Research Article

Jana Lekscha and Milad Mirbabaie*

How to analyze cyberbullying on social media platforms

A systematic literature review in information systems

https://doi.org/10.1515/icom-2025-0005
Received February 21, 2025; accepted August 18, 2025;
published online September 4, 2025

Abstract: The use of social media across all demographic groups has brought the harmful nature of cyberbullying into sharper focus. Detecting cyberbullying on social media platforms has become widely researched in Information Systems (IS) research. However, previous literature has primarily focused on designing technical detection tools to classify cyberbullying, overlooking connected contextual, social, and psychological dynamics between users. Therefore, this research summarizes existing cyberbullying analysis methods, focusing on social and technical aspects. It evaluates existing approaches from a systematic literature review to identify suitable strategies to improve cyberbullying detection. We identified thirty-one fundamental articles in this context. We extracted three categories to address improvement potential: detection methods and technologies, socio-behavioural perspectives for cyberbullying, and regulatory and contextual influences. Based on our findings, we provide a research agenda and recommendations for designing strategies to detect cyberbullying.

Keywords: cyberbullying; cyber harassment; detection; social media; systematic literature review

1 Introduction

Social media has established itself as an integral part of social life.[1,2] More than 60 % of people worldwide are using social media.[3] The increasing interconnectivity allows constant communication, information sharing, and community building.[4] Despite many positive effects, social networks offer enormous scope for abuse phenomena in the digital space, including cyberbullying, hate speech, and cyber harassment.[5] These terms are often used interchangeably but have clear boundaries in intent and scope. First, cyberbullying is defined as "insulting, threatening, exposing or harassing individuals using communication media, such as smartphones, emails, websites, forums, chats and communities".[6] It targets individuals and uses the anonymity and reach of the digital environment to inflict harm, leading to psychological damage. Conversely, hate speech refers to cyberbullying directed at a group of people and is characterized by hostile language against those groups.[7] Third, cyber harassment represents an overarching term. It is understood as an online engagement between a bully and a victim in front of a group. Such activities include, for example, threats, damage to reputation, or harassment via e-mail or direct messaging.[8] All phenomena show the negative side of digital platforms while being different in their scope and intent. Ultimately, it is difficult to control these actions and can result in considerable psychological and social consequences.

More precisely, this psychological harm can lead to depression, concentration issues, reduced self-esteem, or, in extreme cases, suicidal thoughts.[9] Even a single severe accident can harm an individual's psychological well-being. Cyberbullying further has a greater long-lasting impact since social media reaches a larger number of people in a short period of time. In addition, the digital traces of cyberbullying can often remain for a long time, even if the original content has been deleted, which increases the impact on victims.[10] Although especially younger generations are affected by cyberbullying incidents, as around 55 % of students have experienced cyberbullying, adults also become victims, as 41 % have experienced a form of cyber harassment,[11] emphasizing the far-reaching societal impact.

Cyberbullying and its increasing societal impact have raised significant research interest in automatic detection methods of cyberbullying and removal of bullying

***Corresponding author: Milad Mirbabaie**, University of Bamberg, Kapuzinerstraße 16, 96047 Bamberg, Germany,
E-mail: milad.mirbabaie@uni-bamberg.de.
https://orcid.org/0000-0002-9455-5773
Jana Lekscha, University of Bamberg, Kapuzinerstraße 16, 96047 Bamberg, Germany

content, especially in the field of Information Systems (IS).[12] Researchers strive to develop detection methods to identify and classify such content to achieve this goal.[13] Motivated by the growing amount of research on the technical possibilities for detecting cyberbullying[12] and the aforementioned psychological issues caused by these incidents,[9] this paper aims to synthesize existing research approaches. It provides a research agenda for further researchers. So far, IS research focuses on reactive detection measures and understanding cyberbullying messages.[14–16] We aim to systematically analyze prior research to highlight key challenges and identify research avenues to enable reactive as well as proactive measures against cyberbullying. In doing so, we focus on socio-technical systems, acknowledging that social and technical aspects work together. Thus, hardware and software design are influenced by social and organizational factors.[17,18] This forms an understanding of social structures and roles that inform the design of systems and involve the communities of people and technology.[17,18]

In this way, our research paper differs from purely technological viewpoints, as we include human, institutional, and cultural dynamics in evaluating technological solutions. We thus comprise user interactions and organizational aspects of social media in the analysis. Hence, we pose the following research question: *How can socio-technical systems be leveraged to detect cyberbullying on social media using existing methods?*

To address the question, we performed a systematic literature review according to Webster and Watson.[19] In doing so, we provide an overview of suitable measures for cyberbullying detection, which can serve as a guide for future research by (1) highlighting aspects that need more attention because they were neglected so far (e.g., multilingual or multimedia detection) and (2) pointing out problems in prior research. Besides, it contributes to practice by summarizing state-of-the-art tools to develop intelligent detection measures. While making research accessible to practitioners, we narrow the gap between theory and practical application. Thus, on the one hand, we provide insights into the most valuable methods for detecting cyberbullying. We also provide an update for further theoretical research on new solutions for combating cyberbullying.

2 Background

2.1 Cyberbullying characteristics

Scientists have already examined the phenomenon of cyberbullying in numerous studies. However, there is no consensus on how to measure such messages or posts in a standardized way, as some linguistic nuances, cultural differences, and diverse communication styles need to be taken into account.[20,21] This section deals with the definition of cyberbullying.

Existing research identifies a broad definition of cyberbullying. The most comprehensive term, "cyberaggression," describes any offensive or insulting behavior using digital means.[22,23] Based on this, some studies see cyberbullying as a specific subset of cyberaggression.[20,23] The concept is partly based on the concept of traditional bullying.[20,24] The underlying definition refers to repeated and deliberate aggressive behavior to harm or harass others, strengthened by an imbalance of power between the bully and the victim.[20]

While bullying traditionally involves repeated actions, its significance in the virtual realm remains contentious. Specifically, the concept of repetition on social media is not clearly defined.[25,26] While certain research argues that repetition is crucial for the definition of cyberbullying, others suggest that even a single serious incident can be decisive[20] and can jeopardize psychological well-being. In addition, some studies show that even lasting damage and the reach of an online post can be classified as repetition.[26,27] This may also be attributed to the fact that many things on the internet are difficult to remove, and people are therefore repeatedly exposed to the same damage. The characteristic of repetition can also vary between direct and indirect forms. Thus, direct cyberbullying refers to sending harmful content to victims using, for example, direct messages.[26,28] In contrast, indirect cyberbullying intends to humiliate or isolate the other person by distributing false and harmful information, lies, or other damaging content on social media platforms.[26,28]

Further, an action is only considered bullying if it intends to harm, whether physically, psychologically, or socially. A key challenge lies in assessing the intention to harm as it requires an evaluation of the psychological motives for the action and is difficult to comprehend on online platforms. On the one hand, research suggests that individual cases should be assessed subjectively[21] or that the repetition of the harm should measure the intention.[29]

In cyberbullying, the criterion of power imbalance between the victim and the bully can manifest in new forms. In the virtual world, this imbalance can be characterized by factors such as anonymity or popularity.[5,30] Many victims feel especially powerless when the perpetrators are anonymous or particularly popular.

The literature defines additional online context-specific aspects of cyberbullying based on these criteria. In social media, it is particularly easy to remain anonymous, as users

can use fake usernames and addresses and thus conceal their identity. In addition, a we-intention on social media is also increasingly being observed, which shows that people feel safer in groups and tend to participate in collective trolling or cyberbullying.[31] Trolling refers to deliberate, provocative online behavior to disrupt conversations or annoying others. There is often no clearly defined victim. It should be distinguished from cyberbullying, as it is more strategic, and trolling is directed at individuals or groups to influence public opinion or provoke reactions.[32] Such contexts also encourage bullies to harass others without being identified and also foster the environment for cyberbullying incidents, as people do not dare to deal with the victim's reaction.[5,30,33]

On the other hand, feelings of guilt or shame can also lead to those affected not daring to report incidents for fear of reactions.[34,35] Consequently, the lack of monitoring, supervision, or intervention by authorities, such as parents, bystanders, or social media platform administrators, can exacerbate and lead to long-lasting and escalating cyberbullying incidents.[26,36] Therefore, the problem is often unpredictable and can extend beyond the immediate and observable impacts of these incidents on victims.[26]

2.2 Types of cyberbullying

Cyberbullying can have various facets, which differ from one another. To understand these distinctions, the eight common differentiations are briefly summarized.

First, cyberbullying can involve hostile verbal behavior, in the form of insults and taunts, and can occur between two or more people. This form is called *flaming*.[37,38] If a bully sends rude, offensive, or vulgar language with the aim of insulting, then research speaks of *cyber harassment*.[37,38] *Cyberstalking*, on the other hand, differentiates itself from intensive attacks, personal threats, or frightening the victim on a systematic and repeated basis.[39] These include, for example, death threats, requests for serious illnesses or disabilities, and non-consensual sexual acts. Character assassination or spreading rumors is classified in the literature as *denigration*.[38] *Masquerading* involves the bully posing as the victim and sending offensive messages, which creates the impression that these messages originate from the victim.[40,41] This can lead to confusion and damage to the victim's reputation. *Trickery* involves publishing sensitive or personal materials used to the victim's detriment.[38] For example, forwarding personal photos to expose the victim. Finally, *exclusion* and *doxing* can be distinguished. The former refers to actively excluding a person from the group community,[37,42] such as leaving specific online communities through blackmail. The eighth term, *doxing*, contains "dox", which stands for "documents".[43] The term generally describes the publication of identifiable and often private information about a person on, for example, social media.[37,44] This can include personal data or information about family circumstances.

3 Method

To understand which measures can be used to analyze cyberbullying on social media, the authors conducted a systematic literature review.

For the literature search, we included databases from the IS research. More precisely, we used the AIS eLibrary and Web of Science, as they represent the primary outlets for IS research. The first offers articles on IS conferences, and the second comprises the body of knowledge accumulated through academic journals. We focused on the premier journals of the IS discipline.[1] These journals include the eleven most important journals in the IS discipline. Thus, we draw articles from peer-reviewed journals and conference proceedings relevant to the IS field.

To define a search string to identify relevant articles in the databases, we initially reviewed related literature on cyberbullying in social media to find appropriate terms related to our research focus. During this process, we iteratively refine so that the results fit our topic but are broad enough to include suitable works that use less common terminology. From the screening our author team identified the following search strings and searched through the title and abstract: *("Hate speech" OR "Hatespeech" OR "Online hate" OR "Cyberbullying" OR "Cyber bullying" OR "Cyber harassment" OR "Online harassment") AND ("Social media" OR "Online platforms" OR "Social networks" OR "Social platforms" OR "Online communities").* The search term has been adapted to the databases' respective search options. As cyberbullying emerged with the growing popularity of social media in the 2000s,[1] all years since its introduction appear to be considered.

The literature search delivered 62 initial articles matching our search string. We followed the systematic literature review process to define our final literature sample according to Webster and Watson.[19] We employed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)[45] checklist to guarantee a sufficient standard design.

1 Decision Support Systems, European Journal of Information Systems, Information & Management, Information and Organization, Information Systems Journal, Information Systems Research, Journal of the AIS, Journal of Information Technology, Journal of MIS, Journal of Strategic Information Systems and MIS Quarterly.

Table 1: Inclusion and exclusion criteria.

Criterion	Criterion description
IC1	The language of the articles had to be English.
IC2	Articles had to be published in journals or conference proceedings.
IC3	Only research articles were included. Opinions, keynotes and viewpoints were excluded from the sample.
EC1	Literature reviews were excluded from the sample.
EC2	NeuroIS studies were excluded from the sample.
EC3	Articles that stayed at descriptive level were excluded from the sample.
EC4	Articles that do not analyze cyberbullying from detection perspective were excluded from the sample.

In the first step of literature identification, we excluded one duplicate ($n = 1$) found across the databases. In a second step, we assessed the relevance of the literature using the selection criteria to decide on inclusion or exclusion. A detailed description of the criteria can be found in Table 1.

To be included in the screening phase, the articles had to be published in English and presented in journals or conference proceedings as complete research articles or research in progress. Applying these criteria resulted in 57 articles that were part of the screening process. We started our literature screening by excluding literature reviews, NeuroIS studies, and studies on the descriptive level of cyberbullying ($n = 17$). All of the filtered articles ($n = 40$) could be retrieved. In the next step, the assessment for eligibility, we excluded research articles that did not deal with cyberbullying from a detection perspective. The articles should primarily deal with cyberbullying analysis, prevention, or detection measures, and always bring new aspects to the cyberbullying analysis on social media and refer to centrally controlled social media. In cases where it was not certain whether a criterion applied, the research papers were discussed as to whether to include or exclude them. Our final literature sample included 31 articles. We then performed an iterative forward and backward search that followed the snowballing procedure by Wohlin.[46] Thus, we used the citations and references of the selected articles. The snowballing procedure identified 52 additional articles completed in our review sample. In sum, our literature search resulted in 83 articles.

We further analyzed these articles inductively. Regarding our research objective, we focused on different methods for analyzing cyberbullying discussed in the IS literature. To this end, we carried out two steps. First, we read all articles in depth and commented on the analytical approaches presented. Second, our author team discussed the heterogeneous descriptions and standardized the terminology. The selected papers deal with an analysis or detection method of the cyberbullying problem. They therefore provide important results for answering the research question and addressing cyberbullying on social media by analyzing the issue qualitatively or quantitatively. Figure 1 illustrates the results at each step of the process.

4 Findings

We reduced the original number of articles to 31 representative papers, and after our forward and backward search, our database reached 83 articles. To present the current state of the research, we examined when the articles in our database were published. Research on cyberbullying detection has been established in our sample in IS since 2012 and continues to this day. Most of the articles were published in 2016. An overview of the publication years can be seen in Figure 2.

We used an inductive analysis approach to derive the three overarching categories to find commonalities among the selected articles.[19] While we screened the literature, we iteratively sorted the articles based on their thematic focus and aggregated them into three categories. Thus, three approaches emerged from this process. The articles could be categorized as (1) *detection methods and technologies*, (2) *socio-behavioral perspectives* for cyberbullying, or (3) *regulatory and contextual influences*. Since some papers covered multiple dimensions, we allowed multiple sorting.

The first category is characterized by research that involves technological and methodological approaches to detecting cyberbullying. This type of work is empirical, based on data collection and detection measures for cyberbullying, presenting and discussing these methods through quantitative analysis. Most of the papers identified in this approach deal with algorithmic techniques for content classification, data analysis, and platform-specific detection. The second category focuses on social and human behaviors that influence cyberbullying behavior in social media. These articles deal with the analysis of user behavior, psychological factors behind the cyberbullying phenomenon, and social dynamics between users, including

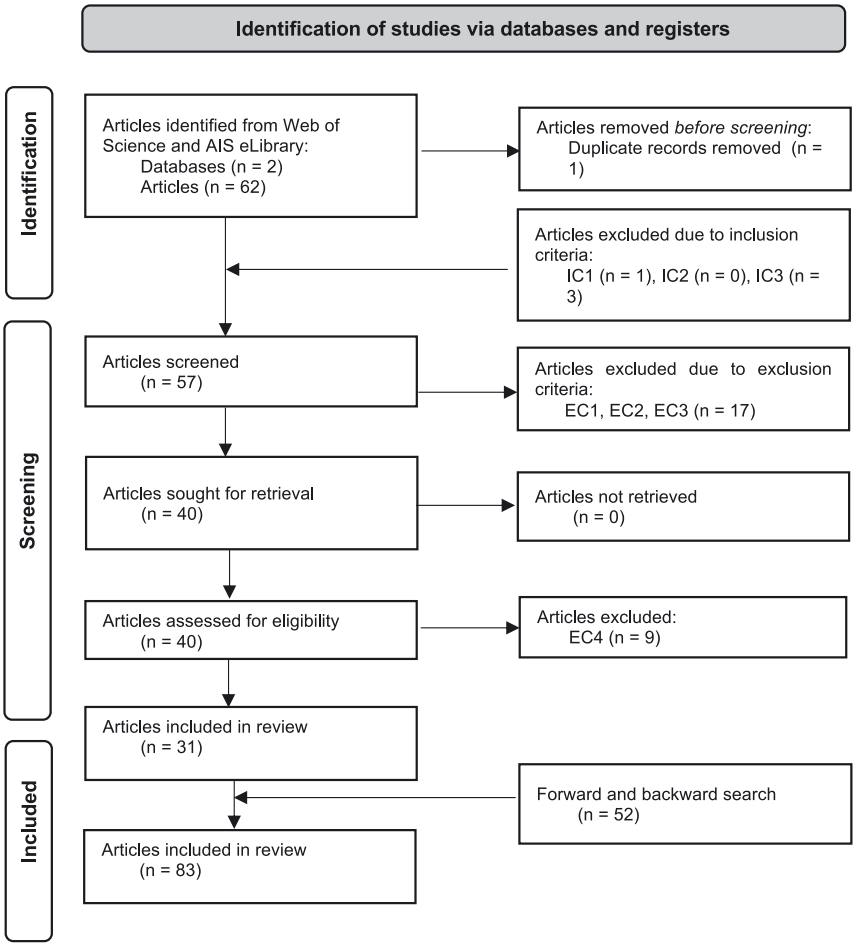

Figure 1: Literature search process.

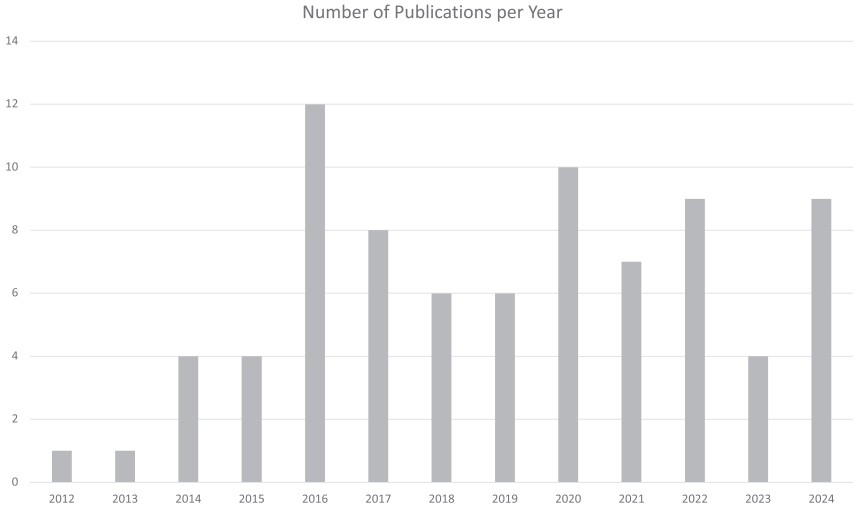

Figure 2: Number of published articles per year.

bystander reactions and peer influences. The third and final category includes studies that address cyberbullying and how legal frameworks, platform policies, and cultural or regional factors influence it. The papers look more precisely at how these aspects affect the prevalence of cyberbullying incidents. The classification provides an all-encompassing and holistic view of cyberbullying and its detection. On the one hand, (1) which technical solutions can be used to take preventive action against it, (2) which social factors encourage these incidents, and (3) which external

regulatory structures play a role. Ultimately, this classification allows for the integration of findings across all three approaches and facilitates the development of a research agenda.

A detailed overview of categorizing all the included articles after forward and backward search into the three defined categories can be found in the Appendix in Table A1.

4.1 Detection methods and technologies

The papers assigned to this focus on technological and methodological approaches to detecting cyberbullying.

In this regard, all the studies focused on analyzing social media platforms. Posts on Twitter were analyzed most frequently, but data from Facebook, Instagram, YouTube comments, and cross-platform data were also evaluated. Table 2 summarizes which platform was used for data extraction in the studies.

The research papers show a wide variety of methods. Deep learning models play a central role in detecting cyberbullying and hate speech. Models using BERT-based contextual embeddings outperform existing methods in several benchmark datasets.[57] Additionally, short-term memory (LSTM) networks[81] have proven robust in detecting complex patterns, especially when combined with more advanced approaches. Nevertheless, research also shows that models work with long-tail data, where the hate speech class is severely underrepresented.[60] On a technical level, the importance of keywords and query-based detection is also highlighted, as keyword-based filters combined with machine learning enable high precision in identifying hate speech.[47,56,66,69,71,85,86,92,93] Studies highlight that metadata-rich, large-scale datasets can significantly improve the performance of models.[59,68,94,95] Automatic annotation, for example, is an efficient approach to processing large amounts of data.[64] In addition, using AI also opens up new possibilities, as Explainable AI (XAI) is increasingly being used for cyberbullying detection.[48]

Table 2: Platforms used for data extraction in the studies.

Platform	Papers
Twitter	47–77
YouTube	75,78–80
Facebook	35,80–83
Instagram	75,84,85
Yahoo!	86
Ask.fm	87
Weibo	88
Perverted-Justice	89
Different platform	4,37,87,90–101

XAI provides insights into how classification models work and strengthens user trust by addressing many AI systems' "black box" problem.[48,61] XAI is increasingly being integrated into tools such as dashboards and supports moderators by improving transparency and increasing user acceptance.[55,84] The technology is also proving advantageous in terms of political and social biases, which can significantly affect data classification.[48,70] XAI can be used to make biases visible and increase the fairness of models.[61]

Some research papers deal with linguistic analysis by proving that emotionally charged language can increase cyberbullying.[82,89] It is also highlighted that analysis of gender-specific language patterns can demonstrably improve the accuracy of classification.[63,78,79] In addition to linguistic features, combining text and image data has great potential, especially in preventing cyberbullying, as different data modalities (e.g., images in social networks) provide important contextual information.[49,50]

With regard to context, the studies also show that it is important to identify specific cultural and linguistic environments.[52,54] For example, a new dataset was developed for Indonesian tweets[62] or the Sinhala language.[4] Arabic content was analyzed, where it was found that informal spellings need to be taken into account.[37]

4.2 Socio-behavioral perspectives

The articles assigned in this approach investigate social and human behaviors that influence cyberbullying behavior on social media.

In this regard, research papers highlight the rising cyberbullying problem, which also affects adults.[102] Especially, anonymity and easy commenting encourage cyberbullying in online contexts.[102] Thus, improving education to reduce cyberbullying incidents and create awareness of online responsibility[24,80,103] is essential. It is shown that control imbalance can contribute to increased cyberbullying behavior.[104] Implementing social media features could actively contribute to reducing cyberbullying through targeted IT design (e.g., visibility of user identities or reporting mechanisms).[104] Further research reports that certain personality traits increase cyberbullying behavior,[105] and analyzes factors influencing the reporting of such postings.[35,106]

However, the studies deal with different perspectives. Some research focuses on the perpetrators and examines fundamental character traits or factors influencing cyberbullying behavior.[24] In contrast, selected papers look more closely at the behavior of individuals not actively involved in cyberbullying incidents.[35,106,107] For example, studies focus on the bystander intervention model, which shows

that each step influences the next and that empathy, attitudes, and awareness correlate positively with effective bystander intervention.[35,108] In addition, bystander silence on social media increases the harm, while visible rejection reduces the impact of hate speech. Therefore, the importance of counter-speech to combat cyberbullying is underlined, and the need for public policies is highlighted.[109] More specifically, the literature also highlights that behavioral intervention in the form of the perceived emergency caused by the cyberbullying incident increases responsibility and awareness, thus the reporting behavior.[35]

Despite the focus on the socio-behavioral perspective, AI can be used to combat cyberbullying. AI-generated prompts can effectively increase bystanders' empathy and motivate proactive reporting behavior, which can ultimately encourage intervention in cases of cyberbullying.[15] While focusing on bystanders, it is suggested that platform owners should raise awareness about the availability of reporting tools and focus on developing tools that positively impact the sub-dimensions of trust in the system's anonymity.[35,106]

Further, studies have found a connection between individual personality traits and spreading hate speech on the Internet. Research approaches show that certain personality traits can promote or increase the spread of harmful content.[110] In addition, connections between an individual's emotional state and their harmful intent suggest that emotional analysis can be a valuable tool in cyberbullying detection.[111]

In addition to classic posts, it is also necessary to analyze contextual characteristics such as user demographics, posting habits, and user activities. For this reason, studies also focus on context-aware, supervised models that can be built for more precise detection of intersectional hate speech.[66]

The importance of peer groups is also explained, as the role of peer influence in spreading cyberbullying behavior is confirmed and has a significant impact on cyberbullying dynamics.[87]

4.3 Regulatory and contextual influences

The papers assigned to this approach focus on external influencing factors such as laws, platform guidelines, and cultural or geographical conditions.

In general, anonymity and the lack of accountability in online environments facilitate cyberbullying. Further, there is a consensus about the need for better education to help prevent inappropriate online behaviors.[103,112] Therefore, victims must understand how to responsibly use digital tools, platforms, and social media. From the offender's point of view, they must understand their obligations as digital citizens and cyber-ethics to change their immoral actions to prevent deviant behaviors such as cyberbullying.[24]

Studies show that legal interventions, such as introducing specific laws, can significantly reduce the intensity of hateful language on social media.[77] However, social media platforms often over-block suspicious content to avoid legal consequences.[113] Nevertheless, success depends heavily on user reports and EU-internal monitoring.

In addition, much research analyzes different cultural and linguistic aspects.[37] For example, sensitive events, such as the pandemic, lead to a rise in hateful posts targeted at specific communities,[72] or tweets in African American English and from self-identified African Americans are more often classified as offensive.[64] This highlights the need to consider dialect-related biases to avoid unintentional racial bias.[37]

When considering cyberbullying in a political context, it is important to assess biases in training data, as they significantly affect the performance of hate speech classifiers.[51] XAI can be used to uncover and fix these biases.[48]

Despite technological advances and the potential uses of AI and XAI, it remains relevant to consider social and political contexts to avoid bias and enable successful detection. Regulatory guidelines and counter-narrative promotion offer additional potential to mitigate cyberbullying in social media.

5 Discussion

Our dataset is limited to articles with "cyberbullying" and "social media" or comparable synonyms in the title or abstract. Although the search criteria were specific, the hit rate was still relatively high after screening and relevance check. The evaluation of our paper selection shows that cyberbullying detection has been a topic of increasing importance in research since 2012 (see Figure 2). This may be due to the drastic increase in the use of social media platforms since then. Average social media usage has almost doubled compared to 2024, resulting in the increased spread of hate speech.[114]

As a first category, we identified an approach of *detection methods and technologies* in our systematic literature review. The given research papers mainly focused on collecting and analyzing Twitter data (see Table 2). Although using Twitter data is quite common in IS research,[115] it also has disadvantages.[93,116,117] Firstly, Twitter only made limited amounts of data available through the API, limiting the scalability of data collection.[117] After the transformation to X, data extraction became limited to data collection,

which may reduce the representativeness and comparison. Furthermore, the platform, and thus also the research papers based on it, are often limited to textual content, which was originally limited to a short format of 140 characters and later expanded to 280 characters.[116] Notwithstanding this simplification of analysis using machine learning techniques,[47,56,66,69,71,85,86,92,93] it excludes cyberbullying in the form of visual content, such as photos, videos, or audio recordings. This implies that the results have limited generalizability, and trickery is excluded from the outset.

The overall problem of cyberbullying across social media platforms cannot be completely examined. For example, Slivko and Andres[77] focused on the right-wing parties in Germany and Austria. While this specialization offers potential as the target group often promotes xenophobic attitudes on social media, it also confines the research scope to a specific demographic subset. Similar generalizability problems may occur when analyzing specific languages[4,37,62,77,118] or a single social media platform.[4,74,77,119] To enhance the validity and relevance of the findings, the findings should be validated across various platforms and broadened to encompass the commonly used social media in today's society.[120]

Combining the data development approach with cyberbullying detection reveals that quantitative research regarding cyberbullying is mainly from a computational perspective and includes machine learning techniques up to AI.[4,101,110,113] These approaches especially enable automated cyberbullying detection, leading to high potential in detecting hate speech reactively.[27,121] However, research should address the reasons behind cyberbullying and elaborate further on prevention measures. The aforementioned restriction to textual content constitutes a disadvantage because most individuals use multimedia platforms, such as Instagram, which have not been considered so far.

Additionally, cyberbullying is challenging to identify due to its context-dependent nature, which can vary across different cultures and languages,[121] but also depends on ongoing conversations. Consequently, distinguishing between legitimate speech and hate speech is particularly difficult for pre-trained models, especially given the significant fluctuations in language use and the evolving patterns for hate speech classification. Besides, as they are associated with context dependency, it is difficult for data models to characterize linguistic metaphors, different language styles, and subjectivity, and to correctly classify them as cyberbullying.[4,74] Such a context dependency can either be in the post itself, if images are included, or in the context of the conversation, making it important to be able to understand conversations.

To solve these challenges and the ongoing data biases, there is a need for continuous updating of the dataset to include evolving language and cultural knowledge. Studies have already recognized the importance of context-specific variables, such as user activities or profiles.[83] Therefore, carefully selecting and considering potential sources of biases is essential to implement more precise automatic cyberbullying detection. Further, a human intermediary is possibly needed to assess the incident individually, particularly in cases where it leads to criminal charges.

We further identified that the social context is being considered in IS research. Thus, socio-behavioral perspectives and regulatory influences can be leveraged to analyze cyberbullying. These articles deal with the background of cyberbullying[103] and also with the role of personality characteristics.[35,90,106]

While Trabelsi et al.[103] also emphasize the complexity of moderation systems due to the contextual nature of social media platforms, they suggest introducing community guidelines on platforms. These should prevent cyberbullying based on societal interests and through more educational opportunities.

Similarly, Kaluarachchi and Trieu[24] provide unique qualitative insights into cyberbullying dynamics by analyzing court case transcripts. In this respect, it is one of the few studies overcoming the issue of data-based bias. However, the research findings highlight that the responsibility for cyberbullying incidents falls firmly on the perpetrators and victims. It possesses certain strengths in creating a general awareness of the use of technology and possibly reducing cyberbullying incidents. However, it is questionable whether targeted cyberbullying attacks can be avoided this way (e.g.[24,103]) or whether increased external monitoring, such as AI approaches, is necessary.[15,113]

Incorporating personality traits, such as openness, conscientiousness, extraversion, agreeableness, and neuroticism, in automatic detection, the accuracy of correctly distinguishing between hate speech and non-hate speech is significantly enhanced.[90] To increase the precision of automatic cyberbullying recognition systems, additional personality characteristics or theories, besides Barrick and Mount[90] can be examined to expand the successful classification further. Thus, research has identified the dark triad (machiavellianism, narcissism, psychopathy) as important factor explaining behavior in online communities.[122,123]

Despite criticism that systematic literature reviews have a very specific focus, our aim with this method was to explore how socio-technical systems can be leveraged to detect cyberbullying on social media platforms by systematically identifying and evaluating existing methods. Thus, a systematic literature

review is a suitable method for this. However, it should be acknowledged that our research findings are derived from a restricted selection of research articles. Consequently, the recommendations and conclusions presented herein predate this study pool.

6 Research agenda

Our systematic literature review provided insights into cyberbullying analysis methods on social media platforms and contributed to our understanding of the current challenges in research. Based on these findings, we present research and practical recommendations on further cyberbullying analyses.

6.1 Detection methods and technologies

First, research primarily used Twitter data to analyze hate speech.[4,74,77,119] Thus, it concentrated on textual analysis methods using machine learning techniques andNatural Language Processing (NLP) to detect cyberbullying content. Only a few studies researched qualitatively.[24,103] Significant advancements and diverse methodologies are evident in automatic detection measures. Research progressively focuses on technical solutions, with studies encompassing a spectrum from traditional linguistic categorization to exploring social dimensions.

Further, the potential of AI in automatic detection and the significance of Generative AI models were underscored.[113] AI-based detection components seem to be suitable for analyzing social media content. Since research on AI has considered issues of explainable and linguistic hate speech and cyberbullying detection in isolation, a gap in system and dataset design can be identified. Therefore, automatic detection should be able to process different types of content and languages. A system should therefore be able to recognize images, texts, and videos and to interpret the context to which the content refers.

The current focus on analyzing public posts on a single platform may underrepresent the true extent of cyberbullying. Private messaging features on social media platforms likely harbor many undetected incidents due to their increased anonymity.

To overcome problems in language changes, it is essential to regularly update the vocabulary of the detection component and incorporate new trending words. Further, additional common personality traits can enhance automatic detection systems to understand cyberbullying behaviors. Since context is crucial to correctly classify cyberbullying, combining machine learning models with human moderation can be recommended to handle ambiguous cases and improve precision. This may be particularly advisable when dealing with serious cases that could have legal consequences. To leverage human expertise and empower social media users to report cyberbullying content, integrating a reporting button on social media platforms is a viable solution.

Research highlights the critical role of education and societal awareness in mitigating cyberbullying incidents.[35,103,106] Additionally, incorporating personality traits into detection measures has enhanced accuracy.[90] Emphasis should also be placed on improving reporting systems and fostering user responsibility. Based on these findings, the following recommendations for action arise:

Possible research questions:

(1) *To what extent can personality traits analysis improve automated cyberbullying detection on social media?*
(2) *How can cyberbullying detection be adapted to various social media platforms?*
(3) *How should AI-based cyberbullying detection be designed to analyze multimedia content on social media?*
(4) *How can cyberbullying detection methods be implemented to analyze private conversations?*
(5) *How can detection methods be adapted to recognize the context of social media posts?*

6.2 Socio-behavioral perspectives

Given the increasing importance of social media for various generations, it is recommended to develop and promote educational initiatives that teach digital citizenship, responsible social media use, and the impact of cyberbullying. Such initiatives could significantly enhance the reporting behavior of bystanders.

While there are already reporting forms for cyberbullying on social media posts in German-speaking countries, these forms are typically found on separate websites rather than being integrated with social media platforms. To improve user-friendliness and encourage the reporting of offensive content, research should focus on developing plugins for mobile apps that can be seamlessly integrated with social media applications. Additionally, strengthening community guidelines on social media platforms and providing robust support systems for victims and bystanders would further enhance the effectiveness of these measures.

Possible research questions:

(1) *How can users' digitally responsible behavior be encouraged?*
(2) *To what extent can individual reporting decrease cyberbullying incidents?*

(3) *How should education initiatives be designed to foster a digitally responsible social media use, considering different age groups?*
(4) *How should social media platform operators integrate reporting possibilities that they will be actively used?*

6.3 Regulatory and contextual influences

Studies in this approach analyze external factors influencing cyberbullying and hate speech. These include laws, various guidelines for platforms, and geographical conditions. Research shows that laws can significantly affect the intensity of hate speech.[77] However, the question remains about how sustainable these effects are and how they vary nationally. Furthermore, platform policies are also considered key elements, but the problem of so-called "over-blocking" is not comprehensively analyzed. Since the systematic literature review only refers to centrally controlled social media, extending the research to decentralized social media would be interesting, as they can differ significantly due to the lack of central authority.

Possible research questions:

(1) *To what extent do legal regulations against hate speech have a long-term impact on users and their behavior?*
(2) *What are the legal and ethical implications of over-blocking practices?*
(3) *What side effects do platform measures against cyberbullying have on freedom of expression on social media?*
(4) *How can harmful content be regulated in decentralized social media networks without undermining user autonomy and freedom of expression?*

7 Conclusion

In this study, we investigated the common cyberbullying detection research. We conducted a systematic literature review to examine how socio-technical systems can be leveraged to detect cyberbullying on social media using existing methods. Our analysis revealed 83 articles that were published during the last two decades. Based on our analysis, we inductively identified three categories of research: (1) *detection methods and technologies*, (2) *socio-behavioral perspectives,* and (3) *regulatory and contextual influences*. These categories present a holistic view of cyberbullying detection and highlight the importance of different lenses. Thus, focusing on the interplay between understanding human behavior and contextual influences and using algorithm-based detection for cyberbullying on social media is important. Our categorization showed how socio-technical systems can be used to understand cyberbullying detection in terms of technical, social, and regulatory aspects.

While these approaches have provided valuable insights, they are not without limitations. Many approaches rely predominantly on Twitter data, limiting the scope of analysis to text content and overlooking platforms with photo or video content and private communication, such as Instagram. In addition, research often focuses on certain linguistic features, which compromises the generalizability of the results.

To address these shortcomings, we have made recommendations that suggest more diverse datasets for further research and addressing different social media platforms. Efforts to improve automatic detection measures should consider evolving language trends and cultural nuances. Incorporating personality traits into detection algorithms can improve accuracy, while human moderation can provide crucial context in ambiguous cases, especially those with legal implications. Further, societal interventions in the fight against cyberbullying also represent a vital solution alongside technical options. In this case, it would be conceivable to prepare younger generations in particular for the responsible use of social media and enable the possibility of reporting content via plugins or reporting forms. Finally, social media operators should strengthen community guidelines and integrate reporting options into mobile applications.

This study is subject to limitations. One limitation of our systematic literature review is that it is conducted from the perspective of IS researchers. Thus, our literature search analyzes leading information systems journals and conferences. Future research should analyze cyberbullying detection at an interdisciplinary level to ensure better generalizability of the results. Second, despite careful selection, there is a particular risk of bias in the articles presented. This is based on our decision not to conduct a separate quality assessment. Instead, we deliberately relied on the methodological quality of the articles, which was ensured by the targeted selection of high-ranking, peer-reviewed publications.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: This manuscript underwent a language check using AI-based tools to improve clarity.

Conflict of interest: The author states no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

Appendix

See Table A1

Table A1: Literature classification.

Authors	Approaches			Method, key finding & cyberbullying type
	Detection methods and technologies	Socio-behavioral perspectives	Regulatory and contextual influences	
Agarwal & Chowdary (2021)[72]	x		x	An adaptive ensemble learning model was used to detect hate speech, showing strong cross-dataset performance on topics like COVID-19 and elections.
Ahmed et al. (2022)[71]	x			A transformer-based ensemble model was developed to detect trait-based cyberbullying on Twitter, achieving high F1-scores on both balanced and imbalanced datasets.
el Akbar et al. (2019)[70]	x		x	Using a Naïve Bayes classifier in WEKA, the study detected political hate speech on Indonesian Twitter with 93.4 % accuracy.
Al-garadi et al. (2016)[69]	x			A supervised machine learning model using tweet, user, and network features was developed to detect cyberbullying on Twitter, achieving high accuracy with an F-measure of 0.936.
Arango et al. (2022)[13]	x			This study critiques hate speech detection models for overestimated performance due to dataset bias and overfitting and highlights the need for better validation and cross-lingual generalization.
Arayankalam et al. (2024)[97]	x		x	A mixed-method analysis of 179 countries shows that centralized online regulation increases political hate speech by amplifying government surveillance and disinformation.
Ashktorab (2016)[84]	x			Using a participatory design approach, the study explores technological interventions for mitigating cyberbullying on Instagram, focusing on tertiary prevention after harm has occurred.
Aivazpour & Beebe (2020)[124]		x		Using a vignette-based experiment, the study found that power imbalance and supportive bystanders increase cyberbullying intention, especially under anonymity.
Ayo et al. (2020)[68]	x			This survey evaluates machine learning methods for hate speech detection on Twitter, proposing a generic metadata architecture that achieves strong classification performance across multiple metrics.
Badjatiya et al. (2017)[67]	x			Using deep learning to classify tweets as racist, sexist, or neither, the study achieved significant performance gains over traditional n-gram methods.
Bretschneider et al. (2014)[74]	x			A pattern-based approach with text normalization and person identification was used to detect online harassment, improving classification performance in social networks.
Bretschneider & Peters (2016)[73]	x			The study introduces a harassment graph to detect cyberbullies and victims in online communities, enabling severity measurement and early intervention.
Burnap & Williams (2016)[66]	x	x		Using typed dependency parsing, the study classifies cyber hate on Twitter across protected characteristics like race and disability, improving detection of intersectional hate speech.

Table A1: (continued)

Authors	Approaches			Method, key finding & cyberbullying type
	Detection methods and technologies	Socio-behavioral perspectives	Regulatory and contextual influences	
Caron et al. (2022)[75]	x			The study applies transfer learning and Transformer models to detect toxic language, aiming to support automated content moderation on social media platforms.
Dadvar & De Jong (2012)[78]	x			The study proposes enhancing cyberbullying detection by incorporating user characteristics and cross-platform behavioral analysis beyond just content.
Dadvar et al. (2013)[79]	x	x		By incorporating user context and profile information, the study improves cyberbullying detection beyond content-only approaches.
Davidson et al. (2017)[64]	x			A multi-class classifier was trained to distinguish hate speech from offensive language on Twitter, revealing that racist and homophobic content is more reliably detected as hate speech than sexist content.
Davidson et al. (2019)[65]	x		x	The study reveals racial bias in hate speech detection datasets, showing that classifiers disproportionately label African-American English tweets as abusive.
Del Bosque & Garza (2014)[63]	x			The study treats aggressive text detection as a regression problem and finds that multi-feature approaches, especially those using profane word detection, perform best on Twitter data.
Del Vigna et al. (2017)[81]	x	x	x	Using annotated Italian Facebook comments, the study compares SVM and LSTM models for hate speech detection, showing strong performance across multiple hate categories.
Ejaz et al. (2024)[95]	x			The study introduces a semi-synthetic dataset for cyberbullying detection that incorporates aggression, repetition, peerness, and intent to harm, enabling more comprehensive model evaluation.
Elisabeth et al. (2020)[62]	x			Using machine learning models, the study detects implicit hate codes in Indonesian tweets, achieving high accuracy for hate-free content but lower performance on identifying hidden hate speech.
Emmery et al. (2022)[92]	x			The study shows that cyberbullying classifiers are highly sensitive to adversarial lexical changes, but robustness can be improved through data augmentation using perturbed samples.
Fernando & Deng (2023)[4]	x			Using class-based feature selection with machine learning, the study improves hate speech detection in Sinhala social media, achieving higher F1-scores and better generalization.
Gemes et al. (2021)[61]	x			The study combines a BERT-based model with a high-precision rule-based system to detect offensive tweets, achieving strong results and highlighting label ambiguities in the HASOC dataset.
Gröndahl et al. (2018)[94]	x			By testing seven hate speech detection models, the study reveals their vulnerability to simple adversarial text manipulations and highlights that data quality matters more than model architecture.
Guo & Gauch (2024)[93]	x			The study improves cyberbullying detection by integrating sarcasm detection into a BERT-based MLP, reducing misclassification of sarcastic abusive content.

Table A1: (continued)

Authors	Approaches			Method, key finding & cyberbullying type
	Detection methods and technologies	Socio-behavioral perspectives	Regulatory and contextual influences	
Haidar et al. (2017)[37]	x	x		Using machine learning and NLP techniques, the study builds a detection system for Arabic-language cyberbullying, filling a gap in multilingual research.
He et al. (2024)[15]		x		An AI-based experimental system shows that empathy-triggering prompts can enhance bystander intervention and reporting in cyberbullying cases.
Hosseinmardi (2015)[85]	x			This study detects cyberbullying incidents on Instagram using classifiers trained on image-comment pairs, highlighting the value of combining visual and textual features.
Hu et al. (2024)[125]		x		Applying situational action theory, the study finds that cyberbullying behavior is shaped by habitual actions, peer influence, and neutralization techniques.
Kaluarachchi & Trieu (2022)[24]			x	Through an exploratory study, the authors suggest that digital literacy for victims and cyber-ethics for offenders are key strategies to prevent cyberbullying.
Kaluarachchi et al. (2022)[126]		x	x	Using qualitative analysis of 75 court cases, the study examines adult cyberbullying by analyzing the roles of offender, victim, technology, and guardianship.
Kaluarachchi et al. (2020)[127]			x	The study proposes a conceptual model for cyberbullying intervention grounded in crime and routine activity theories, focusing on offender motivation and opportunity.
Kaluarachchi et al. (2021)[128]		x		An investigative model for adult cyberbullying is developed and validated using international court cases, offering insight into behavioral and legal patterns.
Khairy et al. (2021)[111]		x		This study evaluates in an online survey whether Egyptian Facebook users pay attention, and act on a reporting system. The extent of users' satisfaction with the reporting systems are measured.
Lee & Ram (2020)[90]	x	x		The PERSONA model uses personality-based deep learning to detect hate speech, showing improved accuracy by incorporating inferred psychological traits.
Lee & Ram (2024)[110]		x		By integrating low-level personality factors into a deep learning framework, the study significantly enhances hate speech detection performance
Lee et al. (2018)[76]	x			The study develops a cyberbullying detection model for Twitter using text features, sentiment, and readability analysis, finding that incorporating readability improves prediction of harassment and ridicule.
Li et al. (2016)[96]	x	x		The study proposes a cyberbullying detection method based on parent-child comment relationships, using third-party reactions to improve accuracy with publicly available data.
Li et al. (2024)[83]	x			The study proposes a low-resource framework for hateful meme detection that maintains high accuracy while reducing computational demands, aiming to support equitable access to AI moderation tools.
Loebbecke et al. (2021)[113]			x	The study conceptualizes hate speech and explores the regulatory context for developing AI-based detection systems, aiming to build and evaluate an AI solution for tackling online hate.

Table A1: (continued)

Authors	Approaches			Method, key finding & cyberbullying type
	Detection methods and technologies	Socio-behavioral perspectives	Regulatory and contextual influences	
Lowry et al. (2016)[102]		x		Using a modified social learning model, the study shows that anonymity and heavy social media use among adults foster cyberbullying through socialization and disinhibition mechanisms.
Lowry et al. (2017)[104]		x		Using control balance theory and a factorial survey with adults, the study finds that certain IT design features can reduce cyberbullying by influencing users' sense of control, accountability, and deindividuation.
MacAvaney et al. (2019)[60]	x			The study proposes a multi-view SVM model for hate speech detection that balances interpretability and performance, addressing key challenges like linguistic subtlety and definitional ambiguity.
Mangaonkar et al. (2015)[59]	x			The study introduces a collaborative computing approach to detect cyberbullying on Twitter, achieving faster and more accurate results compared to stand-alone methods.
Marwa et al. (2018)[58]	x			The study applies deep learning models like LSTM, BLSTM, and CNN to classify online harassment in tweets, showing strong performance on a human-labeled harassment dataset.
Mozafari et al. (2020)[57]	x			The study fine-tunes BERT using a transfer learning approach for hate speech detection on Twitter, achieving strong performance in identifying racism, sexism, and offensive content.
Muneer & Fati (2020)[56]	x			The study compares seven machine learning classifiers for detecting cyberbullying on Twitter without victim input, finding that logistic regression achieves the best overall performance.
Nahar et al. (2014)[112]			x	The study introduces a semi-supervised fuzzy SVM approach for cyberbullying detection in streaming social media data, achieving strong performance with minimal labelled input.
Nandhini & Sheeba (2015)[55]	x			The study uses the Levenshtein algorithm and Naïve Bayes classifier to detect and categorize cyberbullying activities such as flaming, harassment, racism, and terrorism on social networks.
Nickerson et al. (2014)[108]		x		The study applied the bystander intervention model to bullying and sexual harassment. A confirmatory factor analysis with secondary school students confirmed its five-factor structure.
Nobata et al. (2016)[86]	x	x		The study develops a machine learning approach for detecting abusive language in online comments, outperforming deep learning baselines and enabling nuanced analysis across contexts.
Ogunleye & Dharmaraj (2023)[98]	x			The study applies large language models like BERT and RoBERTa to cyberbullying detection, finding that RoBERTa outperforms traditional models on datasets from Formspring and Twitter.
Özçift et al. (2019)[54]	x			The study develops a cyberbullying detection model for Turkish tweets using grid search–optimized Bayesian logistic regression, achieving high accuracy through feature selection and supervised learning.

Table A1: (continued)

Authors	Approaches			Method, key finding & cyberbullying type
	Detection methods and technologies	Socio-behavioral perspectives	Regulatory and contextual influences	
Potha & Maragoudakis (2014)[89]	x			The study applies time series modelling and SVD-based signal analysis to predict the severity of cyberbullying in predator-victim dialogues, using neural networks for classification.
Rajput et al. (2022)[99]	x			The study introduces a CNN-LSTM model for detecting hate-inducing memes in code-switched Indian languages, achieving strong results on a newly created multimodal political meme dataset.
Romsaiyud et al. (2017)[53]	x			The study enhances Naïve Bayes with k-means clustering of word patterns to detect cyberbullying, improving classification accuracy across eight predefined categories.
Sap et al. (2019)[52]	x		x	The study reveals that hate speech detection models exhibit racial bias against African American English due to biased annotations and shows that dialect-priming can mitigate this effect.
Schieb & Preuss (2016)[82]	x			Through literature review and simulation, the study finds that the effectiveness of counterspeech on Facebook depends on the relative size of hate groups and the persuasive power of counter-speakers.
Silva et al. (2016)[80]	x	x		The study presents an automated model for detecting cyberbullying on social networking sites and integrates it into a Facebook app that alerts parents to harmful interactions involving their children.
Singh et al. (2016)[51]	x		x	The study introduces a probabilistic socio-textual information fusion model that accounts for feature confidence and interdependencies, significantly improving cyberbullying detection in social networks.
Slivko & Andres (2021)[77]	x		x	Using a difference-in-differences analysis, the study finds that Germany's NetzDG law significantly reduced hate intensity on Twitter, especially among highly toxic and influential users.
Svetasheva & Lee (2024)[100]	x			The study leverages large language models for hate speech detection, showing that LLM-generated synthetic data can effectively boost performance in low-resource and complex domains.
Squicciarini et al. (2015)[87]	x	x		The study models cyberbullying dynamics by analyzing peer influence and user interactions in social networks, revealing how bullying behavior can spread through social pressure.
Tahmasbi & Fuchsberger (2018)[27]	x	x		Through a systematic literature review, the study identifies key challenges in automated cyberbullying detection and outlines future research directions to improve accuracy and effectiveness.
Tahmasbi & Fuchsberger (2019)[101]	x			The study introduces ChatterShield, a multi-platform self-training cyberbullying detection system that incorporates human moderator feedback to enhance detection across minors' social networks.
Tahmasbi & Rastegari (2018)[107]		x		The study develops a cyberbullying detection model that incorporates social context and user network positions, showing that cyberbullying often occurs without explicit negative language.

Table A1: (continued)

Authors	Approaches			Method, key finding & cyberbullying type
	Detection methods and technologies	Socio-behavioral perspectives	Regulatory and contextual influences	
Trabelsi et al. (2022)[103]		x	x	This study uses focus groups with online content moderators to explore how cyberbullying is conceptualized, revealing that context and platform dynamics play a crucial role in detection and that clearer moderation guidelines are essential for effective intervention.
Vijayakumar et al. (2021)[91]	x			The study uses a CNN-based deep learning model to detect image-based cyberbullying and integrates a chatbot system to alert users and guardians, addressing the growing threat of visual online harassment.
Wang et al. (2020)[50]	x			This study proposes a fine-grained cyberbullying detection model using a Graph Convolutional Network (GCN) trained on a semi-supervised, balanced dataset generated via Dynamic Query Expansion (DQE); results show the GCN model effectively classifies bullying based on age, ethnicity, gender, religion, and other victim traits.
Waseem & Hovy (2016)[49]	x			This study introduces a hate speech detection approach on Twitter using critical race theory-informed annotation and a combination of character n-grams with extra-linguistic features, revealing that user-level attributes can be strong predictors of racist and sexist content.
Wich et al. (2020)[48]	x		x	This study examines how political bias in training data affects hate speech detection, showing that models trained on left-, right-, or neutral-biased datasets perform differently and that explainable ML tools can reveal and mitigate such bias.
Withers et al. (2017)[129]		x		Using personality and addiction scales on 290 young participants, the study finds that narcissism, Machiavellianism, and psychopathy are positively linked to social media addiction, while social anxiety strongly predicts addiction but does not moderate the Dark Triad's effect.
Wong et al. (2016)[106]		x		The study applies social cognitive theory to examine how self-efficacy and response efficacy influence users' willingness to report online harassment on social networking sites.
Wong (2017)[130]		x		The study develops a framework for understanding how victims of online harassment use platform features as protective coping strategies, guided by the extended parallel process model (EPPM).
Wong et al. (2021)[35]	x	x		The study applies a bystander intervention framework to social media harassment, showing that platform design and sociotechnical factors influence users' willingness to report harassment proactively.
Wu et al. (2020)[88]	x	x		The study proposes an improved TF-IDF with position weighting combined with fastText to build a binary classifier, achieving efficient and accurate cyberbullying detection on social networks.

Table A1: (continued)

Authors	Approaches			Method, key finding & cyberbullying type
	Detection methods and technologies	Socio-behavioral perspectives	Regulatory and contextual influences	
Xie et al. (2023)[121]	x			This review examines current NLP and ML approaches to hate speech detection, highlights key challenges like implicit expressions, and suggests integrating interdisciplinary insights to improve deep learning model generalization.
Zapata et al. (2024)[109]		x		This study shows that only majority or unanimous bystander opposition reduces perceived harm of hate speech, underscoring the role of collective norms in shaping responses.
Zhang et al. (2018)[47]	x			This study introduces a Convolution-GRU model for hate speech detection on Twitter, outperforming baselines across most benchmark datasets.
Zhang et al. (2022)[131]			x	Using criminological theory, the study finds that low self-control and digital routines significantly predict workplace cyberbullying, offering organizational insights.
Zhu et al. (2023)[132]		x		This study finds that bystanders engage in celebrity cyberbullying when social capital gains outweigh retaliation risks, shaped by publicity and network ties.

References

1. Kumar, A. M.; Gupta, S. Governance of Social Media Platforms: A Literature Review. *Pac. Asia J. Assoc. Inf. Syst.* **2023**, *15*, 56—86.
2. Ruiz-Bravo, N. V.; Selander, L.; Roshan, M. Preparing, Fostering, and Fallowing: Cultivating Digital Safe Spaces. In *Proceedings of European Conference on Information Systems*; Paphos, Cyprus, 2024.
3. Richter, F. *Infographic: Always on?* Statista Daily Data. https://www.statista.com/chart/14088/frequency-of-internet-usage-in-the-united-states (accessed 2025-02-14).
4. Fernando, E. N.; Deng, J. D. Enhancing Hate Speech Detection in Sinhala Language on Social Media Using Machine Learning. In *Proceedings of Australasian Conference on Information Systems*; Wellington, New Zealand, 2023.
5. Anjum; Katarya, R. Hate Speech, Toxicity Detection in Online Social Media: A Recent Survey of State of the Art and Opportunities. *Int. J. Inf. Secur.* **2024**, *23* (1), 577—608.
6. BMFSFJ. *Was ist Cybermobbing?* https://www.bmfsfj.de/bmfsfj/themen/kinder-und-jugend/medienkompetenz/was-ist-cybermobbing--86484 (accessed 2025-02-14).
7. Kaufhold, M.-A.; Bayer, M.; Bäumler, J.; Reuter, C.; Mirbabaie, M.; Stieglitz, S.; Basyurt, A. S.; Fuchß, C.; Eyilmez, K. CYLENCE: Strategies and Tools for Cross-Media Reporting, Detection, and Treatment of Cyberbullying and Hatespeech in Law Enforcement Agencies. In *Mensch und Computer 2023 — Workshopband*; Gesellschaft für Informatik e.V.: Rapperswil, Switzerland, 2023.
8. Kumar, B.; Mathew, S. K. The Dark Side of Social Networking Sites: A Review of Cybercrime Research. *Pac. Asia J. Assoc. Inf. Syst.* **2024**, *16* (3), 1—31.
9. Patchin, J. W. *Cyberbullying Data*; Cyberbullying Research Center. https://cyberbullying.org/2023-cyberbullying-data (accessed 2025-02-14).
10. Bozyiğit, A.; Utku, S.; Nasibov, E. Cyberbullying Detection: Utilizing Social Media Features. *Expert Syst. Appl.* **2021**, *179*, 1—12.
11. Vogels, E. A. *The State of Online Harassment*; Pew Research Center. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (accessed 2025-02-14).
12. Gandhi, A.; Ahir, P.; Adhvaryu, K.; Shah, P.; Lohiya, R.; Cambria, E.; Poria, S.; Hussain, A. Hate Speech Detection: A Comprehensive Review of Recent Works. *Expert Syst.* **2024**, *41* (8), 1—24.
13. Arango, A.; Pérez, J.; Poblete, B. Hate Speech Detection Is Not as Easy as You May Think: A Closer Look at Model Validation (Extended Version). *Inf. Syst.* **2022**, *105*, 1—11.
14. He, Z.; Li, Y.-J.; Lee, M. K. O. IT Solutions for Tackling Cyberbullying: A Literature Review. In *Proceedings of the Pacific-Asia Conference on Information Systems*: Ho Chi Minh City, Vietnam, 2024.
15. He, Z.; Li, Y.-J.; Lee, M. K. O. Understanding Empathy and Bystander Intervention in Cyberbullying: An Experiment Design of AI System Intervention. In *Proceedings of the International Conference on Information Systems*; Bangkok, Thailand, 2024.
16. Kaluarachchi, C.; Sedera, D.; Warren, M. A Review of Adult Cyberbullying Research from Multi-Disciplinary Archives and Directions for Future Studies. In *Proceedings of the Australasian Conference on Information Systems*: Sydney, Australia, 2021.
17. Emery, F. E.; Trist, E. L. Socio-Technical Systems. *Manag. Sci. Models Tech.* **1960**, *2*, 83—97.
18. Baxter, G.; Sommerville, I. Socio-Technical Systems: From Design Methods to Systems Engineering. *Interact. Comput.* **2011**, *23* (1), 4—17.

19. Webster, J.; Watson, R. T. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Q.* **2002**, *26* (2), xiii—xxiii.
20. Chang, V. Inconsistent Definitions of Bullying: A Need to Examine People's Judgments and Reasoning about Bullying and Cyberbullying. *Hum. Dev.* **2021**, *65* (3), 144—159.
21. Corcoran, L.; Mc Guckin, C.; Prentice, G. Cyberbullying or Cyber Aggression? A Review of Existing Definitions of Cyber-Based Peer-To-Peer Aggression. *Societies* **2015**, *5* (2), 245—255.
22. Elsafoury, F.; Katsigiannis, S.; Pervez, Z.; Ramzan, N. When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection. *IEEE Access* **2021**, *9*, 103541—103563.
23. Mladenović, M.; Ošmjanski, V.; Stanković, S. V. Cyber-Aggression, Cyberbullying, and Cyber-Grooming: A Survey and Research Challenges. *ACM Comput. Surv. (CSUR)* **2021**, *54* (1), 1—42.
24. Kaluarachchi, C.; Trieu, V.-H. Cyberbullying Prevention and Reduction Strategies: An Exploratory Study. In *Proceedings of Australasian Conference on Information Systems*: Melbourne, Australia, 2022.
25. Gaffney, H.; Farrington, D. P.; Espelage, D. L.; Ttofi, M. M. Are Cyberbullying Intervention and Prevention Programs Effective? A Systematic and Meta-Analytical Review. *Aggress. Violent Behav.* **2019**, *45*, 134—153.
26. Teng, T. H.; Varathan, K. D.; Crestani, F. A Comprehensive Review of Cyberbullying-Related Content Classification in Online Social Media. *Expert Syst. Appl.* **2024**, *244*, 1—58.
27. Tahmasbi, N.; Fuchsberger, A. Challenges and Future Directions of Automated Cyberbullying Detection. In *Proceedings of the Americas Conference on Information Systems*; New Orleans, LA, USA, 2018.
28. Sarna, G.; Bhatia, M. P. S. Content Based Approach to Find the Credibility of User in Social Networks: An Application of Cyberbullying. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 677—689.
29. Peter, I.-K.; Petermann, F. Cyberbullying: A Concept Analysis of Defining Attributes and Additional Influencing Factors. *Comput. Hum. Behav.* **2018**, *86*, 350—366.
30. Rawat, A.; Kumar, S.; Samant, S. S. Hate Speech Detection in Social Media: Techniques, Recent Trends, and Future Challenges. *WIREs Comput Stat.* **2024**, *16* (2), e1648.
31. Truong, D.-H.; Chen, J. V. Understanding the We-Intention to Participate in Collective Trolling on Social Networking Sites: The Online Disinhibition Perspective. In *Proceedings of the Pacific-Asia Conference on Information Systems*: Ho Chi Minh City, Vietnam, 2024.
32. Li, Y.-J.; Cheung, C. M. K.; Shen, X.-L.; Lee, M. K. O. When Socialization Goes Wrong: Understanding the We-Intention to Participate in Collective Trolling in Virtual Communities. *J. Assoc. Inf. Syst.* **2022**, *23* (3), 678—706.
33. Canbay, P. Predicting Discriminative Personality Profile of Haters from Digital Texts. *Knowl.-Based Syst.* **2024**, *287*, 1—12.
34. Lin, C. A.; Xu, X. Cyberbullying and Social Media Communication: Spiral of Silence, Relational Aggression and Schadenfreude. *J. Soc. Media Soc.* **2024**, *13* (2), 23—47.
35. Wong, R. Y. M.; Cheung, C. M. K.; Xiao, B.; Thatcher, J. B. Standing up or Standing by: Understanding Bystanders' Proactive Reporting Responses to Social Media Harassment. *Inf. Syst. Res.* **2021**, *32* (2), 561—581.
36. Jenaro, C.; Flores, N.; Frías, C. P. Anxiety and Depression in Cyberbullied College Students: A Retrospective Study. *J. Interpers. Violence* **2021**, *36* (1—2), 579—602.
37. Haidar, B.; Chamoun, M.; Serhrouchni, A. Multilingual Cyberbullying Detection System: Detecting Cyberbullying in Arabic Content. In *2017 1st Cyber Security in Networking Conference (CSNet)*; Rio de Janeiro, Brazil, 2017.
38. Huang, C. L.; Alimu, Y.; Yang, S. C.; Kang, S. What You Think Is a Joke Is Actually Cyberbullying: The Effects of Ethical Dissonance, Event Judgment and Humor Style on Cyberbullying Behavior. *Comput. Hum. Behav.* **2023**, *142*, 1—10.
39. Gautam, A. K.; Bansal, A. Email-Based Cyberstalking Detection on Textual Data Using Multi-Model Soft Voting Technique of Machine Learning Approach. *J. Comput. Inf. Syst.* **2023**, *63* (6), 1362—1381.
40. Ayeni, O.; Owolafe, O.; Ogunjobi, P. A Security System for Detecting Denial of Service (DDoS) and Masquerade Attacks on Social Networks. *J. Inf. Secur. Cybercrimes Res.* **2022**, *5*, 80—86.
41. Irani, D.; Balduzzi, M.; Balzarotti, D.; Kirda, E.; Pu, C. Reverse Social Engineering Attacks in Online Social Networks. In *Detection of Intrusions and Malware, and Vulnerability Assessment*, *Lecture Notes in Computer Science*; Holz, T.; Bos, H., Eds.; Springer: Berlin, Heidelberg, Vol. *6739*, 2011; pp. 55—74.
42. Petter, S.; Wake Forest University; Giddens, L.; University of North Texas. Is it Your Fault? Framing Social Media Inclusion and Exclusion Using Just World Theory. *J. Assoc. Inf. Syst.* **2023**, *24* (5), 1248—1270.
43. Karimi, Y.; Squicciarini, A.; Wilson, S. Automated Detection of Doxing on Twitter. *Proc. ACM Hum. Comput. Interact.* **2022**, *6* (CSCW2), 1—24.
44. Fang, Y.; Risius, M.; Cheung, C. Understanding the Current State of Knowledge and Future Directions of Doxing Research: A Social Cognitive Theory Perspective. In *Proceedings of the Hawaii International Conference on System Sciences*: Hawaii, USA, 2023.
45. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D. G. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *BMJ* **2009**, *339*, 1—8.
46. Wohlin, C. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*; London, UK, 2014.
47. Zhang, Z.; Robinson, D.; Tepper, J. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *The Semantic Web*; Gangemi, A.; Navigli, R.; Vidal, M.-E.; Hitzler, P.; Troncy, R.; Hollink, L.; Tordai, A.; Alam, M., Eds.; Springer International Publishing: Cham, 2018; pp. 745—760.
48. Wich, M.; Bauer, J.; Groh, G. Impact of Politically Biased Data on Hate Speech Classification. In *Proceedings of the 4th Workshop on Online Abuse and Harms*; Akiwowo, S.; Vidgen, B.; Prabhakaran, V.; Waseem, Z., Eds.; Association for Computational Linguistics: Virtually Co-located with EMNLP, 2020; pp. 54—64.
49. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*; Andreas, J.; Choi, E.; Lazaridou, A., Eds.; Association for Computational Linguistics: San Diego, California, 2016; pp. 88—93.
50. Wang, J.; Fu, K.; Lu, C.-T. SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. In *2020 IEEE International Conference on Big Data (Big Data)*; Atlanta, Georgia, USA, 2020; pp 1699—1708.
51. Singh, V. K.; Huang, Q.; Atrey, P. K. Cyberbullying Detection Using Probabilistic Socio-Textual Information Fusion. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and*

Mining (ASONAM); San Francisco, California, USA, 2016; pp 884—887.
52. Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; Smith, N. A. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Korhonen, A.; Traum, D.; Màrquez, L., Eds.; Association for Computational Linguistics: Florence, Italy, 2019; pp. 1668—1678.
53. Romsaiyud, W.; na Nakornphanom, K.; Prasertsilp, P.; Nurarak, P.; Konglerd, P. Automated Cyberbullying Detection using Clustering Appearance Patterns. In *2017 9th International Conference on Knowledge and Smart Technology (KST)*; honburi, Thailand, 2017.
54. Özçift, A.; Kılınç, D.; Bozyiğit, F. Application of Grid Search Parameter Optimized Bayesian Logistic Regression Algorithm to Detect Cyberbullying in Turkish Microblog Data. *Acad. Platform — J. Eng. Sci.* **2019**, *7* (3), 355—361.
55. Nandhini, B. S.; Sheeba, J. I. Cyberbullying Detection and Classification using Information Retrieval Algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET* 2015*)*; ICARCSET '15; Unnao, India, 2015.
56. Muneer, A.; Fati, S. M. A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet* **2020**, *12* (11), 1—20.
57. Mozafari, N.; Weiger, W. H.; Hammerschmidt, M. Trust Me, I'm a Bot—Repercussions of Chatbot Disclosure in Different Service Frontline Settings. *J. Serv. Manag.* **2021**, *33* (2), 221—245.
58. Marwa, T.; Salima, O.; Souham, M. Deep Learning for Online Harassment Detection in Tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*; Tebessa, Algeria, 2018.
59. Mangaonkar, A.; Hayrapetian, A.; Raje, R. Collaborative Detection of Cyberbullying Behavior in Twitter Data. In *2015 IEEE International Conference on Electro/Information Technology (EIT)*; Dekalb, IL, USA, 2015.
60. MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate Speech Detection: Challenges and Solutions. *PLoS One* **2019**, *14* (8), 1—16.
61. Gemes, K. A.; Kovacs, A.; Reichel, M.; Recski, G. Offensive Text Detection on English Twitter with Deep Learning Models and Rule-Based Systems. In *FIRE-WN 2021*; Mehta, P.; Mandl, T.; Majumder, P.; Mitra, M., Eds., Vol. *3159*, 2021; pp. 283—296.
62. Elisabeth, D.; Budi, I.; Ibrohim, M. *Hate Code Detection in Indonesian Tweets Using Machine Learning Approach: A Dataset and Preliminary Study*; Yogyakarta, Indonesia, 2020.
63. Del Bosque, L. P.; Garza, S. E. Aggressive Text Detection for Cyberbullying. In *Human-Inspired Computing and Its Applications*; Gelbukh, A.; Espinoza, F. C.; Galicia-Haro, S. N., Eds.; Springer International Publishing: Tuxtla Gutiérrez, Mexico, 2014; pp. 221—232.
64. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. *Proc. Int. AAAI Conf. Web Soc. Media* **2017**, *11* (1), 512—515.
65. Davidson, T.; Bhattacharya, D.; Weber, I. Racial Bias in Hate Speech and Abusive Language Detection Datasets. arXiv 2019. https://doi.org/10.48550/arXiv.1905.12516.
66. Burnap, P.; Williams, M. L. Us and Them: Identifying Cyber Hate on Twitter across Multiple Protected Characteristics. *EPJ Data Sci.* **2016**, *5* (1), 1—15.
67. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion; WWW '17 Companion;* Perth, Australia, 2017.
68. Ayo, F. E.; Folorunso, O.; Ibharalu, F. T.; Osinuga, I. A. Machine Learning Techniques for Hate Speech Classification of Twitter Data: State-Of-The-Art, Future Challenges and Research Directions. *Comput. Sci. Rev.* **2020**, *38*, 1—34.
69. Al-garadi, M. A.; Varathan, K. D.; Ravana, S. D. Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network. *Comput. Hum. Behav.* **2016**, *63*, 433—443.
70. el Akbar, R. R.; Shofa, R. N.; Paripurna, M. I.; Supratman The Implementation of Naïve Bayes Algorithm for Classifying Tweets Containing Hate Speech with Political Motive. In *Proceedings of the 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC);* Bandung, Indonesia, 2019.
71. Ahmed, T.; Ivan, S.; Kabir, M.; Mahmud, H.; Hasan, K. Performance Analysis of Transformer-Based Architectures and Their Ensembles to Detect Trait-Based Cyberbullying. *Soc. Netw. Anal. Min.* **2022**, *12* (1), 1—17.
72. Agarwal, S.; Chowdary, C. R. Combating Hate Speech Using an Adaptive Ensemble Learning Model with a Case Study on COVID-19. *Expert Syst. Appl.* **2021**, *185*, 1—9.
73. Bretschneider, U.; Peters, R. Detecting Cyberbullying in Online Communities. In *Proceedings of the European Conference on Information Systems*; İstanbul, Turkey, 2016.
74. Bretschneider, U.; Wöhner, T.; Peters, R. Detecting Online Harassment in Social Networks. In *Proceedings of the International Conference on Information Systems*: Auckland, New Zealand, 2014.
75. Caron, M.; Bäumer, F. S.; Müller, O. Towards Automated Moderation: Enabling Toxic Language Detection with Transfer Learning and Attention-Based Models. In *Proceedings of the Hawaii International Conference on System Sciences*: Hawaii, USA, 2022.
76. Lee, P.-J.; Hu, Y.-H.; Chen, K.; Tarn, J. M.; Cheng, L.-E. Cyberbullying Detection on Social Network Services. In *Proceedings of the Pacific Asia Conference on Information Systems*; Yokohama, Japan, 2018.
77. Slivko, O.; Andres, R. Regulation of Hate Speech and Hatefulness on German Twitter. In *Proceedings of the International Conference on Information Systems*; Austin, TX, USA, 2021.
78. Dadvar, M.; de Jong, F. Cyberbullying Detection: A Step toward a Safer Internet Yard. In *Proceedings of the 21st International Conference on World Wide Web*; WWW '12 Companion; Lyon, France, 2012.
79. Dadvar, M.; Trieschnigg, D.; Ordelman, R.; De Jong, F. Improving Cyberbullying Detection with User Context. In *Advances in Information Retrieval*, *Series Eds.; Lecture Notes in Computer Science*; Serdyukov, P.; Braslavski, P.; Kuznetsov, S. O.; Kamps, J.; Rüger, S.; Agichtein, E.; Segalovich, I.; Yilmaz, E.; Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J. M.; Mattern, F.; Mitchell, J. C.; Naor, M.; Nierstrasz, O.; Pandu Rangan, C.; Steffen, B.; Sudan, M.; Terzopoulos, D.; Tygar, D.; Vardi, M. Y.; Weikum, G., Eds.; Springer: Berlin, Heidelberg, Vol. *7814*, 2013; pp. 693—696.
80. Silva, Y. N.; Rich, C.; Hall, D. BullyBlocker: Towards the Identification of Cyberbullying in Social Networking Sites. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*; San Francisco, California, USA, 2016.
81. Del Vigna, F.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; Tesconi, M. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In

Proceedings of the First Italian Conference on Cybersecurity (ITASEC17); Venice, Italy, 2017.
82. Schieb, C.; Preuss, M. Governing Hate Speech by Means of Counterspeech on Facebook. In *Proceedings of the 66th ICA Annual Conference*; Fukuoka, Japan, 2016.
83. Li, Y.; Chan, J.; Peko, G.; Sundaram, D. Towards Resource Inequities in Catching the "Dark Side" of Social Media: A Hateful Meme Classification Framework for Low-Resource Scenarios. In *Proceedings of the Hawaii International Conference on System Sciences*: Hawaii, USA, 2024.
84. Ashktorab, Z. A Study of Cyberbullying Detection and Mitigation on Instagram. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*; San Francisco, USA, 2016; pp. 26—130.
85. Hosseinmardi, H.; Mattson, S. A.; Rafiq, R. I.; Han, R.; Lv, Q.; Mishra, S. Detection of Cyberbullying Incidents on the Instagram Social Network. arXiv 2015. https://doi.org/10.48550/arXiv.1503.03909.
86. Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*; WWW '16; Montreal, Canada, 2016.
87. Squicciarini, A.; Rajtmajer, S.; Liu, Y.; Griffin, C. Identification and Characterization of Cyberbullying Dynamics in an Online Social Network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*; Paris, France, 2015.
88. Wu, J.; Wen, M.; Lu, R.; Li, B.; Li, J. Toward Efficient and Effective Bullying Detection in Online Social Network. *Peer-to-Peer Netw. Appl.* **2020**, *13* (5), 1567—1576.
89. Potha, N.; Maragoudakis, M. Cyberbullying Detection Using Time Series Modeling. In *2014 IEEE International Conference on Data Mining Workshop*; IEEE: Shenzhen, China, 2014; pp. 373—382.
90. Lee, K.; Ram, S. PERSONA: Personality-Based Deep Learning for Detecting Hate Speech. In *Proceedings of the International Conference on Information Systems*: Hyderabad, India, 2020.
91. Vijayakumar, V.; Hari Prasad, D.; Adolf, P. A Novel Approach for Image Based Cyberbullying Detection and Prevention. *Int. J. Comput. Appl.* **2021**, *183* (22), 41—45.
92. Emmery, C.; Kádár, Á.; Chrupała, G.; Daelemans, W. Cyberbullying Classifiers Are Sensitive to Model-Agnostic Perturbations. arXiv 2022. https://doi.org/10.48550/arXiv.2201.06384.
93. Guo, X.; Gauch, S. Using Sarcasm to Improve Cyberbullying Detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING;* Torino, Italia, 2024; pp. 52—59.
94. Gröndahl, T.; Pajola, L.; Juuti, M.; Conti, M.; Asokan, N. All You Need Is "Love": Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security;* Toronto, Canada, 2018.
95. Ejaz, N.; Razi, F.; Choudhury, S. Towards Comprehensive Cyberbullying Detection: A Dataset Incorporating Aggressive Texts, Repetition, Peerness, and Intent to Harm. *Comput. Hum. Behav.* **2024**, *153*, 1—11.
96. Li, Z.; Kawamoto, J.; Feng, Y.; Sakurai, K. Cyberbullying Detection Using Parent-Child Relationship between Comments. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*; iiWAS '16; San Francisco, USA, 2016.
97. Arayankalam, J.; Soral, P.; Khan, A.; Krishnan, S.; Bose, I. Does Centralization of Online Content Regulation Affect Political Hate Speech in a Country? A Public Choice Perspective. *Inf. Manag.* **2024**, *61* (2), 1—18.
98. Ogunleye, B.; Dharmaraj, B. The Use of a Large Language Model for Cyberbullying Detection. *Analytics* **2023**, *2* (3), 694—707.
99. Rajput, K.; Kapoor, R.; Rai, K.; Kaur, P. Hate Me Not: Detecting Hate Inducing Memes in Code Switched Languages. In *Proceedings of the Americas Conference on Information Systems*: Minneapolis, USA, 2022.
100. Svetasheva, A.; Lee, K. Harnessing Large Language Models for Effective and Efficient Hate Speech Detection. In *Proceedings of the Hawaii International Conference on System Sciences*: Hawaii, USA, 2024.
101. Tahmasbi, N.; Fuchsberger, A. ChatterShield — A Multi-Platform Cyberbullying Detection System for Parents. In *Proceedings of the Americas Conference on Information Systems*: Cancun, Mexico, 2019.
102. Lowry, P. B.; Zhang, J.; Wang, C.; Siponen, M. Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model. *Inf. Syst. Res.* **2016**, *27* (4), 962—986.
103. Trabelsi, Z.; Mellouli, S.; Khoury, R. Online Content Moderation and the Challenge of Conceptualizing Cyberbullying. In *Proceedings of the Americas Conference on Information Systems*: Minneapolis, USA, 2022.
104. Lowry, P. B.; Moody, G. D.; Chatterjee, S. Using IT Design to Prevent Cyberbullying. *J. Manag. Inf. Syst.* **2017**, *34* (3), 863—901.
105. Mathew, B.; Kumar, N.; Ravina; Goyal, P.; Mukherjee, A. Analyzing the Hate and Counter Speech Accounts on Twitter. arXiv 2018. https://doi.org/10.48550/arXiv.1812.02712.
106. Wong, R. Y. M.; Cheung, C. M. K.; Xiao, B. Understanding Users' Willingness to Report Online Harassment on Social Media Networking Sites: The Role of Efficacy. In *Proceedings of the Pacific Asia Conference on Information Systems*; Chiayi, Taiwan, 2016.
107. Tahmasbi, N.; Rastegari, E. A. Socio-Contextual Approach in Automated Detection of Cyberbullying. In *Proceedings of the Hawaii International Conference on System Sciences*: Hawaii, USA, 2018.
108. Nickerson, A. B.; Aloe, A. M.; Livingston, J. A.; Feeley, T. H. Measurement of the Bystander Intervention Model for Bullying and Sexual Harassment. *J. Adolesc.* **2014**, *37* (4), 391—400.
109. Zapata, J.; Sulik, J.; von Wulffen, C.; Deroy, O. Bystanders' Collective Responses Set the Norm against Hate Speech. *Humanit. Soc. Sci. Commun.* **2024**, *11* (1), 1—13.
110. Lee, K.; Ram, S. Deep Learning for Hate Speech Detection: A Personality-Based Approach. In *Companion Proceedings of the ACM Web Conference 2024*; WWW '24; Singapore, Singapore, 2024; pp. 1667—1671.
111. Khairy, M.; Mahmoud, T. M.; Abd-El-Hafeez, T.; Mahfouz, A. User Awareness of Privacy, Reporting System and Cyberbullying on Facebook. In *Advanced Machine Learning Technologies and Applications*; Hassanien, A.-E.; Chang, K.-C.; Mincong, T., Eds.; Springer International Publishing: Cham, 2021; pp. 613—625.
112. Nahar, V.; Al-Maskari, S.; Li, X.; Pang, C. Semi-Supervised Learning for Cyberbullying Detection in Social Networks. In *Databases Theory and Applications*; Springer: Cham, 2014; pp. 160—171.

113. Loebbecke, C.; Luong, A. C.; Obeng-Antwi, A. AI for Tackeling Hate Speech. In *Proceedings of the European Conference on Information Systems*; Marrakesh, Morocco, 2021.
114. We Are Social; DataReportal; Hootsuite *Global Daily Social Media Usage 2024*; Statista. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/ (accessed 2025-02-14).
115. Zimmer, M.; Proferes, N. J. A Topology of Twitter Research: Disciplines, Methods, and Ethics. *Aslib J. Inf. Manag.* **2014**, *66* (3), 250—261.
116. X Platform; *Counting Characters when Composing Tweets*. X. https://docs.x.com/resources/fundamentals/counting-characters (accessed 2025-06-16).
117. X Platform; *Rate Limits*. X. https://docs.x.com/x-api/fundamentals/rate-limits (accessed 2025-06-16).
118. Bouliche, A. Detection of Cyberbullying in Arabic Social Media Using Dynamic Graph Neural Network. In *Tunisian-Algerian Joint Conference on Applied Computing*; Constantine, Algeria, 2022.
119. Ho, S. M.; Kao, D.; Chiu-Huang, M.-J.; Li, W.; Lai, C.-J.; Ankamah, B. Charged Language on Twitter: A Predictive Model of Cyberbullying to Prevent Victimization. In *Proceedings of the 2019 Pre-ICIS Workshop on Information Security and Privacy*; Munich, Germany, 2019.
120. Statista. *U.S. Social Network Users 2023, by Age Group*; Statista. https://www.statista.com/statistics/1337525/us-distribution-leading-social-media-platforms-by-age-group/ (accessed 2025-06-04).
121. Xie, H. S.; Namvar, M.; Risius, M. A Review of Hate Speech Detection: Challenges and Innovations. In *Proceedings of the DIGIT Workshop*: Hyderabad, India, 2023.
122. Gaia, J.; Murray, D.; Sanders, G.; Sanders, S.; Upadhyaya, S.; Wang, X.; Yoo, C. The Interaction of Dark Traits with the Perceptions of Apprehension. In *Proceedings of the Hawaii International Conference on System Sciences*: Hawaii, USA, 2022.
123. Tang, W. Y.; Reer, F.; Quandt, T. The Interplay of Gaming Disorder, Gaming Motivations, and the Dark Triad. *J. Behav. Addict.* **2020**, *9* (2), 491—496.
124. Aivazpour, Z.; Beebe, N. L. The Impact of Power Imbalance on Cyberbullying: The Role of Bystanders Intervention. In *Proceedings of the International Conference on Information Systems*; Hyderabad, India, 2020.
125. Hu, S.; Lei, W.; Zhu, H.; Hsu, C. Cyberbullying Perpetration on Social Media: A Situational Action Perspective. *Inf. Manag.* **2024**, *61* (6), 1—11.
126. Kaluarachchi, C. D.; Sedera, D. D.; Warren, M. Cyberbullying Among Adults: A Qualitative Content Analysis of the Legal Responses to a Complex Social Problem. In *Proceedings of the International Conference on Information Systems*: Copenhagen, Denmark, 2022.
127. Kaluarachchi, C.; Sedera, D.; Warren, M. An Intervention Model for Cyberbullying Based on the General Theory of Crime and Routine Activity Theory. In *Proceedings of the Australasian Conference on Information Systems*: Wellington, New Zealand, 2020.
128. Kaluarachchi, C.; Sedera, D.; Warren, M. An Investigative Model of Adult Cyberbullying: A Court Case Analysis. In *Proceedings of the Pacific Asia Conference on Information Systems*: Dubai, UAE, 2021.
129. Withers, K. L.; Terrell, S. R.; Parrish, J. L.; Ellis, T. J. The Relationship between the "Dark Triad" Personality Traits and Deviant Behavior on Social Networking Sites. In *Proceedings of the Americas Conference on Information Systems*: Boston, USA, 2017.
130. Wong, R. Y. M. Dealing with Online Harassment: Understanding Online Protective Coping Strategies on Social Nerworking Sites. In *Proceedings of the Pacific Asia Conference on Information Systems*: Langkawi, Malaysia, 2017.
131. Zhang, S.; Leidner, D.; Cao, X.; Liu, N. Workplace Cyberbullying: A Criminological and Routine Activity Perspective. *J. Inf. Technol.* **2022**, *37* (1), 51—79.
132. Zhu, H.; Hsu, C.; Zhou, Z. Bystander Pro-celebrity Cyberbullying: An Integrated Perspective of Susceptibility to Retaliation and Social Capital Gains. *Inf. Manag.* **2023**, *60* (5), 1—13.