if } S_{(\text{ioc},c)} \geq t \\ 0, & \text{otherwise} \end{cases}$$

As before, we used all three CLIP model architectures and tested three different prompt configurations. The best α, γ, and t values were determined jointly through grid search, again optimizing for the F_β score with $\beta = 0.5$ on the subsample of 700 images, this time including also their captions. The grid search yielded OCR-extracted text weights between 0% and 10%, image feature weights between 70% and 90%, and caption feature weights between 10% and 30%, again highlighting the dominant role of image features. Optimal threshold values in this setting ranged from 0.20 to 0.31 across models.

To benchmark these results, we employed GPT4.1 with $\textit{Configuration}_3$ prompts, based on the assumption that detailed prompts would enhance performance. Including more detail, however, such as the full information from the codebook, performed worse than the $\textit{Configuration}_3$ prompt. We set the temperature parameter to 0, the image detail to "high," and evaluated both image-only inputs and combined image and caption inputs.

3.5 Few-Shot Classification Approach

Several few-shot approaches have been proposed to improve CLIP's performance, including methods that focus on optimizing prompt design, such as Zhou et al.'s [26] Context Optimization method. More recently, a new approach called "Training-free adaption method for CLIP" (Tip-Adapter) [50] has been developed. This method combines CLIP's zero-shot advantage of not needing to retrain the model, while still improving the performance through few-shot learning, matching, or even surpassing methods that require the full retraining of CLIP. Tip-Adapter is a method that constructs an adapter using a cache model derived from a small set of training samples. Instead of retraining the CLIP model, it appends a lightweight adapter to the frozen CLIP backbone. For classification tasks, it retrieves features from the cache and combines them with CLIP's pre-trained representations [50, p. 495].

To further improve Tip-Adapter's performance, Zhang et al. [50] introduced a fine-tuned variant called Tip-Adapter-F, which "unfreeze[s] the cached keys as learnable parameters" [50, p. 495] while keeping the CLIP backbone frozen. This approach requires minimal training effort while achieving state-of-the-art performance, outperforming models that rely on significantly more extensive training. Thus, we chose Tip-Adapter-F for our few-shot classification approach because it yields better results than the original Tip-Adapter.

To create our few-shot training dataset for the CLIP models, we followed the recommendations of prior research [26, 50] that suggests using 16 samples per class yields optimal improvements in performance. To ensure that our training data was as unambiguous as possible, we randomly selected 16 images per class from our dataset that were annotated with a single distinct label. In cases where fewer than 16 single-label samples were available, we appended the set with additional samples that featured the same label, even if they carried a second annotation. We followed the same strategy for image-caption pairs by prioritizing samples where the combined framing of image and text featured a single label. In necessary cases, we completed the set with samples featuring two annotations to reach 16 samples per class.

For fine-tuning the Tip-Adapter-F, we followed the recommendations provided by the authors [50]. Specifically, we used the AdamW optimizer with a learning rate of 0.001 and employed a cosine learning rate scheduler to gradually reduce the learning rate over time. Although we initially set the number of training epochs to 20, we reduced it to 15, as no further improvement in training loss or accuracy was observed beyond that point. Figure 1 shows an example of the training progression. While the original paper suggests a batch size of 256, we opted for a batch size of 15 due to our smaller dataset of 112 samples (7 classes $\times$ 16 samples). For feature fusion of image, OCR-extracted text, and caption embeddings, we applied the weighting scheme that had been optimized in our zero-shot experiments. We then tuned the Tip-Adapter-F hyperparameters δ and ϵ via grid search for each model architecture, prompt configuration, and input type. Across all combinations, δ consistently performed best at 0.5, while ϵ values were optimal at 7.5.

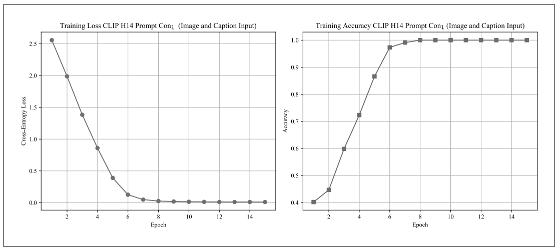


Fig. 1 Exemplary training loss and accuracy of the Tip-Adapter-F H14 model with $\mathit{Configuration}_1$ image and caption input

After fine-tuning, we recalculated the optimal classification thresholds t because Tip-Adapter-F changes the scale and distribution of similarity scores. Using the same subsample, we performed a grid search, optimizing for F-beta scores with $\beta = 0.5$ while keeping the zero-shot feature weightings and the fine-tuned δ and ϵ values fixed. In this few-shot setting, optimal thresholds ranged from 0.01 to 0.17 across models. Finally, we ran each CLIP model with all prompt configurations for both image-only and image-and-caption inputs.

For our few-shot classification approach with GPT4.1, we chose five image-only samples and three image-and-caption samples, since prior work has shown that three to five examples are sufficient for few-shot approaches with GPT models [51]. We restricted the image-and-caption set to just three examples to avoid exceeding OpenAI's internal token limit. In all cases, we prioritized examples annotated with a single label. In cases where three or five single-label examples were not available, we supplemented the few-shot data with samples containing two labels. These examples were incorporated into the prompt alongside our $\mathit{Configuration}_3$ information when providing input to GPT4.1. As in the zero-shot setup, the temperature parameter was set to 0 to ensure deterministic outputs, and the image detail was set to "high."

4 Results

4.1 Overview of Label Distribution Ground Truth

Table 1 displays the distribution of the ground truth, separated by image, caption, and their combination.

Table 1 Distribution of ground truth labels across image, caption, and their combination

Label	Image	Caption	Combination
Riot	488	685	787
Confrontation	528	801	903
Spectacle	1,285	1,273	1,716
Debate	1,558	3,151	3,227
Victimization	570	1,193	1,257
Violent	468	677	780
Peaceful	167	248	315
None	546	52	10
Overall n	5,610	8,080	8,995

Figure 2 presents the distribution of the number of unique labels per post, separated into caption, image, and their combination. A single post may contain multiple labels within both images and captions. The maximum number of possible unique labels is seven, reflecting the five frame categories as well as whether the protest is portrayed as violent and/or peaceful.

4.2 Performance of Zero-Shot Classification

Table 2 shows the overall performance results of all models and prompt configurations in the zero-shot setting. Precision, recall, and F1 score were computed using scikit-learn's MultiLabelBinarizer, which evaluates each metric on a per-sample basis and then averages the scores across all samples. Accuracy was calculated based on the exact match between all predicted and ground truth labels for each sample.

Table 3 presents the individual labels of the two best-performing CLIP models based on the overall F1 score from our zero-shot classification approach and the GPT4.1 results, using image-only and image-and-caption inputs.

To assess which labels are most frequently misclassified by automated methods compared

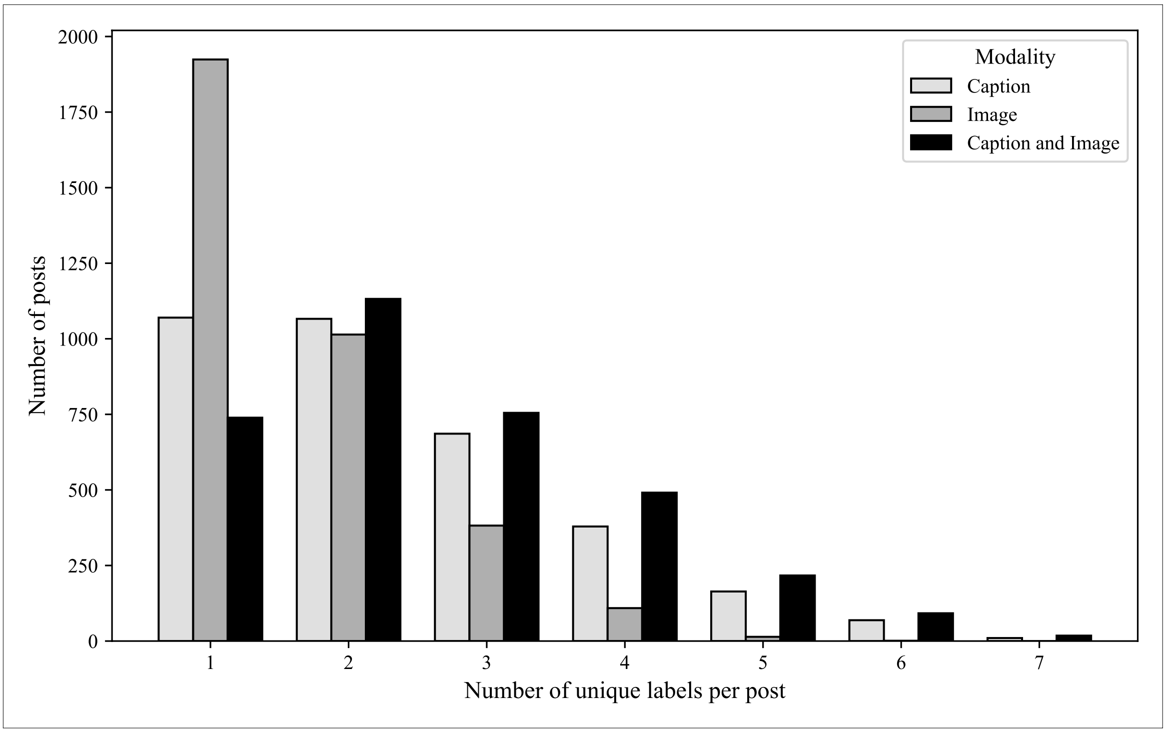


Fig. 2 Distribution of the number of unique labels per post, separated into caption, image, and their combination

to human coding—and thus to identify which frames are easier or harder to predict—Figure 3 presents the best performing CLIP model, L14 with *Configuration*$_2$, using images as the sole input. The figure shows the relative proportions of true positive, false positive, false negative, and true negative predictions for each label.

Figure 4 displays the same breakdown for the CLIP model L14 with *Configuration*$_1$, which achieved the best performance when using both image and caption as inputs.

4.3 Performance of Few-Shot Classification

Table 4 shows the overall performance results of all models and prompt configurations in our few-shot setting. Precision, recall, and F1 score were computed using scikit-learn's MultiLabelBinarizer.

Table 5 presents the individual labels of the two best-performing CLIP models based on the overall F1 score from our few-shot classification approach and the GPT4.1 results, using image-only and image and caption inputs.

To assess how the few-shot setting affects label misclassification compared to human coding, Figure 5 presents results from the best-performing CLIP model in this setting—L14 with *Configuration*$_2$—using images as the sole input. Figure 6 shows the corresponding results for the image-and-caption input condition, where the best-performing model was H14 with *Configuration*$_1$. As in the previous figures, both figures report, for each label, the relative proportion of true positive, false positive, false negative, and true negative predictions.

5 Discussion

After manually labeling the Instagram posts according to the five frames based on the protest paradigm and evaluating whether a protest is portrayed as violent and/or peaceful, we found that the debate, spectacle, and victimization frames were most prominently featured in both images and captions (see Table 1). These findings can be explained as follows. Media outlets on social

Table 2 Overall performances of zero-shot classification; best CLIP and GPT4.1 performances according to F1 score are highlighted in bold (percentages)

Input	Model	Configuration	Accuracy	Precision	Recall	F1 score
Image	L14	1	10.1	33.8	59.0	38.5
		2	**17.5**	**40.2**	**55.6**	**42.5**
		3	20.2	37.9	43.6	37.9
	H14	1	12.4	37.7	57.7	40.8
		2	15.9	34.6	46.0	36.2
		3	14.8	34.7	48.5	37.0
	G14	1	12.6	33.6	51.2	37.2
		2	15.7	34.8	46.3	36.7
		3	16.5	34.6	41.3	35.1
	GPT4.1	**3**	**33.9**	**64.7**	**70.8**	**64.7**
Image and Caption	L14	**1**	**1.1**	**41.7**	**80.2**	**50.5**
		2	3.1	43.5	61.7	44.8
		3	2.4	42.2	66.0	46.1
	H14	1	3.3	46.8	68.7	49.6
		2	1.3	37.7	68.5	43.9
		3	1.5	37.2	53.8	39.4
	G14	1	1.2	42.1	75.6	49.9
		2	1.3	39.8	78.8	48.4
		3	1.4	40.0	77.1	48.4
	GPT4.1	**3**	**25.8**	**75.6**	**81.8**	**75.2**

platforms increasingly compete for user attention, making the spectacle frame a strategic and likely choice, since dramatic or visually striking content is more likely to engage audiences. Similarly, the frequent use of the debate frame can be attributed to posts often including brief context or background information on the protest. The prominence of the victimization frame underscores the need for a nuanced analysis of how violence is portrayed in protest coverage. The prevalence of this frame in our dataset can be explained by the fact that the analyzed posts came from U.S.-centric news outlets and cover periods that included key events such as the deaths of George Floyd and Mahsa Amini, both of which we coded under the victimization frame. The riot and confrontation frames appear less frequently in our dataset, a somewhat surprising finding, as a higher prevalence of these frames would align more closely with the predictions of the protest paradigm. However, prior research has also shown that these frames are less common than others (e.g., [9]) and that news coverage of violent protests on social media, especially in visual content, is less prevalent than assumed (e.g., [40]).

Moreover, it is noteworthy that images most frequently feature a single label, whereas the most common number of labels when combining image and caption is two (see Figure 2). In our dataset, a frequent example of this combination is images depicting a spectacle frame accompanied by captions that provide additional contextual information, thereby representing the debate frame. This again highlights the importance of visually striking content within social media, where visuals play a central role [38] and are often paired with more detailed information as a caption. Overall, we, therefore, find that frames typically associated with the protest paradigm, especially those related to violence committed by protesters or confrontation with state authorities, appear less frequently than might be expected. In contrast, visually striking and emotional content, such as the victimization and spectacle frames, is used frequently. However, these results must be

Table 3 Individual label performance of best zero-shot models using image-only and image-and-caption inputs (percentages)

Input	Model and Configuration	Label	Accuracy	Precision	Recall	F1 score
Image	L14 and *Configuration*$_2$	Riot	82.6	39.7	43.4	41.5
		Confrontation	79.7	41.7	81.1	55.0
		Spectacle	50.1	35.8	42.3	38.7
		Debate	**72.4**	**73.3**	**61.6**	**66.9**
		Victimization	75.2	33.6	51.1	40.6
		Violent	74.2	19.3	28.2	22.9
		Peaceful	74.1	10.4	56.9	17.6
		None	70.3	27.9	55.1	37.0
	GPT4.1 and *Configuration*$_3$	Riot	94.0	76.7	82.8	79.6
		Confrontation	91.7	66.9	90.9	77.1
		Spectacle	69.9	62.4	48.7	54.7
		Debate	**80.2**	**71.3**	**94.1**	**81.2**
		Victimization	91.3	75.6	70.0	72.7
		Violent	94.1	79.3	76.3	77.8
		Peaceful	79.9	16.8	79.6	27.7
		None	85.9	77.2	16.1	26.7
Image and Caption	L14 and *Configuration*$_1$	Riot	75.6	46.8	50.7	48.7
		Confrontation	62.2	39.7	85.2	54.1
		Spectacle	50.0	49.9	87.4	63.5
		Debate	**75.1**	**94.3**	**78.1**	**85.5**
		Victimization	44.1	38.6	89.7	54.0
		Violent	38.5	23.8	79.9	36.5
		Peaceful	24.6	10.0	90.2	17.9
		None	94.5	0.5	10.0	1.0
	GPT4.1 and *Configuration*$_3$	Riot	91.1	82.1	78.1	80.1
		Confrontation	79.7	56.9	93.0	70.6
		Spectacle	63.4	64.1	60.6	62.3
		Debate	**85.4**	**96.6**	**87.4**	**91.8**
		Victimization	82.5	78.5	71.6	74.9
		Violent	90.2	84.5	69.4	76.2
		Peaceful	81.9	29.8	72.4	42.3
		None	99.7	0.0	0.0	0.0

interpreted in the context of the specific protests examined, as discussed earlier.

When evaluating the models' performances in relation to *RQ1*, we observe that multimodal models such as CLIP and GPT4.1, overall, perform considerably well in classifying protest frames. To gain a more nuanced understanding of this finding, however, we need to differentiate between (1) metrics, (2) individual label predictions, (3) model architectures, (4) prompt configurations, and (5) the use of zero-shot versus few-shot learning.

First, we note that overall accuracy tends to be considerably lower than F1 scores (see Tables 2 and 4). This is because accuracy is based on exact matches between predicted and ground truth label combinations for each sample, whereas the F1 score is calculated per label and then averaged. This suggests that while models are effective at identifying individual labels (see Tables 3 and 5), correctly predicting the full set of labels for a given post is more challenging. Given that protest paradigm frames often depend on subtle nuances, which leads to the coding of multiple

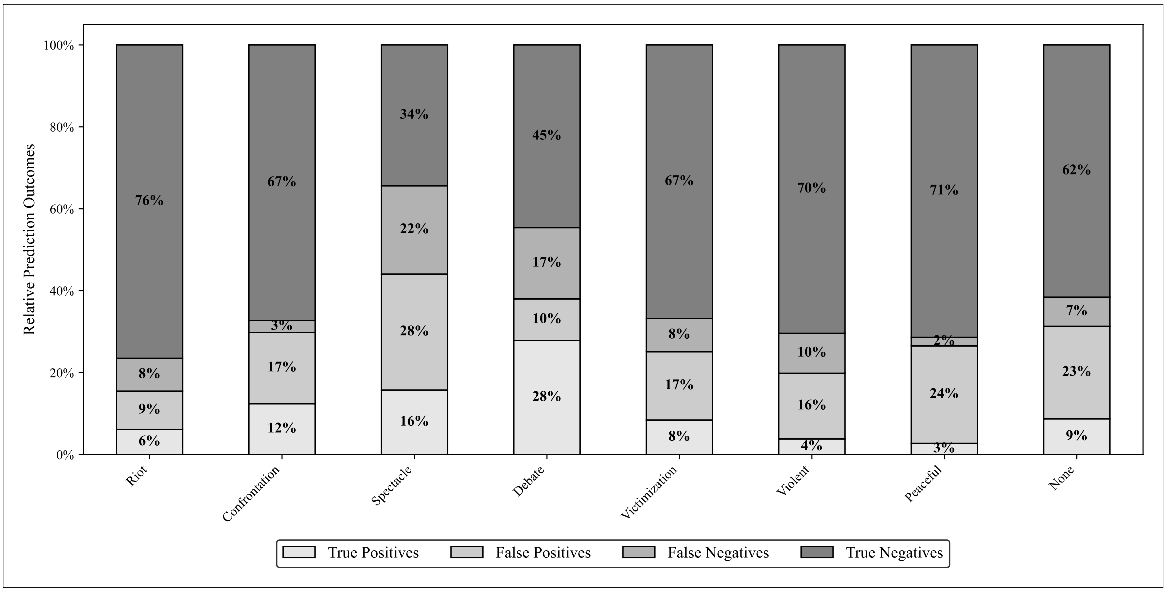


Fig. 3 Prediction outcomes per label, showing the proportion of true positives, false positives, false negatives, and true negatives for the L14 model with *Configuration*$_2$, using image-only input in the zero-shot setting

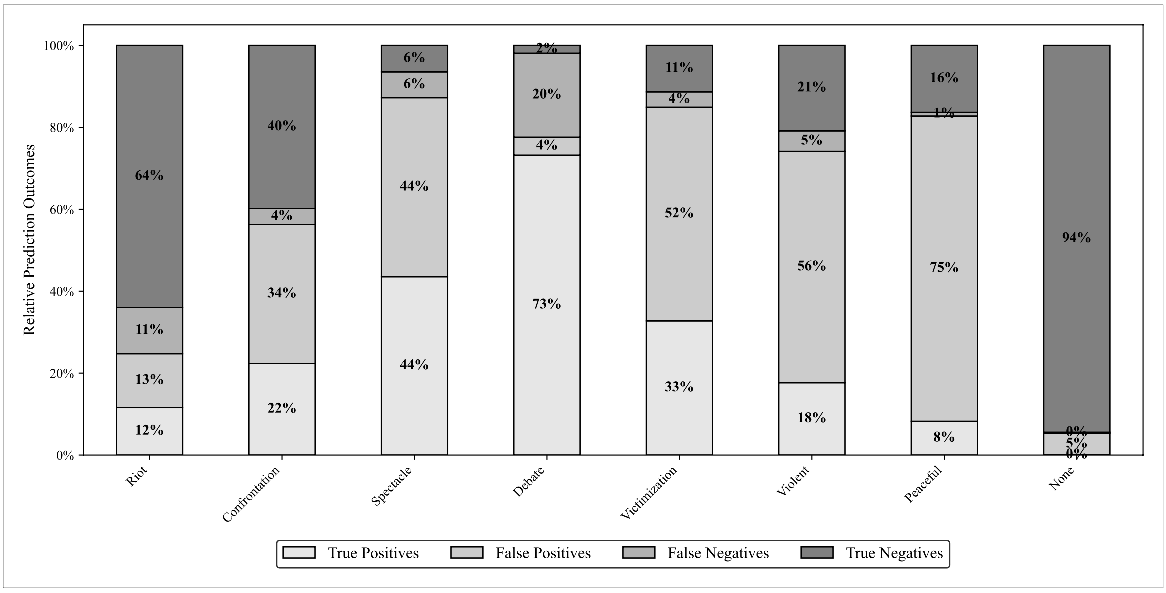


Fig. 4 Prediction outcomes per label, showing the proportion of true positives, false positives, false negatives, and true negatives for the L14 model with *Configuration*$_1$, using image-and-caption input in the zero-shot setting

frames per post (see Figure 2), this result is not entirely unexpected. Nonetheless, the strong performance in predicting individual labels is a promising outcome.

Second, we observe notable differences in the classification performance of individual labels (see Tables 3 and 5). The debate frame is consistently well predicted, with accuracy reaching up

Table 4 Overall performances of few-shot classification; best CLIP and GPT4.1 performances according to F1 score are highlighted in bold (percentages)

Input	Model	Configuration	Accuracy	Precision	Recall	F1 score
Image	L14	1	7.9	40.1	60.8	43.9
		2	**16.5**	**49.9**	**59.8**	**49.9**
		3	20.2	52.3	50.9	48.1
	H14	1	14.9	49.1	53.0	46.7
		2	15.1	46.4	45.6	42.5
		3	17.5	48.2	45.3	43.6
	G14	1	9.0	38.2	44.1	37.4
		2	9.8	40.1	57.9	43.0
		3	13.4	44.5	52.9	44.3
	GPT4.1	**3**	**39.8**	**72.6**	**62.8**	**64.5**
Image and Caption	L14	1	2.3	50.1	74.9	54.0
		2	5.1	55.2	74.9	56.9
		3	4.4	53.0	72.3	55.5
	H14	**1**	**3.2**	**55.4**	**75.3**	**57.2**
		2	2.7	49.8	65.4	49.9
		3	2.4	47.1	65.3	49.5
	G14	1	3.5	53.5	67.2	54.0
		2	3.1	51.9	56.2	47.0
		3	3.5	53.7	57.2	48.9
	GPT4.1	**3**	**28.8**	**85.3**	**68.1**	**72.1**

to 86.6% and F1 scores as high as 92.6%, a particularly strong result given that this frame is the most frequently occurring frame in our dataset. In contrast, labels such as peaceful and spectacle show lower performance, with maximum F1 scores of 58.1% (accuracy 91.6%) and 63.5% (accuracy 71.6%), respectively. This difference may be due to the more complex nature of these labels, which makes them more difficult to detect.

To better understand these individual performance differences, it is useful to examine the results in Figures 3-6, which show how labels were correctly and incorrectly classified across image-only versus image-and-caption inputs, differing between zero-shot and few-shot settings. These figures present the following:

The debate frame shows a consistent improvement when captions are included, in both zero-shot and few-shot settings. Performance increases further when moving from zero-shot to few-shot settings, as seen by comparing Figures 3 and 5 (image only), as well as Figures 4 and 6 (images and caption). This suggests that the additional contextual information provided by captions, combined with limited training examples, substantially benefits the models' ability to detect this frame.

In contrast, the peaceful label performs poorly in all settings, primarily due to high false positive rates. Human coders, following the codebook definition, assigned this label only to clear and explicit signs of peacefulness, whereas the automated models appear to interpret more subtle cues as peaceful. This discrepancy suggests that the operational definition of peaceful protests may need to be further refined and more conservative if automated methods are to achieve more accurate classification results.

The spectacle frame, while achieving high true-positive predictions, particularly in the image-and-caption settings (Figures 4 and 6), also exhibits high false-positive rates, indicating frequent over-prediction. This frame is challenging for both automated methods and human coders, as its definition covers a wide range of features, such as celebrity involvement and the scale of the

Table 5 Individual label performance of best few-shot models using image-only and image-and-caption inputs (percentages)

Input	Model and Configuration	Label	Accuracy	Precision	Recall	F1 score
Image	L14 and *Configuration*$_2$	Riot	82.3	40.5	52.7	45.8
		Confrontation	78.3	39.7	80.3	53.1
		Spectacle	59.5	46.8	62.7	53.6
		Debate	**73.8**	**68.8**	**77.0**	**72.7**
		Victimization	74.9	32.5	47.5	38.6
		Violent	76.5	24.7	35.7	29.2
		Peaceful	73.4	10.7	61.1	18.2
		None	84.1	0.0	0.0	0.0
	GPT4.1 and *Configuration*$_3$	Riot	89.3	71.5	40.6	51.8
		Confrontation	92.6	75.2	77.5	76.3
		Spectacle	71.6	68.2	44.6	53.9
		Debate	81.5	86.6	69.9	77.4
		Victimization	87.6	58.8	83.5	69.0
		Violent	**94.1**	**79.4**	**78.4**	**78.9**
		Peaceful	94.4	44.4	64.1	52.5
		None	86.6	70.9	26.7	38.8
Image and Caption	H14 and *Configuration*$_1$	Riot	80.8	62.5	39.8	46.6
		Confrontation	71.7	47.7	72.1	57.2
		Spectacle	56.5	56.3	56.9	56.6
		Debate	**86.6**	**95.8**	**89.6**	**92.6**
		Victimization	68.5	57.2	53.9	55.5
		Violent	41.6	21.1	57.4	30.8
		Peaceful	38.6	10.8	78.4	18.9
		None	99.7	0.0	0.0	0.0
	GPT4.1 and *Configuration*$_3$	Riot	89.8	85.8	66.6	75.0
		Confrontation	87.0	75.2	75.4	75.3
		Spectacle	65.4	76.3	44.3	56.1
		Debate	**68.3**	**97.8**	**67.6**	**80.0**
		Victimization	83.7	82.8	69.8	75.8
		Violent	86.6	88.6	46.8	61.2
		Peaceful	91.6	53.3	63.8	58.1
		None	99.3	0.0	0.0	0.0

protest. The broad scope of the category makes it more susceptible to misclassification.

When comparing the individual label predictions in the zero-shot and few-shot settings for image-and-caption inputs, classification improves in terms of true negative predictions but declines in true positive predictions across all labels, with the debate frame being the only exception. These patterns suggest that few-shot learning with image-and-caption inputs results in a more conservative classification, reducing wrongly classified labels but also missing some correct labels. In the image-only setting, the pattern is often reversed: few-shot learning increases true positive predictions, while zero-shot learning features more true negative predictions. This suggests that few-shot learning makes the model more liberal in assigning true as well as false labels, while zero-shot learning is more conservative, prioritizing true negatives but missing correct labels.

Overall, there are substantial classification differences between individual labels, which can, at least in part, be mitigated by providing additional

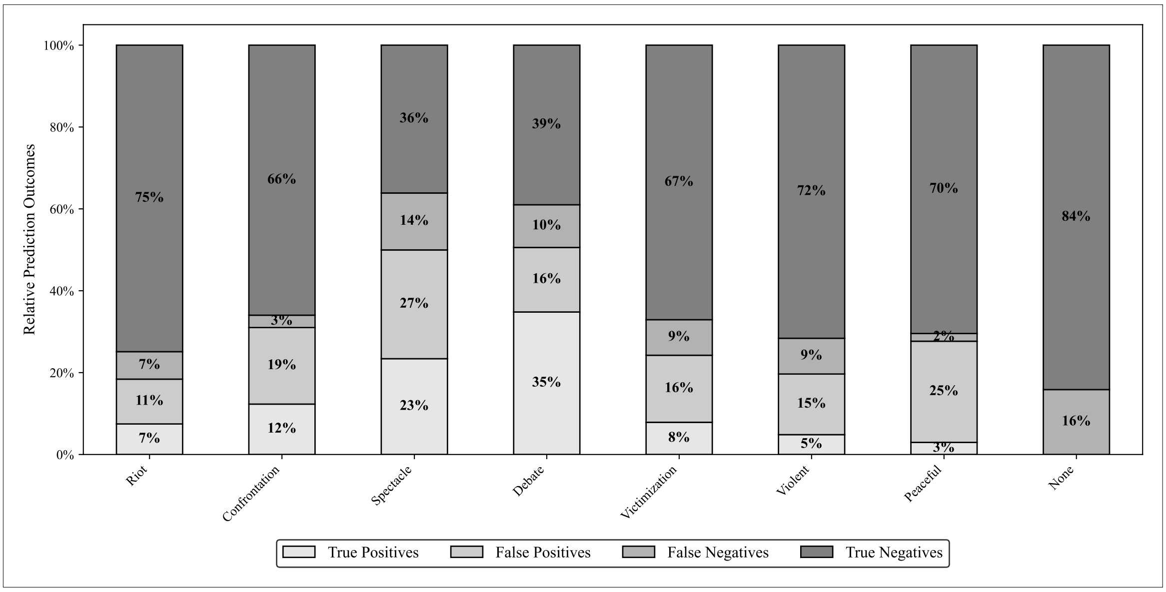


Fig. 5 Prediction outcomes per label, showing the proportion of true positives, false positives, false negatives, and true negatives for the L14 model with *Configuration*$_2$, using image-only input in the few-shot setting

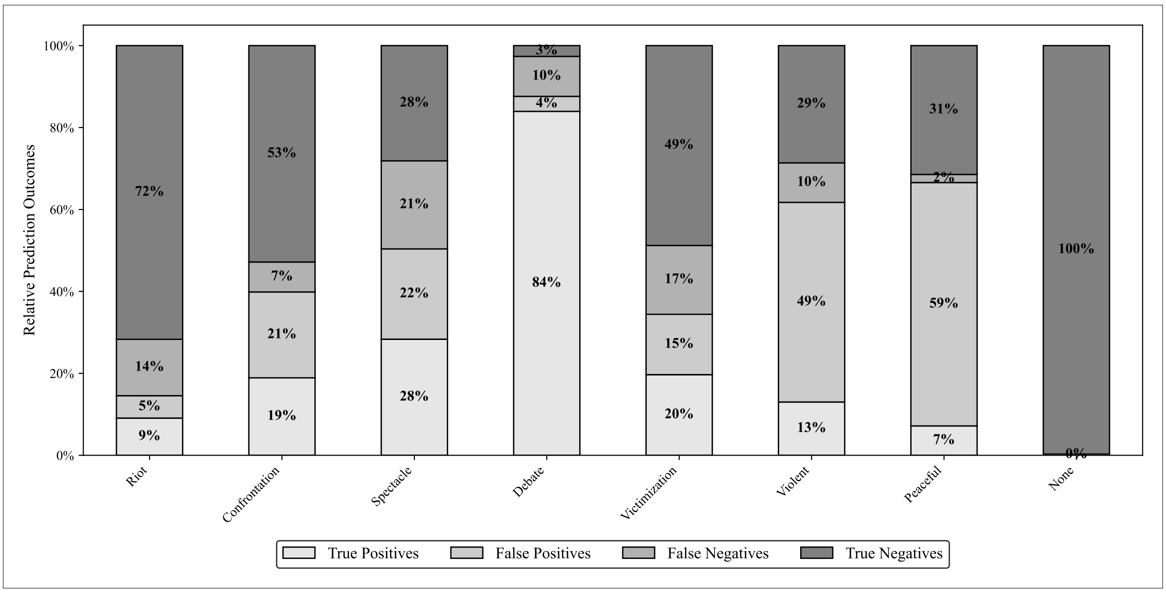


Fig. 6 Prediction outcomes per label, showing the proportion of true positives, false positives, false negatives, and true negatives for the H14 model with *Configuration*$_1$, using image-and-caption input in the few-shot setting

information through captions or by moving from a zero-shot to a few-shot setting.

The classification performance of the individual labels is comparable to or even surpasses previous results in news frame classification using text-only models, such as those based on the GPT3 architecture, which achieved an average accuracy of 72.0% [24]. When comparing individual F1 scores to those reported by Tourni et al. [22], who classified frames in news headlines and images related to gun violence based on CNNs and BERT, we also obtained comparable or better results: while some of their news frame labels achieve F1 scores above 90%, others fall below 40%, depending on the frame and model used. Hence, we conclude that while our models' performances across full label combinations are moderate—likely due to the complexity and latent nature of protest frames related to the protest paradigm—the results for individual frame labels demonstrate strong performance, particularly given the challenges of multimodal input.

Third, we observe that GPT4.1 consistently outperforms every CLIP model architecture across nearly all overall and individual performance metrics. This finding is not surprising, given that GPT4.1 is a significantly larger model trained on more extensive and diverse datasets. Previous research has also demonstrated that GPT models tend to outperform smaller models such as BERT on similar classification tasks, (e.g. [24]). Nevertheless, CLIP models still show promising results. For instance, the H14 model achieved an overall F1 score of 57.2% (few-shot, $\textit{Configuration}_1$, image-and-caption input). Moreover, this model also outperforms GPT4.1 in classifying the debate frame, highlighting that while CLIP models may not surpass GPT4.1 in overall performance metrics, they can achieve better results for specific labels.

When comparing different CLIP model architectures, an interesting pattern emerges: Based on F1 scores, the L14 model outperforms all other models in the zero-shot setting for both image-only and image-and-caption inputs, as well as in the few-shot setting for image-only input. Only in the few-shot setting with combined inputs does the H14 model achieve the highest performance. These findings suggest that larger model sizes of CLIP do not necessarily yield better performance on complex classification tasks. In fact, smaller models such as L14 tend to perform equally well or even better. This observation is in line with findings from Mavridis et al. [52, p. 21], who reported that "shallow models outperform[ed] deep language models" in the context of news frame classification.

Fourth, when evaluating different prompt configurations with CLIP models, we find that less detailed prompts perform better across all settings, with either $\textit{Configuration}_1$ or $\textit{Configuration}_2$ achieving the highest F1 scores. This finding is plausible, as CLIP models are trained on image–text pairs that, while informative, tend to be concise and to the point. Providing excessive detail in prompts does not necessarily improve classification performance and may even confuse the model. Therefore, it seems advisable to use prompts that emphasize the most important aspects of the classification task while avoiding unnecessary detail.

Fifth, when comparing model performance between zero-shot and few-shot settings, we see that few-shot models generally perform better. This improvement is particularly pronounced for CLIP models. For example, the L14 model ($\textit{Configuration}_2$, image and caption input) shows an overall F1 score improvement of up to 12.1% when moving from zero-shot to few-shot classification. This highlights the effectiveness of few-shot approaches, especially when done with methods like Tip-Adapter-F. In contrast, GPT4.1 displays a different pattern. The differences between zero-shot and few-shot performance are relatively minor. This outcome is not surprising, as GPT4.1 has been trained on an extremely large and diverse dataset, giving it a strong foundation even without additional examples. As a result, few-shot examples have less impact on performance. Still, the slight improvements in accuracy and precision suggest that GPT4.1 can benefit from few-shot learning, albeit with trade-offs in recall.

Thus, to answer *RQ1*, we find that multimodal models perform considerably well in solving complex tasks such as news frame classification related to the protest paradigm, achieving results that are comparable, or in some cases even better, than those reported in prior studies, even when applied to more complex data involving both image and text. This is particularly noteworthy given that traditional supervised models typically require large amounts of labeled training data, the production of which is highly labor-intensive. Model performance, however, depends strongly on model architecture, prompt design, and the distinction between zero-shot and few-shot learning, especially for CLIP models. Moreover, performances vary considerably between individual labels. Overall, GPT4.1 outperforms CLIP models by a significant margin, largely due to its extensive training data and model size. Still, given the high cost and limited transparency of GPT models, CLIP represents a promising alternative. CLIP models not only classify certain labels more accurately than GPT-4.1, but, when enhanced with few-shot learning, can deliver substantial performance gains with minimal annotation effort.

In relation to *RQ2*, we find that in both the zero-shot and few-shot settings, the best-performing models (based on overall F1 score) achieved better results when using both image-and-caption inputs compared to image-only inputs. In the zero-shot setting, the best image-and-caption CLIP model, L14, achieved an F1 score of 50.5%, compared to 42.5% for the best image-only CLIP model (GPT-4.1: 75.2% vs. 64.7%). In the few-shot setting, the best-performing CLIP model for image-and-caption input, H14, reached an F1 score of 57.2%, compared to 49.9% for the best image-only model (GPT-4.1: 72.1% vs. 64.5%). This is in line with previous research on news frame prediction, which has also shown that combining textual and visual features improves predictive performance compared to using either modality alone [22]. Additionally, the comparison of individual label performance (Figures 3-6) shows that adding caption information consistently increases true positives across all labels when compared to the image-only condition. This indicates that captions provide complementary context to the visual content, enabling the models to more reliably identify the correct frames.

Moreover, when examining the accuracy scores of all best-performing models, we see a consistent trend: all models achieve higher accuracy when using image-only input. This is likely because fewer input signals simplify the classification task, and our ground truth also contains fewer labels per sample when only image data are used. It may also indicate that on visually centric platforms such as Instagram, expressive images with clear frames are more likely to be featured in news reporting, while captions, which are relatively brief compared to full newspaper articles, tend to contain less-obvious framing.

Overall, these results suggest that combining image and text generally enhances frame classification in terms of F1 score. However, this benefit comes with a trade-off in accuracy, likely due to the increased complexity introduced by the multimodal input and the greater number of possible labels.

6 Conclusion

In this paper, we evaluated the performance of CLIP and GPT4.1 in classifying protest frames in news content based on image and text input, differentiating between zero-shot and few-shot approaches. Our findings show that these multimodal models perform considerably well in this complex classification task. However, performance is influenced by model architecture, the distinction between correctly identifying individual labels versus full label combinations, and prompt configurations. While GPT4.1 significantly outperforms CLIP models, CLIP remains a compelling alternative due to its greater transparency and the absence of token-related costs. Since CLIP models already perform well for certain frames, such as the debate frame, and can be substantially improved through few-shot learning, they represent a practical option for achieving solid results with minimal training effort.

Moreover, our results show that the application of frames in relation to the protest paradigm differs from what would be expected based on research on traditional media framing. We observe an abundance of the debate, spectacle, and victimization frames, whereas the riot and confrontation frames occur less frequently. Most often, submissions feature two frames when analyzing captions and images, with the most common combination in our context being the spectacle and debate frames. As discussed, these distributions can be attributed both to the logic of how information and news are presented on social media and to the specific protest cases examined. We therefore encourage future research to examine in greater depth how the application of frames differs between news on social media and in traditional outlets such as newspapers, and to investigate the mechanisms driving these differences.

While this study offers valuable insights into multimodal frame classification, it should be considered in light of its limitations. First, given the rapid advances in prompt-based models, it is likely that even more powerful models will soon become available. Future research is encouraged to evaluate the performance of these emerging models in handling complex tasks, such as frame classification. Second, expanding the dataset could enhance the generalizability of the findings. This includes incorporating more diverse news sources beyond the United States, exploring different platforms, and integrating additional media types, such as video data. Third, during manual coding, we observed that the alignment between frames and their delegitimizing or legitimizing functions as defined by the protest paradigm was not always consistent. For example, the spectacle frame, typically considered delegitimizing, also presented protests in a positive light. Thus, future research is encouraged to reevaluate the frames proposed by the pro-test paradigm, particularly in terms of their delegitimizing and legitimizing functions.

Declarations

- Funding: No funding was received for conducting this study.
- Competing interests: The authors have no competing interests to declare that are relevant to the content of this article.
- Ethics approval and consent to participate: Our research strictly adheres to ethical standards, ensuring that we do not violate privacy rights or disregard any societal or cultural values.
- Consent for publication: All authors agree with the content and gave consent to submit this manuscript.
- Data availability: We ensure the reproducibility of our results by providing a detailed account of our model configurations, hyper-parameters, and the various models employed. To enhance transparency, we include a comprehensive codebook that clearly explains the ground-truth labeling process. While we thoroughly document our data reduction steps, we are unable to share the raw data itself due to licensing restrictions.
- Materials availability: All materials that do not violate licensing restrictions are included.
- Code availability: May be shared upon request.
- Author contribution: All authors contributed equally to this work.

Appendix A Codebook

A.1 1st Step of Coding: Detection of Relevant Posts

In this coding step, we determine whether (1) or not (0) a post features at least one protest event executed by a protest movement or an individual protester affiliated with a movement. A post qualifies if it substantially references at least one protest event in the image, text, or both. "Substantially" means the image must directly depict a protest event or the text must include at least one full sentence mentioning it, which allows for the framing of the protest. A protest event can be featured either directly or indirectly through people talking about it. Posts that depict multiple protest events, whether carried out by the same or different protest movements or individuals affiliated with a movement, are also included. Posts labeled as "1" are then analyzed for their framing and whether the post emphasizes the violent and/or peaceful nature of the protest. If a post does not feature a reference to a protest event, it is excluded from further analysis. Using protest events as our unit of analysis for subsequent frame analysis and the evaluation of whether the violent and/or peaceful nature of the protest is emphasized aligns with the literature on the protest paradigm [9, 11, 29, 31].

Definition: Generally, a "protest event is defined as a collective, public action by non-state actors that expresses criticism or opposition and is associated with the articulation of a social or political concern" ([53, p. 4], own translation)

Rucht, Hocke, and Ohlemacher [53] specify this definition according to the following four characteristics:

1. "The protest must have an action-oriented character and thus represent a form of explicit opposition, at least calling on others to take action" ([53, p. 4], own translation).
2. "The action must be connected to the formulation of a societal or political issue, or it must be able to be connected to one" ([53, p. 5], own translation).
3. "The protest has a public character, meaning it either takes place in a public space or at least aims for a public impact or targets a person or

institution of public interest" ([53, p. 6], own translation).

4. "The participants of the action are – generally speaking – collective and non-state actors" ([53, p. 6], own translation).

Note that for characteristic (4), "collective" can also refer to an individual who speaks on behalf of a protesting group or actively supports it. In addition, "non-state actors" include party factions and members of parliament but not executive bodies or entire parliaments [53, p. 7].

Exclusion criteria:

- Information event: The post's focus is on an event that provides information related to a protest but does not articulate specific societal or political issues.
  Examples: panel discussions; theater performances; cultural events; information sessions.
- Remembrance: The post's focus is on a protest event as a historical act of remembrance, without formulating current social or political issues.
  Examples: references to Black History Month; historical mentions of the women's suffrage movement; commemorations of the Stonewall Riots.
- Statistical Information: The post's focus is only on statistical information related to a protest movement and is not related to protest events.
  Examples: statistical information on a protest movement; opinion surveys related to a protest movement.
- Consequences: The post's focus is on the consequences in the context of the protest or protest event.
  Examples: police officers being convicted; reports on court proceedings; policy reforms; police defunding.
- Overview: The post provides a summary of various news events over a specific period, rather than centering on a protest event.
  Examples: a recap of last year's major news; a summary of the past week's key events.

A.2 2nd Step of Coding: Frame Detection

In this coding step, we assess the presence (1) or absence (0) of the following frames within a post. We code frames based on all information in a post that contributes to the framing of protest events. This includes all signals related to the featured protest events, regardless of whether the post refers to a single event or multiple events carried out by the same or by different protest movements or individuals affiliated with a movement. The specific frames are applied if certain aspects of protest events are highlighted. Since frames are not mutually exclusive, a post can be assigned multiple frames. We also evaluate whether a post emphasizes protests as violent (1) or not (0), and whether it emphasizes protests as peaceful (1) or not (0). We assign a code to each dimension in the following configurations:

- Visual: the image alone.
- Textual: the description alone.
- Combined: both the image and its accompanying description.

Riot Frame: Apply the riot frame if the post features references that align with the following criteria:

- Definition: "The riot frame is most prominently operationalized through the lens of physical violence and property destruction. However, the concept also extends to threats and implied violence. To explore the riot frame through the lens of threat, coders identified if (1) the post included mentions of the action of physical violence or property destruction, or (2) communication rhetoric or messages insinuated violence or destruction (expressed intentions to harm someone)" [37, p. 1676].
- Further explanation: For the application of this frame, there must be explicit mentions of violence or threats against people or property; simply describing protesters as "angry," for example, does not qualify. Unproven accusations of protest violence in the context of a trial are not part of this frame. Violence against property can include acts of vandalism, such as graffiti. Note that based on this definition, actions such as blocking streets, as in the Canadian trucker protests, do not qualify for this frame.
- Examples: protesters looting stores; setting fire to objects; engaging in violence against individuals or groups; storming buildings; vandalizing property with, for example, graffiti; threats of violent behavior.

Confrontation Frame: Apply the confrontation frame if the post features references that align with the following criteria:

- Definition: "Confrontation focused on clashes, altercations, and disputes between protesters and authorities, including arrests" [29, p. 617].
- Further explanation: The protesters do not necessarily need to be visible in the post; it is sufficient if the clash with the police is implied. Note that if the police march with the protest or police officers embrace protesters, the post is not coded as "confrontation." Reports of protesters being killed by state authorities or being shot with, for example, rubber bullets during a protest indicate a confrontation between the two, thereby classifying it under this frame.
- Examples: protesters barricading themselves to confront the police; protest movements clashing with police or state authorities; mention of tear gas usage; protest movements engaging in confrontational interactions with politicians.

Spectacle Frame: Apply the spectacle frame if the post features references that align with the following criteria:

- Definition: "Spectacle, also known as the circus frame, showed the oddity, drama, and emotion of protest. Articles with the spectacle frame emphasized the size of marches, the influence of celebrities on mobilizations, or portrayed protesters as unusual" [29, p. 617].
- Further explanation: It is not enough for the protest event to be merely spectacular; the post must explicitly emphasize the spectacle of the protest. Note also that art installations need to be exceptional to be coded as a "spectacle."
- Examples: the post's language related to the protest is sensational; protesters wear striking or unconventional clothing to attract attention (note that traditional and indigenous clothing do not qualify); protest movements stage attention-grabbing performances; celebrities are involved in the protest event (note that politicians are also celebrities and are coded within this frame if they are part of the protest; recognizable activists who take part in protests regularly, such as Greta Thunberg, are not considered celebrities; in images, celebrities must be shown during their protest participation for this frame to apply); celebrities directly offer their opinion on the protest (note that legislative decisions related to a protest or talking to a protester do not count); young children are prominently featured as part of the protest; emotions are strongly highlighted; conspiracy theories are prominently featured; protest-related fireworks are shown; extreme violence that can be considered a spectacle is depicted; police and protesters hug or take a knee together; the enormous size of the protest is emphasized (begins with mentioning a thousand protesters); a high number of arrests and/or killings of protesters is highlighted (starting at 50).

Debate Frame: Apply the debate frame if the post features references that align with the following criteria:

- Definition: "[T]he debate frame explained the arguments, demands, and reasons of the protesters, and the larger context and historical background that triggered mobilizations" [29, p. 617].
- Further explanation: References related to this frame may appear in images, such as on protest signs, or within the text of a post. Displaying only the name of a protest movement on a sign or simply naming the protest movement (e.g., climate protest) without mentioning any further information on the protest does not qualify for this frame. When people discuss a protest event by providing background information and a broader context, or when a quote from them related to the event is depicted, this frame is also applied. If an art installation does not contain a concrete message that relates to the grievances or demands, it is not coded as debate.
- Examples: protesters' grievances are emphasized; protesters' viewpoints and demands are highlighted; the context of the protest is featured; the goals of the protesters are articulated; a quote by someone is featured referencing the context of the protest.

Victimization Frame: Apply the victimization frame if the post features references that align with the following criteria:

- Definition: The victimization frame portrays protesters as victims of physical violence by state authorities. This can also include threats and implied physical violence. The victimization frame is also applied when the protest theme highlights physical violence committed by state authorities against protesters.

- Further explanation: Protests with the central theme of physical state violence are also considered within this frame. However, simply displaying the name of a protest movement on a sign or simply naming the protest movement does not qualify. Symbolic actions, such as taking a knee in support of Black Lives Matter, qualify. Displaying the image or name of someone who was subjected to police violence, such as Mahsa Amini, George Floyd, or Alexei Navalny, also qualifies. State violence against the property of protesters in the context of a protest also qualifies. State violence in the context of war, when directed at civilians rather than specifically targeting protesters, does not fall within this frame.
- Examples: protesters are beaten by police; tear gas is used against protesters; protesters are physically harmed by state authorities; the protest theme is linked to police and/or physical state violence against protesters, such as the killing of George Floyd or Mahsa Amini.

Label each post's image, text, and their combination to indicate whether the protest is emphasized as violent (1) or not (0) and whether it is emphasized as peaceful (1) or not (0). This approach is in line with the literature in this field [54, p. 253]. "Violent" and "peaceful" are defined as follows:

Violent Protest:

- Definition: "a form of protest that causes direct harm to people or property" [54, p. 253].
- Further explanation: A post must include direct references to harm against people or property (here, vandalism such as graffiti on statues and buildings also qualifies). Concrete threats of violence are coded as "violent." However, unproven accusations of protest violence in the context of a trial are not coded as "violent."
- Examples: burning houses; the threat of violence against police; protesters physically hurting state authorities.

Peaceful Protest:

- Definition: a form of protest that depicts "peaceful activities such as vigils or silent-pacific marches, or the ... direct reference to the nonviolent course of events" [29, p. 617-618].
- Further explanation: Peaceful protests can also occur when state authorities explicitly engage in peaceful protest with the protesters. Note that the peaceful aspect must be directly connected to the protest itself; featuring only a memorial site, for instance, does not qualify. Protesters only sitting during a protest do not qualify.
- Examples: references to, for instance, "respectful" protests; showing "peace signs" such as two fingers in a "V"; white doves; mentioning "love"; protesters raising their hands in surrender in front of the police; kneeling in front of the police.

Appendix B Prompts

Basic Definitions

- Riot: Focus on physical violence or destruction like fires, looting, or smashed property.
- Confrontation: Clashes with police or authorities, often involving arrests.
- Spectacle: Unusual, emotional, or dramatic protest with costumes or celebrity presence and emphasis on big size of protest.
- Debate: Emphasis on demands, reasons, arguments, grievances, history, and broader context of protest.
- Victimization: Physical police or state violence against protesters.
- Violent: Protest causing physical harm or damaging property.
- Peaceful: Calm protest actions like vigils, peace signs, or silent nonviolent marches.

Basic Definitions and Explanations

- Riot: Focus on physical violence or destruction like fires, looting, or smashed property. Must show explicit violence or threats against people or property-graffiti counts. Anger or unproven claims do not qualify. Blocking roads is not sufficient.
- Confrontation: Clashes with police or authorities, often involving arrests. Protesters do not need to be visible-confrontation can be implied. Rubber bullets or protester deaths count; hugging police does not.

- Spectacle: Unusual, emotional, or dramatic protest with costumes or celebrity presence and emphasis on big size of protest. The spectacle must be clearly emphasized; ordinary art without emphasis does not qualify.
- Debate: Emphasis on demands, reasons, arguments, grievances, history, and broader context of protest. Protest signs or text must explain the cause or give context-naming the protest alone is insufficient.
- Victimization: Physical police or state violence against protesters. Kneeling or raised fists only count in the BLM context. Naming victims of police violence such as George Floyd, Mahsa Amini, or Alexei Navalny qualifies. General laws or war violence do not.
- Violent: Protest causing physical harm or damaging property. Includes real attacks, destruction, or threats. Graffiti counts. Unproven accusations do not.
- Peaceful: Calm protest actions like vigils, peace signs, or silent nonviolent marches. Must clearly relate to protest. Sitting or showing a memorial does not qualify. Peaceful interaction with police can qualify.

Basic Definitions, Explanations, and Examples

- Riot: Focus on physical violence or destruction like fires, looting, or smashed property. Must show explicit violence or threats against people or property-graffiti counts. Anger or unproven claims do not qualify. Blocking roads is not sufficient. Examples: looting stores; setting fires; smashing windows; storming buildings; violent threats.
- Confrontation: Clashes with police or authorities, often involving arrests. Protesters do not need to be visible-confrontation can be implied. Rubber bullets or protester deaths count; hugging police does not. Examples: protesters blocking police; tear gas; clashes with state forces; confrontations with politicians.
- Spectacle: Unusual, emotional, or dramatic protest with costumes or celebrity presence and emphasis on big size of protest. The spectacle must be clearly emphasized; ordinary art without emphasis does not qualify. Examples: protesters in striking outfits; performances; celebrities giving opinions or joining marches; fireworks; hugging police; massive protest size (1000+); 50+ arrests.
- Debate: Emphasis on demands, reasons, arguments, grievances, history, and broader context of protest. Protest signs or text must explain the cause or give context-naming the protest alone is insufficient. Examples: signs or text showing protest goals; reasons for action; historical background; stated grievances.
- Victimization: Physical police or state violence against protesters. Kneeling or raised fists only count in the BLM context. Naming victims of police violence such as George Floyd, Mahsa Amini, or Alexei Navalny qualifies. General laws or war violence do not. Examples: protesters beaten or tear gassed; scenes of state-inflicted harm; naming people who suffered from police violence.
- Violent: Protest causing physical harm or damaging property. Includes real attacks, destruction, or threats. Graffiti counts. Unproven accusations do not. Examples: burning houses; threats against police; protesters injuring authorities.
- Peaceful: Calm protest actions like vigils, peace signs, or silent nonviolent marches. Must clearly relate to protest. Sitting or showing a memorial does not qualify. Peaceful interaction with police can qualify. Examples: peace signs; white doves; mentions of love; showing of victory/-peace sign; raising hands in surrender in front of the police; kneeling in front of police.

References

[1] Brown DK, Mourão RR. Protest Coverage Matters: How Media Framing and Visual Communication Affects Support for Black Civil Rights Protests. Mass Communication and Society. 2021;24(4):576–596. https://doi.org/10.1080/15205436.2021.1884724.

[2] Harlow S, Brown DK. A New Protest Paradigm: Toward a Critical Approach to Protest News Analyses. The International Journal of Press/Politics. 2023;28(2):333–343. https://doi.org/10.1177/19401612231153377.

[3] Masullo GM, Brown DK, Harlow S. Shifting the Protest Paradigm? Legitimizing and

Humanizing Protest Coverage Lead to More Positive Attitudes toward Protest, Mixed Results on News Credibility. Journalism. 2024;25(6):1230–1251. https://doi.org/10.1177/14648849231200135.

[4] McLeod DM, Hertog JK. The Manufacture of 'Public Opinion' by Reporters: Informal Cues for Public Perceptions of Protest Groups. Discourse & Society. 1992;3(3):259–275. https://doi.org/10.1177/0957926592003003001.

[5] Harlow S, Bachmann I. Police, Violence, and the "Logic of Damage": Comparing U.S. and Chilean Media Portrayals of Protests. Mass Communication and Society. 2024;27(2):254–277. https://doi.org/10.1080/15205436.2023.2186247.

[6] Mourão RR, Brown DK. Black Lives Matter Coverage: How Protest News Frames and Attitudinal Change Affect Social Media Engagement. Digital Journalism. 2022;10(4):626–646. https://doi.org/10.1080/21670811.2021.1931900.

[7] Chan JM, Lee CC. The Journalistic Paradigm on Civil Protests: A Case Study of Hong Kong. In: Arno A, Dissanayake W, editors. The News Media in National and International Conflict. Routledge; 1984. p. 183–202.

[8] McLeod DM, Hertog JK. Social Control, Social Change and the Mass Media's Role in the Regulation of Protest Groups. In: Demers D, Viswanath K, editors. Mass Media, Social Control, and Social Change: A Macrosocial Perspective. Iowa State University Press; 1999. p. 305–332.

[9] Harlow S, Salaverría R, Kilgo DK, García-Perdomo V. Protest Paradigm in Multimedia: Social Media Sharing of Coverage about the Crime of Ayotzinapa, Mexico. Journal of Communication. 2017;67(3):328–349. https://doi.org/10.1111/jcom.12296.

[10] Harlow S, Brown DK, Salaverría R, García-Perdomo V. Is the Whole World Watching? Building a Typology of Protest Coverage on Social Media From Around the World. Journalism Studies. 2020 Aug;21(11):1590–1608. https://doi.org/10.1080/1461670X.2020.1776144.

[11] Harlow S, Brown DK. Protest News and Facebook Engagement: How the Hierarchy of Social Struggle Is Rebuilt on Social Media. Journalism & Mass Communication Quarterly. 2021;98(3):665–691. https://doi.org/10.1177/10776990211017243.

[12] Lerman K, Ghosh R. Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 4; 2010. p. 90–97.

[13] PewResearchCenter.: Social Media and News Fact Sheet. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

[14] Corrigall-Brown C. The Power of Pictures: Images of Politics and Protest. American Behavioral Scientist. 2012 Feb;56(2):131–134. https://doi.org/10.1177/0002764211419358.

[15] Cowart HS, Saunders LM, Blackstone GE. Picture a Protest: Analyzing Media Images Tweeted From Ferguson. Social Media + Society. 2016;2(4):2056305116674029. https://doi.org/10.1177/2056305116674029.

[16] Liu S, Guo L, Mays K, Betke M, Wijaya DT. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding US Gun Violence. In: Bansal M, Villavicencio A, editors. Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Hong Kong, China; 2019. p. 504–514.

[17] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies. Minneapolis: Association for Computational Linguistics; 2019. p. 4171–4186.

[18] Jumle V, Makhortykh M, Sydorova M, Vziatysheva V. Finding Frames with BERT: A Transformer-Based Approach to Generic News Frame Detection. Social Science Computer Review. 2025;p. 08944393251338396. https://doi.org/10.1177/08944393251338396.

[19] Jiang Y, Jin X, Deng Q. Short Video Uprising: How #BlackLivesMatter Content on TikTok Challenges the Protest Paradigm. In: Proceedings of the 16th International AAAI Conference on Web and Social Media. Atlanta, Georgia, United States; 2022. .

[20] Joo J, Steinert-Threlkeld ZC. Image as Data: Automated Content Analysis for Visual Presentations of Political Actors and Events. Computational Communication Research. 2022;4(1):11–67. https://doi.org/10.5117/CCR2022.1.001.JOO.

[21] Kim M, Bas O. Seeing the Black Lives Matter Movement Through Computer Vision? An Automated Visual Analysis of News Media Images on Facebook. Social Media + Society. 2023;9(3):20563051231195582. https://doi.org/10.1177/20563051231195582.

[22] Tourni IC, Guo L, Hu H, Halim E, Ishwar P, Daryanto T, et al. Detecting Frames in News Headlines and Lead Images in U.S. Gun Violence Coverage. In: Findings of the Association for Computational Linguistics: EMNLP 2021; 2021. p. 4037–4050.

[23] Gan YS, Hayes S, Whitmarsh L. The Greta Effect: Is There More Public Support for Climate Protesters Who Are Young and Female? Environmental Science & Policy. 2024;162:103924. https://doi.org/10.1016/j.envsci.2024.103924.

[24] Alonso Del Barrio D, Gatica-Perez D. Framing the News: From Human Perception to Large Language Model Inferences. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. Thessaloniki Greece: ACM; 2023. p. 627–635.

[25] Pastorino V, Sivakumar JA, Moosavi NS.: Decoding News Narratives: A Critical Analysis of Large Language Models in Framing Detection. arXiv.

[26] Zhou K, Yang J, Loy CC, Liu Z. Learning to Prompt for Vision-Language Models. International Journal of Computer Vision. 2022 Sep;130(9):2337–2348. https://doi.org/10.1007/s11263-022-01653-1.

[27] Entman RM. Framing: Toward Clarification of a Fractured Paradigm. Journal of Communication. 1993;43(4):51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x.

[28] Gitlin T. The Whole World Is Watching: Mass Media in the Making and Unmaking of the New Left. University of California Press; 2003.

[29] Brown DK, Harlow S, García-Perdomo V, Salaverría R. From #Ferguson to #Ayotzinapa: Analyzing Differences in Domestic and Foreign Protest News Shared on Social Media. Mass Communication and Society. 2018;21(5):606–630. https://doi.org/10.1080/15205436.2018.1469773.

[30] Brown DK. Police Violence and Protests: Digital Media Maintenance of Racism, Protest Repression, and the Status Quo. Journal of Broadcasting & Electronic Media. 2021 Jan;65(1):157–176. https://doi.org/10.1080/08838151.2021.1897821.

[31] Gil-Lopez T. Mainstream Protest Reporting in the Contemporary Media Environment: Exploring (In)Stability and Adherence to Protest Paradigm From 1998 to 2017. Journalism & Mass Communication Quarterly. 2021 Sep;98(3):692–724. https://doi.org/10.1177/1077699020984783.

[32] Gruber JB. Troublemakers in the Streets? A Framing Analysis of Newspaper Coverage of Protests in the UK 1992–2017. The International Journal of Press/Politics. 2023 Apr;28(2):414–433. https://doi.org/10.1177/

19401612221102058.

[33] Harlow S, Brown DK. Perceptions versus Performance: How Routines, Norms and Values Influence Journalists' Protest Coverage Decisions. Journalism. 2022 Feb;23(2):372–390. https://doi.org/10.1177/1464884920983058.

[34] Mosallaei A. Protest Paradigm Revisited: Is Depicting Protestors' (Counter)Violence Really Bad? Digital Journalism. 2024;13(1):159–178. https://doi.org/10.1080/21670811.2024.2329651.

[35] Jiménez-Martínez C. The Instrumental Mediated Visibility of Violence: The 2013 Protests in Brazil and the Limitations of the Protest Paradigm. The International Journal of Press/Politics. 2021 Jul;26(3):525–546. https://doi.org/10.1177/1940161220923984.

[36] Brown DK, Mourao RR, Sylvie G. Martin to Brown: How Time and Platform Impact Coverage of the Black Lives Matter Movement. Journalism Practice. 2019 Apr;13(4):413–430. https://doi.org/10.1080/17512786.2018.1507680.

[37] Brown DK. Media Models for Nonviolence: Instagram Representations of The# WomensMarch Mass Mobilization News and Audience Engagement. International Journal of Communication. 2022;16:19.

[38] Marwick AE. Instafame: Luxury Selfies in the Attention Economy. Public Culture. 2015;27(1):137–160. https://doi.org/10.1215/08992363-2798379.

[39] García-Perdomo V, Magaña MI, Hernández-Rodríguez JC, Ventín-Sánchez JA. Engaging Social Media Audiences with Riots: TV and Newspapers' Coverage of the 2019 Protests in Colombia and Chile. International Communication Gazette. 2024 Dec;86(8):607–632. https://doi.org/10.1177/17480485231206363.

[40] Henn T, Posegga O. Attention-Grabbing News Coverage: Violent Images of the Black Lives Matter Movement and How They Attract User Attention on Reddit. PLoS one. 2023;18(8):e0288962. https://doi.org/https://doi.org/10.1371/journal.pone.0288962.

[41] Rossi L, Neumayer C, Henrichsen J, Beck LK. Measuring Violence: A Computational Analysis of Violence and Propagation of Image Tweets From Political Protest. Social Science Computer Review. 2022 Jan;41(3):905–925. https://doi.org/10.1177/08944393211055429.

[42] Newman N, Fletcher R, Schulz A, Andi S, Robertson CT, Nielsen RK. Reuters Institute Digital News Report 2021; 2021.

[43] Newman N, Fletcher R, Robertson CT, Eddy K, Nielsen RK. Reuters Institute Digital News Reports 2022; 2022.

[44] Newman N, Fletcher R, Eddy K, Robertson CT, Nielsen RK. Reuters Institute Digital News Reports 2023; 2023.

[45] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models from Natural Language Supervision. In: International Conference on Machine Learning. PmLR; 2021. p. 8748–8763.

[46] Yoon Y, Yoon S, Park K. Assessing News Thumbnail Representativeness: Counterfactual Text Can Enhance the Cross-Modal Matching Ability. In: Ku LW, Martins A, Srikumar V, editors. Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 9009–9024.

[47] Ilharco G, Wortsman M, Wightman R, Gordon C, Carlini N, Taori R, et al.: OpenCLIP.

[48] Zhang B, Zhang P, Dong X, Zang Y, Wang J. Long-CLIP: Unlocking the Long-Text Capability of CLIP. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer Vision – ECCV 2024. vol. 15109. Cham: Springer Nature Switzerland; 2025. p. 310–325.

[49] OpenAI.: GPT-4.1. https://platform.openai.com.

[50] Zhang R, Zhang W, Fang R, Gao P, Li K, Dai J, et al. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Proceedings of Computer Vision – ECCV 2022. vol. 13695. Cham: Springer Nature Switzerland; 2022. p. 493–510.

[51] Loukas L, Stogiannidis I, Diamantopoulos O, Malakasiotis P, Vassos S. Making LLMs Worth Every Penny: Resource-Limited Text Classification in Banking. In: 4th ACM International Conference on AI in Finance. Brooklyn NY USA: ACM; 2023. p. 392–400.

[52] Mavridis P, Inel O, Wilcke X, Makhortykh M, de Jong M, Mazepus H, et al. Framing Is Mightier than the Sword: Detection of Episodic and Thematic Framing in News Media. Human Computation. 2024;11(1):1–28. https://doi.org/10.15346/hc.v11i1.132.

[53] Rucht D, Hocke P, Ohlemacher T. Dokumentation Und Analyse von Protestereignissen in Der Bundesrepublik Deutschland (Prodat) Codebuch. Berlin: Wissenschaftszentrum Berlin; 1992.

[54] Shuman E, Goldenberg A, Saguy T, Halperin E, van Zomeren M. When Are Social Protests Effective? Trends in Cognitive Sciences. 2024;28(3):252–263. https://doi.org/10.1016/j.tics.2023.10.003.

Appendix A.6: Paper VI

Bibliographic data	Cordero, J.M., Henn, T., Holtel, F., Sánchez Gómez, J.Á., Arenas D., Šipka, A., Vollmer, S. 2025. Developing Effective and Value-Aligned AI Tools for Journalists: 12 Critical Questions to Reflect upon. Journalism Practice, 1-20.
Abstract	Artificial intelligence (AI) is playing an increasingly prominent role in journalism, with emerging tools like large language models (LLMs) and other AI applications gaining traction. This paper explores the current landscape in adopting AI tools for the information gathering and story development phase in journalistic practice. We identify two contemporary challenges for implementing AI tools in journalism: effectiveness and value alignment. While the first challenge includes the issues of generalizability, costs, user-friendliness, and accuracy, the latter refers to adherence to journalistic values, independence from external actors, transparency, and data privacy issues. We reflect upon these challenges based on the development process of prototyping an open-access AI tool for classifying documents and prioritizing information-rich pages in collaboration with the Organized Crime and Corruption Reporting Project (OCCRP), a globally recognized investigative journalism network. Based on the literature and our experience, we derive twelve critical questions that we deem important to guide the development of future AI tools for journalism.
DOI	10.1080/17512786.2025.2465894

JOURNALISM PRACTICE
https://doi.org/10.1080/17512786.2025.2465894

Developing Effective and Value-Aligned AI Tools for Journalists: 12 Critical Questions to Reflect upon

José Miguel Cordero[a†], Theresa Henn[b†], Frederik Holtel[c†], José Á. Sánchez Gómez[d†], Diego Arenas[e], Andrea Šipka[e,f] and Sebastian Vollmer[e,f]

[a]Department of Computer Science, University of Chile, Santiago, Chile; [b]Department of Information Systems and Social Networks, University of Bamberg, Bamberg, Germany; [c]Department of Economics, Rhineland-Palatinate Technical University at Kaiserslautern-Landau, Kaiserslautern, Germany; [d]Department of Statistics, University of California at Riverside, Riverside, CA, USA; [e]Department for Data Science and its Applications, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany; [f]Department of Computer Science, Rhineland-Palatinate Technical University at Kaiserslautern-Landau, Kaiserslautern, Germany

ABSTRACT

Artificial intelligence (AI) is playing an increasingly prominent role in journalism, with emerging tools like large language models (LLMs) and other AI applications gaining traction. This paper explores the current landscape in adopting AI tools for the information gathering and story development phase in journalistic practice. We identify two contemporary challenges for implementing AI tools in journalism: effectiveness and value alignment. While the first challenge includes the issues of generalizability, costs, user-friendliness, and accuracy, the latter refers to adherence to journalistic values, independence from external actors, transparency, and data privacy issues. We reflect upon these challenges based on the development process of prototyping an open-access AI tool for classifying documents and prioritizing information-rich pages in collaboration with the Organized Crime and Corruption Reporting Project (OCCRP), a globally recognized investigative journalism network. Based on the literature and our experience, we derive twelve critical questions that we deem important to guide the development of future AI tools for journalism.

ARTICLE HISTORY
Received 13 May 2024
Accepted 2 February 2025

KEYWORDS
Journalistic AI; generative AI; investigative journalism; computational journalism; document classification; machine learning

Introduction

Artificial intelligence (AI) tools have undeniably transformed the work of journalists (e.g., de Haan et al. 2022; Helberger et al. 2022; Stray 2019), leading to a "structural transformation of making news and engaging with the audience" (Helberger et al. 2022, 1606). In recent years, the field has been profoundly reshaped in particular due to the widespread public use of large language models (LLM) such as OpenAI's ChatGPT and Meta's LLaMA (Pavlik 2023; Shi and Sun 2024). These generative AI models enable journalists to

CONTACT Theresa Henn theresa.henn@uni-bamberg.de
†These authors contributed equally to this work.

automate tasks related to content creation in unprecedented ways (Nishal and Diakopoulos 2024). Already before LLMs, many non-generative models had been developed, aiding journalists in addressing challenges such as linking records, automating classification tasks, and extracting meaningful information from heterogeneous documents (Quinonez and Meij 2024; Stray 2019). In general, AI tools are quickly increasing their presence in the journalistic practice, realizing their potential to "disrupt the media industry" (de-Lima-Santos, Yeung and Dodds 2024, 1).

Despite the significant attention that AI models have received in the past years, many journalists remain skeptical, and several journalistic institutions are hesitant to adopt these models (Cools and Diakopoulos 2024; Nishal and Diakopoulos 2024; van Dalen 2024), leading to limited implementation, especially in countries outside of Europe and the US (Gondwe 2023; Pinto and Barbosa 2024). While these tools have been available for several years, research has identified multiple reasons that hinder or complicate their successful implementation in journalism (e.g., de Haan et al. 2022; Stray 2019).[1]

We argue in this article that the issues associated with the adoption of generative and non-generative AI tools can be summarized in two conflicting challenges. The first challenge relates to the *effectiveness* of AI tools for journalists. Research indicates that the widespread adoption of AI tools is often constrained by their limited generalizability, which hampers their ability to support the creation of diverse and in-depth stories (Stray 2019). Additional obstacles include high costs, particularly for small news organizations (Borchardt 2022; Stray 2019), and the lack of user-friendly interfaces, which may require specialized computational expertise and thus limit accessibility for many journalists (Borchardt 2022; Stray 2019). Although recent publicly available generative AI models have overcome many of these challenges, they still experience accuracy problems. In particular, these models are prone to hallucinations (Opdahl et al. 2023; Quinonez and Meij 2024; Shi and Sun 2024), which in turn can contribute to the spread of misinformation.

The second challenge in adopting AI tools in journalism pertains to *value alignment*, encompassing concerns related to the values embedded in the tools, as well as issues of independence, transparency, and privacy (de Haan et al. 2022; Gutierrez Lopez et al. 2023; Lin and Lewis 2022; Opdahl et al. 2023; Stark and Diakopoulos 2016). In this context, tool ownership is a critical issue, as many AI tools today are owned by "outside elite actors, such as tech companies" (de Haan et al. 2022, 1776). Often, these organizations do not disclose the tools' internal details and might design them primarily to maximize engagement and revenue rather than aligning them with journalistic values (Borchardt 2022; de Haan et al. 2022; Diakopoulos 2019; Pinto and Barbosa 2024; Simon 2022). Furthermore, privacy and confidentiality have become especially prevalent concerns in the case of generative AI tools as these models can expose classified and sensitive data (Carlini et al. 2021; de-Lima-Santos, Yeung, and Dodds 2024).

Consequently, contemporary journalism faces the challenge of developing AI tools that are not only effective by addressing the issues related to the first challenge but also value-aligned with regard to the second challenge. Following Helberger et al. (2020, 1607), who proclaimed that "[t]he technological change AI brings may be inevitable; how AI reshapes journalism is not," we aim to contribute to this line of research by asking: *What are critical questions to consider when developing AI tools that address the needs and values of the journalistic community?*

To address this question, we provide a detailed exploration of the two challenges touched upon above. This exploration is not only informed by reviewing the literature, but also by our experience of developing an AI tool for journalists. More concretely, we prototype an exemplary, open-access, and easy-to-use AI tool that aids journalists in automatically classifying and prioritizing pages in large document databases. To develop our tool, we collaborated with the Organized Crime and Corruption Reporting Project (OCCRP), a worldwide investigative reporting network of independent media centers and journalists in over 40 countries and one of the most important players in international investigative journalism. We use the insights obtained during the development process to derive twelve critical questions that can guide the development of effective and value-aligned AI tools for journalists in the future.

Our work contributes to the literature on the role of AI tools in journalism. First, we consolidate the contemporary discussion of challenges in the implementation of AI in journalism which previous literature has identified (see also Simon 2023; Stray 2019). Second, we advance the discussion by exploring the two challenges drawing from the practical experience of developing an AI tool prototype. We thereby follow a practice-based research methodology, which "emphasizes the study of real-world problems and practices, rather than solely theoretical or abstract concepts" (Fridman, Krøvel, and Palumbo 2023, 4). Our findings offer practical assistance for academics and practitioners alike who seek to create effective AI tools and align the developed product with journalistic values. We view our approach as addressing concerns that the "current discussion on [AI] guidelines … lacks a crucial element: how these principles translate into action for news companies" (de-Lima-Santos, Yeung, and Dodds 2024, 16). Finally, we also reflect on the role that institutional factors can play in the development of such tools for the journalistic community.

This paper is structured as follows: In our theoretical background, we provide an overview of the status of AI tools within journalism by elaborating on both challenges mentioned above. Our methodology section provides a description of the development of our prototype's application to OCCRP's data platform Aleph and its performance results. In our discussion section, we explore how our prototype relates to the debate concerning effective and value-aligned AI tools and derive twelve critical questions for future AI tool development for the journalistic community. Finally, in our conclusion, we summarize our results, highlight the limitations of our approach, and outline potential future research on effective and value-aligned AI tool integration for the journalistic community.

Theoretical Background

The literature (Lin and Lewis 2022) distinguishes three stages where AI enters journalistic practice: (1) information gathering, (2) selection and production, and (3) distribution and consumption. In our study, we focus on AI tools in the first and second stage. However, note that AI tools have become an integral part of each stage of the journalistic process (see, e.g., de Haan et al. 2022; Opdahl et al. 2023; Shi and Sun 2024). In the following sections, we will delve deeper into the issues related to effectiveness and value alignment, drawing on insights from the literature.

Effective AI in Journalism

We identify four issues related to effectiveness that impede and complicate the successful adoption of AI tools in journalism.

The first issue concerns the *generalizability* of AI tools. Research indicates that journalism tasks are often tailored to the specific story being reported on (Stray 2019), resulting in the development of story-specific AI tools (e.g., Poston, Rubin, and Pesce 2015). While these tools may be effective for particular cases, they inherently lack broader applicability and necessitate the collection and curation of new datasets for each story. Conversely, tools that are general enough to be applicable beyond a single story but specific to the problems of journalism (such as fact-checking or document analysis tools) are much less adopted (de Haan et al. 2022, 1783).

Second, developing and implementing AI tools often involves high *costs*. This issue includes various sub-aspects. First, AI development expertise is costly to obtain in the labor market (Broussard et al. 2019; Hamilton 2016), as newsrooms compete with tech companies and other well-funded major players for hiring and retaining skilled talent (Simon 2022, 1838). Illustratively, by May of 2021, the average yearly U.S. salary for news analysts, reporters, and journalists was $63,230, while the average software developer salary was nearly twice as much, at $120,990 (Bureau of Labor Statistics 2021). Second, other inputs for training sophisticated AI tools, such as computational resources and data, are expensive (Simon 2022, 1838). For example, it took investigative reporter Julia Angwin and her team at ProPublica over a year to collect the data required to publish the story "Machine Bias" (Angwin et al. 2016). Third, given that "[t]he financial outlook for most of the news industry internationally remains difficult" (Simon 2022, 1838), especially for local newsrooms (Borchardt 2022), integrating AI tools specifically developed for the journalistic practice will likely continue to be challenging.

Third, AI tools specifically designed for journalists often lack *user-friendly* design because they are frequently developed in-house by newsrooms. Unlike major tech companies, which can leverage vast user data to optimize their platforms through methods like A/B testing (see, e.g., Lin and Lewis 2022, 1636), tools for newsrooms and journalists typically rely on much smaller user bases, which limits their ability to achieve similar levels of optimization. This smaller user base and the overall lower availability of resources for user interface optimization introduce friction for the user, as they impair the user experience, making adoption less likely. Several studies recognize that journalists are more inclined to use user-friendly tools, such as social media and online search engines, than tools designed explicitly for the journalistic research process (de Haan et al. 2022).

Recently, generative AI tools such as ChatGPT have become publicly available online. These new models are highly successful in mitigating issues related to effectiveness, as they are cost-effective (Castillo-Campos, Varona-Aramburu, and Becerra-Alonso 2024), designed with user-friendly interfaces (Pinto and Barbosa 2024), and versatile across a range of text generation tasks (Shi and Sun 2024). However, these LLMs have problems regarding a fourth issue which we call *accuracy*. As it is well known, generative AI models like ChatGPT are prone to hallucinations, i.e., often produce answers that contain outright false or misleading information (Breazu and Katsos 2024; Opdahl et al. 2023; Shi and Sun 2024). While accuracy is a concern for all AI models, hallucinations are particularly problematic in journalism, as false information can have serious consequences if not identified before publication (Nishal and Diakopoulos 2024).

Hallucination issues are among the top reasons for newsrooms' reluctance to adopt generative AI models. A 2023 study in Brazil, for example, has found that no newsroom in the country had formally adopted ChatGPT (Pinto and Barbosa 2024). Due to the challenges, Bloomberg has reverted to developing and training their own AI models in-house (Quinonez and Meij 2024) to minimize hallucination errors by utilizing more tailored and specific training data. While large news organizations like Bloomberg can afford to develop in-house AI models, smaller newsrooms often lack the resources and capacity to undertake such measures (Borchardt 2022; Shi and Sun 2024, 584; Simon 2022, 1838). This demonstrates that despite the benefits of LLMs, these models can fall short, and vast amounts of resources are still required to adapt these technologies to satisfy the specific needs of journalistic organizations.

Value-aligned AI in Journalism

We identify four issues pertaining to journalistic values, collectively termed value alignment, which impede and complicate the successful adoption of AI tools in journalism.

First, a variety of journalism scholars agree that *journalistic values* such as reporting accuracy, relevance, and diversity (Lin and Lewis 2022) "serve as the ethical backbone of newsrooms" (de-Lima-Santos, Yeung, and Dodds 2024, 4). As an endeavor grounded in these values, journalism is frequently described as the "fourth estate" within democracies (Helberger et al. 2020). In this context, it is crucial to recognize that AI systems are not value-neutral, as they inherently embody their designers' and developers' value judgments, biases, and perspectives. If news organizations and journalists do not incorporate their ideas and values into AI tools, the choices in AI design will be made by other stakeholders like big tech companies (Broussard et al. 2019; Simon 2023). However, the values and interests of these actors might not align with those of journalists (de Haan et al. 2022). For instance, big tech companies tend to prioritize increasing user engagement over other goals by emphasizing personalization and virality.

Moreover, research indicates that these journalistic values can often be compromised in the context of AI models due to biased training data. This is a particular issue in the context of LLMs like ChatGPT, as these models are trained on vast datasets, particularly web-based data, which is inherently biased. Consequently, they tend to reinforce existing stereotypes, perpetuate prejudices, and further marginalize underrepresented groups and regions of the world (Breazu and Katsos 2024; Gondwe 2023; Opdahl et al. 2023; Shi and Sun 2024). Since the output of generative AI models resembles human-written text, biases in the underlying data can have serious implications, as detecting these biases can be challenging (Castillo-Campos, Varona-Aramburu, and Becerra-Alonso 2024). In a journalistic context, the spread of such biases and misinformation is particularly concerning (Gondwe 2023). Thus, to overcome this issue and incorporate journalistic values into AI tool development, Diakopoulos (2019, 181) suggests that news organizations should consider "designing and building the technology themselves so that it better aligns their values with the implementation of the technology."

Second and relatedly, tool ownership and *independence* are crucial issues related to the adoption of AI tools in journalism. The literature has pointed out the importance of developing AI tools in journalism that are independent of major tech corporations such as Google, Meta, and OpenAI (e.g., de Haan et al. 2022; de-Lima-Santos, Yeung, and Dodds 2024; Simon 2023). On the one hand, journalists might have unique operational needs, necessitating specialized software tailored to their specific domain. Generalist "one-size-fits-all" solutions, like widely applicable search engines (de Haan et al. 2022; Nishal and Diakopoulos 2024), could be inadequate for journalists' more specific needs. On the other hand, when tools are not owned by journalistic organizations, changes in the models and tools do not have to be reviewed and approved by journalists, effectively leaving them out of control (Simon 2023). For instance, tech corporations often modify their algorithms, as exemplified by the constant changes in Google's PageRank algorithm or modifications of social media feed algorithms. When such alterations occur, they inevitably impact the work of journalists, yet their ability to influence or mitigate these effects is limited (Simon 2022, 1843–1844).

This issue has implications for journalism's role in democracy. If major journalistic non-profit organizations and newsrooms are not able to avoid "infrastructure capture" (Simon 2022, 1835) by actors such as big tech corporations, the integrity of quality journalism could be at stake (Helberger et al. 2022). Thus, by increasingly relying on AI tools that are owned by big tech companies, such as search engines and generative AI models like ChatGPT,

> news organizations risk becoming even more tethered to platform companies in the long-run, potentially limiting their autonomy and, by extension, contributing to a restructuring of the public arena as news organizations are re-shaped according to the logics of platform businesses. (Simon 2022, 1833)

One response to this issue is to foster "effective collaboration between journalists and AI specialists" (de-Lima-Santos, Yeung, and Dodds 2024, 15). In particular, collaboration of newsrooms and academic researchers can provide major benefits, specially to local organizations (see Pinto and Barbosa 2024; Rinehart and Kung 2023).

Third, the literature raises the concern that AI tools are often "black boxes" that lack *transparency* (de-Lima-Santos, Yeung, and Dodds 2024), impeding a comprehensive understanding of their internal processes (de Haan et al. 2022; Diakopoulos, Zhang, and Salway 2013). This issue is inherent to various machine learning techniques, such as deep learning, and is particularly pronounced for LLMs like ChatGPT, which consist of billions of parameters. This enormous size makes it extremely challenging to gain a thorough understanding of how such models arrive at specific outputs or decisions (Simon 2022). In addition to this lack of technical transparency, much of the code and algorithms utilized by prominent tech companies are proprietary and closely guarded secrets inaccessible to external scrutiny. This lack of transparency impedes external parties from evaluating thoroughly whether biases and errors occur, potentially biasing the journalistic research process for writing stories (de Haan et al. 2022). Only when source code is accessible and organizations can independently inspect an AI tool's internal architecture is it possible to effectively assess whether it aligns with journalistic values.

The fourth issue centers around *data privacy* and confidentiality. For example, the data or queries shared with online AI tools are often preserved and exploited by the companies that own them (de-Lima-Santos, Yeung, and Dodds 2024), in many cases with the purpose

Table 1. Overview of the challenges we identified.

Challenge	Issues
Effectiveness	Generalizability
	Costs
	User-friendliness
	Accuracy
Value alignment	Journalistic values
	Dependency
	Transparency
	Data privacy

of model retraining. While the preservation of this data may be problematic in general, it is especially concerning when the shared information is sensitive, which is frequently the case in the journalistic context. While all online services pose risks in this regard, generative AI tools are especially vulnerable to this type of privacy concern. Recent work has identified the risk that generative AI can memorize training data and share it with users with the right choice of prompts (Carlini et al. 2021). Thus, journalistic AI tools must have a clear guideline on which data may be preserved for retraining, as well as mechanisms to ensure that the privacy of journalists and the confidentiality of their data are safeguarded in cases where sensitive information is involved.

In summary, journalists face two main challenges when adopting AI tools in their practice (see Table 1). Regarding effectiveness, the development of AI tools can be costly, and the resulting product can often lack features such as user-friendliness and general applicability, which impedes their successful adoption. While some recent generative AI models overcome these problems, they introduce new issues, such as accuracy problems due to model hallucinations. Thus, no AI tool fully embodies the ideal features of effectiveness. Second, AI tools face challenges regarding their alignment with journalistic values. These issues are especially pronounced when these AI tools are owned by large tech companies, which complicates the integration of journalistic values into these models. Generative AI models face particular issues related to value alignment, as they are especially affected by issues related to data privacy and confidentiality.

Methodology

Problem Statement and Research Context

To address the two challenges identified within the literature, we discuss the prototyping of an AI tool for the automatic classification and prioritization of document pages from large databases. We developed our prototype in cooperation with OCCRP, one of the key organizations in the global field of investigative journalism. Since 2011, through OCCRP's investigative work of unearthing evidence of crime and corruption, 7.5 billion dollars of assets have been seized, and more than 380 official investigations were initiated, leading to 672 government actions and more than 580 arrests (OCCRP 2023). We created our prototype to be integrated into their investigative data platform Aleph, which depicts an archetype of investigative journalism data platforms.

Developing the AI tool prototype presented in this paper is motivated by the following data problem experienced by OCCRP. Their platform, Aleph, provides investigative journalists across the globe access to over 350 million documents of leaked, scraped, and otherwise collected data (equaling roughly 37 terabytes). Documents come in various file extensions (like pdf, jpeg, or tiff), are of varying page lengths, originate from more than 140 countries, and are thus multilingual (Aleph 2023). Since the archive of documents hosted by Aleph is updated by adding large databases of documents at once (e.g., data leaks), it is currently impossible to systematically capture the document type of all documents in Aleph. Furthermore, due to the heterogeneity and size of the archive, Aleph hosts an unknown variety of document types. Consequently, journalists must manually explore large document databases to extract the valuable document types they need for developing stories, which can be highly time-consuming. This paper's prototype tackles this issue by classifying documents according to their information value and document type, laying the foundation for working with these documents more efficiently. Thus, due to its document diversity, the unknown number of document types, and the large amount of data, Aleph represents a suitable data platform for training and applying our AI tool prototype.

Also, note that given OCCRP's use of highly sensitive data, integrating AI tools from external providers like Google or OpenAI into their pipeline raises confidentiality and privacy concerns. Thus, collaborating with researchers to develop AI tools depicts an approach that overcomes this issue, underscoring the value of partnerships between journalistic organizations and researchers, as highlighted in previous studies (de-Lima-Santos, Yeung, and Dodds 2024; Pinto and Barbosa 2024).

Automatized Document Classification with AI Tool Prototype

This subsection provides an overview of how our AI tool prototype (1) automatically distinguishes between document types and (2) prioritizes between information-rich and information-poor pages. A visual summary of the different process steps within our AI tool is depicted in Figure 1. Note that the AI tool prototype's code is freely accessible at the following GitHub repository: https://github.com/DSSGxDFKI/occrp-document-classifier?tab=readme-ov-file.

For our AI tool prototype, we chose an image-based classification approach. Our reasons for this decision are twofold. First, the documents in OCCRP's database are multilingual, yet

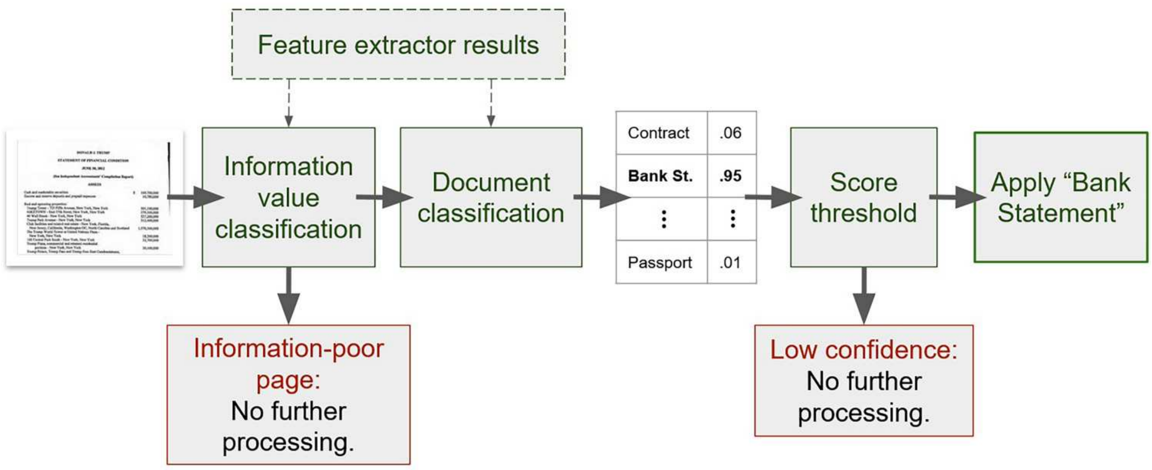


Figure 1. Visualization of the processing steps of our prototype.

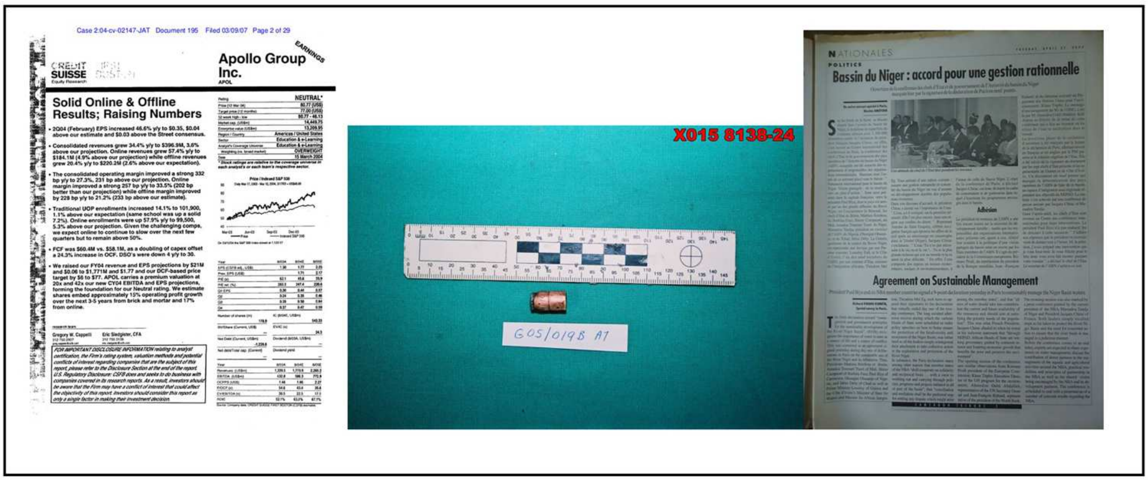


Figure 2. Example of documents hosted on Aleph. (a) An exhibit from a US federal court archive (b) A picture of evidence of an International Criminal Tribunal for the former Yugoslavia hearing (c) A photo of a newspaper article.

training data for this variety of languages is scarce. Hence, a language-based classification approach may only provide reliable classifications for documents in the languages most present in the training data and perform poorly in languages not present. Conversely, the visual structure of documents belonging to a given document type often follows a "pre-defined form or standard set of components" (Shin, Doermann, and Rosenfeld 2001, 232) and is unaffected by language heterogeneity. Thus, image-based classification approaches tend to be more generalizable. Second, the format of the documents in OCCRP's database is highly heterogeneous, as illustrated in Figure 2. This heterogeneity may pose a challenge when extracting the written content of documents.

The first processing step of our AI tool prototype is to preprocess the document data. More specifically, we automatically split the document into individual pages, transform each page into JPG image format, and resize the page image to a consistent pixel resolution of 227 × 227.

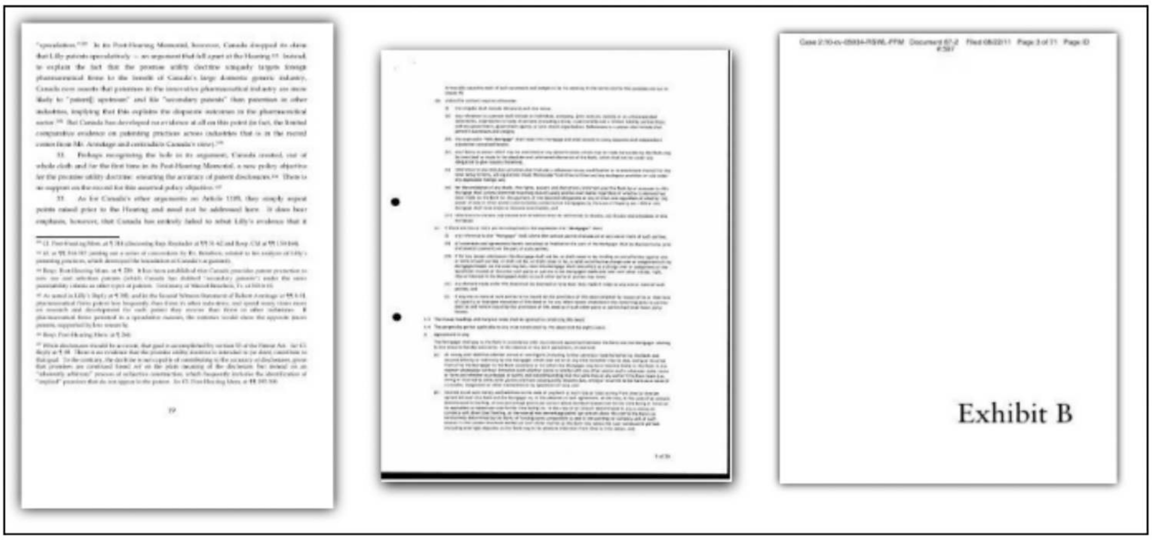


Figure 3. Examples of information-poor pages.

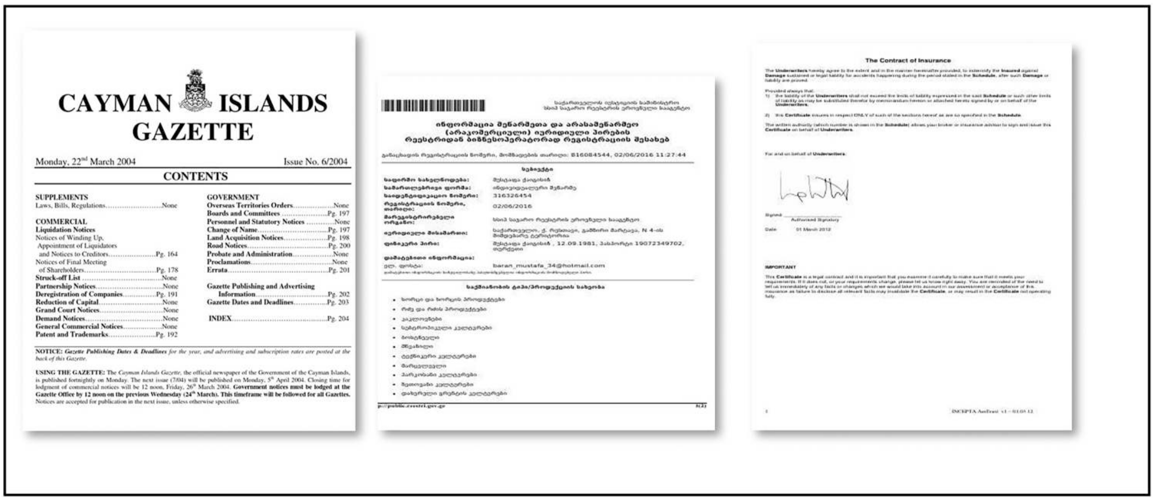


Figure 4. Examples of information-rich pages.

After preprocessing, the second step of our prototype is to identify information-rich pages with an information value classification model. For this classification, we define an information-poor page as a page containing long blocks of text or near-empty pages (see examples of information-poor pages in Figure 3). All other pages, like a person's bank details, are considered information-rich (see examples of information-rich pages in Figure 4). If a document page is identified as information-rich in this step, it will continue to the following classification step. If a page is deemed information-poor, it is omitted in the subsequent steps.

Once information-rich pages are identified, the third step in our AI tool prototype is to classify each page according to its document type. To this end, we train a classification model that classifies nine different document types useful for the journalistic community: bank statements, company registries, contracts, court documents, gazettes, invoices, passports, receipts, and shipping receipts. In this step, each page is assigned nine numeric scores, where each score corresponds to a different document type. The scores represent how well the given page matches each document type, where a higher value for the score of a specific type corresponds to a higher likelihood that the page belongs to this document type.

As previously discussed, the number of document types in the Aleph database is unknown due to its high document heterogeneity. Therefore, it is essential that our classification model recognizes if a document does not belong to one of the predefined classes. This scenario of "dynamic, open environments where some new/test documents may not belong to any of the training classes" (Shu, Xu, and Liu 2017, 2911) is widespread in complex datasets and referred to in the literature as the "open-world assumption." In our last step, we incorporate this consideration in our prototype by contrasting each page's highest score to a threshold value that controls for the quality of our classified document type and can dynamically adjust when adding more document classes. This safety thresholding step ensures that document pages are only classified if there is strong evidence that they are part of one of the document classes the model was trained on. This improves model performance and prevents previously unseen document classes from being wrongfully classified. For more information regarding the control threshold, see Shu, Xu, and Liu (2017).

Data Overview

Data is necessary to train our AI tool prototype and evaluate its performance appropriately. We use a combination of publicly available datasets and newly curated datasets sourced from the Aleph document archive. Through cooperation with OCCRP, we received data from their data platform Aleph corresponding to the nine document types mentioned above.

To improve the quality of our classification results, we performed several cleaning and preprocessing steps on the data provided by OCCRP. First, through several rounds of open coding, we developed a document taxonomy describing the characteristic visual features of each document type. Based on this shared understanding of each document's visual structure, we were able to define and label each document type consistently. We conducted the same process for creating an additional taxonomy to define information-rich and information-poor pages.

Based on our document type taxonomy, we manually selected 100 documents per type from the original data provided by OCCRP. As a result, we curated 900 documents from our nine distinct document types. We refer to this dataset as the "document type dataset." Additionally, we extracted pages from the selected 900 multi-page documents fitting our definition of information-rich and information-poor pages. As a result of this cleaning process, we received 900 information-rich pages (100 pages per document type) and 1233 information-poor pages. We call this dataset the "information value dataset."

Since image-based approaches need large training datasets to generate accurate classification results, we decided to boost our data with image augmentation techniques and pre-training. Image augmentation is a commonly used method that "enhance[s] the size and quality of training datasets" (Shorten and Khoshgoftaar 2019, 1) by, e.g., random zooming, rotating, changing width and height, and modifying the brightness and angle of the images (as an example, see Figure 5).

Another technique to enhance the performance of image-based classification approaches is the pre-training on large publicly available image datasets (Afzal et al. 2017). Pre-training helps an image-based model learn standard image features across contexts and thus aids the model in enhancing its understanding of the image characteristics on which it is trained. This is especially useful when working with limited training data for image classification. We used the widely known Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) dataset[2] for pre-training, previously considered in multiple works (Afzal et al. 2017; Kölsch et al. 2017). The RVL-CDIP dataset contains scans of 400,000 document pages, divided into 16 document types. This dataset of 37 GB is highly curated and available in a consistent image format. Therefore, no further processing steps were necessary to use this data to train our AI tool. An exemplary overview of the different classes present in the RVL-CDIP dataset can be seen in Figure 6.

Model Architecture and Training the AI Tool Prototype

For our AI tool prototype's information value assessment and document classification steps, we use convolutional neural network (CNN) models, widely applied in the literature for the visual classification of images and documents (Afzal et al. 2017). We use the

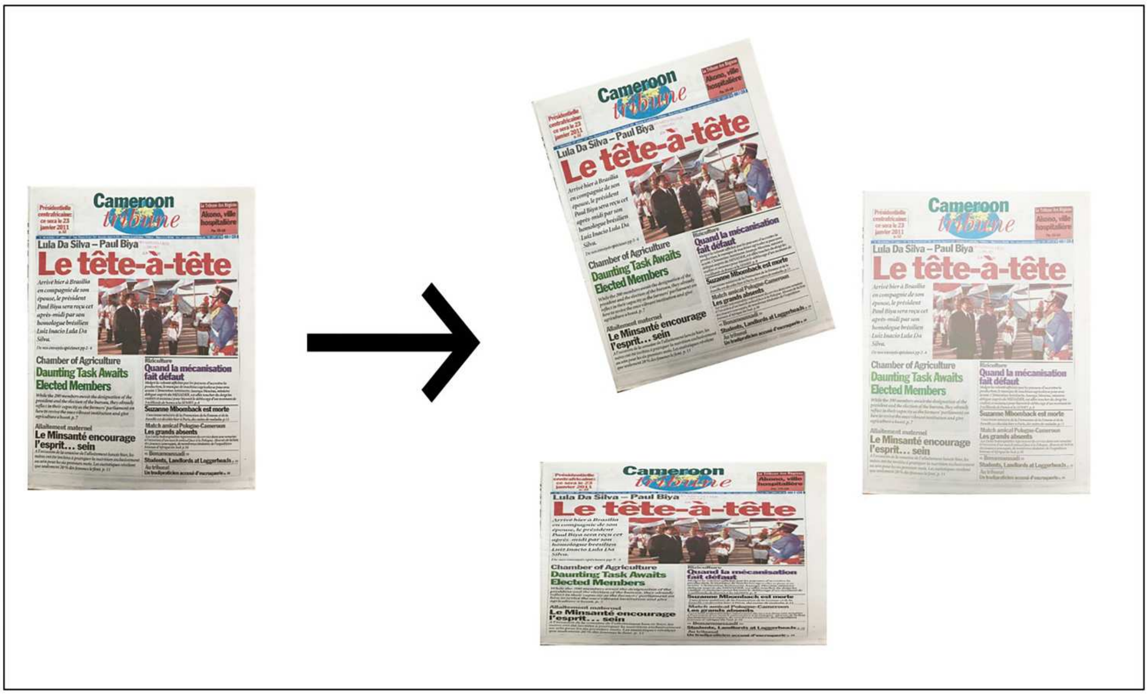


Figure 5. Illustration of three different augmentations.

EfficientNetB4 model architecture, as it performed best among the five architectures we considered.

The first step in our training process is to pre-train the CNN model architecture with the RVL-CDIP dataset, enhancing the prototype's overall performance. The pre-trained models provide the basis for the further training of our two classifiers: the information value classifier and the document type classifier. We retrain the information value classifier on 60% of the observations in the information value dataset and the document type classifier on 60% of the document type dataset. The remaining 40% of each dataset is used for testing and validation. To enhance the variety of images used for training, all document page images used in this training process are augmented as described in the section "Data Overview."

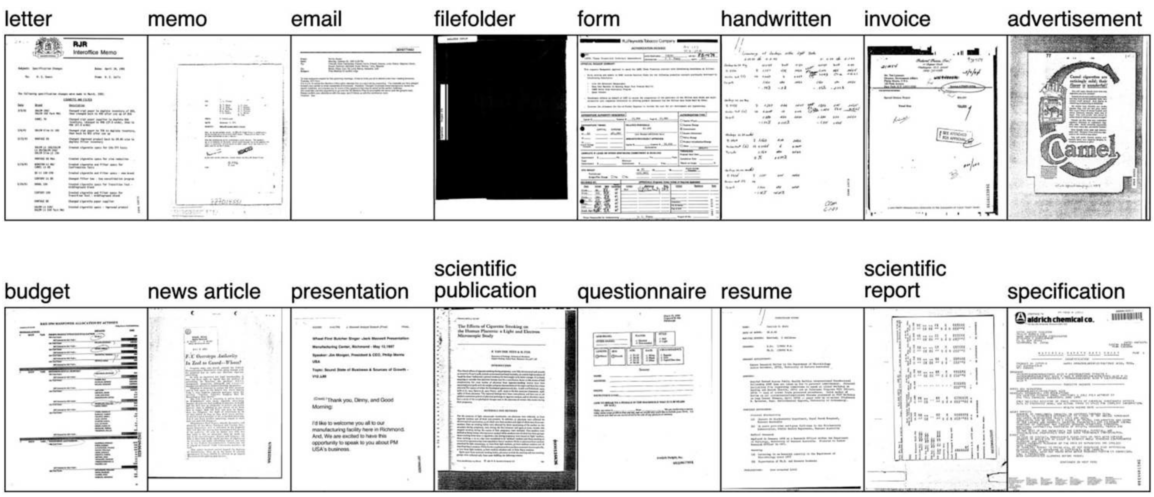


Figure 6. Exemplary overview of RVL-CDIP dataset (Harley, Ufkes, and Derpanis 2023).

Table 2. Accuracy levels of the nine document classes.

Document class	Level of accuracy (in %)
Bank statements	83.33
Contracts	41.37
Company registries	83.33
Court documents	83.33
Gazettes	86.67
Invoices	70.00
Other	100.00
Passports	86.67
Receipts	96.67
Shipping receipts	90.00

Document Classification Results

With the EfficientNetB4 architecture we obtain an overall accuracy of 91.71%. Thus, out of 100 documents, nearly 92 are correctly classified as information-rich or information-poor pages and are subsequently labeled with their correct document class. However, we observe disparate accuracy results depending on the document class (see Table 2). Receipts, gazettes, and passports are classified correctly in nearly all cases, whereas contracts are classified accurately in only about 41 out of 100 cases. One explanation for this observation could be that contracts are quite heterogeneous in their composition since they are utilized in diverse application fields (e.g., financial or human resources contexts). However, since all other classes receive accuracy levels above or equal to 70.00%, and the overall performance is comparable with previous results in this context (e.g., Afzal et al. 2017), our AI tool generates sound results useful for investigative journalists.

Discussion

As demonstrated in the previous section, our prototype can effectively assist journalists in extracting valuable information from large datasets of unlabeled documents. While the previous section focused on the technical development and results of our prototype, the following discussion explores the insights gained from our development experience, particularly concerning the two identified challenges above, thereby addressing our research question.

Creating Effective AI Tools in Journalism

First, when considering the issue of *generalizability* (Stray 2019), our prototype is a valuable tool for writing multiple stories since it currently classifies nine different document types that are important to journalists. As our prototype can be easily extended when trained with new data, it can be updated for new and more general tasks. If retrained on new document data, our AI tool's integrated threshold (Shu, Xu, and Liu 2017) for document label decisions constitutes a significant advantage as it dynamically adjusts to new data, rendering our tool applicable to new contexts.

Second, through our prototype development process, we encountered several strategies that we deem particularly useful for newsrooms with limited financial resources. More specifically, we found that pre-training our prototype with general datasets like the RVL-CDIP dataset can significantly increase its classification results. Thus, we suggest that organizations like OCCRP could use these publicly available and free-of-charge datasets when creating new AI applications based on computer vision. Moreover, we observed that image augmentation is an effective method to easily enlarge small datasets that can then be used to train AI tools, yielding good results. Lastly, we believe it is essential to explore innovative options for data gathering that go beyond augmentation and pre-training. One promising avenue could be a crowdsourced "human-in-the-loop" approach. Organizations like OCCRP could integrate an interactive labeling system into their data platforms that allows users to give feedback on the model's predictions (for similar applications, see, e.g., Nadj et al. 2020). This would allow journalists to take part in the creation of meaningful data and improve the cost-effectiveness of the development process of AI tools.

Third, the level of *user-friendliness* of an AI tool is essential for its adoption by a broad range of journalists and practitioners with varying technical skills (Simon 2023). To address this issue, we advocate that journalistic organizations could use their existing infrastructure and incorporate tools as functionalities of these interfaces. For instance, OCCRP's Aleph is widely used by journalists worldwide as a user-friendly interface to search large databases of documents. Thus, integrating an AI tool like the one presented in this paper into Aleph's functionalities enables seamless usage while leveraging the already present infrastructure.

Fourth, model *accuracy* is a crucial aspect of any AI model. Our prototype is not designed to generate text and, therefore, does not experience hallucinations. Nevertheless, our development process has provided us with valuable insights into enhancing accuracy in other regards. First and foremost, using a well-curated dataset and ongoing model fine-tuning, our prototype has demonstrated impressive performance, achieving an overall accuracy of over 91%. Second, our integrated classification threshold serves as a "technical safeguard" for our prototype, filtering out data that it cannot classify based on its training set. This mechanism helps prevent erroneous results that may occur due to insufficient or inappropriate training data.

Creating Value-aligned AI Tools

First, our prototype addresses the integration of *journalistic values* such as reporting accuracy, relevance, and diversity in several ways. In terms of accuracy, our prototype demonstrates great performance in correctly classifying documents, as outlined above. With respect to relevance, our first-page classification model effectively differentiates between important and non-important information. Regarding diversity, we have created a prototype that can be easily adapted to new document classes and based on its open-source code, can be inspected, modified, and extended to adhere to organization-specific requirements and preferences.

To incorporate these journalistic values into AI tools, we consider it essential to directly engage with journalists (see, e.g., de Haan et al. 2022 for the merit of conducting interviews with journalists)—the end users of the products. We initiated this process in our project by consulting OCCRP-associated journalists who already use Aleph. Our inquiries aimed to gain insights into their typical workflows, understand how the tool integrates into their tasks, identify appreciated features, and discern any functionalities they found lacking. Moreover, regular communication with OCCRP staff, who possess a nuanced understanding of the needs of journalistic end users, provided valuable guidance in shaping the tool's design decisions. Soliciting feedback from journalists in areas like document classification enables the incorporation of their interpretations into subsequent tool output. By keeping journalists in the loop, the responsibility for subjective decisions is shared rather than resting solely with developers, enabling both parties to leverage their expertise effectively. This collaboration could foster a "human-in-the-loop" approach, as previously mentioned, where journalists participate in the data labeling process and are thereby able to address issues related to data bias. Hence, we agree with de-Lima-Santos, Yeung, and Dodds (2024, 15) who state that the "successful adoption of AI technologies in newsrooms hinges on effective collaboration between journalists and AI specialists."

Second, our project uncovered principles related to *independence* (see, e.g., de Haan et al. 2022; Simon 2022, 2023). As a major non-profit journalistic organization, OCCRP is arguably more committed to traditional journalistic values than big tech companies. By developing a prototype for OCCRP, we help to reduce journalists' dependence on these corporations.

Generative AI tools such as ChatGPT are, in many cases, owned by a major tech company, such as OpenAI, and integrated into the products of other big tech firms like Microsoft. Since these tools are not open-source, journalists who rely on them are constrained to accept any alterations in design or availability. As Pinto and Barbosa (2024, 336–337) argue, for journalistic organizations in Global South countries, such developments increase the dependency on technology provided by actors in the Global North. In contrast, our openly available prototype provides flexibility, allowing adoption and modification by any journalistic organization into its data processing pipeline.

Third and relatedly, our prototype exhibits high *transparency* compared to many tools commonly relied upon by journalists and other users. In contrast to proprietary tools, including ChatGPT, our prototype is open-source (see GitHub repository). The provided transparency allows anyone to scrutinize the tool's training process and behavior, enabling the identification of potential biases and errors. Organizations can then customize the tool to align with the specific needs of their user base. Thus, this transparency facilitates error detection and supports the development of future functionalities, rendering the software more generalist (Stray 2019).

Fourth, we addressed *data privacy* concerns during the development of our prototype in two ways. First, as our prototype is open-source, it can be set up on a local machine. Journalistic organizations can, therefore, process sensitive documents without the concern that those would be shared with any external actor. Second, as a discriminative rather than a generative tool, our prototype does not generate new data, ensuring that no information is inadvertently shared during its use. This privacy concern associated with LLMs does, therefore, not apply to our prototype.

Implications for the Role of Organizations

The effectiveness and value alignment of AI tools depend heavily on the institutions that are involved in their development and ownership. While large journalistic organizations like Bloomberg News (Quinonez and Meij 2024) have the resources to develop value-aligned and highly effective AI tools, many small news organizations lack this infrastructure. Therefore, small newsrooms are limited to adopting AI tools owned by large technology companies like OpenAI, Facebook or Google, whose models suffer major shortcomings in their alignment with journalistic values.

With this context in mind, we believe that large journalistic non-profit organizations such as OCCRP, ICIJ, and AP have a unique opportunity to develop and share effective, value-aligned AI tools with the journalistic community. By lending their expertise and collaborating with these non-profit organizations, we consider that academic researchers can play a pivotal role in developing these tools.

Non-profit organizations have great potential for creating effective AI tools. First, these organizations often host vast databases, like OCCRP's Aleph, which can be exploited to develop sufficiently generalist tools that are useful for multiple reporting stories. Second, data from these databases can be exploited cost-effectively for training. As exemplified through our work, even small sets of labeled data can provide accurate modeling results, which can be gradually improved as more data becomes available. Third, some of these organizations presently host platforms with friendly user interfaces and have significant user traction (e.g., Aleph 2023). Major journalistic non-profit platforms can more easily integrate AI tools into their interfaces, allowing their current users to enjoy the benefits of user-friendliness.

Major journalistic non-profit organizations are also exceptionally well suited to create tools aligned with journalists' values. First, these organizations' direct contact with journalists facilitates them obtaining first-hand suggestions from relevant professionals to consider in their AI tool development process. For example, the Associated Press Local News AI initiative (Rinehart and Kung 2023) gathered survey results from more than 200 news professionals in the US and used the provided feedback to design five AI tools for small newsrooms. Second, the independence of non-profit organizations

Table 3. Critical questions for effective and value-aligned AI tool development in journalism.

Challenge	Sub issue	Critical questions
Effective AI	Generalizability	1. Is the AI tool applicable to multiple contexts?
	Costs	2. Are there open-source datasets, AI models, or software provided by, for instance, journalistic organizations that can be used to build or enhance the AI tool?
		3. Can a crowd-sourced approach be applied to enhance the quality and quantity of the data used for training the AI tool?
	User-friendliness	4. Is the AI tool designed in a user-friendly way, making it useful for users with different levels of technical skills?
	Accuracy	5. Does the AI tool consistently produce reliable and accurate results in line with journalistic standards?
		6. Are there technical or human safeguards in place for the AI tool that minimize erroneous results, and, to the extent possible, mitigate their impact?
Value-aligned AI	Journalistic values	7. Does the AI tool reflect journalistic values such as reporting accuracy, relevance, and diversity?
		8. Were journalists involved and consulted in the development process?
	Independence	9. To what extent do journalists and newsrooms own the AI tool they use (data, code, infrastructure)?
		10. How are journalists affected by changes to components that they do not own?
	Transparency	11. Is the AI tool's source code publicly available and well-documented, making it easy to understand and open to scrutiny?
	Data Privacy	12. Have precautions been taken to prevent or mitigate the risk of the model sharing sensitive information from users with actors external to the journalistic organization?

allows them to align their tools with the needs of journalists, avoiding conflicts of incentives commonly faced by other actors such as tech companies.

While non-profit organizations possess data, infrastructure, and incentives for value alignment, the computational training and maintenance of AI tools can constitute a challenge. In this regard, academic researchers can intervene by sharing their technical knowledge for these tasks. While our prototype serves as an example, similar partnerships can be found elsewhere. The Associated Press Local News AI initiative (Rinehart and Kung 2023) created collaborations between university researchers and small newsrooms to aid the development of AI tools for journalists across the US. In Brazil, a review of the literature found that academia is the third-largest contributor to AI adoption in journalism (Pinto and Barbosa 2024). We hope academic researchers and non-profit journalistic organizations can further recognize this opportunity and seek collaborations in developing AI tools for a more robust and independent journalistic practice.

Critical Questions for the Development of AI Tools in Journalism

Based on these observations, we derive the following twelve questions that we deem important when developing further effective and value-aligned AI tools for journalists (see Table 3). Likewise, these questions can be used to evaluate the usefulness of existing AI tools and their alignment with journalistic values.

Conclusion

In this article, we develop a set of critical questions that may guide future development to create effective and value-aligned AI tools in journalism. Our analysis is grounded in the development process of an AI tool prototype designed to categorize documents by type and to distinguish between information-rich and information-poor pages within a journalistic context. Regarding the effectiveness aspect of AI tools for journalism, we suggest that tools should be generalist, cost-effective, user-friendly, and accurate. Concerning the value alignment aspect, we propose that the tools should reflect journalistic values, be independent of companies that do not prioritize journalistic values, be transparent, and address data privacy issues. Regarding the type of organization, we argue that major journalistic non-profit organizations like OCCRP are in the best position to lead the development of such tools.

We see several avenues for future research. First, regarding our prototype's technical advancements, it seems promising to create a document classifier based on large language models able to handle multiple types of inputs (i.e., PDFs and images). For example, given that GPT4o has demonstrated exceptional performance in other domains, it might be incorporated into a document classification AI tool. While pursuing such an approach might result in high effectiveness (being accurate while not costly to set up), note that this would raise several of the concerns related to value alignment outlined in this paper (being non-transparent and dependent on a large tech firm). Our objective in this study was not to create a competitive product but to draw general insights from the experience of developing a concrete, effective, and value-aligned AI tool prototype. We believe that the insights about which aspects to consider and which critical questions to ask when developing an AI tool for the journalistic community are generally applicable.

Our main contribution, which is on the conceptual level, remains valid despite the most recent technological developments.

Second, we see great potential to more rigorously analyze institutions' role in developing successful AI tools in journalistic practice (for a similar discussion, see Simon 2023). The results of our discussion show that major journalistic non-profit organizations like OCCRP could be the leading actors in implementing AI tools in journalism. As a final remark, we want to re-emphasize the role that researchers can play in developing AI tools for journalism. There are great opportunities for non-profit journalistic organizations to collaborate with interested academics. If researchers and journalistic organizations work together, it might be possible to overcome the "gravitational force" that pulls news organizations into the "orbit" of major tech companies (Simon 2022, 1850).

Notes

1. For recent work on design strategies that should be considered when implementing AI tools in the journalistic workflow, please see Gutierrez Lopez et al. (2023).
2. For accessing the dataset, see https://adamharley.com/rvl-cdip/

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported as part of the Data Science for Social Good fellowship hosted by the Rhineland-Palatinate Technical University Kaiserslautern-Landau in cooperation with the German Research Centre for Artificial Intelligence (DFKI GmbH).

Data Availability Statement

The data supporting this study's findings are available from the Organized Crime and Corruption Reporting Project's data platform Aleph. Restrictions apply to the availability of these data, which were used under license for this study. Data is available from the authors upon request with the permission of the Organized Crime and Corruption Reporting Project.

References

Afzal, Muhammad Zeshan, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. 2017. "Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification." In *Proceedings of the 14th International Conference on Document Analysis and Recognition*, 883–888. Kyoto, Japan: IEEE.

Aleph. 2023. "OCCRP Aleph: The Global Archive of Research Material for Investigative Reporting." *Aleph*. https://aleph.occrp.org.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." In *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Borchardt, Alexandra. 2022. "Go, Robots, Go! The Value and Challenges of Artificial Intelligence for Local Journalism." *Digital Journalism* 10 (10): 1919–1924. https://doi.org/10.1080/21670811.2022.2149584.

Breazu, Petre, and Napoleon Katsos. 2024. "ChatGPT-4 as a Journalist: Whose Perspectives Is It Reproducing?" *Discourse & Society* 35 (6): 09579265241251479.

Broussard, Meredith, Nicholas Diakopoulos, Andrea L. Guzman, Rediet Abebe, Michel Dupagne, and Ching-Hua Chuan. 2019. "Artificial Intelligence and Journalism." *Journalism & Mass Communication Quarterly* 96 (3): 673–695. https://doi.org/10.1177/1077699019859901.

Bureau of Labor Statistics. 2021. "Occupational Employment and Wage Statistics." May 2021 Report. Bureau of Labor Statistics. https://www.bls.gov/oes/2021/may/oes_stru.html

Carlini, Nicholas, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, et al.. 2021. "Extracting Training Data from Large Language Models." In *Proceedings of the 30th USENIX Security Symposium*, 2633–2650. Vancouver, B.C., Canada: USENIX Association.

Castillo-Campos, Mar, David Varona-Aramburu, and David Becerra-Alonso. 2024. "Artificial Intelligence Tools and Bias in Journalism-Related Content Generation: Comparison between Chat GPT-3.5, GPT-4 and Bing." *Tripodos* 55:99–115.

Cools, Hannes, and Nicholas Diakopoulos. 2024. "Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities." *Journalism Practice*, 1–19. https://doi.org/10.1080/17512786.2024.2394558.

de-Lima-Santos, Mathias-Felipe, Wang Ngai Yeung, and Tomás Dodds. 2024. "Guiding the Way: A Comprehensive Examination of AI Guidelines in Global Media." *AI & Society*, 1–19.

de Haan, Yael, Eric van den Berg, Nele Goutier, Sanne Kruikemeier, and Sophie Lecheler. 2022. "Invisible Friend or Foe? How Journalists Use and Perceive Algorithmic-Driven Tools in Their Research Process." *Digital Journalism* 10 (10): 1775–1793. https://doi.org/10.1080/21670811.2022.2027798.

Diakopoulos, Nicholas. 2019. "Towards a Design Orientation on Algorithms and Automation in News Production." *Digital Journalism* 7 (8): 1180–1184. https://doi.org/10.1080/21670811.2019.1682938.

Diakopoulos, Nicholas, Amy X. Zhang, and Andrew Salway. 2013. "Visual Analytics of Media Frames in Online News and Blogs." In *Proceeding of the IEEE InfoVis Workshop on Text Visualization*, 15–16. Atlanta, Georgia, USA: IEEE.

Fridman, M., R. Krøvel, and F. Palumbo. 2023. "How (Not to) Run an AI Project in Investigative Journalism." *Journalism Practice*, 1–18. https://doi.org/10.1080/17512786.2023.2253797.

Gondwe, Gregory. 2023. "CHATGPT and the Global South: How Are Journalists in Sub-Saharan Africa Engaging with Generative AI?" *Online Media and Global Communication* 2 (2): 228–249. https://doi.org/10.1515/omgc-2023-0023.

Gutierrez Lopez, Marisela, Colin Porlezza, Glenda Cooper, Stephann Makri, Andrew MacFarlane, and Sondess Missaoui. 2023. "A Question of Design: Strategies for Embedding AI-Driven Tools into Journalistic Work Routines." *Digital Journalism* 11 (3): 484–503. https://doi.org/10.1080/21670811.2022.2043759.

Hamilton, James. 2016. *Democracy's Detectives: The Economics of Investigative Journalism*. Cambridge, Massachusetts, USA: Harvard University Press.

Harley, Adam W, Alex Ufkes, and Konstantinos G. Derpanis. 2023. "The RVL-CDIP Dataset." https://adamharley.com/rvl-cdip/.

Helberger, Natali, S. J. Eskens, M. Z. van Drunen, M. B. Bastian, and J. E. Möller. 2020. "Implications of AI-Driven Tools in the Media for Freedom of Expression." In *Artificial Intelligence–Intelligent Politics: Challenges and Opportunities for Media and Democracy*, 1–34. Nicosia, Cyprus: Council of Europe.

Helberger, Natali, Max van Drunen, Judith Moeller, Sanne Vrijenhoek, and Sarah Eskens. 2022. "Towards a Normative Perspective on Journalistic AI: Embracing the Messy Reality of Normative Ideals." *Digital Journalism* 10 (10): 1605–1626. https://doi.org/10.1080/21670811.2022.2152195.

Kölsch, Andreas, Muhammad Zeshan Afzal, Markus Ebbecke, and Marcus Liwicki. 2017. "Real-time document image classification using deep CNN and extreme learning machines." In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, 1318–1323. Kyoto, Japan: IEEE.

Lin, Bibo, and Seth C. Lewis. 2022. "The One Thing Journalistic AI Just Might Do for Democracy." *Digital Journalism* 10 (10): 1627–1649. https://doi.org/10.1080/21670811.2022.2084131.

Nadj, Mario, Merlin Knaeble, Maximilian Xiling Li, and Alexander Maedche. 2020. "Power to the Oracle? Design Principles for Interactive Labeling Systems in Machine Learning." *KI-Künstliche Intelligenz* 34 (2): 131–142.

Nishal, Sachita, and Nicholas Diakopoulos. 2024. "Envisioning the Applications and Implications of Generative AI for News Media." In *CHI'23 Workshop on Generative AI and HCI*. Hamburg, Germany: ACM.

OCCRP. 2023. "Impact to Date." https://www.occrp.org/en/impact-to-date.

Opdahl, Andreas L., Bjørnar Tessem, Duc-Tien Dang-Nguyen, Enrico Motta, Vinay Setty, Eivind Throndsen, Are Tverberg, and Christoph Trattner. 2023. "Trustworthy Journalism through AI." *Data & Knowledge Engineering* 146:102182. https://doi.org/10.1016/j.datak.2023.102182.

Pavlik, John V. 2023. "Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education." *Journalism & Mass Communication Educator* 78 (1): 84–93. https://doi.org/10.1177/10776958221149577.

Pinto, Moisés Costa, and Suzana Oliveira Barbosa. 2024. "Artificial Intelligence (AI) in Brazilian Digital Journalism: Historical Context and Innovative Processes." *Journalism and Media* 5 (1): 325–341.

Poston, Ben, Joel Rubin, and Anthony Pesce. 2015. "LAPD Underreported Serious Assaults, Skewing Crime Stats for 8 Years." *Los Angeles* Times. http://www.latimes.com/local/cityhall/la-me-crime-stats-20151015-story.html

Quinonez, Claudia, and Edgar Meij. 2024. "A New Era of AI-Assisted Journalism at Bloomberg." *AI Magazine* 45 (2): 187–199. https://doi.org/10.1002/aaai.12181.

Rinehart, Aimee and Ernest Kung (2023, October). "Associated Press' Local News AI Initiative." Associated Press. https://www.ap.org/solutions/artificial-intelligence/local-news-ai/

Shi, Yi, and Lin Sun. 2024. "How Generative AI Is Transforming Journalism: Development, Application and Ethics." *Journalism and Media* 5 (2): 582–594.

Shin, Christian, David Doermann, and Azriel Rosenfeld. 2001. "Classification of Document Pages Using Structure-Based Features." *International Journal on Document Analysis and Recognition* 3 (4): 232–247. https://doi.org/10.1007/PL00013566.

Shorten, Connor, and Taghi M. Khoshgoftaar. 2019. "A Survey on Image Data Augmentation for Deep Learning." *Journal of Big Data* 6 (1): 60. https://doi.org/10.1186/s40537-019-0197-0.

Shu, Lei, Hu Xu, and Bing Liu. 2017. "DOC: Deep Open Classification of Text Documents." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, edited by Martha Palmer, Rebecca Hwa, and Sebastian Riedel, 2911–2916. Copenhagen, Denmark: Association for Computational Linguistics.

Simon, Felix M. 2022. "Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy." *Digital Journalism* 10 (10): 1832–1854. https://doi.org/10.1080/21670811.2022.2063150.

Simon, Felix M. 2023. "Escape Me If You Can: How AI Reshapes News Organisations' Dependency on Platform Companies." *Digital Journalism* 12 (2): 149–170. https://doi.org/10.1080/21670811.2023.2287464.

Stark, Jennifer A, and Nicholas Diakopoulos. 2016. "Towards Editorial Transparency in Computational Journalism." *Computation + Journalism Symposium* 5:1–5.

Stray, Jonathan. 2019. "Making Artificial Intelligence Work for Investigative Journalism." *Digital Journalism* 7 (8): 1076–1097. https://doi.org/10.1080/21670811.2019.1630289.

van Dalen, Arjen. 2024. "Revisiting the Algorithms Behind the Headlines. How Journalists Respond to Professional Competition of Generative AI." *Journalism Practice*, 1–18. https://doi.org/10.1080/17512786.2024.2389209.

Appendix A.7: Paper VII

Bibliographic data	Henn, T., Posegga, O. 2023. Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit. PLOS One, (18)8, e0288962.
Abstract	Portrayals of violence are common in contemporary media reporting; they attract public attention and influence the reader's opinion. In the particular context of a social movement such as Black Lives Matter (BLM), the portrayal of violence in news coverage attracts public attention and can affect the movement's development, support, and public perception. Research on the relationship between digital news content featuring violence and user attention on social media has been scarce. This paper analyzes the relationship between violence in online reporting on BLM and its effect on user attention on the social media platform Reddit. The analysis focuses on the portrayal of violence in images used in BLM-related digital news coverage shared on Reddit. The dataset is comprised of 5,873 news articles with images. The classification of violent images is based on a VGG19 convolutional neural network (CNN) trained on a comprehensive dataset. The results suggest that what significantly affects user attention in digital news content is not the display of violence in images; rather, it is negative article titles, the news outlet's political leanings and level of factual reporting, and platform affordances that significantly affect user attention. Thus, this paper adds to the understanding of user attention distributions online and paves the way for future research in this field.
DOI	10.1371/journal.pone.0288962

PLOS ONE

RESEARCH ARTICLE

Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit

Theresa Henn, Oliver Posegga*

Department of Information Systems and Social Networks, University of Bamberg, Bamberg, Germany

* oliver.posegga@uni-bamberg.de

OPEN ACCESS

Citation: Henn T, Posegga O (2023) Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit. PLoS ONE 18(8): e0288962. https://doi.org/10.1371/journal.pone.0288962

Editor: Fabio Saracco, 'Enrico Fermi' Research Center, ITALY

Received: July 14, 2022

Accepted: July 6, 2023

Published: August 9, 2023

Data Availability Statement: The complete overview of the data used in this study is provided in the data availability statement included in this submission. The data used in this study comprises three parts: 1. Image data: "UCLA Protest Image Dataset" (training data) The dataset is available from the original curators of the "UCLA Protest Image Dataset" [1]. All information to obtain and prepare the dataset can be obtained from their public GitHub repository [2]. 2. Image data: Test data to evaluate the performance of the image classifier The data cannot be shared publicly

Abstract

Portrayals of violence are common in contemporary media reporting; they attract public attention and influence the reader's opinion. In the particular context of a social movement such as Black Lives Matter (BLM), the portrayal of violence in news coverage attracts public attention and can affect the movement's development, support, and public perception. Research on the relationship between digital news content featuring violence and user attention on social media has been scarce. This paper analyzes the relationship between violence in online reporting on BLM and its effect on user attention on the social media platform Reddit. The analysis focuses on the portrayal of violence in images used in BLM-related digital news coverage shared on Reddit. The dataset is comprised of 5,873 news articles with images. The classification of violent images is based on a VGG19 convolutional neural network (CNN) trained on a comprehensive dataset. The results suggest that what significantly affects user attention in digital news content is not the display of violence in images; rather, it is negative article titles, the news outlet's political leanings and level of factual reporting, and platform affordances that significantly affect user attention. Thus, this paper adds to the understanding of user attention distributions online and paves the way for future research in this field.

Introduction

It is indisputable that "[t]he mainstream adoption of social media applications has caused a paradigm shift in how people communicate, collaborate, create, and consume information" [1]. This paradigm shift has also affected the consumption of and confrontation with news [2], as users are increasingly "reading the news through social [media] platforms" [3]. During the last ten years, social media platforms have risen to one of the most prominent sources for news consumption [3,4], with nearly half of the U.S. population receiving news from social media websites in 2021 [5]. Thus, it becomes increasingly important to understand how news attract user attention in a social media context since the exposure to news shared on social media can have severe offline consequences [6], and news shared online is an integral part of understanding today's media system [7,8]. The analysis of news shared online has thus become an

because of copyright issues. We provide a complete description of how to recreate the dataset using publicly available resources and the UCLA Protest Image Dataset in the data availability statement. 3. NewsMTSC data The dataset is available from Hugging Face [3, 4]. 4. Reddit data: Reddit submissions and the final dataset The primary dataset underlying our models is available via OSF (https://osf.io/4dr2x/?view_only=39ff5cb39c254b6c93eb12f9c205a941) for peer review purposes and will be made publicly available in case of acceptance. [1] Won D, Steinert-Threlkeld ZC, Joo J. Protest activity detection and perceived violence estimation from social media images. MM 2017 - Proc. 2017 ACM Multimed. Conf., 2017. https://doi.org/10.1145/3123266.3123282. [2] https://github.com/wondonghyeon/protest-detection-violence-estimation [3] Hamborg F, Donnay K. NewsMTSC: a dataset for (multi-) target-dependent sentiment classification in political news articles. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021. pp. 1663–1675. doi:10.5167/uzh-207183 [4] NewsMTSC dataset. In: Hugging Face. 2022 [cited 6 Jun 2023]. Available: https://huggingface.co/datasets/fhamborg/news_sentiment_newsmtsc.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

essential precursor to improving our understanding of how social media news consumption influences "civic awareness and engagement in civic and political life" [9].

With this paper, we aim to contribute to this line of research by analyzing how the *violent* coverage of social movements affects user attention on social media. The portrayal of violence in traditional media types such as "television, radio newspapers, magazines . . . or in other words, any form of mass communication available before the advent of digital media" [10] has been studied thoroughly [11,12]. Research finds that traditional media frequently feature violent content as a way of entertaining their audiences and, thus, attracting public attention [11–13], which yields greater revenues [14,15]. It remains uncertain, though, whether this link between the portrayal of violence and the attraction of attention can also be transferred to the online sphere. Initial analyses concluded that sharing images featuring violence by individual users on social media did not lead to greater user engagement [16,17], but the effects of online news featuring violence on social media user attention remain unclear. Therefore, we see the need to further our understanding of the relationship between violence featured in the digital news content shared on social media platforms and its effects on user attention.

To address this research gap, we analyze the portrayal on social media of the social movement Black Lives Matter (BLM) and how that attracts user attention. According to Diani [18], a movement can be defined as a social movement if it's based on "networks of informal interactions between a plurality of individuals, groups and/or organisations, engaged in a political and/or cultural conflict, on the basis of a shared collective identity." We characterize BLM as a social movement because it is a decentralized structure of networked individuals [19] fighting for political, structural, and cultural change [19] while collectively identifying and organizing under the umbrella of symbols such as the hashtag #BlackLivesMatter [20].

The BLM movement began with a social media hashtag after the acquittal of the man charged in the shooting death of Black teenager Trayvon Martin in Florida in 2012 [21], grew nationally in 2014 after the deaths of Michael Brown in Missouri and Eric Garner in New York [19], and was reignited by the death of George Floyd, a Black man murdered by a white police officer on May 25, 2020, in Minneapolis [21]. People all over the world went into the streets to show solidarity and join the movement's fight against "racial injustice and police brutality" [21].

We analyze the news coverage of this specific social movement for two reasons. First, the BLM movement gained a considerable amount of media attention that "delegitimized and racialized the movement by accentuating conflict and violence" [22], which aligns with an established line of research that stresses that media coverage often focuses on violent and disruptive aspects of a movement's protest rather than its goals and social criticism [22–25]. Since public opinion and support of a social movement depend highly on how that movement is portrayed in news reporting [26–28], especially when violence is involved [23], it is particularly interesting to analyze the news portrayal of the BLM movement, as both protestors and police have committed acts of violence [29]. Second, the BLM movement "has deep ties to social media" [21], with hashtags such as #BlackLivesMatter sparking controversial online discussions and fueling worldwide protests and political unrest [21]. The consequence of the public being drawn to violent media content online could be a negative public bias toward the movement, which may escalate the tense political situation surrounding that movement [29].

In particular, we examine the effect of images that depict violence and nonviolence and are used in news articles shared on the social media platform Reddit, as images play a significant role in how news articles shape "the public's understanding of issues and events" [30] (for a more detailed description of subreddits and the types of submission we analyze, see section "Social media platform Reddit"). These images are either photos or still images from videos that are part of the news articles linked in submissions posted on Reddit (henceforth, we refer to both as "images"). These linked news articles are either hosted by the digital extension of traditional media formats, such as on the website of the *New York Times* newspaper or are hosted by digital-born media, such as blogs or pure online news outlets [31,32].

As it is unclear whether earlier findings from media research can be transferred to social media environments [33], we focus not only on the effects on user attention of images featuring violence but also evaluate the impact of other attributes of shared digital news content that research has shown to be significant predictors of user engagement (for more information, see section "Theory and hypotheses"). We structure these attributes into three categories: (1) attributes related to the news item (in our case, the submission featuring a linked news article), such as the sentiment expressed in a news article's title; (2) attributes of the news provider (in our case, the digital extension of traditional media formats or digital-born news media), such as the outlet's political leanings or level of factual reporting; and (3) attributes concerning the community rules and platform affordances of the subreddits analyzed, such as the flagging of submissions that violate submission guidelines. By considering this multitude of attributes that can affect user attention, our analysis accounts for the complexity of social media landscapes, with their various, mutually dependent elements [1].

As our objective is to elucidate whether the portrayal of violence or other attributes of digital news content shared on Reddit have a significant impact on user attention, we pose the following two research questions specific to BLM, which is the social movement on which we focus in our analysis:

RQ1: *How does the visual portrayal of violence in BLM-related news articles shared on Reddit affect user attention*?

RQ2: *How do other attributes of digital BLM-related news content shared on Reddit impact user attention*?

Thus, our paper contributes to the understanding of the link between the type of online news coverage (violent and nonviolent) of a popular social movement [29] and the attention paid to it on the social media platform Reddit. By analyzing several attributes of online news reporting that impact user attention, our work not only tests the effect of images featuring violence on user attention but also accounts for attributes related to the news item, news provider, and community rules and platform affordances that might influence user attention. Since news consumption through social media continues to grow [3,4], this work contributes to the understanding of news consumption in an increasingly digitized world.

This paper is structured as follows. First, we briefly summarize the main steps and findings of the research to highlight our study's contribution. Second, we dive deeper into previous research analyzing the relationship between news coverage of social movements and attention distributions while differentiating between traditional, digital, and social media. Third, we present our theoretical background and hypotheses in detail, grounded in the "negativity bias" and the "picture superiority effect." Fourth, we detail our methodology, including the operationalization of our key variables, a description of our data collection on Reddit, and the transfer learning and performance of the VGG19 convolutional neural network (CNN) that we trained to classify violence in images. After conducting our data analysis, we discuss our findings, highlight the limitations of our work, and provide an outlook for future research.

Summary of main steps and findings

The paper's main goal is to analyze the relationship between news images that feature violence and user attention in the context of BLM on Reddit (RQ1), as there has been only scarce research on the relationship between digital news content featuring violence and user attention on social media. Based on the theoretical concepts "negativity bias" [34,35] and the "picture superiority effect" [36], we assume that there should be a positive effect between images featuring violence and user attention. In addition, we test for several other attributes of online news reporting that potentially impact user attention (RQ2) and subdivide them into (1) attributes related to the news item, (2) attributes of the news provider, and (3) attributes concerning community rules. For all attributes, we theoretically derive and test hypotheses empirically (Table 1 is an overview of the hypotheses and their acceptance or rejection).

For the empirical analyses, we proceeded as follows. We used the Pushshift submission dumps to download submissions from the three popular news-sharing subreddits, r/politics, r/news, and r/worldnews. We filtered the data to match our topic (BLM), period of analysis (May 25, 2020, the day of George Floyd's murder, to May 25, 2021), and submissions that received a minimum of attention (removal of submissions with zero comments). To complete the dataset with variables necessary for testing our hypotheses, we undertook three methodological steps: (1) for the attributes of the news provider, we added information regarding the linked news outlet's political leaning, factual reporting, traffic, and type of media outlet from the commonly used website mediabiasfactcheck.com, which is widely applied in research [37]; (2) for each submission, we downloaded, if possible, the first image of the linked news outlet and automatically classified it as violent or nonviolent with a VGG19 CNN that we trained on a comprehensive and context-specific dataset; and (3) to detect the sentiment of a news title (attribute related to the news item), we trained a Bidirectional Encoder Representation from Transformers (BERT) model and classified each title as either positive or negative. We received information on all other variables related to attributes of community rules from the Pushshift submission dumps. We conducted a negative binomial regression to analyze the relationship between user attention (measured in number of comments) and our independent and control variables that we derived theoretically from the existing literature.

Based on the results of our analysis, we can answer the two research questions. With respect to RQ1, it is not images featuring violence that affect user attention in the case of BLM reporting on Reddit. Rather, with respect to RQ2, we find that (1) the political leaning of a news

Table 1. Overview of hypotheses and their acceptance or rejection.

Hypotheses	Acceptance	Rejection
H1: *Users will pay greater attention to submissions that feature images depicting violence than submissions that do not.*		X
H2: *A submission with a title conveying a negative sentiment will receive greater user attention than a submission with a positive title.*	X	
H3: *A submission's linked image will have a more substantial effect on user attention than the sentiment of a submission's title.*		X
H4: *A submission that includes a link to a conservative news outlet will receive greater user attention than a submission that includes a link to a liberal or neutral news outlet.*		X
H5: *A submission that includes a link to a less-factual news outlet will receive greater user attention than a submission that includes a link to a highly factual news outlet.*		X
H6: *A submission that includes a link to a popular news outlet with high traffic will receive greater user attention than a submission with a link to a less-popular news outlet with minimal or medium traffic.*		X
H7: *A submission posted in r/politics will receive greater user attention than a submission posted in r/news or r/worldnews.*		X
H8: *Users will pay greater attention to submissions flagged as "NSFW" than to nonflagged submissions.*		X
H9: *A submission with a flair tag will receive less user attention than a submission without a flair tag.*	X	
H10: *The number of a submission's cross-posts will positively affect user attention.*	X	

https://doi.org/10.1371/journal.pone.0288962.t001 outlet (politically neutral news outlets have a positive effect compared to conservative or liberal outlets); (2) its level of factual reporting (high factual reporting is positively associated, low factual reporting is negatively associated compared to mixed factual reporting); (3) Reddit community rules and affordances (the subreddits r/news and r/worldnews are positively associated compared to r/politics, link flair is negatively correlated, the number of cross-posts has a positive effect), and (4) the negative titles of news significantly correlate with user attention in the case of BLM reporting on Reddit.

In the following, we provide a thorough explanation of our theoretical background and methodological procedure and discuss our results in detail.

Theoretical background

Related literature

In the past, traditional media such as newspapers, television, and radio– known for their gatekeeping function, agenda-setting power, and one-way communication– have been the predominant source for news consumption [38]. The media system, however, has changed [8], and a digital transformation has led traditional media to enter the online space and distribute news through their own digital channels (e.g., the online version of *The New York Times*). These traditional types of media have been complemented by digital-born media such as blogs or pure online newspapers that "produce a wide variety of content, from in-depth investigative journalism to eyeball-catching clickbait" [8].

The introduction of social media networks such as Reddit, Twitter, and Facebook, which function as intermediaries for news exposure and consumption [39], is another addition to this growing digital side of the media system. On these platforms, digital news can be produced and shared by both individual users and large news organizations, emphasizing the significant shift in power structures in news production. Thus, social media networks "are becoming more and more integral to how people find information [and news]" [40] by also enabling people to produce, share, and discuss information and news with their networks of other people.

This interplay between digital media (comprised of the digital extensions of traditional media and digital-born media; henceforth referred to as "digital media") and social media platforms are the focus of our analysis. Together with traditional media, they constitute the so-called "hybrid media system" [7].

The relationship between media and their reporting on social movements has a long history in research [27,28,41–44]. When reporting on social movements, both digital and traditional media outlets need to decide which events to include in their news coverage and how to portray the social movement [see, e.g., 27,41,42]. Since some events are regarded as more "newsworthy" than others [42], as they "resonate with more general social concerns" [43], media outlets report on specific events rather than others, which is also referred to as "selection bias" in media research [42,43]. A particular selection criterion in this context is the reporting on negative events, especially in a political or social movement context [see, e.g., 44,45], as negativity, in general, has long been prevalent in the news coverage of both traditional and digital media [46,47].

Reporting about violence is a specific type of negative news coverage [48]. Social movements, in particular, have been subjected to this type of negative news coverage [e.g., 49,50] "emphasizing violence and deviant behaviour" [22], which tends to increase the audience size as it "contributes to the media spectacle . . . around protest events" [16]. Yet, this delegitimizing news coverage and focus on violence that amplifies the sensationalist aspects of protest is "often to the detriment of the social cause or movement . . ." [26]; it is referred to in research as the "protest paradigm" [23,25]. Thus, media reporting constructs a certain interpretation of the protest event, also called "description bias" [see, e.g., 28], "that differs from both the objectives of protestors and interpretations of other observers" [28], which in turn highly impacts the public perception of the social movement [e.g., 23,25,26]. Therefore, it is important to further our knowledge of media reporting on social movements featuring violence and the attention it attracts.

To thoroughly understand the effects of violence in media reporting and its impact on attention, we need to differentiate between traditional and digital media news reporting. Whereas in a social movement context the role of violence in traditional media coverage has been studied thoroughly, though little attention has been paid to its effects in digital news reporting (see next paragraphs). Thus, our paper analyzes whether findings from traditional media research can be replicated for the digital news coverage.

The literature shows that traditional media's news coverage frequently features violent aspects of social movements [see, e.g., 23,41,51] to attract attention, as "[b]urning buildings and burning tires make better television than peaceful vigils and orderly marches" [41]. For instance, Boykoff [52] showed that influential newspapers, in particular, typically apply framing that features violent aspects of social movements because it serves as a prominent device for attracting readers' attention. McFarlane and Hay [51] evaluated the newspaper reporting on the protests against the 3rd World Trade Organization (WTO) Ministerial Conference in 1999, which portrayed protests ". . . within an 'anarchy and violence' narrative structure in which repeated reference was made to property destruction by protestors, to their acts of violence, and to the presence of 'masked anarchists dressed in black'" [51]. This particular framing was applied to resonate with the news' readership and receive attention while delegitimizing the protest and guiding "readers towards a positioning supporting the WTO" [51].

In a similar vein, in his study about the press reporting on the 1968 London anti-war rally, Murdock [50] showed that the press framed the movement to a viewpoint "already familiar to the reader" [50] by portraying the protestors as militant extremists or anarchists ready to use violence. This framing made the news coverage more appealing to the audience and therefore added to its received attention. Likewise, Boyle et al. [23] found that newspapers often emphasize the violent actions of social movements disproportionally when reporting on them, regardless of country and across time. Such violent framing of a movement "can have a powerful influence on how audience members perceive issues and parties to a conflict, including perceptions of protestors" [23]. McLeod and Detenber [53] were among the first to analyze the impact of television coverage on violent protests and the effects on public identification with a movement. The authors showed that news stories shared on television have the power to legitimize or delegitimize the cause that motivates protest and thus sway the public in favor of or against a social movement. Vliegenthart and Walgrave [54] summed up the findings of several different authors who concluded, among other findings, that a social movement protest is more likely to receive traditional media coverage if it includes acts of violence; this leads to greater audience exposure and thus greater attention.

The notion of violence and social movements also plays a role on social media. Some studies have analyzed how activists themselves use violence in their social media posting (tweets, videos, and photos) as a means of drawing attention to their cause [55]. Since the sharing of images affects user attention online [22,56], and such images are decisive for how social movements are perceived [57,58], it is important to analyze the visual portrayal of violence in the social media communication of social movements.

Neumayer and Rossi [59] contrasted how different actors, such as police, journalists, and activists, shared images with varying levels of violence on the social media platform Twitter as a way of communicating their issues. They concluded that activists often struggle to garner visibility for their grievances because images posted by other actors (e.g., the police) that feature violence overshadow the cause of the movement. Through a manual image content analysis on Twitter, Kharroub and Bas [17] found that in the 2011 Egyptian revolution, it was not emotionally arousing images (featuring, e.g., violence) posted by actors such as activists or the police that compelled users to retweet messages but rather efficacy-eliciting images (featuring, e.g., national or religious symbols). Rossi et al. [16] undertook the first large-scale study in this field, evaluating the correlation between images featuring violence posted by different types of users, such as activists, news agencies, and the police, and their propagation on Twitter during the G20 summit protests in Hamburg, Germany, in 2017. Within the five days of their study context, they found that the visual portrayal of violence did not affect the number of retweets or mentions.

In the aforementioned studies, the portrayal of violence on social media did not affect user engagement to the expected extent. These studies, however, evaluated images featuring violence that had been shared by various user types, such as activists, police, and journalists. Our analysis, by contrast, focuses exclusively on how digital *news coverage* of social movements and violence affects user attention on social media platforms. In addition, we control for other potential attributes of digital news reporting that could influence user attention. Thus, our paper offers three unique contributions. First, we evaluate whether previously established findings from traditional media research can be transferred to the online world, as research suggests with respect to offline news habits [60,61]. Thus, we test whether shared online news articles featuring visual displays of violence receive increased user attention on Reddit. Second, we analyze the case of a prominent and contemporary social movement with a strong digital presence over a prolonged period of time (one year) on the social media platform Reddit, which hosts subreddits with an explicit focus on sharing and discussing online news content. Third, since social media is a complex communication environment [1] and the effect of digital news coverage of violence on user attention is unclear, we extend our image analysis by analyzing other attributes of the shared digital news content (attributes of news item, news provider, and community rules and platform affordances). By doing so, we create an extended statistical model that accounts for several plausible factors that affect online user attention.

Theory and hypotheses

So-called "negativity bias" is one broadly acknowledged concept used to explain people's tendency to be drawn to negative news, which is attributed to a psychological remnant of human evolution [see, e.g., 35]. Rozin and Royzman [34] describe the concept as "a general bias, based on both innate predispositions and experience, in animals and humans, [which] give[s] greater weight to negative entities." Hence, negative information is experienced more intensely, leaving a more persistent impression in people's memories than positive information [62]. This mechanism partially explains why we are naturally drawn to negative [47] and also violent media content: our "brain is simply built with a greater sensitivity to unpleasant news" [Prof. John T. Cacioppo cited in 63]. This is, in part, an explanation for the predominance of negative news reporting in traditional media [64]. Based on these findings and the suggestion that offline news habits can be transferred to digital news consumption [60,61], we hypothesize that digital news content depicting violence in the context of social movements should also receive greater user attention when shared on social media.

Further, in line with the saying that "a picture is worth a thousand words," we argue that the visual display of violence should garner even greater attention [30]. Pictures have a more significant impact on the human brain than textual representations of information, which can be absorbed more rapidly through visual stimuli; the content of images is more likely than text to be remembered [36]; and emotionally charged images, in particular, tend to attract greater attention than those lacking in emotion [65]. This phenomenon is also referred to as the "picture superiority effect" [36]. Social media platforms have increasingly become a "visual medium" [66], a sign of the growing importance of images in online contexts [67]. Plus, it has been shown that the willingness of social media users to share (and thus engage with) a digital news article about social movements increases once visuals are included [22]. Thus, we analyze the effect on user attention of images featuring violence since sharing images highly correlates with user attention online [22,56].

This paper defines violence as "intentionally caused or carelessly accepted damage to/destruction of property or the injuring/killing of people" [68]. By applying this strict definition of violence by Rucht [68], our analysis focuses on the display of physical violence and intentionally excludes other forms of violence, such as verbal assault. We decided on this relatively conservative definition of violence because we aim to measure the effect of "extreme" forms of violence, which we assume receives more user attention than more "minor" types of violence (see section "Operationalization of variables" for further information).

Based on this theoretical background, our first hypothesis is:

H1: *Users will pay greater attention to submissions that feature images depicting violence than submissions that do not.*

As stated before, we analyze the effects of other BLM-related news content attributes shared on Reddit. We focus first on attributes related to the news item. We examine submissions to subreddits that consist of the title of a news article and a link to a digital source news outlet (henceforth, we refer to both digital extensions of traditional media and digital-born media as "news outlet") and conclude that each submission's title is an integral part of a submission's content and can, thus, affect user attention [69]. Because an article's tone can influence users' perceptions of and reactions to the content [70], we control for the sentiment of titles. Since previous research established a positive relationship between negative titles and news readership online [71], we also assume that negative titles will be more attractive than positive ones, which also aligns with the negativity bias. From this, we derive our second hypothesis:

H2: *A submission with a title conveying a negative sentiment will receive greater user attention than a submission with a positive title.*

In line with the picture superiority effect, Hessel et al. [67] have shown that "image features always outperform text features" in their effects on user attention (see also [22]). Therefore, we argue that the impact of images featuring violence outweighs the effect of titles conveying a negative sentiment. Thus, our third hypothesis states:

H3: *A submission's linked image will have a more substantial effect on user attention than the sentiment of a submission's title.*

Other critical attributes that might influence user attention can be related to the news provider. Thus, we evaluate the political leanings of the news outlet referenced by a submission. There is an extensive body of research analyzing the relationship between political ideology and susceptibility to the negativity bias [72,73]. It is argued that people who identify as "conservatives" (as understood in the U.S. context) are more receptive to negativity compared to those who identify as liberals (as understood in the U.S. context) [72], although some recent studies have challenged these findings [73]. Based on these considerations, we assume that conservative news outlets cover greater amounts of violent and negative content than other outlets since they align with their audience's expectations of news reporting [74] (openness to negative news). In line with the negativity bias, we assume that this type of reporting should lead to a higher level of user attention compared to liberal or neutral reporting. Hence, our fourth hypothesis is:

H4: *A submission that includes a link to a conservative news outlet will receive greater user attention than a submission that includes a link to a liberal or neutral news outlet.*

Likewise, a news article's level of truthfulness has a major effect on user engagement and attention on social media. Research has shown that false news, in particular, "diffused significantly farther, faster, deeper, and more broadly" [75] on social media, thus reaching a broader audience and receiving greater user attention. Based on this discrepancy between true and false news, we assume that a news outlet's level of factual reporting will influence user attention. Thus, news outlets that are known for a less-factual reporting style ("false news") will receive greater user attention than outlets that adhere to a more factual style of reporting ("true news"). Our fifth hypothesis is:

H5: *A submission that includes a link to a less-factual news outlet will receive greater user attention than a submission that includes a link to a highly factual news outlet.*

Moreover, a news outlet's general level of prominence could affect the level of attention a user pays to a submission that references it. Therefore, we also test for the linked news outlet's traffic by assuming that those with more traffic are more popular and thus receive greater user attention. Hence, our sixth hypothesis states:

H6: *A submission that includes a link to a popular news outlet with high traffic will receive greater user attention than a submission with a link to a less-popular news outlet with minimal or medium traffic.*

Finally, we examine attributes related to Reddit-specific community rules and platform affordances "that shape how people engage with [online] environments" [76] and thus could have an effect on a submission's visibility and, in turn, user attention. In our analysis, we evaluate data from the three subreddits r/politics, r/news, and r/worldnews. Of these, only r/politics allows the posting of thumbnail images, and so we assume–in line with our assumptions regarding the picture superiority effect–that submissions in this subreddit receive greater attention than submissions posted in the other two (for more information regarding the subreddits analyzed, see section "Social media platform Reddit"). By controlling for the subreddit in which a submission was posted, we also account for different community sizes, and that posting in more popular subreddits could lead to greater exposure and, therefore, greater user attention–or, in contrast, to less visibility due to greater competition for attention with other posts [69].

We also include whether a submission is flagged as "Not Safe for Work (NSFW)," which, in our context, likely implies the depiction of a higher degree of (physical) violence within the attached news link. Based on our previous assumptions regarding the effect of violence on user attention, we expect a positive relationship between "NSFW" submissions and user attention.

Further, we add whether a submission receives what Reddit calls a "flair tag," which, in our case, indicates that the subreddit's moderators commented on the submission. Considering that our three subreddits are known to follow strict submission guidelines (see the front page of each subreddit), we assume that submissions with a flair tag have somehow violated the submission guidelines by, for example, being out of date or irrelevant to the subreddit. Based on these considerations, we assume that submissions with a flair tag receive less user attention.

Finally, we control how often a submission has been cross-posted, that is, posted as a new submission to other subreddits. It seems plausible to assume that exposure to a broader audience also attracts greater user attention to the initial submission. We, therefore, assume a positive relationship between the number of cross-posts and user attention.

Summarizing these assumptions regarding the Reddit-specific affordances and community rules, our final four hypotheses are:

H7: *A submission posted in r/politics will receive greater user attention than a submission posted in r/news or r/worldnews.*

H8: *Users will pay greater attention to submissions flagged as "NSFW" than to nonflagged submissions.*

H9: *A submission with a flair tag will receive less user attention than a submission without a flair tag.*

H10: *The number of a submission's cross-posts will positively affect user attention.*

We include additional control variables to control for contextual factors that could affect user attention online. We control for the timing and date of a submission, as Reddit is a dynamic environment with varying attention distribution over time [67]. Further, we test whether the type of news outlet referenced in a submission influences user attention. We differentiate between digital-born news outlets (websites) and the digital extension of traditional news media (e.g., TV stations and newspapers) since–generally speaking–both types of media likely display a different level of professionalism and follow a different logic of news reporting (see section "Related literature"), which might impact user attention differently.

Materials and methods

Operationalization of variables

Attention is a broad concept but can be defined generally as: "Notice taken of someone or something; the regarding of someone or something as interesting or important" [77]. There are several ways to interact with posts and "take notice" of a particular submission on social media. In the case of Reddit, we operationalize our dependent variable, "user attention," with the number of comments posted on each submission. This proxy offers an indication of whether users interacted substantially with a submission [22] that seems to be particularly popular in a political news context [78] and thus paid attention to it. Other attention measures, such as up- or downvoting on a submission, can be seen as an expression of whether a submission's content aligns with the community rules of a subreddit [79] and thus indicates quality control rather than being a measure of user attention. Therefore, we operationalize our dependent variable, "user attention," with the discrete variable, "number of comments." (To sustain our results, however, we also ran a negative binomial regression model that takes into account the number of up-and downvotes as an expression of user attention. Yet, also with this measure, the main effects remain the same. For more information, please see Supporting Information, Negative binomial regression models).

We use the conservative definition by Rucht introduced in section "Theory and hypotheses" to capture the complex concept of violence (*H1*). Hence, to be classified as "violent," an image must show physical harm to property or people (in the case of the BLM movement, e.g., looting shops or using pepper spray against protestors). Images showing only "minor" violence, such as verbal assault or provocative gestures, are not coded as "violent." As stated before, we intentionally apply this conservative definition of violence because, first, in the BLM context, there were significant acts of violence by protestors who looted shops and police officers attacking the protestors with "pepper spray, tear gas, and rubber bullets" [29]. Hence, we are optimistic about gathering enough meaningful data. We also anticipate that the conservative definition of violence will lead to more precise results by assuming that a more extreme visual display of violence attracts greater user attention. For our automated image classification, we dichotomously code a violent image as "1" and a nonviolent image as "0."

As *H2* indicates, titles of submissions with negative connotations could positively impact user attention. To classify the connotation of each submission's title as either "negative" or "positive," we utilize a BERT model [80] as it depicts one of the most prominent and state-of-the-art language models for sentiment analysis [see, e.g., 81]. BERT is a transformer model that can capture the context of a sentence by considering the bidirectional dependencies between words within a sentence [80]. Thus, the model outperforms simpler approaches such as bag-of-word models that are dictionary-based and do not account for the context of the classified sentence [see, e.g., 81]. For our analysis, we use the "BERT base model uncased," which we derived from the Hugging Face platform [82]. The model is trained on English Wikipedia and the BookCorpus data [80] and contains 110 million parameters. To customize the model for classifying the sentiment of news articles, we trained the model with the NewsMTSC dataset [83], a labeled dataset for sentiment classification of political news articles containing the categories positive, negative, and neutral. Since we are interested in positive and negative sentiment classification, we dropped the neutral samples from the training, validation, and test dataset. After training, the model receives a training accuracy of 99.17%, a validation accuracy of 87.66%, and a test accuracy of 87.40% (for more information on the data reduction, the training, and validation of the model, please see Supporting Information, Sentiment model). In addition, to evaluate the effect of sentiment on user attention even more granularly, we ran another negative binomial regression model with different sentiment variables, which, however, yields almost similar results (for more information, please see Supporting Information, Negative binomial regression models).

To label the categorical variables related to linked news outlets, we use information retrieved from the website mediabiasfactcheck.com, which is widely applied in research [37]. We manually label each news outlet's "political leaning" (*H4*), "factual reporting" (*H5*), and "traffic" (*H6*) according to the respective classifications by mediabiasfactcheck.com. If we do not find all information on the news outlet, we code the observation as missing and drop the submission.

After coding our data, we merge some labels to avoid small *n* distributions within a category and obtain more sound results. Thus, our variable "political leaning" is comprised of the labels "liberal" (merging the "left," "left-center," and "far left" categories), "conservative" (merging the "right," "right-center," "far right," "extreme right," and "far right conspiracy pseudoscience" categories), "conspiracy" (merging the "conspiracy," "junk news," and "conspiracy pseudoscience" categories), and "neutral" (merging the "least bias" and "pro-science" categories). The variable "factual reporting" consists of the labels "high" (merging "high" and "very high"), "mixed" (merging "mixed" and "mostly"), and "low" (merging "low" and "very low"). Finally, the variable "traffic" is split into the labels "high," "medium," and "minimal."

The remaining Reddit-specific community and affordance variables are coded as follows: the categorical variable "subreddit" (*H7*) is divided into the three subreddits, "r/politics," "r/news," and "r/worldnews"; both variables "NSFW" (*H8*) and "link flair" (*H9*) are binary and coded as "0" if the submission was not flagged and as "1" if it was; and the variable "number of cross-posts" (*H10*) is discrete.

To operationalize our time-control variables, we proceed as follows: Reddit submissions are timestamped in the UTC format when retrieved from the Pushshift submission dumps, but we recode submissions with our time-controlling categorical variables "weekday/weekend" and "time of day" using U.S. Pacific time. We do so for two reasons: a disproportionately high number of Reddit users live in a U.S. time zone [84]; and several violent BLM events happened on the West Coast, especially in Portland, Oregon, marking it an important area for (violent)

BLM protest [85]. For the variable "weekday/weekend," we classify Saturday and Sunday as "weekend" and all other days of the week as "weekday." We code the variable "time of day" as "morning" for submissions posted between 4am and 12pm, as "afternoon/evening" for those posted between 12pm and 8pm, and as "night" for those posted between 8pm and 4am.

For coding our other control variable, "type of news outlet," we also use information retrieved from the website mediabiasfactcheck.com and apply the categorical labels "newspaper," "magazine" (merging "magazine" and "journal"), "news agency," "organization/foundation," "radio station" (merging "radio" and "radio station"), "TV station" (merging "TV station" and "TV station/website"), and "website" (merging "website," "app/website," "website/newspaper," and "website/video"). In our analysis, the latter one is considered to be digital-born media, and the others are considered to be digital extensions of traditional media. (For more information regarding the variable distributions, please see Supporting Information, Descriptive overview of variables.)

Data collection and VGG19 image classifier

Social media platform Reddit. Although traditional media still serve as important news sources, consuming news through social media websites has become increasingly important as sources of news [5,86], even beginning to replace traditional media [87]. Thus, analyzing the relationship between digital news consumption and attention distributions on social media websites is essential for understanding one integral part of today's hybrid media system [7,8]. In our analysis, we focus on Reddit because it is often regarded as "the front page of the Internet" [88] and is one of the most popular social media news sources [5], known as a place to share new social trends, political opinions, and innovative ideas [89]. Moreover, Reddit has a unique structure compared to other social media platforms. It is divided into various subreddits in which "users actively choose to participate in specific discussion groups that interest them . . . rather than creating friend networks" [89]. Thus, "social connections are less salient on Reddit, which seems more centered on the content" [67], making the platform a particularly good choice as the foundation for our analysis of the discourse on and attention paid to the BLM movement.

When evaluating Reddit's communication and interaction landscape, it is critical to remember that its audience of users is skewed in certain ways that could affect results; for example, the social composition is biased toward American men of middle age with left-leaning political views [84]. To account for these biases to a degree we deem possible, we analyze three relatively politically neutral subreddits that are highly popular for sharing news on Reddit and are commonly used as subreddits of analysis in this context [see, e.g., 90]. Moreover, this study aims to analyze the effect of mainstream news reporting on user attention to capture the general relationship between shared news and user attention. Therefore, we intentionally focus on the most prominent subreddits for sharing news with strict community rules, mainly featuring mainstream media outlets [91]. We disregard more radical or fringe news subreddits on Reddit as they attract a different audience, operate according to different community rules, share different types of content, and, therefore, contain a different logic of attention distribution [92].

The first two subreddits are r/worldnews, with 28.6 million users (as of April 16, 2022), and r/news, with 24.5 million subscribers; both are premier subreddits for exchanging and discussing international news. We also include the subreddit r/politics, with 8 million members; it is dedicated exclusively to U.S.-specific news. These three subreddits have similar community rules and standards for user submissions (regarding title length, title content, and timeliness of news articles; see submission guidelines of each subreddit). Submissions posted in all three subreddits consist of a link to a news article and a title matching the title of the news article, which allows for a direct comparison; only one, r/politics, allows the posting of thumbnails. Since all three subreddits are dedicated to sharing news articles only, a similar type of image attached to a submission's linked article can be expected. This, in turn, enhances the comparability of all three subreddits and the submissions to them and paves the way for a correct categorization of the image classifier, which is also trained on protest-related news images.

Reddit BLM dataset. We use the monthly Pushshift submission dumps [88] to gather the data of the three subreddits, which is regularly done for collecting Reddit data [e.g., 93]. We downloaded the Pushshift submission dumps for May 2020 to May 2021 (105GB) on May 9, 2023, and filtered the data regarding the three subreddits of interest r/worldnews, r/politics, and r/news (830MB), which led to 1,887,202 submissions. With our objective of evaluating the BLM movement reignited in 2020, we reduced the data toward the period of analysis from May 25, 2020, the day of George Floyd's murder, to May 25, 2021. This yielded 1,763,659 submissions. We removed all submissions whose titles did not match one of the three keywords, "black lives matter," "george floyd," or "blm," leaving a total of 17,083 submissions relevant to our analysis.

Note that we also tested how an increased keyword set related to our analysis topic would increase our data corpus. We, therefore, added the BLM-specific keywords "police violence," "breonna taylor," and "rayshard brooks" to our keyword set; this, however, did not result in a significant increase in submissions and added more noise to the dataset (for more information, see Supporting Information, Data availability statement). We thus decided to continue working with the initial keyword set.

For our data cleaning process, we removed a particular type of submission from our dataset, so-called "megathreads," which summarizes a multitude of submissions about highly popular events and typically receives a disproportionate amount of attention; this reduced the number of submissions to 17,069. After dropping all missings regarding our control variables from mediabiasfactcheck.com ("political leaning" (*H4*), "factual reporting" (*H5*), "traffic" (*H6*), and "type of news outlet"), we received 11,813 submissions.

Given our goal to analyze the impact of images displaying violence, the next step is to collect the images from the news articles referenced in our dataset's submissions. We use the images from the linked news outlets because the three subreddits allow only submissions with news titles and news article links, thus setting the focus on the originating news outlet and its images. We assume that users who engage with the submission also click on the URL link, as it is the primary purpose of the submission itself. We use the Python library "Newspaper3k" to extract the main images (also used as thumbnails in r/politics) from each news article in our dataset. The "main image" of an article is defined within the source code of each website and typically corresponds to the first image on the website. Note that if Newspaper3k could not extract an image due to, for example, captchas of a news outlet's website, we manually extracted the main image in the cases possible. To test how well Newspaper3k extracted the correct main image of a news article, we manually evaluated 200 random images retrieved by Newspaper3k. Of 200 images, 152 were correctly detected as the main image (76.0%). As nearly all other images that were not labeled as the main image were still part of the news article and most often the second or third image on the website, we deemed this result acceptable and continued working with those images. After downloading the images detected by Newspaper3k and manually added by us, we retrieved 9,373 images for our dataset (no image was available for 2,440 submissions).

For our analysis, we are interested in submissions that received at least a minimum of attention and thus have been recognized by Reddit users. We, therefore, removed all submissions that received zero comments which reduced the dataset to 5,876 submissions (yet, to sustain our results, we also ran a negative binomial regression model that included submissions with zero comments, yielding similar results for the main effects; for more information, see Supporting Information, Negative binomial regression models). After dropping all missing values of the other variables related to our analysis, our final dataset consists of 5,873 observations.

Training dataset for image classifier

Prelabeled training and validation datasets are necessary to build an automated image classifier that can label an image as violent (= 1) or nonviolent (= 0). Therefore, we use images from the "UCLA Protest Image Dataset" of Won et al. [94], a prelabeled protest-related dataset of 40,764 images. In this dataset, "violence" is operationalized as a continuous variable; we transfer that labeling to our dichotomous coding scheme by manually applying Rucht's [68] violence definition to the images. Using their dataset's most- and least-violent images, we end up with 2,566 images in the training dataset and 700 in the test dataset. Moreover, to train our image classifier with general protest images *and* BLM-specific images, we automatically download images via the search engine "Bing," employing 25 BLM and protest-related related keywords in English and German (for the entire keyword set, see Supporting Information, Data availability statement). We classify them following the same procedure outlined above. Finally, we enlarge our dataset using image augmentation, resulting in 7,972 violent images and 8,986 nonviolent images in the training dataset and 2,854 violent images and 4,088 nonviolent images in the validation dataset (for more information related to the dataset creation, please see Supporting Information, Data availability statement).

Transfer learning & performance of VGG19 image classifier

We use a convolutional neural network architecture to classify violent images for two reasons: the computational power and memory space of CNNs are relatively inexpensive due to parameter sharing; and CNNs autonomously select important image features based on automatized feature extraction and without human supervision, thus achieving state-of-the-art results while avoiding human biases [95] (see Supporting Information, Image classifiers, for different models that we tested but that performed worse than CNNs).

Research has repeatedly shown that using a CNN pre-trained on general image features increases the model performance once finetuned to the specific classification task through transfer learning [see, e.g., 96]. We follow this advice and apply the three state-of-the-art CNN architectures VGG19, ResNet50, and EfficientNetB4 [97,98] pertained on the ImageNet-1k dataset [99] from the Python deep learning libraries "TensorFlow" and "Keras." In addition, we create our own CNN architecture as a fourth model for comparison. As the VGG19 model performed best after finetuning and running the different models, we continue utilizing this CNN architecture for our analysis. In the following, we go into more detail regarding the training and performance of the VGG19 model (for more information regarding the architecture, training, and evaluation of the other models, please see Supporting Information, Image classifiers).

We use transfer learning to adjust the VGG19 model to our classification context of classifying images as violent (1) or nonviolent (0). Therefore, we freeze all layers except for the last fully connected layer of the pre-trained model, as unfreezing more layers does not yield better training results. We replace the last fully connected layer with a dense layer that applies a sigmoid function as an activation, which returns a number between 0 and 1 and is, therefore, suitable for our binary image classification. We compile the model with the adam optimizer, a learning rate of 0.001, and a binary cross-entropy (BCE) loss function, resulting in a binary classification corresponding to our classification of "violent" and "nonviolent." We split the training data into 70% training images and 30% validation images. We also implement an early stop callback function with a minimum delta of 0.1% and a patience parameter of 10 within our training process to prevent overfitting. Thus, our training will stop if our validation loss does not improve within ten epochs by at least 0.1% compared to its best value. We start training the model with 40 epochs and a batch size of 32. As the model's validation loss continuously improves by at least 0.1%, the training ends after 40 epochs.

Our VGG19 image classifier achieves an accuracy value of 91.65% on the training data and 90.29% on the validation data. When assessing the performance on the test dataset, the classifier achieves an accuracy of 91.84%, which is comparable to the training and validation accuracy. The classifier achieves a precision score of 87.78%, a recall score of 93.13%, and an F1 score of 90.37% on the test dataset. Moreover, when assessing the performance of our BCE loss function, both the training and validation loss continuously decreased, which is a sign that the model picks up essential image features without overfitting the data (in epoch 40, a loss of 0.22 on our training data and 0.24 on our validation data). Thus, the classifier's performance is sound and allows for a valid image classification (for more information on the VGG19 model performance, please see Supporting Information, Image classifiers).

Results

As a first step, we classify our BLM images with our VGG19 image classifier, resulting in 5,486 images classified as "nonviolent" (93.4%) and 387 as "violent" (6.6%). To validate whether our image classifier performs as well on our BLM dataset as on the training, validation, and testing set, we manually label 200 random BLM images and compare our classification result with the classifier's results. Of 200 images, 180 are correctly labeled (90.0%), which aligns with the previous validation metrics and therefore supports the classifier's performance in our BLM protest context. Moreover, when calculating the sentiment of a submission's title with the BERT model, we receive 1,253 (21.3%) positively connoted titles and 4,620 (78.7%) negatively connotated titles. To validate these classification results, we draw a random sample of 200 submission titles and manually label them as "positive" or "negative." The manual and automatic sentiment classification matches in 168 out of 200 cases (84.0%). This value is in line with the previous validation and test accuracy of the BERT model, supporting the model's sound performance on our Reddit BLM data. We thus have all the necessary data to run our data analysis (for a description of all variables, see Supporting Information, Descriptive overview of variables).

The count nature of our dependent variable, "number of comments," makes a Poisson or negative binomial regression model most suitable for the analysis. To test whether a Poisson or negative binomial regression would be more appropriate for our data, we conduct a likelihood ratio test for overdispersion in count data ("odTest" function from the R package "pscl"). The likelihood ratio test estimates whether the data's variance equals its mean, which is the assumption of a Poisson model. The test results show that our model's data's variance is greater than the mean (which is also supported by the positively skewed data distribution of our dependent variable "number of comments"), implying a significant data overdispersion. Thus, a negative binomial regression, which relaxes the assumption regarding variance equaling mean, is a more suitable model choice for our data [for more information, see 100] (for more information regarding our model choice, please see Supporting Information, Negative binomial regression models).

To run the negative binomial regression, we use the "glm.nb" function from the R package "MASS." Note that we did not standardize the variables before running the models, as we did not include any interaction terms and have only one numeric control variable, "number of cross-posts." Likewise, standardizing the variables by subtracting the mean and dividing them by the standard deviation before running a negative binominal regression impedes the interpretability of the results. Thus, it is often omitted when performing this type of analysis [for more information, see, e.g., 101]. To further check the suitability of our model, we control for multicollinearity within our model through the variance inflation factor (VIF). Doing so, we detect no worrisome correlation between our variables (all values of VIF < 5, with the variables "type of news outlet" displaying the highest values of 1.8). Moreover, we check for outliers in our data which could bias the results. Therefore, we apply Cook's distance but detect no data point that surpasses the critical value of 1 [102]. Thus, our model is suitable for our analysis by providing meaningful results.

Table 2 shows the results of our negative binomial regression analysis (to test the robustness of our model results, we ran different negative binomial regression models based on variations of the dataset and variables; when comparing the results, only small deviations from our results can be detected, sustaining our findings; for more information, see Supporting Information, Negative binomial regression models).

Note that the following interpretation of the individual coefficients assumes that all other variables are held constant. Also, the nature of social media data only allows for correlative associations between the variables and cannot be causal. Thus, the following result will be interpreted against this background.

When evaluating the coefficient of our main independent variable, "violent image" (-0.012), we find–contrary to our assumption–that images featuring violence are negatively associated with the number of comments, yet not on a statistically significant basis (*H1*). Therefore, this finding cannot be generalized. When analyzing the results of attributes related to the news item, the coefficient of our title's sentiment analysis with BERT turns out to be significant (*H2*), with a p-value of 0.01, demonstrating that negative titles increase the logged count of number of comments by 0.132. Thus, our results show that a submission's title not only has a significant (see p value) but also a greater (see coefficient) effect on user attention compared to images (*H3*).

When evaluating the effects of attributes related to the news provider, conservative and liberal news outlets significantly negatively impact our dependent variable compared to neutral news outlets ($p = 0.01$ and $p = 0.1$, respectively, coefficient of -0.211 and -0.099, respectively) (*H4*). Both high and low factual reporting of a news outlet significantly impact the dependent variable compared to mixed factual reporting (*H5*). High factual-reporting news outlets positively affect the logged count of number of comments by 0.083 compared to mixed factual-reporting outlets. In contrast, low factual-reporting news outlets are negatively correlated with a coefficient of -1.005 (p value = 0.05 and 0.001, respectively) compared to mixed factual reporting outlets. Analyzing the effects of a news website's traffic (*H6*), both news outlets with high and minimal traffic are negatively associated with the logged count of number of comments compared to news outlets with medium traffic (both with a p value of 0.001 and a coefficient of -0.246 and -1.198, respectively).

When examining the variables related to Reddit-specific affordances and community rules, the type of subreddit (*H7*) plays a significant role. Both r/news and r/worldnews positively correlate with the logged count of number of comments compared to r/politics (both a p value of 0.001 and a coefficient of 0.977 and 0.764, respectively. Whether a submission is NSFW (*H8*) does not substantially impact the dependent variable. Yet, link flair (*H9*) and the number of cross-posts (*H10*) turn out to be highly significant, with a p-value of 0.001. If a submission is tagged with a link flair, it negatively correlates with the logged count of number of comments by -0.566, whereas the number of cross-posts is positively associated with the dependent variable with 0.896.

Table 2. Negative binomial regression model (standard errors in parentheses, p values in square brackets).

	dependent count variable: number of comments
violent image	-0.012 (0.072) [0.866]
BERT sentiment negative	0.132** (0.044) [0.003]
political leaning conservative	-0.211** (0.073) [0.004]
political leaning conspiracy	-0.215 (0.990) [0.828]
political leaning liberal	-0.099• (0.058) [0.087]
factual reporting high	0.083* (0.042) [0.046]
factual reporting low	-1.005*** (0.124) [0.000]
traffic high	-0.246*** (0.066) [0.000]
traffic minimal	-1.198*** (0.218) [0.000]
subreddit news	0.977*** (0.048) [0.000]
subreddit worldnews	0.764*** (0.069) [0.000]
NSFW	-1.264 (0.804) [0.116]
link flair	-0.566*** (0.037) [0.000]
number of cross-posts	0.896*** (0.014) [0.000]
weekend US pacific	0.249*** (0.042) [0.000]
morning US pacific	-0.005 (0.039) [0.898]
night US pacific	-0.097• (0.050) [0.054]
type of news outlet magazine	0.263** (0.086) [0.002]

(*Continued*)

Table 2. (Continued)

	dependent count variable: number of comments
type of news outlet news agency	-0.766*** (0.102) [0.000]
type of news outlet organization/foundation	-0.123 (0.182) [0.501]
type of news outlet radio	-0.375* (0.154) [0.015]
type of news outlet TV station	0.052 (0.047) [0.266]
type of news outlet website	0.096• (0.064) [0.064]
Constant	3.276*** (0.093) [0.000]
Observations	5,873
Null deviance	14,121 on 5,872 df
Res. deviance	6,903 on 5,849 df
AIC	50,416
Theta	0.557
Std. Error	0.009

Note: •p<0.1; *p<0.05; **p<0.01, ***p<0.001.

https://doi.org/10.1371/journal.pone.0288962.t002

Finally, when evaluating the effects of our control variables, it turns out that if a submission is posted on a weekend compared to a weekday, the logged count number of comments is positively affected by 0.249 (p value = 0.001). The timing of posting a submission also shows a significant effect. Submissions published at night negatively impact the dependent variable by -0.097 (p value of 0.1) compared to submissions posted during the afternoon/evening (note that for UTC, U.S. Central, U.S. Eastern times, and Central European times, we receive the same effects for the weekend). When examining the effect of the type of news outlet, digital-born media types–namely websites–and digital extensions of traditional media, namely, magazines, news agencies, and radio stations–have a significant effect compared to the reference category newspaper (which belongs to the category digital extension of traditional media outlets). Websites have a significant positive association with 0.096 and a p value of 0.1. Submissions that feature links to news agencies or radio stations significantly negatively correlate with the dependent variable (coefficient of -0.766 and -0.375, respectively, and a p value of 0.001 and 0.05, respectively). In contrast, submissions featuring links to magazines have a positive impact of 0.263 and a p value of 0.01 compared to newspapers (note that newspaper was chosen as a reference category because it has the greatest number of observations).

When comparing the AIC (Akaike Information Criterion) value of our model with those of the other control models (see Supporting Information, Negative binomial regression models), our model is either as suitable or far more suitable for representing the data (AIC value is considerably smaller compared to two models and nearly as big as the AIC of another model). Thus, our model fits the data well compared to the other models, making it a reliable choice for analyzing our data. Also, as the residual deviance is significantly smaller than the null deviance, we can infer that our model explains a significant amount of the variation in the data compared to the null model.

Discussion

Based on an evaluation of our empirical results, we can summarize the answer to our research questions as follows: violent images *do not* affect user attention on Reddit, whereas other attributes related to news reporting–such as the political leanings and factual reporting of a news outlet, a news outlet's title, Reddit community rules and subreddits, the timing of a submission, and a news outlet's media type–are *significantly* correlated. In this section, we discuss these results in the context of our hypotheses by providing theoretical and practical implications.

In light of our empirical findings, we cannot verify that a violent visual representation leads to greater user attention and must therefore reject *H1*. Our findings align with previous studies that could not detect a significant relationship between online images featuring violence and attention in a protest context [16,17]. It may be that violent images depict a form of negativity that is too intensive for social media users who prefer less negativity in their news consumption [3] or prefer shallow social content for entertainment purposes [66]. Thus, our results confirm that image virality on social media–such as the images of cats that enjoy such tremendous popularity [67]– is highly context-sensitive.

In addition, we need to keep in mind that in our current "attention economy" [103], there is an abundance of social media content competing for user attention, which might diminish the effect of images featuring violence. Thus, our analysis shows that it is difficult to replicate existing findings from traditional media research regarding the effects of news featuring violence on attention to the online world. Although previous research has shown that offline news consumption habits can be transferred online [60], one must remember that previous research focused on the type of news outlet readers prefer. Thus, our findings should not be seen as a contradiction to earlier findings but rather as a supplement focusing on the type of reporting.

Moreover, one must note that news consumption on social media follows an entirely different logic than the linear news exposure of traditional media [104]. This could partially explain the divergent effects of news featuring violence in different media settings. Future research could expand these insights by combining quantitative and qualitative approaches, as proposed by Mitchelstein and Boczkowski [61].

Beyond these theoretical explanations for our non-confirmative findings, we also need to consider some empirical research design decisions that may have affected our results. First, the operationalization of "violence" according to Rucht's [68] definition is quite strict. When manually assessing the BLM image data, we find that several images could be identified as "violent" were we to adopt a reasonably strict but more inclusive definition.

Second, a binary operationalization is prone to classification errors. The portrayal of violence in images is a nuanced and complex concept and operationalizing it by utilizing a continuous variable would preserve additional information, which might improve our results.

Third, we analyzed the main images used in the news articles submitted to Reddit. While we consider this a reasonable choice, essentially restricting our analysis to the strongest available signal for the visual portrayal of violence in images, extending our research to additional visual depictions of violence on websites linked in submissions would be interesting. Likewise, the subreddits utilized for our analysis prohibit directly posting images within submissions. As these subreddits are the most prominent for news sharing and, thus, are important for analyzing user attention in relation to news articles, we consider this choice appropriate. Yet, based on the subreddits' restrictions on directly posting images, we are likely underestimating the effect of (violent) images. We invite future research to extend our analysis by including subreddits related to the sharing of news that allow for the direct posting of images.

Finally, the operationalization of user attention could be extended by, for example, evaluating the content of comments made to the analyzed submissions. That would allow for the analysis of interaction dynamics triggered by violent content.

While we do not find a significant effect of the portrayal of violence in images, several of our other news-related attributes show promising results. Foremost among these, we find when analyzing attributes related to the news item that negative titles correlate significantly with the attention paid to a submission on Reddit in the context of our study. Hence, we can confirm *H2* and the implications of the negativity bias hold for the relationship between text with a negative sentiment and user attention. This finding aligns with previous results on the positive relationship between negative emotional language and information spread online [3,105]. In addition, it emphasizes the importance of "negativity" when reporting news [64]. Yet, when evaluating these findings, the internal structure of the considered subreddits needs to be taken into account. According to the rules of the subreddits we analyzed, a valid submission features a title and a link to a news outlet's website. Submissions not following these rules are removed immediately, indicating that Reddit and its communities mediate and restrict the information exchanged. These moderating rules emphasize a submission's title, potentially enhancing its effect on user attention. This finding again highlights the importance of considering the logic of a subreddit's affordances and community rules to analyze user attention [69].

In contrast to our assumptions, we cannot confirm that images significantly impact user attention more than written text. Considering that our study shows that a submission title's negative sentiment has a greater coefficient and more significant effect on user attention than violence featured in the accompanying images, the picture superiority effect does not hold in our analysis. Thus, we must reject *H3*. Other studies have come to similar conclusions that although news shared on a "news aggregator website" featuring images positively affects the readership, images negatively affect the number of comments an article receives due to their distracting character [106]. Since Reddit is a social media website known for encouraging the posting of comments and the discussion amongst users [107], it operates entirely differently than news aggregator websites. Thus, we consider our results as complementary to previous research. As an extension of our work, future research could analyze subreddits with diverging community guidelines, which could impact the prominence of images and, thus, the attention paid to them.

We observe several interesting effects when analyzing the impact on user attention of attributes related to the news provider. In contrast to previous research that found no relationship, in the context of negative news, between a news outlet's political stance and user engagement [105], we are able to show that conservative and liberal news outlets correlate with user attention negatively compared to neutral reporting–contrary to our assumptions. Thus, we must reject *H4*. Therefore, our findings indicate that users interested in BLM favor a politically neutral reporting style when informing themselves about news posted in the three subreddits of analysis. Since a tense political climate surrounds BLM [29], one reason for this finding could be that users prefer their information consumption to be less politically charged.

Unlike previous research [75], we did find a significant positive correlation between high factual reporting ("true news") and a negative correlation between low factual reporting ("false news") and user attention compared to mixed factual reporting. Consequently, we must reject *H5*. One possible explanation is that previous research focused on the relationship between the rapid spread of tweets featuring misinformation and user attention [75], whereas we focus on the effects on user attention of a digital news outlet's reporting style. Thus, it seems that the audience for news about BLM is more interested in factual reporting, which could be attributed to the topic's political relevance and the desire to counter existing polarized positions [29]. This finding also aligns with the impression that users interested in BLM on Reddit favor less politically charged news content. Future research could build upon these initial findings and dive deeper into reporting's level of factuality and attention dynamics on social media.

In *H6*, we assumed a positive relationship between the popularity of a news website and the attention received by submissions to it. Our data, however, reveal a negative correlation between very popular and very unpopular news websites with user attention. Therefore, we reject *H6*. This finding can be seen as in line with the findings regarding factual reporting and political leaning: When looking into our data, popular news outlets often feature a low- to mixed factual reporting style and tend to be politically charged (e.g., Fox News, Daily Mail, or The Guardian). Thus, we observe a tendency that users prefer medium popular outlets that are visible (not minimal traffic), with factual and politically unbiased reporting (e.g., Bring Me The News or KMSP Fox 9).

The effects of our attributes related to Reddit-specific variables also lead to interesting findings that highlight the importance of a platform's affordances and community moderating rules. While we cannot confirm that the attachment of a NSFW tag (*H8*) significantly correlates with user attention, we see that a subreddit in which a submission is posted (*H7*), the flair tag (*H9*) attached to a submission, and the number of cross-postings (*H10*) does correlate with user attention. Based on these findings, we again cannot confirm that more violence (NSFW) leads to greater user attention. We refrain, though, from overstating the significance of this finding due to the variable's highly skewed distribution (see Supporting Information, Descriptive overview of variables).

Likewise, the picture superiority effect also does not hold for the posting of thumbnails (r/politics), which highlights again the prominence of a submission's title and the particular structure of the subreddits we considered. It is interesting to observe that r/news and r/worldnews correlate positively with user attention compared to r/politics. This could be a sign that users interested in BLM are rather looking for an international perspective on BLM (r/news, r/worldnews) as it is a worldwide protest [21] compared to solely reading BLM news related to the US context (r/politics).

Finally, a subreddit's community rules and moderation (link flair) significantly affect whether a post will be visible to Reddit users and thus receive attention. The positive relationship between cross-posts and user attention also confirms this effect of visibility on user attention. Therefore, understanding Reddit's specific affordances and community rules (which differ highly between the multitude of subreddits hosted on Reddit) is important for capturing attention distributions on the platform. Summarizing these findings, we must reject *H7* and *H8*, but we confirm *H9* and *H10*.

When evaluating the effects of our time controls, "weekend" and "time of day" turn out to be significant. Our results show that Reddit users are more active during the weekend and the afternoon/evening, resulting in higher attention paid to submissions posted on the weekend and during that period (12pm to 8pm) than those posted at night. The first finding is unsurprising, as users tend to have more time on weekends to engage with social media than on weekdays. In contrast, it is somewhat surprising that users tend to consume and comment on news more often during the day compared to the night (negative effect of submissions posted during the night). Yet, this finding aligns with our impression that users of the subreddits we analyzed who are interested in BLM prefer to consume their news in an organized fashion (significance of regular working hours and not at night). However, these results based on U.S. time zones must be treated with caution since BLM is an international movement with an audience that likely engages with BLM news content from different time zones across the globe. Nevertheless, these U.S.-centered findings are still interesting to take into account, especially given that most Reddit users are U.S.-based and the BLM movement originates from the United States.

Regarding the type of news outlet, we also find significant effects on user attention. In our analysis, websites considered digital-born news outlets and magazines, news agencies, and radio stations considered digital extensions of traditional media significantly correlate with user attention compared to newspapers. We can observe that easy-to-read news outlets such as magazines (e.g., *Newsweek*, *Rolling Stone*) and specific websites (e.g., Buzzfeed News) are positively associated with user attention compared to newspapers. Moreover, news agencies such as Reuters are negatively correlated with user attention, which is not surprising given that the primary role of news agencies is to distribute news to their customers compared to publishing news themselves. Likewise, users seem to prefer written content compared to audio content (negative association with radio stations). These findings indicate that the types of news outlets users favor when informing themselves about BLM on Reddit should be factual but not too cumbersome to read. Users have a positive relationship with digital-born news outlets, whereas they have mixed feelings regarding the digital extension of traditional media, as news outlets correlate positively and negatively with user attention.

We want to stress that these findings relate solely to the three subreddits analyzed within this study and are, therefore, not generalizable to other subreddits or social media platforms. Reddit's subdivision into different subreddits makes its structure highly different than other social media platforms such as Instagram and TikTok. Thus, one must keep in mind that each subreddit and social media platform is differently structured, operates according to different community rules, and entails divergent platform affordances [108], which impacts the salience of different news topics shared and the user engagement and attention paid to them [109].

Conclusion and limitations

Our work contributes to research on the dynamics of user attention on social media in the context of digital news reporting on the BLM movement on Reddit. This work is a first important step for improving our general understanding of attention dynamics on social media in the context of digital news reporting, especially violent news reporting, about social movements, which in turn can influence the perception of and support for such a movement [22–24,26]. Since the consumption of news on social media platforms has become more important than ever before [3–5], this analysis provides valuable insights into a growing news market that is an essential part of the hybrid media system [7,8].

Our six key takeaways are:

- Findings from traditional media research cannot be directly replicated within the online world. Although research suggests that offline and online news consumptions often follow similar patterns [60,61], one must take into account that both environments follow different logics [104]. On social media, in particular, an abundance of content competes for user attention [103], which differs from the linear news exposure in traditional media.
- In our analysis, it is not a singular effect (violence in images) but the interplay of several attributes related to the news item, news provider, and community rules and platform affordances that are essential for capturing attention distributions on Reddit. This highlights the complexity of the social media landscape.
- Submission titles with negative connotations correlate positively with user attention compared to titles with positive sentiment. Thus, we find indications for a negativity bias with respect to written content.

- The picture superiority effect cannot be confirmed in our analysis, given that a submission's headline has a greater coefficient and a significant impact on user attention compared to submissions featuring violent images. In our discussion, we point to several potential explanations for this finding.
- Users consuming BLM-related news tend to exhibit special characteristics by preferring negatively connotated yet politically non-biased, factual, and easy-to-read news published during the day and on weekends. Thus, taking a news topic's audience into account is highly important for understanding the popularity of news content.
- Platform-inherent affordances and community rules make essential contributions to the understanding of user attention dynamics, demonstrating that "both crowds and systems work together to make decisions on what information is important" [3].

These findings emphasize the need to expand our understanding of news consumption and attention distributions in an increasingly digitalized world since "an operational and useful democracy relies on an educated and well-informed population" [110]. This paper contributes another element to this field of research and paves the way for potentially exploring broader questions, such as whether the detected effects of this paper also hold for the digital news coverage of other social movements on other social media platforms.

While the findings presented in this study offer important theoretical and practical insights, they also have to be seen in the light of their limitations—which open up several promising avenues for future research. First, it would be interesting to measure the dynamics of user attention in a news context over time. Doing so would allow for evaluating social media's effect on digital news consumption dynamics by revealing what kind of news attracts the greatest user attention and when.

Second, while we deem our choice of subreddits as a justifiable first step for analyzing the relationship between news and user attention since they represent very important subreddits for sharing news on Reddit, they are not free of biases that affect our study. As the three subreddits do not allow for the direct posting of images, we recognize that we are underestimating the effect of (violent) images on user attention. Future research is invited to add further news-related subreddits with less strict community rules allowing for the direct posting of images. Further, while we intentionally focused on mainstream news outlets in our analysis to capture the general relationship between shared news and user attention, it would be interesting to expand our analysis to more radical subreddits with different community rules, audiences, and content compared to mainstream subreddits, including those that encourage the sharing of conspirational and fringe news. This extension would provide additional and interesting insights regarding the relationship between news and user attention on Reddit. Beyond including more and diverse subreddits, adding other social media platforms and data about different social movements, as well as including additional explanatory attributes for user attention, would extend the scope of this research and add to its generalizability.

Third, analyzing the comment structure and user-generated content in response to a submission could be a welcome addition to investigate user attention further. Focusing on interaction dynamics between users in the context of content featuring violence could deepen our understanding of user reactions to violence portrayed online.

Supporting information

S1 Fig. Histogram of monthly distribution of Reddit BLM submissions.
(TIF)

S2 Fig. Histogram of dependent variable number of comments.
(TIF)

S3 Fig. Training and validation accuracy of BERT model.
(TIF)

S4 Fig. Accuracy of CNN models.
(TIF)

S5 Fig. Accuracy of VIT models.
(TIF)

S6 Fig. VGG19 model loss function.
(TIF)

S7 Fig. VGG19 model precision, recall, and F1-Score.
(TIF)

S8 Fig. VGG19 model receiver operating characteristic (ROC) curve.
(TIF)

S9 Fig. VGG19 model confusion matrix of test dataset relative values.
(TIF)

S10 Fig. VGG19 model confusion matrix of test dataset absolute values.
(TIF)

S1 Table. Descriptive overview of variables.
(DOCX)

S2 Table. Own convolutional neural network architecture.
(DOCX)

S3 Table. Negative binomial regression results of models with different datasets and variables (standard error in parentheses, p values in square brackets).
(DOCX)

S1 File. Appendix–Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit.
(DOCX)

S2 File.
(ZIP)

Author Contributions

Conceptualization: Theresa Henn, Oliver Posegga.

Data curation: Theresa Henn, Oliver Posegga.

Formal analysis: Theresa Henn, Oliver Posegga.

Funding acquisition: Theresa Henn, Oliver Posegga.

Investigation: Theresa Henn, Oliver Posegga.

Methodology: Theresa Henn, Oliver Posegga.

Project administration: Theresa Henn, Oliver Posegga.

Resources: Theresa Henn, Oliver Posegga.

Software: Theresa Henn, Oliver Posegga.

Supervision: Oliver Posegga.

Validation: Theresa Henn, Oliver Posegga.

Visualization: Theresa Henn, Oliver Posegga.

Writing – original draft: Theresa Henn.

Writing – review & editing: Theresa Henn, Oliver Posegga.

References

1. Stieglitz S, Dang-Xuan L, Bruns A, Neuberger C. Social Media Analytics: An Interdisciplinary Approach and Its Implications for Information Systems. Business & Information Systems Engineering. 2014;6. https://doi.org/10.1007/s12599-014-0315-7
2. Newman N, Fletcher R, Levy DAL, Nielsen RK. Reuters Institute Digital News Report 2016. Oxford: Reuters Institute for the Study of Journalism; 2016 p. 124.
3. Horne BD, Adalis S. The Impact of Crowds on News Engagement: A Reddit Case Study. The Workshops of the Eleventh International AAAI Conference on Web and Social Media. AAAI Technical Report; 2017. pp. 751–758. https://doi.org/10.1609/icwsm.v11i1.14977
4. Mitchell A, Simmons K, Matsa KE, Silver L. Publics Globally Want Unbiased News Coverage, but Are Divided on Whether Their News Media Deliver. Pew Research Center. 2018 pp. 3–43.
5. Walker M, Matsa KE. News Consumption Across Social Media in 2021. In: Pew Research Center [Internet]. 2021 [cited 6 Jun 2023]. Available from: https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/.
6. Bridgman A, Merkley E, Loewen PJ, Owen T, Ruths D, Teichmann L, et al. The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. Harvard Kennedy School Misinformation Review. 2020;1. https://doi.org/10.37016/mr-2020-028
7. Chadwick A. The Hybrid Media System: Politics and Power. Oxford University Press; 2013.
8. Jungherr A, Posegga O, An J. Discursive Power in Contemporary Media Systems: A Comparative Framework. International Journal of Press/Politics. 2019. https://doi.org/10.1177/1940161219841543
9. Boulianne S. Online news, civic awareness, and engagement in civic and political life. New Media and Society. 2016;18. https://doi.org/10.1177/1461444815616222
10. Catalan-Matamoros D, Peñafiel-Saiz C. How is communication of vaccines in traditional media: a systematic review. Perspectives in Public Health. 2019;139. https://doi.org/10.1177/1757913918780142 PMID: 29877122
11. Scharrer E. Media exposure and sensitivity to violence in news reports: Evidence of desensitization? Journalism and Mass Communication Quarterly. 2008; 85: 291–310. https://doi.org/10.1177/107769900808500205
12. Bartholow BD, Dill KE, Anderson KB, Lindsay JJ. The Proliferation of Media Violence and Its Economic Underpinnings. Media violence and children: A complete guide for parents and professionals. Westport: Greenwood Publishing; 2003. pp. 1–18.
13. Barber N. Why the News is Loaded with Violence. In: Psychology Today [Internet]. 2016 [cited 6 Jun 2023]. Available from: https://www.psychologytoday.com/us/blog/the-human-beast/201601/why-the-news-is-loaded-violence.
14. Abubakar AT. News Values and the Ethical Dilemmas of Covering Violent Extremism. Journalism and Mass Communication Quarterly. 2020; 97: 278–298. https://doi.org/10.1177/1077699019847258
15. Hamilton JT. Channeling violence: The economic market for violent television programming. Princeton University Press; 2000.
16. Rossi L, Neumayer C, Henrichsen J, Beck LK. Measuring Violence: A Computational Analysis of Violence and Propagation of Image Tweets From Political Protest. Social Science Computer Review. 2022; 41: 905–925. https://doi.org/10.1177/08944393211055429
17. Kharroub T, Bas O. Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. New Media and Society. 2016; 18: 1973–1992. https://doi.org/10.1177/1461444815571914

18. Diani M. The concept of social movement. The Sociological Review. 1992; 40: 1–25. https://doi.org/10.1111/j.1467-954X.1992.tb02943.x
19. Chernega J. Black Lives Matter: Racialised Policing in the United States. Comparative American Studies. 2016; 14: 234–245. https://doi.org/10.1080/14775700.2016.1267322
20. Ray R, Brown M, Fraistat N, Summers E. Ferguson and the death of Michael Brown on Twitter: #BlackLivesMatter, #TCOT, and the evolution of collective identities. Ethnic and Racial Studies. 2017; 40: 1797–1813. https://doi.org/10.1080/01419870.2017.1335422
21. Patnaude L, Lomakina CV, Patel A, Bizel G. Public Emotional Response on the Black Lives Matter Movement in the Summer of 2020 as Analyzed Through Twitter. International Journal of Marketing Studies. 2021; 13: 69. https://doi.org/10.5539/ijms.v13n1p69
22. Mourão RR, Brown DK. Black Lives Matter Coverage: How Protest News Frames and Attitudinal Change Affect Social Media Engagement. Digital Journalism. 2022; 10: 626–646. https://doi.org/10.1080/21670811.2021.1931900
23. Boyle MP, McLeod DM, Armstrong CL. Adherence to the protest paradigm: The influence of protest goals and tactics on news coverage in U.S. and international newspapers. International Journal of Press/Politics. 2012; 17: 127–144. https://doi.org/10.1177/1940161211433837
24. Hertog JK, McLeod DM. Anarchists Wreak Havoc In Downtown Minneapolis: A Multi-level Study of Media Coverage of Radical Protest. Journalism & Mass Communication Monographs. 1995;151.
25. Chan JM, Lee C-C. Journalistic paradigm on civil protests: A case study of Hong Kong. The news media in national and international conflict. Westview Press; 1984. pp. 183–202.
26. McCurdy P. Social movements, protest and mainstream media. Sociology Compass. 2012; 6: 244–255. https://doi.org/10.1111/j.1751-9020.2011.00448.x
27. Koopmans R. Movements and media: Selection processes and evolutionary dynamics in the public sphere. Theory and society. 2004; 33: 367–391. https://doi.org/10.1023/B:RYSO.0000038603.34963.de
28. Smith J, McCarthy JD, McPhail C, Augustyn B. From Protest to Agenda Building: Description Bias in Media Coverage of Protest Events in Washington, D.C. Social Forces. 2001; 79: 1397–1423. https://doi.org/10.1353/sof.2001.0053
29. Dave D, Friedson A, Matsuzawa K, Sabia J, Safford S. Black Lives Matter Protests, Social Distancing, and COVID-19. Cambridge: National Bureau of Economic Research; 2020.
30. Corrigall-Brown C. The Power of Pictures: Images of Politics and Protest. American Behavioral Scientist. 2012; 56: 131–134. https://doi.org/10.1177/0002764211419358
31. Nicholls T, Shabbir N, Nielsen RK. Digital-born News Media in Europe—Digital News Project 2016. Oxford: Reuters Institute; 2016.
32. Vara-Miguel A. Cross-national similarities and differences between legacy and digital-born news media audiences. Media and Communication. 2020; 8: 16–27. https://doi.org/10.17645/mac.v8i2.2733
33. Meraz S. The fight for "how to think": Traditional media, social networks, and issue interpretation. Journalism. 2011; 12: 107–127. https://doi.org/10.1177/1464884910385193
34. Rozin P, Royzman EB. Negativity bias, negativity dominance, and contagion. Personality and Social Psychology Review. 2001; 5: 296–320. https://doi.org/10.1207/S15327957PSPR0504_2
35. Shoemaker PJ. Hardwired for news: Using biological and cultural evolution to explain the surveillance function. Journal of Communication. 1996; 46: 32–47. https://doi.org/10.1111/j.1460-2466.1996.tb01487.x
36. Stenberg G. Conceptual and perceptual factors in the picture superiority effect. European Journal of Cognitive Psychology. 2006; 18: 813–847. https://doi.org/10.1080/09541440500412361
37. Stefanov P, Darwish K, Atanasov A, Nakov P. Predicting the Topical Stance and Political Leaning of Media using Tweets. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. pp. 527–537. https://doi.org/10.18653/v1/2020.acl-main.50
38. Ahlers D. News consumption and the new electronic media. Harvard International Journal of Press/Politics. 2006; 11: 29–52. https://doi.org/10.1177/1081180X05284317
39. Scharkow M, Mangold F, Stier S, Breuer J. How social network sites and other online intermediaries increase exposure to news. Proceedings of the National Academy of Sciences of the United States of America. 2020; 117: 2761–2763. https://doi.org/10.1073/pnas.1918279117 PMID: 31988122
40. Nielsen RK. News media, search engines and social networking sites as varieties of online gatekeepers. Rethinking Journalism Again: Societal Role and Public Relevance in a Digital Age. Routledge; 2016. pp. 93–108.

41. Gamson WA, Wolfsfeld G. Movements and media as interacting systems. The Annals of the American Academy of Political and Social Science. 1993; 528: 114–125.

42. McCarthy JD, McPhail C, Smith J. Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991. American Sociological Review. 1996; 61: 478–499. https://doi.org/10.2307/2096360

43. Earl J, Martin A, McCarthy JD, Soule SA. The use of newspaper data in the study of collective action. Annu Rev Sociol. 2004; 30: 65–80. https://doi.org/10.1146/annurev.soc.30.012703.110603

44. Oliver PE, Maney GM. Political processes and local newspaper coverage of protest events: From selection bias to triadic interactions. American Journal of Sociology. 2000; 106: 463–505. https://doi.org/10.1086/316964

45. Wouters R. From the street to the screen: Characteristics of protest events as determinants of television news coverage. Mobilization: An International Quarterly. 2013; 18: 83–105. https://doi.org/10.17813/maiq.18.1.y6067731j4844067

46. Knobloch-Westerwick S, Mothes C, Polavin N. Confirmation Bias, Ingroup Bias, and Negativity Bias in Selective Exposure to Political Information. Communication Research. 2020; 47: 104–124. https://doi.org/10.1177/0093650217719596

47. Lengauer G, Esser F, Berganza R. Negativity in political news: A review of concepts, operationalizations and key findings. Journalism. 2012; 13: 179–202. https://doi.org/10.1177/1464884911427800

48. Soroka SN. Negativity in democratic politics: Causes and consequences. Cambridge University Press; 2014.

49. Craig G. The spectacle of the street: an analysis of media coverage of protests at the 2000 Melbourne World Economic Forum.[Article based on a paper given at the Conference of the Australian and New Zealand Communication Association (2001: Perth, WA).]. Australian Journal of Communication. 2002; 29: 39–52.

50. Murdock G. Political deviance: The press presentation of a militant mass demonstration. The Manufacture of News: Deviance, Social Problems and the Mass Media London: Constable. London: Constable; 1973. pp. 206–225.

51. McFarlane T, Hay I. The battle for Seattle: protest and popular geopolitics in The Australian newspaper. Political Geography. 2003; 22: 211–232. https://doi.org/10.1016/S0962-6298(02)00090-2

52. Boykoff J. Framing dissent: Mass-media coverage of the global justice movement. New Political Science. 2006; 28: 201–228. https://doi.org/10.1080/07393140600679967

53. McLeod DM, Detenber BH. Framing effects of television news coverage of social protest. Journal of Communication. 1999; 49: 3–23. https://doi.org/10.1111/j.1460-2466.1999.tb02802.x

54. Vliegenthart R, Walgrave S. The interdependency of mass media and social movements. In: Scammell M, Semetko HA, editors. The SAGE Handbook of Political Communication. SAGE Publications; 2012. pp. 387–398. https://doi.org/10.4135/9781446201015.n31

55. Poell T, Borra E. Twitter, YouTube, and Flickr as platforms of alternative journalism: The social media account of the 2010 Toronto G20 protests. Journalism. 2012; 13: 695–713. https://doi.org/10.1177/1464884911431533

56. Keib K, Espina C, Lee YI, Wojdynski BW, Choi D, Bang H. Picture This: The Influence of Emotionally Valenced Images, On Attention, Selection, and Sharing of Social Media News. Media Psychology. 2018; 21: 202–221. https://doi.org/10.1080/15213269.2017.1378108

57. Mattoni A, Teune S. Visions of Protest. A Media-Historic Perspective on Images in Social Movements. Sociology Compass. 2014; 8: 876–887. https://doi.org/10.1111/soc4.12173

58. Della Porta D, Diani M, Doerr N, Mattoni A, Teune S. Visuals in Social Movements. The Oxford Handbook of Social Movements. 2014. https://doi.org/10.1093/oxfordhb/9780199678402.013.48

59. Neumayer C, Rossi L. Images of protest in social media: Struggle over visibility and visual narratives. New Media and Society. 2018; 20: 4293–4310. https://doi.org/10.1177/1461444818770602

60. Lee AM, Carpini MXD. News Consumption Revisited: Examining the Power of Habits in the 21st Century. 11th International Symposium on Online Journalism. Austin, Texas; 2010. pp. 1–32.

61. Mitchelstein E, Boczkowski PJ. Online news consumption research: An assessment of past work and an agenda for the future. New Media and Society. 2010; 12: 1085–1102. https://doi.org/10.1177/1461444809350193

62. Vaish A, Grossmann T, Woodward A. Not All Emotions Are Created Equal: The Negativity Bias in Social-Emotional Development. Psychological Bulletin. 2008; 134: 383. https://doi.org/10.1037/0033-2909.134.3.383 PMID: 18444702

63. Marano HE. Why We Love Bad News. In: Psychology Today [Internet]. 2003 [cited 16 Apr 2022]. Available: https://www.psychologytoday.com/us/articles/200305/why-we-love-bad-news.

64. Harcup T O'Neill D. What Is News? Galtung and Ruge revisited. Journalism Studies. 2001; 2: 261–280. https://doi.org/10.1080/14616700118449
65. Lang A, Newhagen J, Reeves B. Negative video as structure: Emotion, attention, capacity, and memory. Journal of Broadcasting & Electronic Media. 1996; 40: 460–477. https://doi.org/10.1080/08838159609364369
66. Marwick AE. Instafame: Luxury selfies in the attention economy. Public Culture. 2015; 27: 137–160. https://doi.org/10.1215/08992363-2798379
67. Hessel J, Lee L, Mimno D. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. 26th International World Wide Web Conference. 2017. pp. 927–936. https://doi.org/10.1145/3038912.3052684
68. Violence Rucht D. and New Social Movements. International Handbook of Violence Research. Dordrecht: Springer Netherlands; 2003. pp. 369–382. https://doi.org/10.1007/978-0-306-48039-3_20
69. Lakkaraju H, McAuley J, Leskovec J. What's in a name? Understanding the interplay between titles, content, and communities in social media. Proceedings of the 7th International Conference on Weblogs and Social Media. 2013. pp. 311–320. https://doi.org/10.1609/icwsm.v7i1.14408
70. Uhl MW. Explaining U.S. consumer behavior with news sentiment. ACM Transactions on Management Information Systems. 2011; 2: 1–18. https://doi.org/10.1145/1985347.1985350
71. Gu J, Tian J, Wang X, Ling H. Does negative news travel fast? exploring the effect of news sentiment on interactive spiral. International Conference on Human-Computer Interaction. 2017. pp. 435–442. https://doi.org/10.1007/978-3-319-58750-9_60
72. Hibbing JR, Smith KB, Alford JR. Differences in negativity bias underlie variations in political ideology. Behavioral and Brain Sciences. 2014; 37: 297–307. https://doi.org/10.1017/S0140525X13001192 PMID: 24970428
73. Fournier P, Soroka S, Nir L. Negativity Biases and Political Ideology: A Comparative Test across 17 Countries. American Political Science Review. 2020; 114: 775–791. https://doi.org/10.1017/S0003055420000131
74. Iyengar S, Hahn KS. Red media, blue media: Evidence of ideological selectivity in media use. Journal of Communication. 2009; 59: 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x
75. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science. 2018; 359: 1146–1151. https://doi.org/10.1126/science.aap9559 PMID: 29590045
76. Boyd D. Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. Networked Self: Identity, Community, and Culture on Social Network Sites. Routledge; 2010. pp. 47–66.
77. Lexico Dictionaries. Attention | Definition of Attention by Oxford Dictionary on Lexico.com. Lexico Dictionaries. 2022. Available from: https://www.lexico.com/definition/attention.
78. Tenenboim O, Cohen AA. What prompts users to click and comment: A longitudinal study of online news. Journalism. 2015; 16: 198–217. https://doi.org/10.1177/1464884913513996
79. Reddit. Reddiquette [Internet]. 2021 [cited 30 Apr 2022]. Available from: https://reddit.zendesk.com/hc/en-us/articles/205926439-Reddiquette.
80. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. https://doi.org/10.48550/arXiv.1810.04805
81. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review. 2022; 55: 5731–5780. https://doi.org/10.1007/s10462-022-10144-1
82. Hugging Face. bert-base-uncased. In: Hugging Face [Internet]. 2023 [cited 2 Jun 2023]. Available from: https://huggingface.co/bert-base-uncased.
83. Hamborg F, Donnay K. NewsMTSC: a dataset for (multi-)target-dependent sentiment classification in political news articles. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021. pp. 1663–1675. https://doi.org/10.5167/uzh-207183
84. Lin Y. 10 Reddit Statistics Every Marketer Should Know In 2021. In: Oberlo [Internet]. 2021 [cited 16 Apr 2022]. Available from: https://www.oberlo.com/blog/reddit-statistics.
85. Armed Conflict Location & Event Data Project. Demonstrations & Political Violence in America. US Crisis Monitor [Internet]. 2021 [cited 30 Apr 2022]. Available from: https://acled.github.io/US/.
86. Bergström A, Jervelycke Belfrage M. News in Social Media: Incidental consumption and the role of opinion leaders. Digital Journalism. 2018; 6: 583–598. https://doi.org/10.1080/21670811.2018.1423625
87. Bruhn M, Schoenmueller V, Schäfer DB. Are social media replacing traditional media in terms of brand equity creation? Management Research Review. 2012; 35: 770–790. https://doi.org/10.1108/01409171211255948

88. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The pushshift reddit dataset. 14th International Conference on Web and Social Media. 2020. pp. 830–839. https://doi.org/10.1609/icwsm.v14i1.7347
89. Barthel M, Stocking G, Holcom J, Mitchell A. Seven-in-Ten Reddit Users Get News on the Site. In: Pew Research Center [Internet]. 2016 [cited 16 Apr 2022]. Available from: https://www.pewresearch.org/journalism/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/.
90. Duguay PA. Read it on Reddit: Homogeneity and ideological segregation in the age of social news. Social Science Computer Review. 2022; 40: 1186–1202. https://doi.org/10.1177/08944393211001053
91. Zannettou S, Caulfield T, De Cristofaro E, Kourtelris N, Leontiadis I, Sirivianos M, et al. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. Proceedings of the 2017 internet measurement conference. 2017. pp. 405–417. https://doi.org/10.1145/3131365.3131390
92. Soliman A, Hafer J, Lemmerich F. A characterization of political communities on reddit. Proceedings of the 30th ACM conference on hypertext and Social Media. 2019. pp. 259–263. https://doi.org/10.1145/3342220.3343662
93. Nadiri A, Takes FW. A large-scale temporal analysis of user lifespan durability on the Reddit social media platform. Companion Proceedings of the Web Conference 2022. 2022. pp. 677–685. https://doi.org/10.1145/3487553.3524699
94. Won D, Steinert-Threlkeld ZC, Joo J. Protest activity detection and perceived violence estimation from social media images. MM 2017—Proceedings of the 2017 ACM Multimedia Conference. 2017. pp. 786–794. https://doi.org/10.1145/3123266.3123282
95. Liu T, Fang S, Zhao Y, Wang P, Zhang J. Implementation of Training Convolutional Neural Networks. arXiv; 2015. https://doi.org/1506.01195
96. Hussain M, Bird JJ, Faria DR. A study on cnn transfer learning for image classification. Advances in Computational Intelligence Systems. Nottingham, UK: Springer; 2019. pp. 191–202. https://doi.org/10.1007/978-3-319-97982-3_16
97. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning. PMLR; 2019. pp. 6105–6114.
98. Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems. 2022; 33: 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827 PMID: 34111009
99. Ridnik T, Ben-Baruch E, Noy A, Zelnik-Manor L. ImageNet-21K Pretraining for the Masses. arXiv; 2021. https://doi.org/10.48550/arXiv.2104.10972
100. Hilbe JM. Negative binomial regression. Cambridge University Press; 2011.
101. Cameron AC, Trivedi PK. Regression analysis of count data. Cambridge university press; 2013.
102. Cook RD, Weisberg S. Criticism and Influence Analysis in Regression. Sociological Methodology. 1982; 13: 313–361. https://doi.org/10.2307/270724
103. Huberman BA. Social Computing and the Attention Economy. Journal of Statistical Physics. 2013; 151: 329–339. https://doi.org/10.1007/s10955-012-0596-5
104. Trilling DC. Following the news: Patterns of online and offline news consumption. Universiteit van Amsterdam. 2013.
105. Bellovary AK, Young NA, Goldenberg A. Left- and Right-Leaning News Organizations Use Negative Emotional Content and Elicit User Engagement Similarly. Affective Science. 2021; 2: 391–396. https://doi.org/10.1007/s42761-021-00046-w PMID: 34423311
106. Ma Y. Can More Pictures Bring More Readership?: An Examination of the "Picture Superiority Effect" in the News Consumption Process. Procedia—Social and Behavioral Sciences. 2016; 236: 34–38. https://doi.org/10.1016/j.sbspro.2016.12.012
107. Shrestha P, Sathanur A, Maharjan S, Saldanha E, Arendt D, Volkova S. Multiple social platforms reveal actionable signals for software vulnerability awareness: A study of GitHub, Twitter and Reddit. PLoS ONE. 2020;15. https://doi.org/10.1371/journal.pone.0230250 PMID: 32208431
108. Bucher T, Helmond A. The affordances of social media platforms. The SAGE handbook of social media. 2018. pp. 233–254.
109. Aldous KK, An J, Jansen BJ. View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. Proceedings of the International AAAI Conference on Web and Social Media. 2019. pp. 47–57. https://doi.org/10.1609/icwsm.v13i01.3208
110. Lewandowsky S, Ecker UKH, Cook J. Beyond misinformation: Understanding and coping with the "post-truth" era. Journal of Applied Research in Memory and Cognition. 2017; 6: 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

S1 File. Appendix – Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit

Theresa Henn[1], Oliver Posegga[1*]

[1]Department of Information Systems and Social Networks, University of Bamberg, Bamberg, Germany

[*]Corresponding author: E-mail: oliver.posegga@uni-bamberg.de

Table of Contents

Data availability statement

In the following, we will describe in detail the different datasets utilized within our paper to make our analysis transparent and replicable.

The paper is based on **four** different datasets:

- The first dataset is the **training dataset** for the image classifier and consists of 7,972 violent images and 8,986 nonviolent images. We obtained these images partially from Won et al.'s [1] "UCLA Protest Image Dataset," a prelabeled protest-related dataset comprising more than 40,764 images. Besides this data source, we enriched our training dataset with pictures from a Black Lives Matter (BLM) related keyword search on Bing. This dataset was then used to train our image classifier to classify images as violent (1) or nonviolent (0).
- The second dataset is the **test dataset** and was likewise obtained through a combination of images of the dataset provided by Won et al. [1] and images from a keyword search related to the Black Lives Matter movement. Combining both data sources, this dataset contains 2,854 violent and 4,088 nonviolent pictures. It was used to evaluate the performance of the image classifier.
- The third dataset, the **NewsMTSC dataset**, was introduced by Hamborg and Donnay [2] as a labeled dataset for sentiment classification in political news articles. We used their dataset's real-world (rw) sentiment distribution subset, which contains 8,740 train samples, 343 validation samples, and 803 test samples of news sentences experts annotated as positive, negative, or neutral. We used this dataset to train our BERT model to classify submission titles as positive or negative.
- The fourth dataset is the **Reddit BLM dataset** obtained from Reddit via the Pushshift submission dumps. After filtering the data and downloading images attached to the submissions' URLs, we received a final dataset of 5,873 observations. Based on this dataset, we conducted the negative binomial regression analysis presented in the manuscript.

To receive these final datasets, we proceeded as follows:

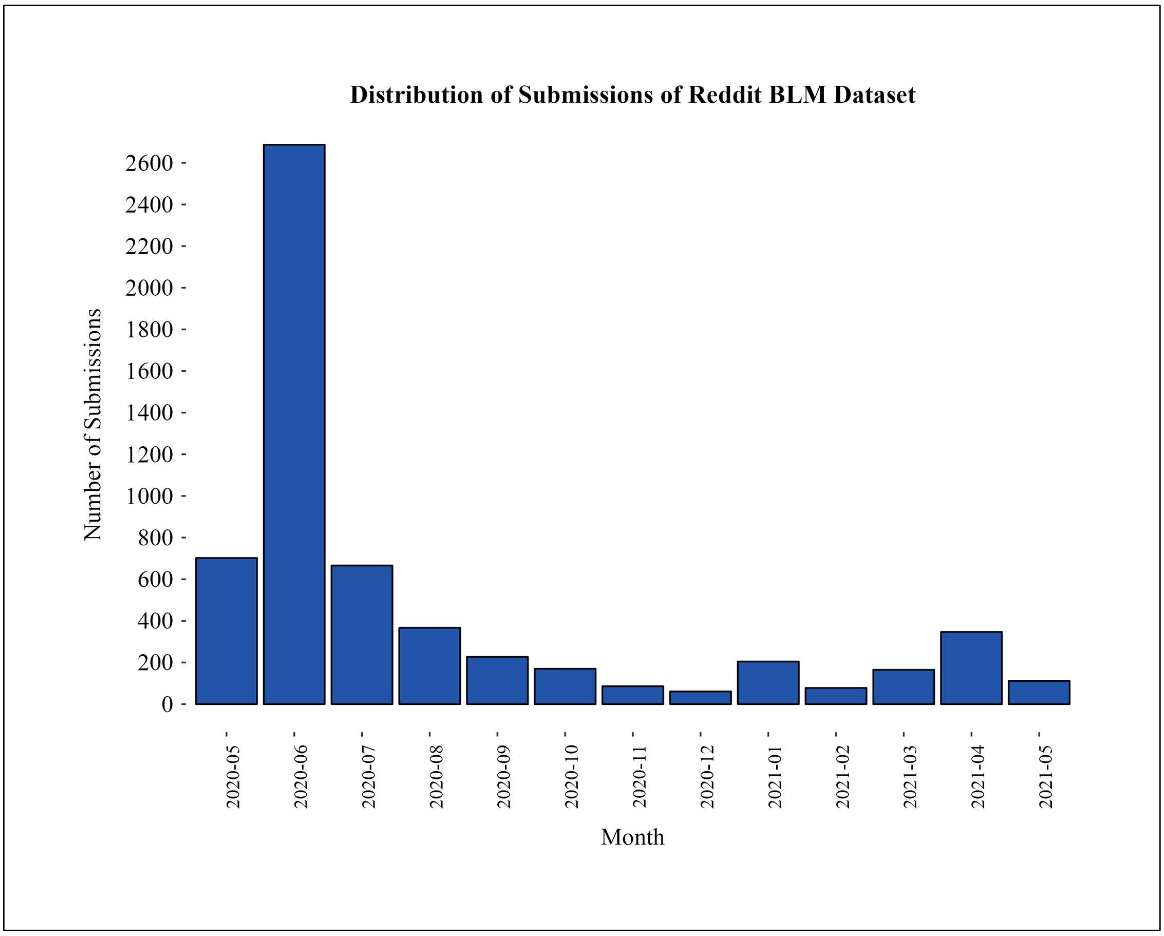
- To reduce Won et al.'s [1] protest dataset to a training dataset appropriate for our analysis, we first transformed their continuous violence codes (ranging from 0 "nonviolent" to 1 "very violent") into a binary code (0 "nonviolent" and 1 "violent"). We classified every original protest picture score above 0.5 as "violent." Using all images classified as violent, we created a balanced sample by adding the same number of nonviolent pictures (with the smallest original violence scores, all close to 0). Thus, our training dataset contained 1,283 violent and 1,283 nonviolent images. We manually reviewed all images and adjusted the violence score if necessary to guarantee that this binary classification conforms with our definition of violence (adopted from Rucht [3]). We extended this sample and included additional images identified via the search engine Bing based on the following 25 keywords in English and German: Black Lives Matter; BLM violen*; police brutality; protest violence; police violence; Protest Gewalt; Black Lives Matter demonstration; Polizei Gewalt; George Floyd; Rayshard Brooks; Donald Trump; Joe Biden; burning; protest shooting; violent demonstration; Hong Kong violence; Portland protest; politicians; mug shot; American politicians; property destruction protest; USA violent protest; protest hurt people; bleeding protest; dead protestors. Based on our definition of violence, we manually extracted 173 additional violent and 692 nonviolent images from the search results. Combining these images with our previous images obtained from the UCLA Protest Image Dataset, we ended up with 1,456 violent and 1,975 nonviolent images. We applied image augmentation to increase our sample size, resulting in 7,972 violent images and 8,986 nonviolent images as our final training dataset.

- The test dataset was created similarly. After selecting the most and least violent images from the test UCLA Protest Image Dataset, we received 350 violent and 350 nonviolent images. Again, we validated and adjusted the violence classification of the extracted images based on Rucht's [3] violence definition and added pictures from the Bing search described above. At this stage, the dataset comprised 309 violent and 478 nonviolent images. After augmenting this data, the final test dataset contained 2,854 violent and 4,088 nonviolent images.

- Since we are interested in differentiating between positive and negative sentiment in news titles, we reduced the NewsMTSC dataset [2] to news sentences labeled as negative or positive. Doing so, we received 3,316 negative and 2,395 positive samples for our training data, 143 negative and 73 positive samples for the validation data, and 295 negative and 189 positive samples for our test data.

- For the Reddit BLM dataset, we downloaded the Pushshift submission dumps for May 2020 until May 2021 (105GB) on May 9, 2023, and filtered the data regarding the three subreddit of interest r/worldnews, r/politics, and r/news (830MB) which led to 1,887,202 submissions. After reducing the data towards the period of analysis from May 25, 2020, the day of George Floyd's murder, until May 25, 2021, we received 1,763,659 submissions, of which r/news contained 878,812 submissions, r/politics 445,051 submissions, and r/worldnews 439,796 submissions. We removed all submissions whose titles did not match one of the three keywords, "black lives matter," "george floyd," and "blm," resulting in a total of 17,083 submissions relevant to our analysis. Note that we also tested how an increased keyword set related to our analysis topic would increase our data corpus. We therefore added the keywords "police violence," "breonna taylor," and "rayshard brooks" to our keyword set. As we only received 3,008 additional submissions after doubling the keyword set while also introducing more noise to our data (e.g., police violence related to Covid-19 protests), we decided to continue working with our initial keyword set, as it showed to capture the relevant submissions for our analysis. For our data cleaning process, we removed a particular type of submission from our dataset, so-called "megathreads," reducing the number of submissions to 17,069. For these submissions, we attempted to receive information regarding the categorical control variables "political leaning" (*H4*), "factual reporting" (*H5*), "traffic" (*H6*), and "type of news outlet" from the website mediabiasfactcheck.com. If we did not find all information on the news outlet, we coded the observation as missing and dropped the submission. Out of the remaining 11,813 submissions, we obtained 9,373 images attached to the linked news outlets through the Python library "Newspaper3k." Thus, we dropped 2,440 submissions for which we could not get an image or from which the image was broken. As we were interested in submissions that at least received a minimum of attention, we deleted all submissions that received zero comments ending up with 5,876 observations. After dropping all missing values of the other variables related to our analysis, we received 5,873 submissions. Thus, our final Reddit BLM dataset comprised 5,873 submissions, including one image per submission. These 5,873 submissions were posted by 2,091 unique users (note that 1,975 submissions were posted by authors that were deleted by the time of data retrieval) and included 4,833 unique URLs; 4,324 submissions were posted in r/politics, 1,114 in r/news, and 435 in r/worldnews. Figure S1 shows a monthly distribution of submissions within our Reddit BLM dataset.

S1 Figure. Histogram of monthly distribution of Reddit BLM submissions.



The primary Reddit BLM dataset underlying our model is available via OSF (https://osf.io/4dr2x/?view_only=39ff5cb39c254b6c93eb12f9c205a941). Since the data (violent and nonviolent images) used in our training and test dataset is considered third-party data belonging to Won et al. [1], we cannot share this data publicly. Yet, all information to obtain and prepare the dataset can be obtained from their public GitHub repository [4]. By applying the different data transformation steps mentioned above, our datasets can be reconstructed, and results can be replicated. The NewsMTSC dataset from Hamborg and Donnay [2] is available via Hugging Face [5].

Descriptive overview of variables

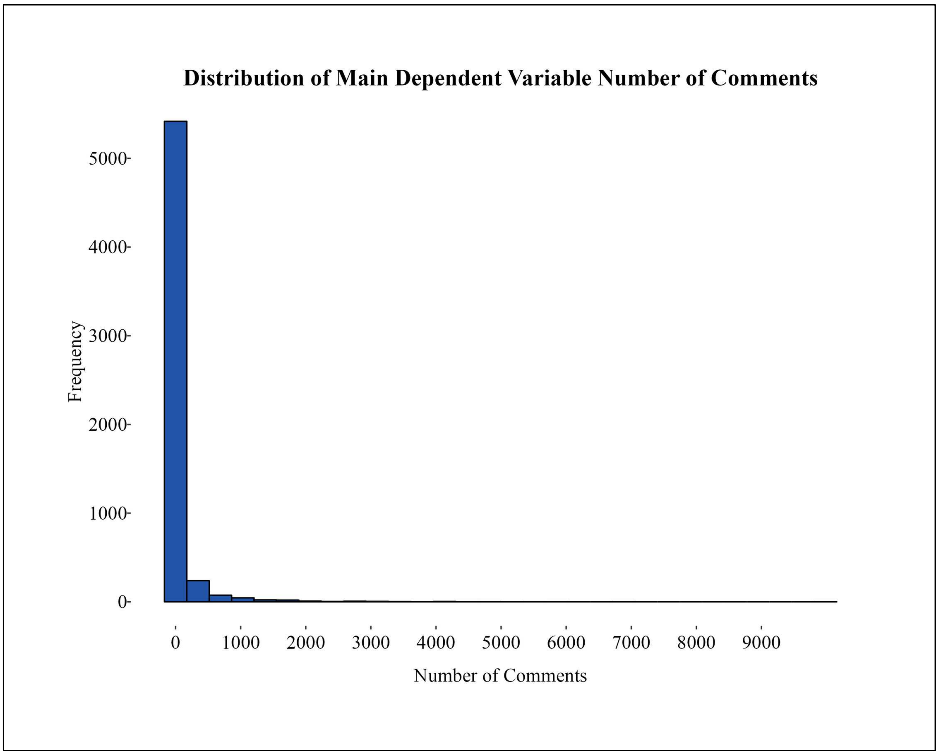
This section provides a descriptive overview of the variables used within our negative binomial regression analysis (see S1 Table). Additionally, we show the distribution of our dependent variable number of comments, which is highly positively skewed (see S2 Figure).

S1 Table. Descriptive overview of variables.

Name	Kind	Type	Categories	Distribution	Retrieved
total number of comments	dependent	interval	numeric	min = 1	Reddit
				max = 9,985	
				median = 8	
violent image	independent	binary	0 = nonviolent	nonviolent = 5,486	VGG19
			1 = violent	violent = 387	
BERT sentiment	control	binary	0 = positive	positive = 1,253	sentiment analysis with BERT model
			1 = negative	negative = 4,620	
political leaning	control	categorical	0 = neutral	neutral = 793	media-biasfactcheck.com
			1= conservative	conservative = 1,000	
			2 = conspiracy	conspiracy = 2	
			3 = liberal	liberal= 4,078	
factual reporting	control	categorical	0 = mixed	mixed = 2,886	media-biasfactcheck.com
			1 = high	high = 2,830	
			2 = low	low = 157	
traffic	control	categorical	0 = medium	medium = 584	media-biasfactcheck.com
			1 = high	high = 5,244	
			2 = minimal	minimal = 45	
subreddit	control	categorical	0 = politics	politics = 4,324	Reddit
			1 = news	news = 1,114	
			2 = worldnews	worldnews = 435	
NSFW	control	binary	0 = no NSFW tag	no NSFW tag = 5,870	Reddit
			1 = NSFW tag	NSFW tag = 3	
link flair	control	binary	0 = no link flair	no link flair = 2,637	Reddit
			1 = link flair	link flair = 3,236	

number of cross-posts	control	interval	numeric	min = 0 max = 23	Reddit
weekday/ weekend (US Pacific time)	control	binary	0 = weekday 1 = weekend	weekday = 4,505 weekend = 1,368	Reddit
time of day (US Pacific time)	control	categor-ical	0 = after-noon/evening 1 = morning 2 = night	afternoon/evening = 2,395 morning = 2,417 night = 1,061	Reddit
type of news outlet	control	categor-ical	0 = newspaper 1 = magazine 2 = news agency 3 = organiza-tion/foundation 4 = radio station 5 = TV station 6 = website	newspaper = 1,986 magazine = 308 news agency = 240 organization/founda-tion = 63 radio station = 85 TV station = 1,953 website = 1,238	media-biasfactcheck .com

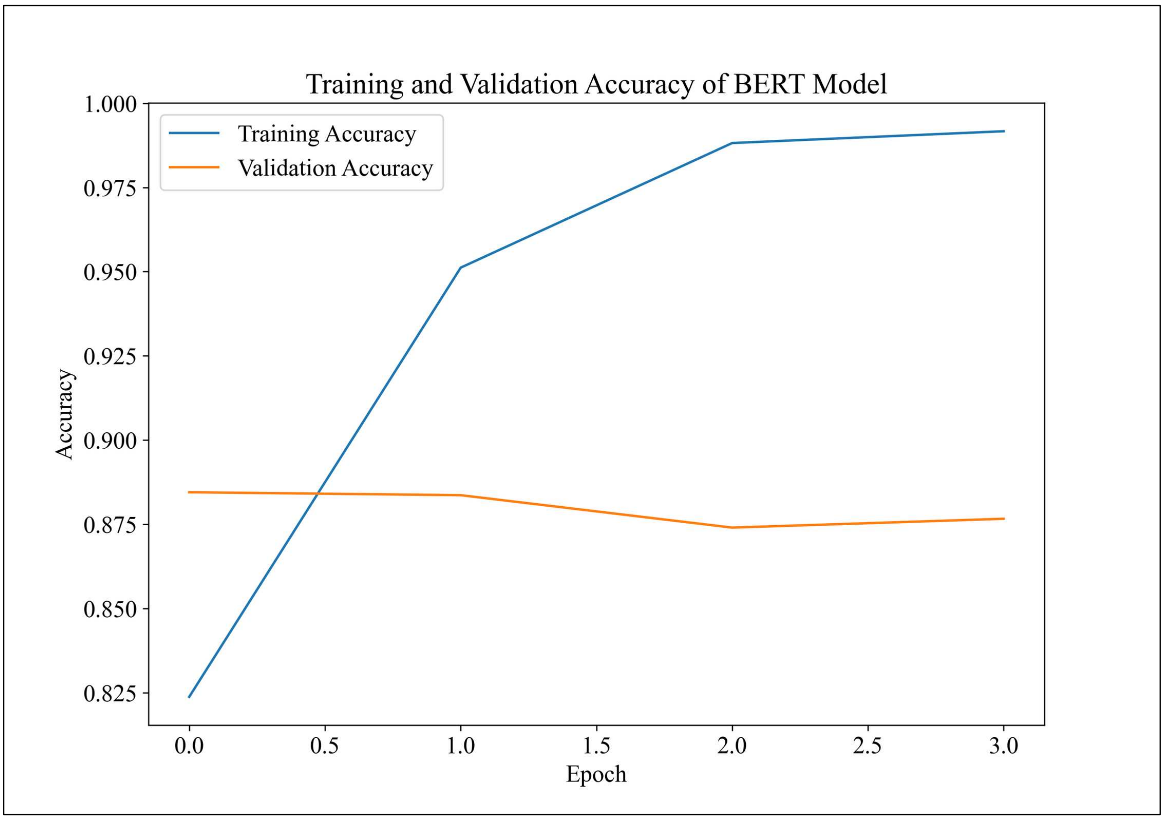
S2 Figure. Histogram of dependent variable number of comments.



Sentiment model

For our sentiment analysis of submission titles, we used the “BERT base model uncased,” derived from the Hugging Face platform [6]. The model is trained on English Wikipedia and the BookCorpus data [7] and contains 110 million parameters. To customize the model for classifying the sentiment of news articles, we trained the model with the NewsMTSC dataset [2], a labeled dataset for sentiment classification in political news articles containing the categories positive, negative, and neutral. Since we were interested in positive and negative sentiment classification, we dropped the neutral samples from the training, validation, and test dataset. We used the remaining 5,711 samples as training data and the remaining 207 samples as validation data. We compiled the model with the adam optimizer, a sparse categorical cross-entropy loss function, and a learning rate of 0.00002. We implemented an early stop callback function with a patience parameter of three, monitoring the validation accuracy. Thus, the model stops the training if the validation accuracy does not improve for three consecutive epochs. We started the training process with ten epochs, but due to the early stop callback function, the training process stopped after three epochs. The model received a training accuracy of 99.17% and a validation accuracy of 87.66% (see S3 Figure). When applying the model to the test dataset, it received an accuracy of 87.40%.

S3 Figure. Training and validation accuracy of BERT model.



To evaluate the performance of the trained BERT model on our Reddit BLM dataset, we drew a random sample of 200 submission titles. We manually labeled them as either positive or negative. The manual and automatic classification matched 168 out of 200 cases (84.0%), which aligns with the previous results on the testing and validation data.

To contrast this result with another sentiment analysis approach, we also tested the performance of the Python library "Valence Aware Dictionary and Sentiment Reasoner" (VADER) on our data. We chose VADER as a second sentiment analysis approach because it has been shown to perform exceptionally well on news data and Reddit [8]. VADER bases its annotations on prelabeled dictionaries with sentiment scores for each listed record. By relating the polarity scores (the amount of negative and positive sentiment within the text) to the computed intensity scores (the strength of the respective emotion), VADER calculates negative, positive, and neutral values for each tested observation. It then blends these values into a single compound score ranging from -1 (very negative) to 1 (very positive). We used this compound value as a threshold by dichotomously labeling a title with a score <0 as "negative" and each title with a score >=0 as “positive.” When comparing the sentiment prediction of VADER on a random sample of 200 submission titles of our Reddit BLM dataset with our manual labeling, 154 out of 200 samples (77.0%) were correctly classified.

Since the BERT model outperformed the VADER classification while also considering the context of a sentence compared to a bag-of-words approach, we utilized our trained BERT model for the analysis in the manuscript.

Image classifiers

For our image analysis, we trained and validated different state-of-the-art convolutional neural network (CNN) and vision transformer (VIT) models with our training dataset to detect the best-fitting architecture for our classification context.

For the CNN models, we used four different architectures: (1) a self-created CNN architecture (see S2 Table) which we built with the Python deep learning libraries “TensorFlow” and “Keras” and trained from scratch, (2) a VGG19 architecture, (3) a ResNet50 architecture, and (4) an EfficientNetB4 architecture which were derived from TensorFlow and which were pre-trained on the ImageNet-1k dataset which comprises 1,000 different image classes and more than 1.2 million images [9]. We decided on these models as the latter three represent state-of-the-art architectures for image classification with CNNs [see, e.g., 10,11]. We decided on the former as it achieved convincing classification results (see S4 Figure) while utilizing fewer parameters than the other models, resulting in a faster training process (see S2 Table for self-created CNN’s architecture and the number of parameters).

S2 Table. Own convolutional neural network architecture.

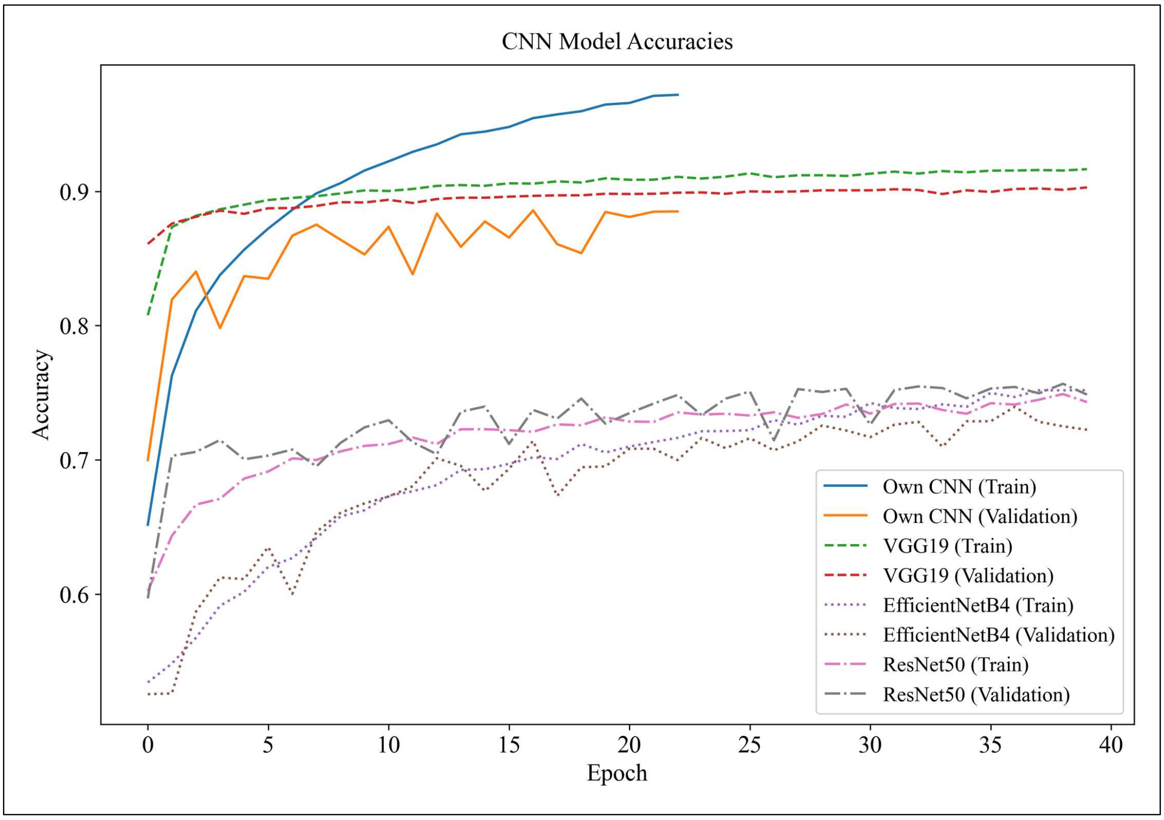
Model: Sequential		
Layer	Output Shape	Parameter
Convolutional	(None, 253, 253, 32)	896
Max Pooling	(None, 126, 126, 32)	0
Convolutional_1	(None, 124, 124, 64)	18,496
Max Pooling_1	(None, 62, 62, 64)	0
Convolutional_2	(None, 60, 60, 128)	73,856
Dropout	(None, 60, 60, 128)	0
Max Pooling_2	(None, 30, 30, 128)	0
Convolutional_3	(None, 28, 28, 128)	147,584
Max Pooling_3	(None, 14, 14, 128)	0
Flatten	(None, 25,088)	0
Dropout_1	(None, 25,088)	0
Dense	(None, 512)	12,845,568
Dense_1	(None, 1)	513
Total parameter: 13,086,913		
Trainable parameters: 13,086,913		
Non-trainable parameters: 0		

For the self-created CNN architecture, we activated the last dense layer with a sigmoid function, which returns a number between 0 and 1 and is, therefore, suitable for our binary image classification. For compiling the self-created CNN, we deployed the optimizer root mean square prop (RMSprop) algorithm with a learning rate of 0.0001 and a binary cross-entropy (BCE) loss function. We started the model's training by splitting the training dataset into 70% training images and 30% validation images. We also implemented an early stop callback function with a minimum delta of 0.1% and a patience parameter of 10 within our training process to prevent overfitting. Thus, the training stops if the validation loss does not improve within ten epochs by at least 0.1% compared to its best value. We initially implemented 40 training epochs with a batch size of 32, stopping at epoch 23 due to our early stop callback function (see S4 Figure).

For the VGG19 model, we froze all layers but the last fully connected layer of the pre-trained model, as unfreezing more layers did not yield better training results. We replaced the last fully connected layer with a dense layer that applies a sigmoid function as an activation. We compiled the model with the adam optimizer, a learning rate of 0.001, and a BCE loss function. We split the training data into 70% training images and 30% validation images. Likewise, we implemented an early stop callback function with a minimum delta of 0.1% and a patience parameter of 10. We started training the model with 40 epochs and a batch size of 32. As the model's validation loss continuously improved by at least 0.1%, the training ended after 40 epochs (see S4 Figure).

For the ResNet50 and EfficientNetB4 architectures, we applied the same steps and parameters described for the VGG19 model. We solely unfroze the last 50 layers of the EfficientNetB4 model, as this approach yielded the best accuracy results when training this model (for an overview of the model accuracies, see S4 Figure).

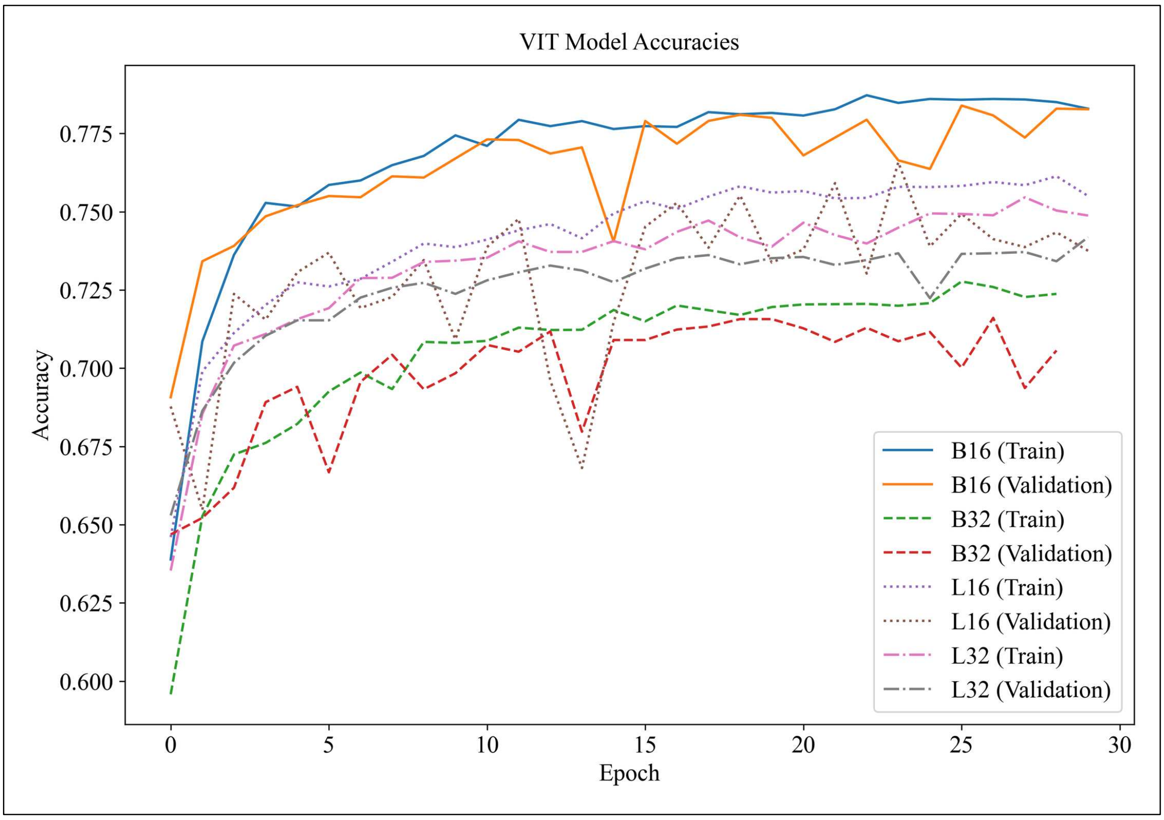
S4 Figure. Accuracy of CNN models.



VIT models are considered one of the latest developments in computer vision by achieving equivalent or better results compared to state-of-the-art CNNs [see, e.g., 12,13]. Thus, we utilized the Keras implementation of four commonly applied and pre-trained VIT architectures and used our training data for the transfer learning [12]. The four architectures are (1) a VIT base model with an input patch size of 16x16 (B16), (2) a VIT base model with an input patch size of 32x32 (B32), (3) a VIT large model with an input patch size of 16x16 (L16), and (4) a VIT large model with an input patch size of 32x32 (L32). Note that these models' configurations are directly adopted from the BERT model architecture [for more information, please see 12]. All four models are pre-trained on the ImageNet-21k dataset, which comprises more than 14 million images and nearly 22,000 different image classes [9]. We removed the last dense layer for all four models and replaced it with a dense layer activated by a sigmoid function. We also included an early stop callback function monitoring the validation loss with a patience parameter of 10 and a minimum delta of 0.1% to prevent overfitting. We compiled all four models with the adam optimizer and a BCE loss function. For training all four models, we split the training data into

70% training and 30% validation data and initiated the training process with 30 epochs and a batch size of 32. As the validation loss continuously improved by at least 0.1% during training for models B16, L16, and L32, the training ended at epoch 30; for model B32, the training ended at epoch 29 (see S5 Figure).

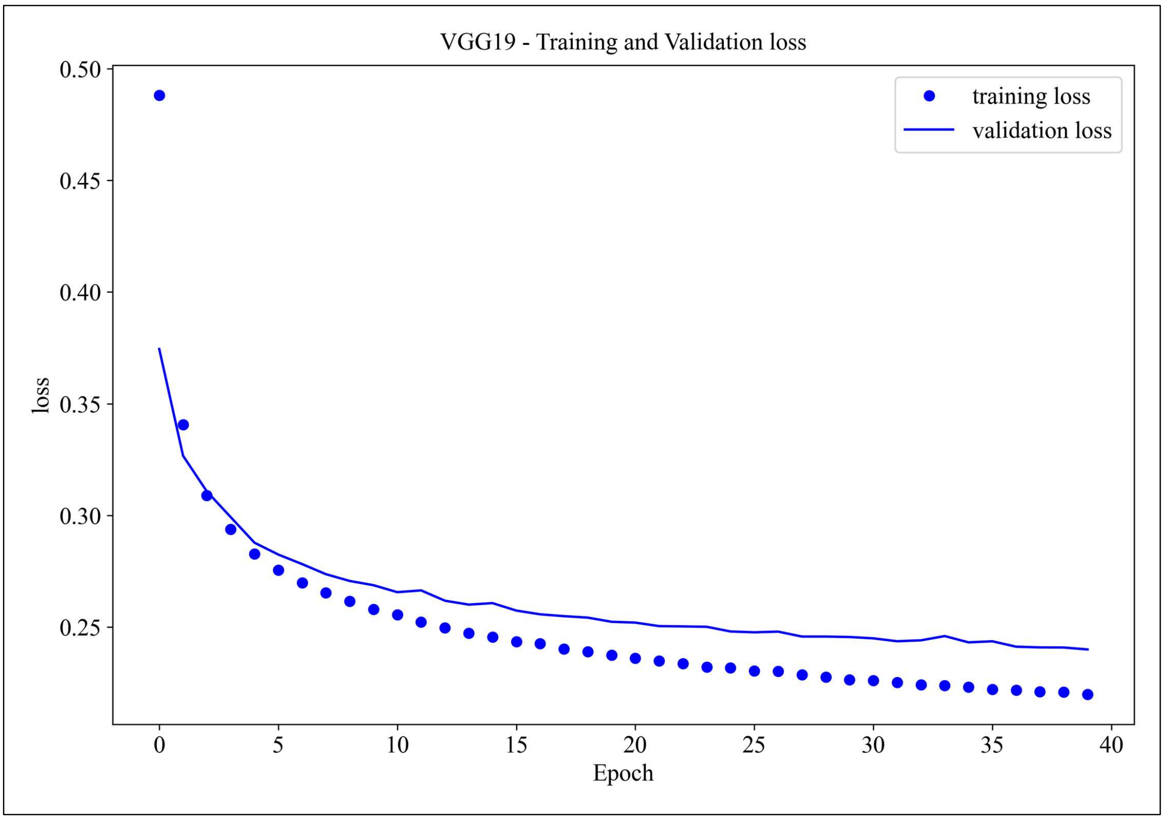
S5 Figure. Accuracy of VIT models.



When comparing the accuracy levels of the training and validation data of the four CNNs and the four VIT models (see S4 Figure and S5 Figure), it becomes evident that our self-created CNN and the VGG19 architecture outperform the remaining CNNs, EfficientNetB4 and ResNet50 as well as all four VIT models. We thus continued working with those two models and tested how well they performed on our image data from the Reddit BLM dataset. We, therefore, classified 200 randomly selected images with both models. When manually evaluating the classification accuracy of the self-created CNN with our manual coding, out of 200 images, 171 (85.5%) were correctly classified as either violent or non-violent. In comparison, when manually evaluating the performance of the VGG19 model on 200 randomly drawn samples from our Reddit BLM dataset, 180 out of 200 (90.0%) images were correctly classified. We, therefore, chose the VGG19 model as our image classifier for the classification of the Reddit BLM dataset.

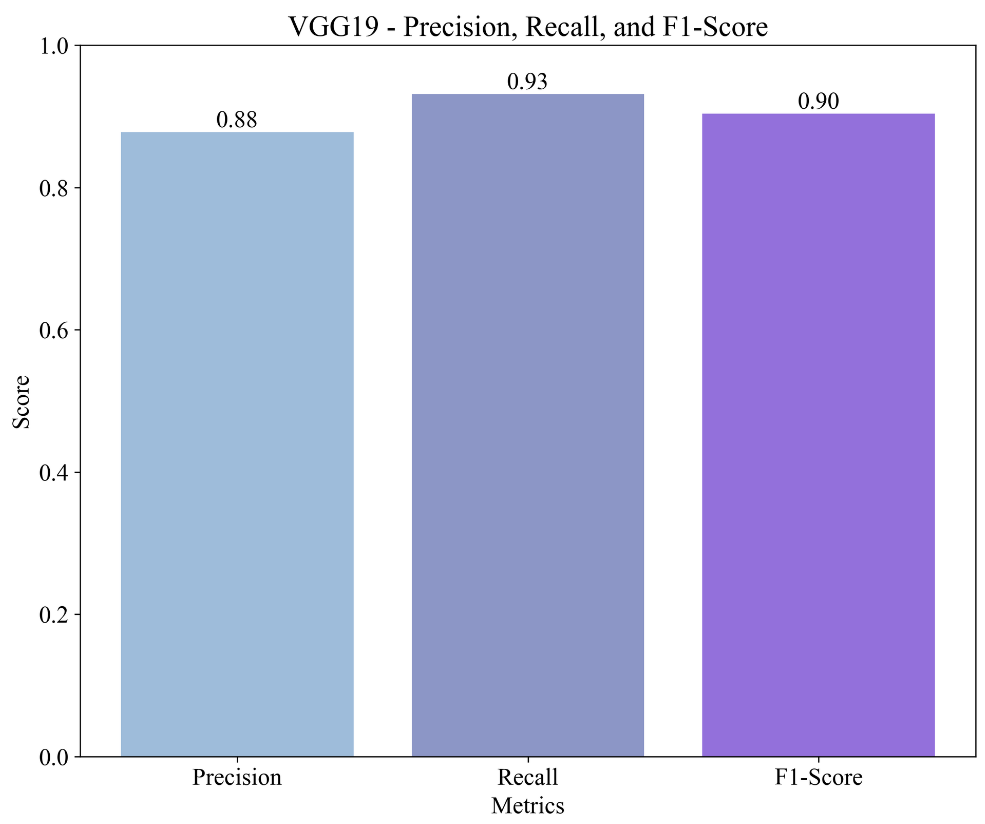
Since the VGG19 model is our model of choice, we also evaluated the model's performance regarding additional performance indicators. The model achieved an accuracy value of 91.65% on the training data and 90.29% on the validation data (see S4 Figure). When assessing the performance on the test dataset, the classifier achieved an accuracy of 91.84%. When evaluating the model's training and validation loss during the training process (see S6 Figure), both curves continuously decreased, converging towards the x-axis. This finding indicates that the model picks up essential image features without overfitting the data.

S6 Figure. VGG19 model loss function.



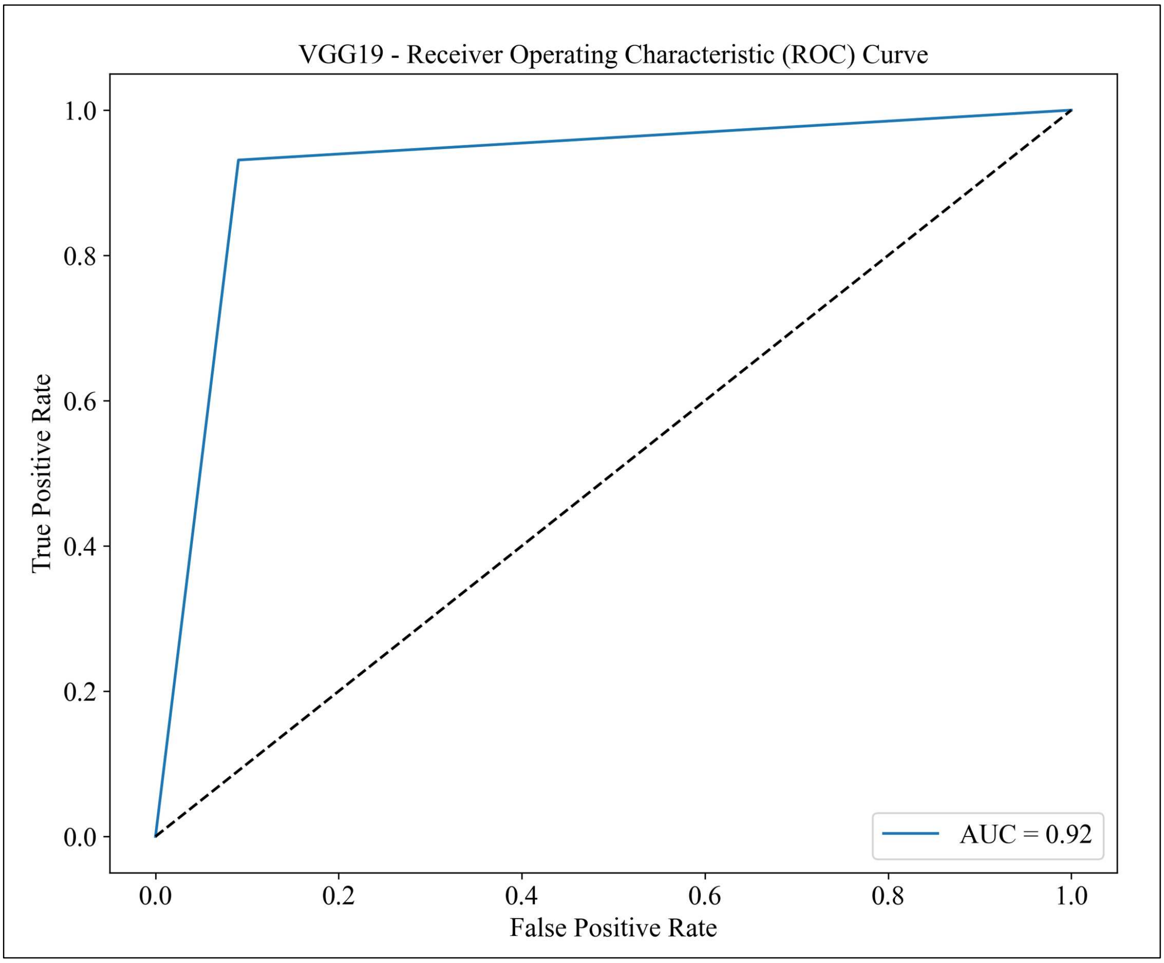
The VGG19 model received a precision score of 87.78%, a recall score of 93.13%, and an F1-score of 90.37% on the test dataset (see S7 Figure). Thus, the model can correctly classify violent images (precision) and detect most of the violent images within the dataset (recall). This outstanding performance on both levels is expressed by the very good F-1 score of 90.37%.

S7 Figure. VGG19 model precision, recall, and F1-Score.



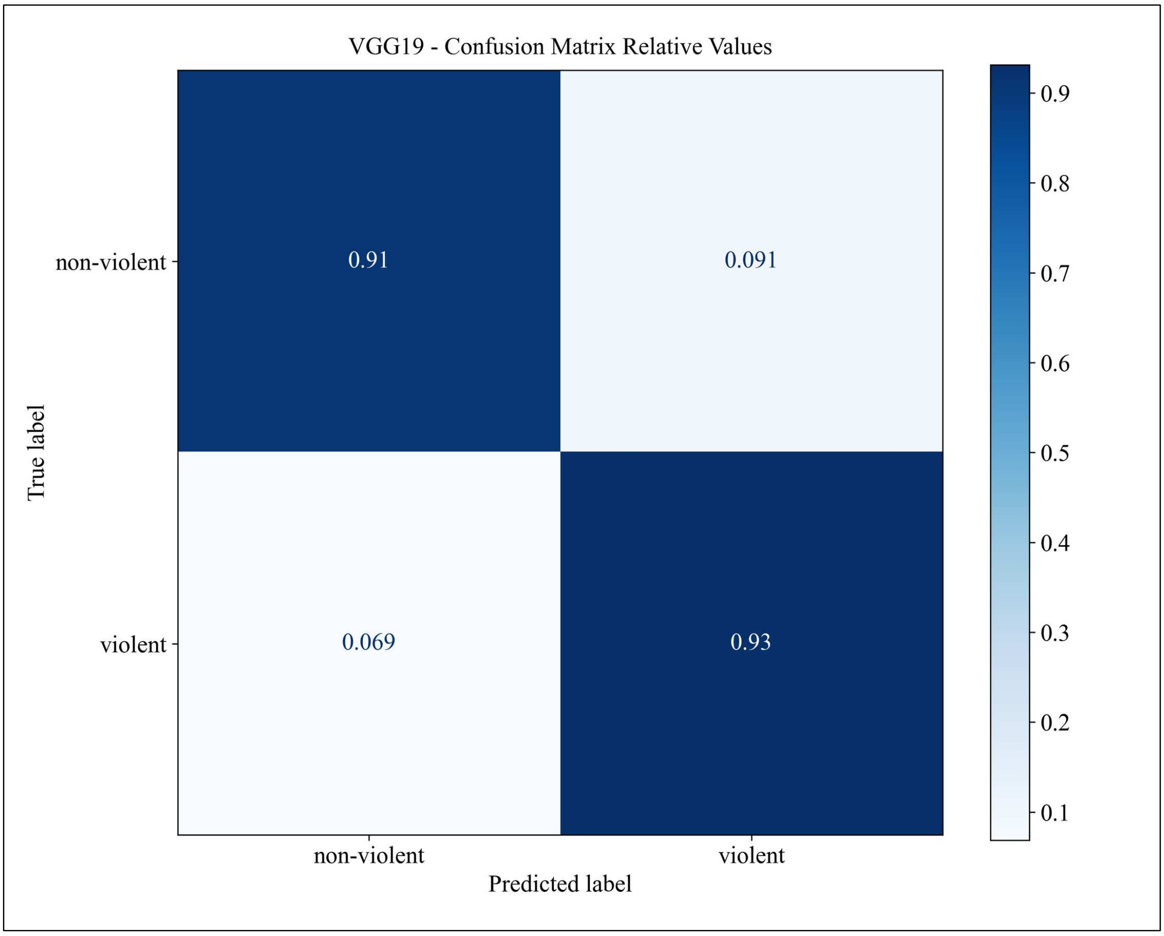
When analyzing the model's relationship between the true positive rate and the false positive rate with the receiver operating characteristic (ROC) curve, we received an area under the curve (AUC) value of 92.0%. This finding shows that the VGG19 model can correctly distinguish between violent and non-violent images in 92.0% of the cases (see S8 Figure).

S8 Figure. VGG19 model receiver operating characteristic (ROC) curve.

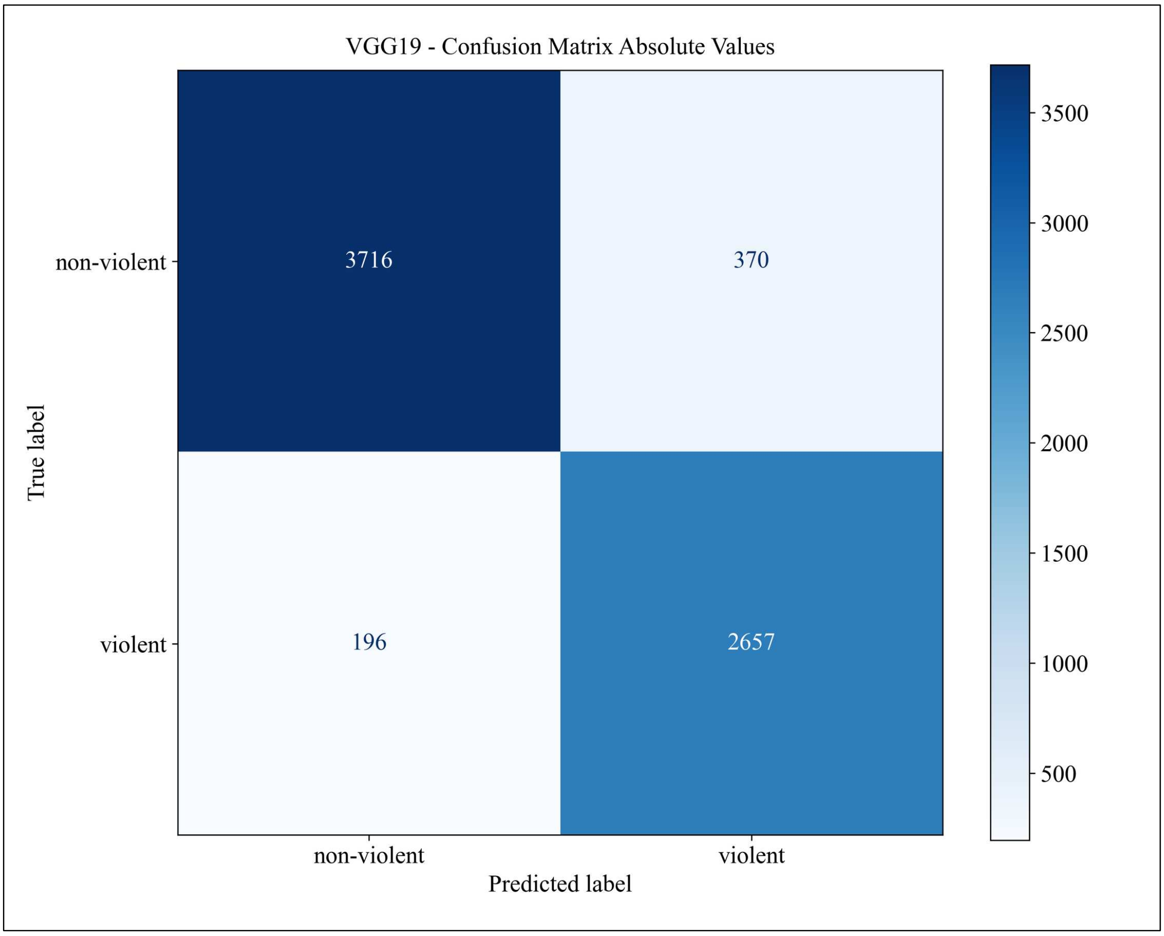


Lastly, we evaluated the VGG19 model's performance on the test dataset in more detail by creating a confusion matrix of relative and absolute values of classified violent and nonviolent images from our test dataset (see S9 Figure and S10 Figure). The model correctly predicts nonviolent images in 91.0% of the cases (3,716 images) and violent images in 93.0% of the cases (2,657 images).

S9 Figure. VGG19 model confusion matrix of test dataset relative values.



S10 Figure. VGG19 model confusion matrix of test dataset absolute values.



All these metrics show that the VGG19 model's performance is sound and, thus, allows for a valid image classification within our analysis.

Negative binomial regression models

To test the robustness of the results of our negative binomial regression model, we ran multiple negative binomial regression models with different datasets and variables (see S3 Table). For all models, we used the “glm.nb” function from the R package “MASS.” In the following, we will provide an overview of the different model results and explanations for diverging effects compared to the model used in the paper. Note that we did not standardize the variables before running the models, as we did not include any interaction terms and only have one numeric control variable (number of cross-posts). Likewise, standardizing the variables by subtracting the mean and dividing them by the standard deviation before running a negative binominal regression impedes the interpretability of the results. It is, therefore, often omitted when performing this type of analysis [for more information, see, e.g., 14]. Note that all other variables are held constant when interpreting the individual coefficients.

As our dependent variable is a count variable, a Poisson or negative binomial regression model is suitable for our analysis. To test whether a Poisson or negative binomial regression would be more appropriate for our data, we conducted for all models a likelihood ratio test for overdispersion in count data (“odTest” function from the R package “pscl”). The likelihood ratio test estimates whether the data’s variance equals its mean, which is the assumption of a Poisson model. The test results showed that the data’s variance was greater than the mean for all models, implying a significant data overdispersion. Thus, a negative binomial regression, which relaxes the assumption regarding variance equaling mean, was a more suitable model choice for our data [for more information, see 15]. Note that while our initial data featured a magnitude of zero comments, a zero-inflated negative binomial regression would not be a suitable model choice as this model assumes a divergent mechanism responsible for creating zero values compared to other values [for more information, see 16]. In the case of Reddit, various factors could be accountable for submissions receiving zero comments, like the novelty of the submission, its compliance with community rules, or its visibility due to cross-posts within different subreddits. Yet, as all submissions underly the same environmental conditions, no separate mechanism is responsible for creating zero comments. Also, note that the distribution of our dependent variable number of comments is very common for most social media content, as there is an abundance of content competing for user attention, reducing the likelihood for individual posts to receive much attention [see, e.g., 17].

To nonetheless test how zero comments might affect our model results, we reran the negative binomial regression with the Reddit BLM dataset, including submissions with zero comments (**model 1**). Only a few differences become visible when comparing the model 1 results with the results reported in our paper. In the following, we will mention significant coefficients that changed their effect direction or became significant/insignificant when comparing both models.

- Compared to the model reported in the paper, model 1's "minimal traffic" coefficient is no longer significant. This change in significance can likely be attributed to the additional data added to model 1, which mainly included submissions featuring news websites with high traffic, thus overshadowing the effect of news websites with minimal traffic.
- The coefficients of r/news and r/worldnews remain significant but negatively correlated with the number of comments in model 1. This direction shift can be explained when looking into the distributions of submissions between the three subreddits and the two datasets. When excluding submissions with zero comments from the data, r/politics comprises 4,324 submissions, r/news 1,114 submissions, and r/worldnews 435 submissions. If we add submissions with zero comments to the dataset, r/politics comprises 4,593 submissions, r/news 4,052 submissions, and r/worldnews 1,084. Thus, especially r/news and r/worldnews feature submissions with zero comments, which then affect the number of comments negatively when added to the model.
- The coefficient of "NSFW" turned out to be significant. Yet, as only 15 submissions are labeled as "NSFW," we refrain from overstating the significance of this finding due to the variable's highly skewed distribution.
- The variable "US pacific night" is no longer significant when rerunning the model with submissions featuring zero comments. This change in significance can be explained by the fact that especially submissions posted during the morning or afternoon/evening receive zero

comments. Thus, when including these submissions and rerunning the model, more noise is added that marginalizes the negative effect of submissions posted during the night.

- When comparing the coefficients of different types of news outlets between the models, the following becomes apparent: Model 1's submissions featuring links to magazines no longer significantly correlate with the number of comments compared to submissions featuring links to newspapers, whereas submissions featuring URLs to organizations/foundations and websites have a significant negative association. Submissions featuring links to TV stations significantly positively affect the number of comments compared to submissions featuring links to newspapers. Except for websites, the coefficients of the mentioned variables only change the significance level but not the effect direction, showing that the association between the variables remains the same, yet due to different data distributions, not in a significant way. The reversed effect of submissions featuring websites on the number of comments compared to submissions featuring a link to newspapers can be explained by the added submissions with zero comments, which nearly doubled the number of submissions featuring a link to websites. Thus, in model 1, these submissions have a negative impact on the number of comments compared to submissions featuring links to newspapers.
- When comparing the model fit measure AIC between both models, model 1 performs worse than the original model included in the paper (AIC value of 60,451 compared to 50,416).

We wanted to dive deeper into the effects of a submission's title's sentiment by not only evaluating how positive or negative sentiment affects the number of comments but also if it matters whether Black Lives Matter (BLM) was directly addressed within the title. In line with the negativity bias [18,19], we assumed that combining a negative sentiment with mentioning BLM in a submission's title should positively affect the number of comments. Note that we did not differentiate whether the negative sentiment directly targeted BLM, as based on the negativity bias, we assumed that negative titles generally receive more attention than positive connotated titles. We, therefore, coded the binary variable "BLM in title negative," which is "1" if BLM appears in the submission's title and if the BERT sentiment model detected a negative sentiment within the title. Otherwise, it was coded "0." In 3,848 cases, the variable was coded as “0,” and in 2,025 cases, it was coded as “1.” To also test whether a positive sentiment combined with the mentioning of BLM in a submission’s title affected the number of comments, we coded another binary variable called “BLM in title positive,” which was “1” if BLM appeared in the submission’s title and the BERT sentiment model detected a positive sentiment for the tile. Otherwise, it was coded as “0.” In 5,192 cases, the variable was coded as “0,” and in 681 cases, as “1.” We reran the model with the same dataset used within the paper while dropping the BERT sentiment variable as it highly correlated with the newly added variables “BLM in title negative” and “BLM in title positive” (**model 2**). When comparing the results between model 2 and the model reported in the paper, we find the following:

- Both variables, “BLM in title negative” and “BLM in title positive,” have a significant positive effect on the number of comments. Thus, we find evidence that it does not matter whether BLM is associated with a positive or negative sentiment within a title. Instead, referencing BLM in the title positively correlates with the users’ attention. This finding shows that it is not only the sentiment of a title but also naming specific keywords like BLM that attracts the attention of users interested in news targeting BLM.
- Only marginal differences can be detected when comparing the remaining results between the two models. Compared to the model in the paper, model 2’s “political leaning liberal” coefficient and type of news outlet “website” are no longer significant. However, the direction of the coefficients remains the same, showing a similar association.
- The model fit of both models only marginally differs. Model 2 receives an AIC score of 50,386 compared to a score of 50,416 from the model used in the paper.

Lastly, we wanted to sustain the findings of our model by using a different variable to capture user attention. The only other variables available through Reddit that provide insights into how often users engaged with a submission are “score” and “upvote ratio.” The score of a submission is determined by the difference between up- and downvotes for each submission. The upvote ratio is the ratio by which submissions have been upvoted. As mentioned in our paper, up- or downvoting a submission depicts rather an indicator of whether a submission’s content aligns with the community rules of a subreddit than being an expression of attention [20]. Yet, as this information is the only available measure besides the number of comments that provides insights into user engagement, we continued with this variable as an alternative operationalization for user attention. As both upvoting and downvoting are expressions of engaging with a submission, we needed to calculate the sum of upvotes and downvotes for all submissions. Note that it would therefore be inappropriate to use the score variable as it blends both variables by subtracting them from each other. We used the "score" and "upvote ratio" variables to calculate the number of upvotes and downvotes per submission. Inspired by [21], we utilized the following formulas for receiving the number of upvotes and downvotes where $u = upvotes$, $d = downvotes$, $s = score$, $r = upvote\ ratio$, and $a = downvote\ ratio$ with the following conditions $s = u - d,\ d = u - s$, $u = s + d$, and $a = 1 - r$:

Formulas for calculating the upvotes:

$$u = r \times (u + d)$$
$$u = r \times (u + (u - s))$$
$$u = r \times (2u - s)$$
$$u = 2ur - sr$$
$$sr = 2ur - u$$
$$sr = u \times (2r - 1)$$
$$u = \frac{sr}{(2r - 1)}$$

Formulas for calculating the downvotes:

$$d = a \times (u + d)$$
$$d = a \times ((s + d) + d)$$
$$d = a \times (2d + s)$$
$$d = 2da + sa$$
$$-sa = 2da - d$$
$$-sa = d \times (2a - 1)$$
$$d = \frac{(-sa)}{(2a - 1)}$$
$$d = \frac{-s \times (1 - r)}{(2 \times (1 - r) - 1}$$

After calculating upvotes and downvotes for all submissions, we added them to receive the new dependent variable "total votes." We used this variable as an alternative operationalization of user attention. Yet, the number of comments still depicts a more substantial and reliable measure of attention (see arguments mentioned within the paper). Similar to our data reduction before, we dropped all submissions with a score of 0 and performed a negative binomial regression with "total votes" as the dependent variable in **model 3**. Most of the results of the model used in the paper could be replicated with only some effect changes, which we will mention in the following.

- In model 3, the negative sentiment of a news article's title has a significant but negative association with the dependent variable. One possible explanation for this reversed effect could be that up- or downvoting a submission indicates whether a submission's content aligns with the community rules. Since news titles that were classified as "negative" often contain derogatory content, they sometimes disregard the strict community rules of the subreddits. Thus, by downvoting them compared to rather positive connotated titles, users are able to sanction the submission if it does not follow the community rules.
- In model 3, the variable "political leaning conspiracy" is significantly negatively correlated with the dependent variable "total votes." This negative relationship indicated that users did not like to up- or downvote a submission that featured a link to a news outlet with conspiratorial content compared to a news outlet with a neutral political leaning. Yet, we treated this result cautiously, as this category only comprises eight observations.
- In model 3, r/news and r/worldnews have a significant but negative effect on the dependent variable "total votes" compared to r/politics. Similarly to model 1, we detect a distributional shift in the number of submissions per subreddit. When dropping all submissions with a score of 0, r/politics comprised 4,006 submissions, r/news 3,890 submissions, and r/worldnews 1,021 submissions compared to 4,324 submissions in r/politics, 1,114 submissions in r/news, and 435 submissions in r/worldnews used within our model reported in the paper. Thus, the reduction of submissions with a score of zero does not remove as much noise (submissions with little engagement) as the reduction of submissions with zero comments. As the number of submissions in r/news and r/worldnews of model 3 is nearly three times higher compared to

the data distribution within the model of the paper, the reversed effect can be attributed to a distributional shift within the data.

- In model 3, the variable “night US pacific” no longer has a significant neagtive effect compared to submissions posted during the afternoon/evening. This finding can be explained through shifts in the data distribution.
- In model 3, both types of news outlets, “organization/foundation” and “website,” have a significant negative effect compared to the reference category “newspaper.” The category “TV station,” on the other side, has a positive correlation. Compared to the original model, one possible explanation for these divergent findings is that up- or downvoting a submission indicates whether a submission’s content aligns with the community rules. As the newspapers included in our dataset are most often very prominent news outlets compared to less prominent organizations/foundations and websites, it seems plausible to assume that users rather know whether a linked newspaper aligns with the community rules of a respective subreddit compared to more unfamiliar linked organizations/foundations or websites. Therefore, we assume that users up- or downvote submissions featuring links to organizations/foundations or websites less often than submissions featuring links to newspapers. TV stations, in contrast, seem to align more with the community rules compared to newspapers (positive correlation).
- When comparing the model fit of both models, the original model outperforms model 3 by far (model 3’s AIC is 86,704, compared to 50,416).

S3 Table. Negative binomial regression results of models with different datasets and variables (standard error in parentheses, p values in square brackets).

	model 1: complete data including zero comments	**model 2: data without zero comments & with BLM in title variable**	**model 3: data without zero scores & total votes as dependent variable**
	dependent count variable: number of comments	dependent count variable: number of comments	dependent count variable: total votes
VGG19 violent image	-0.096 (0.086) [0.264]	0.014 (0.072) [0.845]	-0.006 (0.079) [0.936]

BERT sentiment negative	0.104● (0.065) [0.035]		-0.283*** (0.052) [0.000]
BLM in title negative		0.302*** (0.040) [0.000]	
BLM in title positive		0.212*** (0.059) [0.000]	
political leaning conservative	-0.628*** (0.091) [0.000]	-0.203** (0.073) [0.006]	-0.911*** (0.089) [0.000]
political leaning conspiracy	-1.213 (0.776) [0.118]	-0.080 (0.990) [0.935]	-1.509* (0.724) [0.037]
political leaning liberal	-0.254*** (0.072) [0.001]	-0.048 (0.058) [0.406]	-0.130● (0.067) [0.051]
factual reporting high	0.233*** (0.052) [0.000]	0.091* (0.041) [0.029]	0.565*** (0.048) [0.000]
factual reporting low	-0.709*** (0.136) [0.000]	-1.009*** (0.124) [0.000]	-1.246*** (0.131) [0.000]
traffic high	-0.541*** (0.083) [0.000]	-0.242*** (0.066) [0.000]	-0.639*** (0.079) [0.000]
traffic minimal	-0.200 (0.269) [0.459]	-1.221*** (0.217) [0.000]	-0.665** (0.257) [0.010]
subreddit news	-0.756*** (0.051) [0.000]	0.996*** (0.048) [0.000]	-1.064*** (0.048) [0.000]
subreddit worldnews	-0.299*** (0.076) [0.000]	0.619*** (0.071) [0.000]	-0.852*** (0.070) [0.000]
NSFW	-1.399* (0.586) [0.017]	-1.174 (0.798) [0.141]	-0.447 (0.525) [0.395]
link flair	-0.317*** (0.049) [0.000]	-0.569*** (0.037) [0.000]	-0.814*** (0.046) [0.000]
number of crossposts	1.377*** (0.022) [0.000]	0.901*** (0.014) [0.000]	1.732*** (0.019) [0.000]

weekend US pacific	0.394*** (0.054) [0.000]	0.237*** (0.042) [0.000]	0.534*** (0.051) [0.000]
morning US pacific	-0.004 (0.049) [0.940]	-0.016 (0.039) [0.691]	-0.021 (0.046) [0.643]
night US pacific	0.070 (0.063) [0.262]	-0.085• (0.050) [0.090]	0.081 (0.058) [0.167]
type of news outlet magazine	0.142 (0.115) [0.215]	0.251** (0.086) [0.003]	0.380*** (0.108) [0.000]
type of news outlet news agency	-0.252• (0.133) [0.058]	-0.741*** (0.102) [0.000]	-0.495*** (0.124) [0.000]
type of news outlet organization/foundation	-0.736** (0.236) [0.002]	-0.186 (0.182) [0.307]	-0.862*** (0.227) [0.000]
type of news outlet radio	-0.350• (0.197) [0.076]	-0.338* (0.154) [0.028]	-0.761*** (0.179) [0.000]
type of news outlet TV station	0.475*** (0.057) [0.000]	0.044 (0.047) [0.341]	0.139* (0.054) [0.010]
type of news outlet website	-0.258*** (0.064) [0.000]	0.061 (0.052) [0.236]	-0.263*** (0.060) [0.000]
Constant	3.358*** (0.121) [0.000]	3.218*** (0.089) [0.000]	6.067*** (0.113) [0.000]
Observations	9,729	5,873	8,917
Null deviance	15,007.1 on 9,728 df	14,208.2 on 5,872 df	21,675 on 8,916 df
Res. deviance	9,859.7 on 9,705 df	6,894.7 on 5,848 df	11,726 on 8,893 df
AIC	60,451	50,386	86,704
Theta	0.213	0.561	0.263
Std. Error	0.003	0.009	0.003

Note: •p<0.1; *p<0.05; **p<0.01, ***p<0.001

Thus, when comparing all three models with our original model reported in the paper, the main effects remain robust with minor but explainable deviations. This finding supports the robustness of the model and its results presented in our paper. Likewise, the model reported in the paper exhibits an equal or substantially better model fit than the other models, again highlighting the model's suitability for our analysis.

References

1. Won D, Steinert-Threlkeld ZC, Joo J. Protest activity detection and perceived violence estimation from social media images. MM 2017 - Proceedings of the 2017 ACM Multimedia Conference. 2017. pp. 786–794. doi:10.1145/3123266.3123282

2. Hamborg F, Donnay K. NewsMTSC: a dataset for (multi-)target-dependent sentiment classification in political news articles. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021. pp. 1663–1675. doi:10.5167/uzh-207183

3. Rucht D. Violence and New Social Movements. International Handbook of Violence Research. Dordrecht: Springer Netherlands; 2003. pp. 369–382. doi:10.1007/978-0-306-48039-3_20

4. wondonghyeon. Protest Activity Detection and Perceived Violence Estimation from Social Media Images. In: github [Internet]. 2018 [cited 5 Jun 2023]. Available from: https://github.com/wondonghyeon/protest-detection-violence-estimation

5. fhamborg. NewsMTSC dataset. In: Hugging Face [Internet]. 2022 [cited 6 Jun 2023]. Available from: https://huggingface.co/datasets/fhamborg/news_sentiment_newsmtsc

6. Hugging Face. bert-base-uncased. In: Hugging Face [Internet]. 2023 [cited 2 Jun 2023]. Available from: https://huggingface.co/bert-base-uncased

7. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. doi:10.48550/arXiv.1810.04805

8. Horne BD, Adalis S. The Impact of Crowds on News Engagement: A Reddit Case Study. The Workshops of the Eleventh International AAAI Conference on Web and Social Media. AAAI Technical Report; 2017. pp. 751–758. doi:10.1609/icwsm.v11i1.14977

9. Ridnik T, Ben-Baruch E, Noy A, Zelnik-Manor L. ImageNet-21K Pretraining for the Masses. arXiv; 2021. doi:10.48550/arXiv.2104.10972

10. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning. PMLR; 2019. pp. 6105–6114.

11. Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems. 2022;33: 6999–7019. doi:10.1109/TNNLS.2021.3084827

12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv; 2020. doi:10.48550/arXiv.2010.11929

13. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence. 2022;45: 87–110. doi:10.1109/TPAMI.2022.3152247

14. Cameron AC, Trivedi PK. Regression analysis of count data. Cambridge university press; 2013.

15. Hilbe JM. Negative binomial regression. Cambridge University Press; 2011.

16. Yau KK, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical Journal: journal of mathematical methods in biosciences. 2003;45: 437–452. doi:10.1002/bimj.200390024

17. Feng L, Hu Y, Li B, Stanley HE, Havlin S, Braunstein LA. Competing for attention in social media under information overload conditions. PloS one. 2015;10: e0126090. doi:10.1371/journal.pone.0126090

18. Shoemaker PJ. Hardwired for news: Using biological and cultural evolution to explain the surveillance function. Journal of Communication. 1996;46: 32–47. doi:10.1111/j.1460-2466.1996.tb01487.x

19. Rozin P, Royzman EB. Negativity bias, negativity dominance, and contagion. Personality and Social Psychology Review. 2001;5: 296–320. doi:10.1207/S15327957PSPR0504_2

20. Reddit. Reddiquette [Internet]. 2021 [cited 30 Apr 2022]. Available from: https://reddit.zendesk.com/hc/en-us/articles/205926439-Reddiquette

21. How to calculate individual upvotes and downvotes of a post (Approximate). In: r/TheoryOfReddit [Internet]. 2018. Available from: www.reddit.com/r/TheoryOfReddit/comments/a0yt70/how_to_calculate_individual_upvotes_and_downvotes/

Zusammenfassung (German Summary)

Diese Dissertation beschäftigt sich mit der Frage, wie online Kollektive, insbesondere soziale Bewegungen, im digitalen Raum trotz niedriger Ein- und Austrittshürden, die fluktuierende Akteurskonstellationen begünstigen, als klar abgrenzbare und dauerhafte Akteure fortbestehen können. Zur weiteren Kontextualisierung, Einordnung und theoretischen Fundierung dieses Untersuchungsgegenstandes werden neben sozialen Bewegungen auch noch interessensbasierte online Kollektive und Protestbewegungen untersucht. Zur Bearbeitung dieses Phänomens greift die Arbeit auf das theoretische Konzept der kollektiven Identität in einem relationalen Verständnis zurück. Diese theoretische Perspektive ist besonders relevant, da in prominenten Diskussionen zur Organisation sozialer Bewegungen im digitalen Raum, insbesondere im Kontext der Unterscheidung zwischen *connective* und *collective action*, die Rolle kollektiver Identität häufig marginalisiert wurde. Empirische Befunde und neuere Arbeiten im online Kontext sowie etablierte Ansätze der Bewegungsforschung legen jedoch nahe, dass kollektive Identität ein zentraler Bestandteil sozialer Bewegungen ist und wesentlich dazu beitragen kann, diese auch im digitalen Raum als erkennbare und beständige Akteure sichtbar zu machen.

Im übergeordneten Rahmenpapier dieser Arbeit werden zunächst die theoretischen Grundlagen zur Organisation sozialer Bewegungen im digitalen Raum gelegt. Im Zentrum steht dabei nicht nur die Auseinandersetzung mit Konzepten wie Soziomaterialität und Affordanzen, sondern insbesondere die Darstellung des bestehenden Forschungsstands zur Formierung und Organisation sozialer Bewegungen online, mit besonderem Fokus auf eine kritische Diskussion von Bennett und Segerbergs *logic of connective action*. Aufbauend auf diesen Erkenntnissen sowie den identifizierten Widersprüchen und Forschungslücken, insbesondere im Hinblick auf die Formierung und Ausprägung kollektiver Identität online, formuliert die Dissertation die folgende übergeordnete Forschungsfrage, die den zentralen Ausgangspunkt für die weiteren theoretischen und methodischen Schritte der Dissertation bildet:

Wie können sich soziale Bewegungen online als klar abgrenzbare und beständige Akteure etablieren?

Im weiteren Verlauf des Rahmenpapiers werden verschiedene Definition kollektiver Identität einander gegenübergestellt und kritisch reflektiert. Auf Basis dieser Auseinandersetzung kommt die Dissertation zu dem Schluss, dass ein relationales Verständnis kollektiver Identität geeignet ist, um die kollektive Identitätsbildung sozialer Bewegungen im online Raum zu analysieren. Darauf aufbauend wird das dieser Arbeit zugrundeliegende relationale Verständnis kollektiver Identität näher erläutert. Es lässt sich entlang zweier Dimensionen beschreiben: (1) der Annahme, dass sich kollektive Identität sozialer Bewegungen in Relation zu anderen relevanten Akteursgruppen herausbildet, und (2) der Idee, dass kollektive Identität sowohl in kulturellen Ausdrucksformen als auch in sozialen Strukturen verankert ist. Zur Präzisierung der ersten Dimension wird das Konzept der *identity fields* herangezogen, welches spezifiziert, dass kollektive Identität in Relation zu drei zentralen Akteursgruppen entsteht: *protagonist, antagonist* und *audience*. Aufbauend auf dieser konzeptionellen Differenzierung und der theoretischen Abgrenzung der drei Akteursgruppen werden folgende vier untergeordnete Forschungsfragen formuliert. Sie tragen gemeinsam zur Beantwortung der übergeordnete Forschungsfrage dieser Dissertation bei und bilden zugleich die Grundlage für die systematische Zusammenfassung der Beiträge der einzelnen Artikel.

(1) Wie kreieren soziale Bewegungen im online Raum ein gemeinsames Wir-Gefühl?

(2) Wie grenzen sich soziale Bewegungen online von anderen Akteuren ab?

(3a) Wie können neue Technologien sowohl die Analyse als auch die Erstellung medialer Berichterstattung unterstützen, die sozialen Bewegungen online Identitätsmerkmale zuschreiben?

(3b) Wie beeinflusst die mediale Zuschreibung von Identitätsmerkmalen an soziale Bewegungen online die Aufmerksamkeit von Usern sozialer Medien?

Zur Konkretisierung der zweiten Dimension des relationalen Verständnisses kollektiver Identität werden die kulturelle und soziale Dimension kollektiver Identität näher erläutert. Zu diesem Zweck wird das Konzept der *identity work* eingeführt, das beschreibt, wie kollektive Identität durch das Teilen symbolischer Ressourcen Ausdruck findet. Dieses zunächst kulturell verstandene Konzept wird im Sinne des dieser Dissertation zugrunde liegenden dualen Identitätsverständnisses um eine soziale Dimension erweitert. In dieser Dissertation zeigen sich konkrete Ausprägungen der symbolischen Ressourcen in geteilten Narrativen und Memes, während die soziale Dimension anhand von Netzwerkstrukturen analysiert wird. Beide Dimensionen spiegeln eine *interne* Perspektive auf kollektive Identitätsbildung wider, da sie aus dem Handeln der Bewegung heraus entstehen. Da jedoch in der Literatur mehrfach betont wird, dass auch eine externe Anerkennung kollektiver Identität, etwa durch Medienakteure, eine zentrale Rolle spielt, wird in dieser Dissertation eine weitere Dimension berücksichtigt: die der externen Identitätszuschreibungen. Diese zeigen sich zum einen in der medialen Nutzung von Frames. Zum anderen werden diese beeinflusst von den Voraussetzungen, unter denen neue Technologien, wie KI-gestützte Tools, von Medienakteuren in der Berichterstattung eingesetzt werden. Diese Zuschreibungen bilden die komplementäre *externe* Perspektive zur zuvor beschriebenen internen Perspektive kollektiver Identitätsbildung.

Neben diesen theoretischen Zugängen zur kollektiven Identität hebt diese Dissertation die empirische Relevanz digitaler Spurdaten für die Untersuchung sozialer Bewegungen im digitalen Raum hervor, da sie die sozialen und kulturellen Spuren von Usern sichtbar machen, die zentral für das relationale Verständnis von kollektiver Identität sind. Diese Dissertation kombiniert das Potential digitaler Spurendaten mit einem Mixed-Methods-Ansatz, der sowohl computergestützte Verfahren als auch qualitative Methoden umfasst. Diese Kombination ermöglicht es, multimodale Ausprägungen digitaler Spurendaten differenziert zu analysieren und gleichzeitig eine nuancierte und umfassende Untersuchung kollektiver Identitätsbildungsprozesse im digitalen Raum durchzuführen. Darüber hinaus hilft diese methodische Verknüpfung, kombiniert mit einer soliden theoretischen Konzeptualisierung dabei, interpretative Herausforderungen digitaler Spurendaten zu bewältigen, indem methodische Schwächen ausgeglichen werden und die Ergebnisse theoriegeleitet sinnvoll interpretiert werden können.

Vor diesem Hintergrund werden sieben Papiere innerhalb dieser Dissertation vorgestellt, die sich einerseits den Forschungsfragen zuordnen lassen und andererseits systematisch entlang eines konzeptionellen Frameworks gegliedert sind. Dieses basiert auf zwei Dimensionen: Zum einen orientiert es sich an den drei Akteursgruppen der identity fields, zum anderen berücksichtigt es, welche Ausprägungen von identity work und externe Zuschreibungen in den einzelnen Artikeln analysiert werden, um die kollektive Identitätsnildung sozialer Bewegungen im digitalen Raum zu untersuchen.

Die einzelnen Beiträge der Dissertation lassen sich im Hinblick auf die zugrunde liegenden Forschungsfragen wie folgt zusammenfassen.

Artikel I bildet das theoretische und konzeptionelle Fundament dieser Dissertation. In diesem Artikel wird eine Konzeptualisierung kollektiver Identität für soziale Bewegungen online entwickelt, die die Dualität von Identität durch die Berücksichtigung sowohl kultureller als auch sozialer Dimensionen abbildet. Diese Konzeptualisierung wird mithilfe empirischer Indikatoren operationalisiert, indem konkret Narrative und Netzwerkstrukturen der Bewegung analysiert werden. Angewandt wird der Ansatz auf das Fallbeispiel der Querdenken Bewegung auf der Plattform X (ehemals Twitter). Auf Basis der empirischen Analyse werden theoretische Propositionen formuliert, die zur Erklärung kollektiver Identitätsbildungsprozesse sozialer Bewegungen beitragen. Artikel II und III untersuchen, wie das Teilen von Memes als zentrale symbolische Ressource zur Identitätsbildung beiträgt. In beiden Beiträgen wird das Konzept der *meme languages* eingeführt, die als eine einzigartige Identitätssignatur einzelner online Kollektive verstanden werden. Anhand dieses Konzepts wird untersucht, wie sich verschiedene online Kollektive, etwa Communities, die sich der Alt-Right Bewegungen zuordnen lassen, in der Ausprägung solcher meme languages unterscheiden. Im Fokus stehen dabei insbesondere der Effekt unterschiedlicher ideologischer Grenzen, Community Zwecke und Moderationspraktiken. Basierend auf den gewonnenen Erkenntnissen lässt sich nachvollziehen, wie Memes als Ausdruck kollektiver Identität fungieren und wie sich online Kollektive in ihrer jeweiligen Ausprägung von meme languages ähneln oder voneinander unterscheiden. Alle drei Artikel leisten somit einen zentralen Beitrag zum Verständnis der ersten Frage, wie soziale Bewegungen im digitalen Raum ein geteiltes Wir-Gefühl entwickeln, dieses als Zeichen ihrer kollektiven Identität sichtbar machen und sich somit auch von anderen Kollektiven abgrenzen.

Artikel I leistet zudem einen wichtigen Beitrag zur Beantwortung der zweiten Forschungsfrage, da im Rahmen der Konzeptualisierung und Operationalisierung kollektiver Identität sozialer Bewegungen im digitalen Raum insbesondere die Dimension der *distinctiveness*, also der Abgrenzung gegenüber anderen Akteuren, behandelt wird. Darüber hinaus widmet sich Artikel IV dezidiert dem Konzept der *boundary work* anhand geteilter Memes innerhalb einer Community, die der Alt-Right-Bewegung zugeordnet werden kann. Basierend auf der Social Identity Theory werden Memes qualitativ in Hinblick auf ihre Text- und Bildelemente hinsichtlich referenzierter Grenzen analysiert. Diese Analyse wird durch eine thematische, induktive Inhaltsanalyse sowie eine deduktive Rollenanalyse ergänzt. Beide Beiträge leisten somit einen Beitrag zur Beantwortung der zweiten Frage, indem sie das Verständnis der Abgrenzungsprozesse sozialer Bewegungen im digitalen Raum vertiefen.

Artikel V befasst sich mit der automatisierten Erkennung medialer Frames in der Berichterstattung über Proteste auf der Plattform Instagram. Im Mittelpunkt steht dabei die Erkennung von Frames, die mit dem theoretischen Konzept des *protest paradigm* in Verbindung stehen, sowie die Erkennung von gewaltvollen und friedlichen Protesten. Getestet werden multimodale Modelle, die eine automatisierte Auswertung von Bild- und Textdaten ermöglichen. Die Leistungsfähigkeit dieser Modelle wird anhand eines manuell gelabelten Benchmark Datensatzes evaluiert. Die Ergebnisse zeigen, dass neue technologische Verfahren einen vielversprechenden Beitrag dazu leisten können, automatisiert zu analysieren, wie Medien sozialen Bewegungen Identitätsmerkmale zuschreiben. Artikel VI liefert komplementäre Erkenntnisse, indem er untersucht, wie KI-basierte Modelle gestaltet und implementiert werden müssten, sodass sie von journalistischen Akteuren effizient genutzt werden können. Auf Basis der Literatur werden zwei zentrale Herausforderungen identifiziert, die die aktuelle Implementierung von KI-basierte Modellen im journalistischen Kontext verlangsamen. Vor diesem Hintergrund wurde ein KI-basierter Prototyp entwickelt, der mittels automatisierter Bildanalyse Dokumente aus umfangreichen Datenbanken nach ihrem Typ klassifiziert und inhaltlich relevante Beiträge priorisiert. Der Prototyp entstand in Zusammenarbeit mit der investigativen Journalismus Organisation OCCRP. Die im Entwicklungsverlauf gewonnenen Erkenntnisse werden in Bezug auf die identifizierten Herausforderungen diskutiert, woraus konkrete Fragestellungen für die zukünftige Gestaltung KI-gestützter Technologien in der journalistischen Praxis abgeleitet werden. Damit leistet der Artikel einen wichtigen Beitrag zum Verständnis der Frage, wie KI-basierte Systeme in der medialen Berichterstattung konzipiert und implementiert werden sollten. Beide Artikel leisten somit einen zentralen Beitrag zur Beantwortung Frage 3a.

Artikel VII dieser Dissertation untersucht, wie Medienakteure sozialen Bewegungen online Identitätsattribute zuschreiben und wie sich diese Framing auf die Aufmerksamkeit von Usern auswirken. Im Fokus steht die Berichterstattung über die BLM-Bewegung auf der Plattform Reddit und die Frage, wie sich gewaltvolle Bilder und negative Titel auf die Aufmerksamkeit von Usern auswirken. Zur Beantwortung der Fragestellung wird zunächst eine automatisierte Bildanalyse mithilfe von CNNs durchgeführt, um zu bestimmen, ob die in der Berichterstattung enthaltenen Bilder die Bewegung als gewaltvoll darstellen. Ergänzend werden die Beitragstitel mittels automatisierter Textanalyse auf ihr Sentiment hin klassifiziert. Die Effekte weiterer Attribute, wie etwa die politische Ausrichtung des jeweiligen Nachrichtenanbieters, werden ebenfalls analysiert. Theoretisch stützt sich der Artikel auf Konzepte wie den *negativity bias* und den *picture superiority effect*. Die statistischen Ergebnisse zeigen, dass gewaltvolle Bilder keinen signifikanten Einfluss auf die Aufmerksamkeit der User haben, während negativ formulierte Titel die Aufmerksamkeit signifikant erhöhen. Die Ergebnisse verdeutlichen, wie unterschiedliche von medialen Akteuren zugeschriebene Identitätsmerkmale sozialer Bewegungen online die Aufmerksamkeit der User beeinflussen. Diese Erkenntnisse sind von zentraler Bedeutung, da die mediale Zuschreibung von Identitätsmerkmalen an soziale Bewegungen die öffentliche Wahrnehmung sowie Unterstützung dieser Bewegung maßgeblich beeinflussen kann. Damit leistet dieser Artikel einen wichtigen Beitrag zur Beantwortung der Frage 3b.

Zusammenfassend zeigt diese Dissertation, dass soziale Bewegungen durchaus in der Lage sind, eine kollektive Identität im digitalen Raum auszubilden. Dies gelingt ihnen (1) durch die Entwicklung eines gemeinsamen Wir-Gefühls, das sich in der Nutzung symbolischer Ressourcen wie Narrative und Memes sowie in der Ausprägung konsolidierender Netzwerkstrukturen ausdrückt; (2) durch die Abgrenzung von Gegnern mittels der Zuschreibung negativer Merkmale und dem Ziehen klarer Grenzen; und (3) durch die externe Zuschreibung spezifischer Identitätsmerkmale seitens medialer Akteure, die diese Merkmale in der öffentlichen Kommunikation verbreiten. Zwar wird die Ausprägung kollektiver Identität innerhalb dieser Identitätsfelder von verschiedenen Faktoren beeinflusst, etwa durch die ideologischen Grenzen und Zielsetzungen einer Community, doch zeigen die Ergebnisse, dass soziale Bewegungen imstande sind, eine kohärente kollektive Identität zu formen. Diese kollektive Identität ermöglicht es ihnen, im fluiden und dynamischen online Kontext als erkennbare und beständige Akteure zu agieren, was somit die übergeordnete Forschungsfrage beantwortet. Die Ergebnisse widersprechen damit der Annahme von Bennett und Segerberg, dass kollektive Identität im digitalen Raum an Bedeutung verliere, und leisten einen Beitrag zum vertieften Verständnis von online Organisierung im Kontext digitalen Aktivismus.

Die der Dissertation beigefügten Papiere wurden in einschlägigen Fachzeitschriften und auf renommierten Konferenzen veröffentlicht. Lediglich Artikel V ist aktuell zur Veröffentlichung eingereicht.

Publications

Henn, T., Posegga, O. 2023. Attention-grabbing news coverage: Violent images of the Black Lives Matter movement and how they attract user attention on Reddit. PLOS One, (18)8, e0288962.

Henn, T., Posegga, O. 2023. What Do They Meme? Exploring the Role of Memes as Cultural Symbols of Online Communities. International Conference on Information Systems (ICIS). Hyderabad, India.

Henn, T. 2024. Follow the Memeing: Analyzing the Cultural Diffusion between Mainstream and Alt-Right Communities based on Shared Memes. European Conference on Information Systems (ECIS). Paphos, Cyprus.

Henn, T., Posegga, O. 2024. Peeking behind the Memes: Evaluating the Boundary Work of Online Communities through Shared Memes. International Conference on Information Systems (ICIS). Bangkok, Thailand.

Cordero, J.M., Henn, T., Holtel, F., Sánchez Gómez, J.Á., Arenas D., Šipka, A., Vollmer, S. 2025. Developing Effective and Value-Aligned AI Tools for Journalists: 12 Critical Questions to Reflect upon. Journalism Practice, 1-20.

Henn-Latus, T., Tell, S., Polenz, P., Kern, T., Posegga, O. 2025. In Search of a “Style:” Capturing the Collective Identity of Social Movements Based on Digital Trace Data. Journal of the Association for Information Systems, 26(6), 1566-1629.

Henn-Latus, T., Posegga, O. Submitted. Classifying Protest Frames in Online News Reporting Using Multimodal Models with Zero-Shot and Few-Shot Learning. Journal of Computational Social Science.