Automatic Analysis of Political Debates and Manifestos: Successes and Challenges

Tanise Ceron[1(✉)], Ana Barić[2], André Blessing[1], Sebastian Haunss[3], Jonas Kuhn[1], Gabriella Lapesa[4,5], Sebastian Padó[1], Sean Papay[1], and Patricia F. Zauchner[3]

[1] IMS, University of Stuttgart, Stuttgart, Germany
tanise.ceron@ims.uni-stuttgart.de
[2] FER, University of Zagreb, Zagreb, Croatia
[3] SOCIUM, University of Bremen, Bremen, Germany
[4] GESIS Cologne, Cologne, Germany
[5] DIID, HHU Düsseldorf, Düsseldorf, Germany

Abstract. The opinions of political actors (e.g., politicians, parties, organizations) expressed through *claims* are the core elements of political debates and decision-making. Political actors communicate through different channels: parties publish manifestos for major elections, while individual actors make statements on a day-to-day basis as reflected in the media. These two channels offer different approaches for analysis: Manifestos, on the one hand, are useful to characterize the parties' positions at a global ideological level over time. In contrast, individual statements can be collected to analyze debates in particular policy domains on a fine-grained level, in terms of individual actors and claims. In this article, we summarize a series of studies we have carried out. We apply NLP-driven (semi-)automatic analyses on these two channels and compare their potentials and challenges. The fine-grained analysis yields rich insights into the communication but comes at the cost of three challenges: (a) a substantial hunger for manual annotation, introducing practical hurdles for analysis both within and across languages; (b) difficulties in claim classification arising from the uneven frequency distribution over the theory-based annotation schemas; (c) the need to map actor mentions onto canonical versions. Manifesto-based analysis avoids these challenges to a substantial extent when a more coarse-grained analysis of party positions is sufficient. We highlight the benefits and challenges of both approaches, and conclude by outlining perspectives for addressing the challenges in future research.

Keywords: Claim identification · discourse network analysis · party positioning · argument mining

1 Introduction

Political decision-making in democracies is generally preceded by political debates taking place in parliamentary forums (committees, plenary debates),

P. Cimiano et al. (Eds.): RATIO 2024, LNAI 14638, pp. 71–88, 2024.
https://doi.org/10.1007/978-3-031-63536-6_5

different public spheres (e.g., newspapers, television, social media), and in the exposition of political ideologies in party manifestos [44,46]. In these debates, various actors voice their positions and beliefs, make claims and try to advance their agendas. Political scientists have therefore developed a range of methods to analyze these debates in the dual goal of understanding democratic decision-making and identifying influential actors and important arguments driving the development of these debates. Two prominent ones are as follows:

(a.) To obtain a maximally informative picture, we can identify the claims and actors involved in a given debate, combining political claims analysis [23] and network science, and represent them as *discourse networks* [26,27]. This permits researchers to capture structural aspects of political debates, investigating and reconstructing debates in a fine-grained manner and understanding the reasons why some claims prevail and others fail.

(b.) The more traditional approach in the political science tradition is to abstract away from the details of a given debate and assess positions and beliefs of political actors at the aggregate level of party positions, namely analyzing manifestos. This provides much less detail but focuses on the arguably most important group of political actors and their respective ideologies. Shifts in ideology allow understanding the change of opinions within a party and their electorate [3]. This approach also allows for direct access to actors' opinions as in comparison with news that goes through a selection of actors and decisions when reported in the media outlets.

In this article, we present an overview of the main contributions from a series of studies that aimed at assessing whether these two approaches can be conducted more efficiently using methods from natural language processing (NLP). We start in Sect. 2 with the more complex approach (a), conceptualizing discursive exchanges as discourse networks. Our goal here is to assess how NLP can help to overcome the roadblocks that studies in this perspective are facing because of the time- and labor-intensive annotation required by detailed analyses of political discourse. Then, in Sect. 3, we switch perspective to approach (b), adopting instead the goal of characterizing party positions at the global, ideological level. We demonstrate that this task does not require a full-fledged discourse network analysis, can do with very coarse-grained content categories, and that hardly any manual annotation is necessary. We highlight the benefits and challenges of both approaches, and conclude by outlining perspectives for addressing the challenges in future research.

2 Fine-Grained Analysis of Political Discourse

Our starting point for the first approach is political debates as they are represented in newspaper articles. In these articles, journalists report on claims and positions of all kinds of actors participating in public debates. We conceptualize these discursive interactions as discourse networks [26] — (dynamic) bipartite

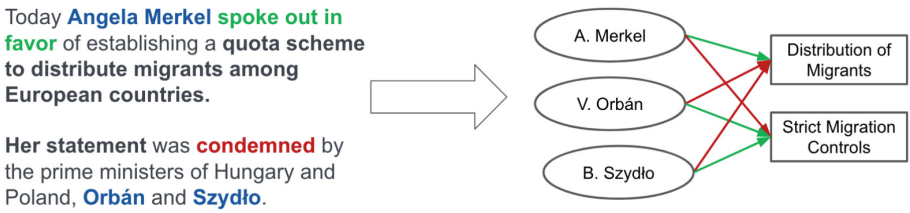

Fig. 1. Discourse Network Example

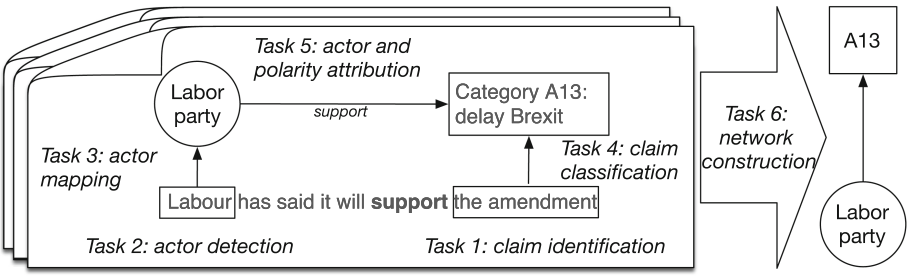

Fig. 2. From newspaper articles to affiliation networks (adapted from [32])

graphs with two types of nodes, namely (a) actors (politicians, parties, organizations, but also groups of citizens such as protesters); and (b) fine-grained categories of claims (purposeful communicative acts in the public sphere by which an actor tries to influence a specific policy or political debate). Edges link actor nodes with the claim nodes that they communicate about and are tagged with a polarity: actors can either support or oppose specific claims. Figure 1 shows an example where actors are ovals, claim categories are rectangles, and green and red edges denote support and opposition. Figure 2 presents a step-by-step guide to developing such a network based on newspaper articles: Given a document, we need to detect text spans that express claims and actors (Tasks 1 and 2), we need to map these text spans onto canonical actors (e.g., "Merkel", "the chancellor", "Mrs. Merkel" are mentions of the canonical actor *Angela Merkel*) and claim categories, respectively (Tasks 3 and 4), and finally we need to establish actor-claim dyads with correct polarities (Task 5) and construct the actual network (Task 6). Until recently, to construct these networks, one needed to meticulously perform these tasks by hand; which costs time and hence money. Therefore, we aim to use NLP to develop predictive models capable of automating this process. This results in a fairly complex computational setup which gives rise to three main challenges:

(1) **Annotation takes long and is costly.** Traditional supervised learning demands a substantial number of annotated datapoints, but annotation of actors and claims calls for expert annotation. This leads to a 'slow start' situation: a sizable amount of manual annotation has to be carried out before computational modeling can proceed. Once models are in place, they can speed up future annotation, but this comes with its own set of challenges [18].

In practice, this means that a combination of time, money, and expertise is necessary to reach that point which might not always be available. Furthermore, carrying out comparative studies requires annotation to be available for multiple languages, even if only for evaluation purposes.

(2) Political claims are difficult to process on a fine-grained level. The codebooks developed by political science experts to describe the relevant claim categories in societal debates need to be sufficiently fine-grained to permit the characterization of competing positions in terms of the discourse network. This consideration often leads to codebooks with anywhere between 50 and over 100 claim categories [4,17,22]. As usual for language data, a few categories are frequent, while the majority are rare. This further exacerbates the problem mentioned in point (1) when learning claim identifications and claim classifiers (cf. Tasks 1 and 4 in Fig. 2): even a relatively large corpus will hardly provide enough examples of the infrequent categories for straightforward learning.

(3) Actor mentions are difficult to aggregate. Most of the mentions of actors in any discourse do not use their canonical name ("Angela Merkel"), but instead short versions ("Mrs. Merkel"), roles ("the chancellor"), or even just personal or possessive pronouns ("she", "her", compare Fig. 1). The mapping of such mentions onto the right actor node in the discourse network is essentially equivalent, in the general, to coreference resolution which is known to be a hard task. While shortcuts exist for some instances, notably the use of entity linking [36] for actors which are represented in some database, there are many actors for which this is not the case – including politicians at the local or regional levels as well as 'ad-hoc' actors such as "several ministers".

In the following Sects. (2.1–2.4), we discuss a series of studies addressing tasks 1–4 from Fig. 2 and responding to these challenges. As gold standard for our studies we use DEbateNet [4], which is a large corpus resource that we created for the analysis of the German domestic debate on migration in 2015. After domain experts from political science developed a codebook for the policy domain, roughly 1000 newspaper articles from the German left-wing quality newspaper 'taz - die tageszeigung' with a total of over 550.000 tokens were annotated for actors, claims, and their relations, and finally used for computational modeling.

2.1 Less Annotation Is More: Few-Shot Claim Classification

As noted in Challenge 1, NLP models that (partially) automate claim detection and classification traditionally require relatively large manually annotated data sets for training or fine-tuning, since the required domain-specific semantic distinctions are hard to recover directly from plain text. Since for most political topics no annotated data exists, research projects usually needed to start with a substantial amount of classical qualitative text analysis. The situation has changed substantially in the last two years, with the advent of large language models and their capacity for transfer learning and few-shot learning [5], that is, the ability to learn new tasks ad hoc, from very small numbers of examples.

To assess the potential of few-shot learning, we have carried out a study to assess whether we are able to replicate the manual annotation in one policy domain – the debate about the exit from nuclear energy in Germany in the year 2013 [19] – based on our models trained on migration debates and with a minimal amount of additional training data [17]. We thus try to process claims on the exit from nuclear energy use like "The Greens want to introduce a bill in the Bundestag for the immediate and final decommissioning of Germany's seven oldest nuclear power plants" with a model trained on claims from the migration debate like "The basic right to asylum for politically persecuted persons knows no upper limit, Merkel also announced in an interview". In this overview, we focus on the tasks of claim identification and claim classification (cf. Figure 2).

We work with a corpus of articles sampled with a keyword-based approach which still contains about as many relevant as irrelevant articles. Claims are identified by a binary sentence classifier. We start by calculating sentence embeddings using a sentence-BERT model (paraphrase-multilingual-mpnet-base-v2; [35]). We then use the manually annotated DEbateNet dataset (cf. Section 2) to train a multi-layer perceptron as claim identifier. Even though trained on data from a completely different topic area, our classifier obtains an F1 score of 0.78 on nuclear energy claims (precision: 0.77, recall: 0.79). This is remarkable, especially considering the large number of irrelevant articles in the corpus.

For claim classification, the model requires some information about the relevant claim categories. In our case, we use the category labels (i.e., names) from the codebook that was used to annotate the claims in the original study as minimal input for a few-shot learning approach. Again, each sentence is embedded with an SBERT model (using the same model as for claim identification). Analogously, the category labels from the codebook are encoded by SBERT. We then compute cosine similarity between all claim candidates and all category labels. Manually checking the top-ranked sentences for each label leads to seed sentences for each category. In the next step, we classify each claim candidate by assigning it to the category of the most similar seed sentence. To control the precision of claim classification, we introduce a threshold for similarity scores: Claim candidates with higher similarity scores are retained, while those below it are filtered out as potentially irrelevant.

When we evaluate whether the model correctly predicts categories for individual claims by computing F1 scores for each category, the model reaches F1 scores ranging from 0.23 to 0.45 for the more frequent claim categories ($n > 20$). Results for infrequent categories are unreliable. When evaluating whether the model correctly predicts the claims in the eight n-core networks of the original study, the results are better, with F1 scores between 0.29 and 0.69. In both cases, the variation of F1 scores across categories shows that especially infrequent categories pose a major challenge to our few-shot approach of generating discourse networks. In the next section, we therefore discuss options to increase the precision of predicting infrequent claim categories.

200 Residency

201 Emergency accommodation/1st adm.	209 Restricted residency obligation
202 Refugee accommodation	210 Subsidiary protection
203 Centralised accommodation	211 Right of residency
204 Provision of living spaces	212 In-kind in contributions
205 Forced occupancy of private housing	213 Church asylum
206 Private accommodation	214 Naturalization
207 Deportation	299 General

Fig. 3. Codebook excerpt: Supercategory residency

Table 1. Claim classification: Precision, Recall, F-Scores on DebateNet newspaper corpus. Simplified from [13].

Freq band	Base			HLE			CRR			HLE+CRR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Overall	61.2	41.9	47.0	75.2	52.2	59.0	70.4	49.0	55.2	76.5	54.3	**60.8**
Low	10.2	9.7	9.6	58.3	30.6	37.4	31.2	16.1	18.7	54.8	29.0	**35.8**
Mid	58.0	36.0	41.8	77.4	55.3	62.2	75.8	49.1	55.8	85.1	58.8	**66.2**
High	73.1	50.8	56.7	77.8	55.6	62.3	76.4	55.9	62.6	77.7	57.9	**64.0**

2.2 Improving Claim Classification with Hierarchical Information

As Challenge 2 formulated above, the many infrequent claim categories are very difficult to recognize accurately. We now describe a study on how to combat this challenge [13]. We start from the observation that political science codebooks are typically structured hierarchically into (at least) two levels, with an upper level that corresponds to broad political supercategories and a lower level that defines specific policies (claim categories). Figure 3 shows an example of the supercategory of residency, which is split up into about a dozen specific claim categories. When we have insufficient amounts of training data available, we can use this information to formulate a prior for the claim category embeddings that are learnt by our model's classifier: in the embedding space, claim categories should be located closer to other categories within the same supercategory than to claim categories of other supercategories. This leads us into the general area of hierarchy-aware classification methods.

While we evaluated various methods [13], we focus here on two approaches. The first one is hierarchical label encoding (HLE) [37] which decomposes the parameters for the specific claim categories into a (shared) supercategory part and a category-specific part, and a regularization approach which we call Class Representation Regularization (CRR) and which encourages the model to minimize the distances among the representations for each specific claim category within the same supercategory. For an experiment, we set up a base classifier based on BERT and combined it with HLE and CRR, training and testing on the DEbateNet gold standard (Sect. 2). The results in Table 1 shows that both

Table 2. Cross-lingual modeling: F1 scores for Claim class for claim identification (Id), macro average for claim classification (Class) on two datasets: DebateNet and Guardian. Simplified from [45].

Model	Train	Test	DebateNet		Guardian	
			Id F_1	Class F_1	Id F_1	Class F_1
Baseline (mono)	de	de	56.2	70.5	–	–
Translate-test	de	de (via en)	55.8	69.5	20.6	53.4
Translate-train	en	en	57.3	67.8	25.5	51.0
Multilingual	de	en	45.8	50.3	20.0	39.0

strategies, HLE and CRR, lead to a clear improvement in overall micro-F1 score over the base classifier (F_1=47.0) to F_1 scores of 59.0 and 55.2, respectively. A combination of the two leads to a further improvement to F_1=60.8. The improvement is most striking for the low frequency band (corresponding to the lowest frequency tercile), improving from F_1=9.6 to F_1=35.8. The developments for the two other frequency terciles are less dramatic, but still substantial (mid: 41.8 → 66.2, high: 56.7 → 64.0). A second study shows similar, albeit smaller, effects for claim classification on party manifestos with categories from MARPOR, a domain-independent claim classification schema [41] discussed in more detail below in Sect. 3.1. This bolsters the interpretation that our improvements are not tied to the specific codebook we used. We conclude that there is considerable space to improve the prediction quality for infrequent claim categories with dedicated methods.

2.3 Multilingual Claim Processing

When we move to another language while staying in the same policy domain – for example, for the purpose of comparative analyses across countries – we find ourselves faced with a specific case of Challenges 1 and 2: Do we have to start over with creating manual annotations? For argument mining, for which the identification of claims is a core task, the potential of machine translation for cross-lingual projection has already been established [16].

We report on a pilot study in claim identification and classification in other languages [45], machine-translating the German DEbateNet articles into English and French (this overview focuses on English). We compared three strategies: (a) backtranslating the foreign-language texts into German and analyzing them with a monolingual German BERT-based claim identifier and classifiers ('translate-test'); (b) building monolingual BERT-based foreign-language models from the translated DebateNet and using them to analyze the data in the respective languages ('translate-train'); (c) training multilingual models based on multilingual BERT on the original German data and then applying it to translated data ('multilingual').

The 'DebateNet' column of Table 2 shows that dealing with multilingual claims with machine translation works well: results are almost identical to the

Table 3. Actor mentions and their canonicalizations in newswire article (https://shorturl.at/WZ159)

	Local mention of actor	Canonical version
1	*President Joe Biden* pleaded with Republicans ...	Joe Biden
2	*Biden* signaled a willingness to make significant changes ...	Joe Biden
3	"We can't let Putin win", *he* said	Joe Biden

monolingual setup. In contrast, using multilingual embeddings incurs a substantial performance penalty. This is in line with previous analyses arguing that multilingual embeddings attempt to solve a harder, more open-ended task than MT systems do [2,34]. Also, claim identification in the multilingual embedding setup drops only ≈10 points F_1 compared to the baseline, while claim classification drops 20 points F_1 – the limiting factor seems to be the embeddings' (in-)ability to account for fine-grained topic distinctions consistently across languages.

This looks like machine translation is, indeed, sufficient to transfer political claims analysis across languages. However, the question is whether machine-translated text is a reasonable proxy for original text in a language. To test for this effect, we annotated a small sample of English reporting from the Guardian on the German migration debate. The results in the 'Guardian' column of Table 2 are much lower than those for the machine translated text. Again, we see an advantage for the MT-based approaches over multilingual embeddings, but less clearly. Particularly striking is the drop for claim identification with the MT approach from 56%-57% to 20-26% F_1. Indeed, a British newspaper is likely to report on German domestic affairs differently from a German newspaper, which leads to differences in claim form and substance: They tend to focus on the internationally most visible actors and report claims on a more coarse-grained level. Beyond the linguistic differences that NLP has so far focused on, therefore, working with newspaper reports from different countries necessitates bridging the cultural differences in framing [42], which may require some amount of manual labeling, or at least few-shot learning (cf. Section 2.1) after all.

2.4 Robust Actor Detection and Mapping

As outlined above, a central but difficult part of discourse network analysis is detecting actors for claims and mapping their textual mentions onto canonical forms (Tasks 2 and 3 in Fig. 2 and Table 3). We now describe a study comparing the two currently dominating approaches for this task [1]: (1) a pipeline of traditional NLP models, and (2) an end-to-end approach based on prompting a large language model (LLM). Once more, DebateNet, which provides a canonicalized representation for each actor, serves as dataset.

The pipeline approach comprises two steps. First, a CRF-based model identifies actor mention spans from the text, given the article with a marked claim as input. Since each claim has (at least) one actor, we constrain our CRF to always

Table 4. Prompt template instruction paraphrases used for robustness check for zero- and few-shot setting.

#	Instruction templates
1	*"Extract only the entity that made the claim in the article. The claim is surrounded with <claim> and <\claim> tags. Output only the entity without any additional explanation. Article: [ARTICLE]"*
2	*"Extract and standardize only the entity that made the marked claim in the article. The claim is surrounded with <claim> and <\claim> tags. Output only the standardized entity without any additional explanation. Article: [ARTICLE]"*
3	*"Retrieve the party or parties responsible for the statement in the given article, contained within <claim> and <\claim> tags. Output only the entity without further elaboration. Article: [ARTICLE]"*
4	*"Identify and output the entity or entities that made the claim within the specified article, enclosed by <claim> and <\claim> tags. Do not include any supplementary information. Article: [ARTICLE]"*

predict at least one actor mention per claim [33]. The second step of our pipeline canonicalizes these actors mentions through classification. We define the classes of this classifier to be (the string representations of) all canonicalized actors which appear at least twice in the training set (229, in our case), complemented by a special class 'keep-as-is' which covers all remaining actors mentions and which – true to its name – does not change the input. This heuristic approach works since infrequent actors are typically expressed by a linguistic expression that can serve well as a canonicalized version (either a full name, or a definition description such as 'the government secretaries'). The input to the classifier is the mention text and its article context. For both steps of our pipeline, we use a pre-trained XLM model [11] as an encoder, which we fine-tune during training.

In the LLM approach, we build on the pre-trained LLama 2 language model [40], directly predicting canonicalized actor strings as a text generation task, conditioned on a prompt containing the target claim. We compare zero- and few-shot prompting settings for base- and instruction-tuned model variants. For both settings, we construct the prompt following the current best practices [24,28,29]. The few-shot approach involves in-context learning, where the prompt contains a number of claim-actor pairs from the training set chosen by the cosine similarity score obtained from SBERT embeddings [35]. In our zero-shot approach, we do not include any claim-actor pairs in our prompt. Instead, we prompt our model with a short English-language description of the task. We experiment with various automatically constructed prompt paraphrases using ChatGPT shown in Table 4.

We evaluate our models via F_1-score. To better comprehend the strengths and weaknesses of both models, we use three evaluation settings. In our strictest *exact-match* setting, predictions are considered correct only if they exactly match the gold-standard actor string. *Correct-up-to-formatting* setting is more lenient

Table 5. Results for the LLM, traditional pipeline and hybrid models in the different evaluation settings.

	Evaluation	Pr	Re	F_1
LLM	exact match	42.66	43.46	43.06
	up to formatting	43.56	44.39	43.98
	up to canonic.	62.39	63.55	62.96
dedicated pipeline	exact match	48.66	59.35	53.47
	up to formatting	48.66	59.35	53.47
	up to canonic.	54.79	66.82	60.21
hybrid approach	exact match	54.33	64.49	58.97
	up to formatting	54.33	64.48	58.97
	up to canonic	64.96	79.39	70.21

by ignoring formatting differences (e.g. whitespaces, capitalization, punctuation) in the predictions. Lastly, in our *correct-up-to-canonicalization* setting, predictions are considered correct if they identify the correct entity, allowing variations in referring expressions For instance, both "the chancellor" and "Merkel" would be counted as correct predictions for the gold-standard actor "Angela Merkel".

Table 5 summarizes the results. While, under the strict exact-match setting, our traditional pipeline outperforms the LLM-based model, the LLM outperforms the pipeline when only evaluating up to canonicalization. This implies that the LLM is actually better than the pipeline at identifying the correct political actor, but struggles to canonicalize these actors consistently.

Motivated by this observation, we introduce a hybrid model that is structurally similar to our traditional pipeline model but includes the LLM prediction as an additional input. In this way, the pipeline can learn to delegate to the LLM when deciding which actor made the claim, and only has to properly canonicalize the LLM's prediction. Table 5 shows the hybrid model's performance under the same three evaluation settings. We find a substantial increase in performance across all settings. This suggests that our hybrid approach is able to leverage additional synergies between our two model architectures, improving upon the constituent models' abilities to both identify and canonicalize actors for claims.

3 Coarse-Grained Analysis of Political Discourse

We now proceed to the second approach, the analysis of manifestos to characterize parties. Party competition is a crucial mechanism in democracies. It creates an arena where a plurality of political viewpoints are given voice, enabling individuals to select one that aligns with their own beliefs. Analyzing this phenomenon is fundamental to understanding voters' choices during elections as well as the decisions taken by governing parties [3]. Researchers analyse party competition by, for example, placing them in a low-dimensional political space:

a one-dimensional left-right or libertarian-authoritarian, or conservative-liberal scale, or in a two-dimensional space formed by combining these scales [20].

We investigate the extent to which the positioning of parties can be captured through their manifestos – the electoral programs in which parties articulate their perspectives, plans, and objectives. Manifestos are crafted with the double intention of conveying information and persuading potential voters [8].

Political researchers analyze party manifestos to explore aspects such as level of similarity among parties concerning different policies [8], party alliances [15], and the alignment between voters' decisions with their worldviews [30]. By offering direct access to the parties' viewpoints, they serve as a robust foundation for comprehending the parties' ideologies regarding different policies. In contrast to the newspaper-based approach, manifesto-based analysis does not provide specific information about what types of decisions were made or articulated. On the other hand, it is arguably the most direct way of accessing the ideologies shared by members of the same party. It also avoids the filtering of information (via actors and their claims) through the lens of media.

3.1 Ideological Characterization

Traditionally, political science has approached the task of identifying party positioning by manually assigning a label to each sentence of a given manifesto. The Manifesto Project (MARPOR, [6]) is a well-known example that follows this method. Its annotations follow a codebook that classifies each sentence into a broader policy domain such as 'external relations' or 'freedom and democracy' as well as assigning a fine-grained label related to a specific category of the policy domain, such as 'freedom and human rights'. The category sometimes also encodes the stance, e.g., 'Constitutionalism: Positive or Negative'. These labels are then analyzed in terms of saliency, assuming that the most frequently mentioned policies are the most important ones for a party. A simplified version of saliency-based analysis is the RILE index, which is a coarse-grained measure that defines lists of 'left' and 'right' policies and simply calculates parties' position on the left-right scale as the relative mention frequency of left vs. right policies [7].

Manual annotations come with a substantial cost and must be carried out for each country and election. We ask whether we can alleviate this burden with unsupervised methods drawn from recent advancements in NLP. In [9], we empirically investigate the following questions with manifestos from Germany: 1) How to create embeddings for parties from their manifestos that yield robust between-party similarities estimates? 2) What aspects of document structure can be exploit for this purpose? 3) How well can these embeddings be computed in a completely unsupervised fashion?

We carry out experiments with six sentence embedding models, all of which estimate party positions on the basis of sentence similarity. These models range from a classic distributional model (fasttext) to transformers [35], applying whitening to ameliorate anisotropy [39] and comparing vanilla and fine-tuned versions. The results are shown in Table 6. Since we hypothesize that using only sentences expressing *claims* (cf. Section 2) might be more informative of the

Table 6. Correlation between our unsupervised scaling method and the ground truth (Wahl-o-Mat). Adapted from [9].

	Only claims		Entire manifestos	
Embeddings	**Domain**	**No domain**	**Domain**	**No domain**
fasttext$_{avg}$	*0.54	0.35	*0.44	0.41
BERT$_{german}$	0.37	*0.47	0.36	*0.48
RoBERTa$_{xml}$	0.39	*0.51	*0.46	*0.54
SBERT$_{vanilla}$	***0.57**	*0.50	***0.53**	*0.57
SBERT$_{domain}$	*0.44	*0.45	0.41	*0.52
SBERT$_{party}$	*0.53	***0.70**	*0.50	***0.69**

positioning of the parties, we introduce an experimental condition which considers only automatically identified claims ('only claims'). Finally, we test whether computing sentence similarity within each MARPOR domain improves results ('Domain'). See [9] for details.

We evaluate our unsupervised scaling method against similarities according to parties' answers to the German Wahl-o-Mat questionnaire, a voting advice application (VAA) [43], generally considered a reliable estimation of party distances. Table 6 shows the results. The correlations are similar between the setup with *only claims* and *entire manifestos*, suggesting that claims are not much more informative than all sentences, at least when they are automatically recognized. The best model overall is based on fine-tuned SBERT$_{party}$ embeddings (which was fine-tuned to make statements by the same party more similar to one another) and computes similarities on an overall level instead of separately for each domain (column 'No domain'). The lack of benefit from domain information might be surprising. One possible explanation is that voters prioritize different domains and do not simply 'average' across them [21].

We believe that these results hold promise for computational political science: leveraging document structure could potentially reduce the need for domain experts to annotate extensive amount of data. Our study has clear limitations, though: first, our experimentation was limited to a single language and dataset. In a follow-up study [31], we have established that classifiers based on state-of-the-art multilingual representations perform robustly in this task across countries and over time. Secondly, we have only considered a few document structure-based cues for fine-tuning. The range of available cues however is enormous and more research is needed in order to better understand the strengths and limitations of sentence embeddings.

3.2 Policy-Domain Characterization

The study from the previous subsection primarily considered the *aggregated* level of overall party positions [12,38]. Political scientists are, however, often interested in specific policy domains. We therefore ask, in [10], how well we can extend the approach presented above to the level of individual *policy domains*.

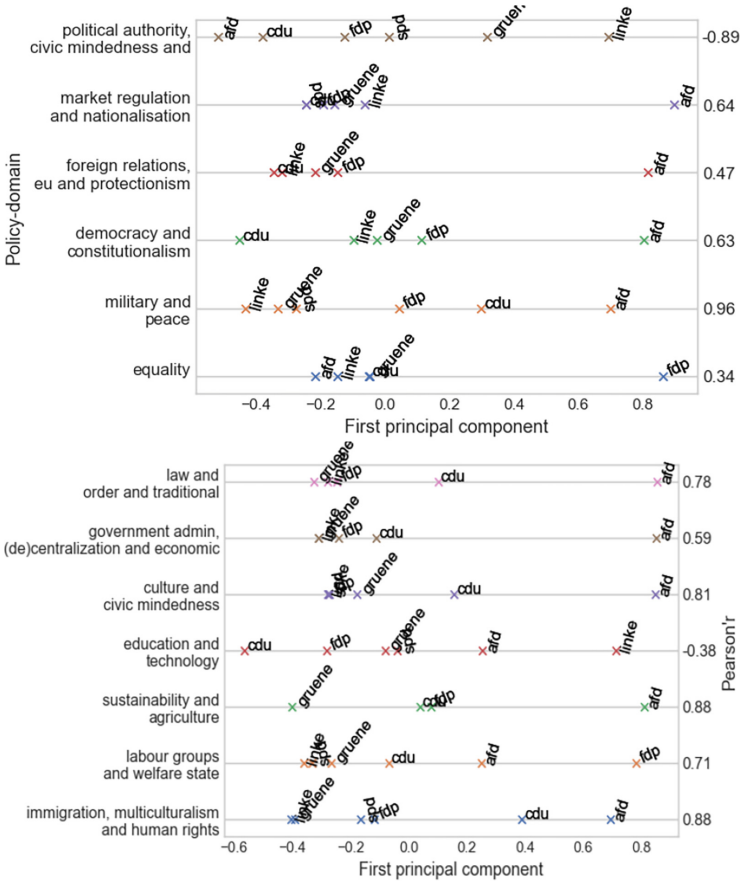

Fig. 4. Automatic prediction of German party positioning within policy domains (right-hand numbers: correlation with RILE scale).

Our approach computes distances between parties at the policy domain level by first training a *policy-domain labeller* which classifies the sentences of unannotated documents and then computing pairwise distances among sentences of the same policy domains across parties. We interpret the first principal component of the aggregated similarity matrix as a policy domain-specific scale.

Our experiments reveal that while the top-performing policy-domain labeller's accuracy is moderate (64.5%), the correlation between the predicted sentences and the ground truth – the RILE index (mentioned in 3.1) – remains remarkably high (r=0.79) in comparison with the annotated scenario (r=0.87). Figure 4 shows the positioning of parties per policy domain. In line with established observations about the German political landscape, a majority of policy domains exhibit a strong correlation to the RILE index, indicating a consistent adherence to the left-right scale. Where this is not the case (EU, market,

government), a cluster of 'established' parties is clearly separated from the populist AfD. When evaluating the predicted setups against manual annotation, we find that the higher the accuracy of the policy-domain labeller within a class, the higher the correlation with the annotated results (Pearson r=0.59, p=0.03). This suggests that the accuracy of the labeller can be used as an indicator of which policy domains to reliably include in the analysis of unannotated manifestos.

This study verifies again that our methods perform well at an aggregated level of information by correlating highly with the RILE index. Moreover, our proposed workflow supplements the previous studies of party positioning with further detailed analysis within the sphere of policy domains. The predictions we obtain align closely with expert assessments, indicating that our workflow provides a reliable method to automatically compute the similarity between parties across some policy domains.

4 Conclusions

This paper considered the challenges of applying NLP methods for a text-based analysis of political debates. We compared two approaches: the first one aims at a fine-grained representation, taking individual statements (claims) and the political actors who made them as its building blocks, with the final goal of extracting discourse network representations from raw texts; the second one targets a coarse-grained representation of the debates at hand, with parties as the actors and their positions expressed in manifestos as its building blocks, with the final goal identifying global ideological positions, across languages and time.

As regards the fine-grained approach, our experiments and analyses show that current transformer-based language models have the potential to fundamentally change the way social scientists can use large text corpora to analyze political discourse. Whereas so far fine-grained analyses of political discourse have mostly been limited to short time spans, single countries or had to employ far-reaching sampling strategies to reduce the amount of texts to be processed. Following the pipeline from Fig. 1 we now know that *claim identification* (Task 1) needs to be preceded by a preparatory task to discard irrelevant documents, but after that, detection models work very well even on topics outside the original training data. *Actor detection* and *mapping* (Tasks 2 and 3) can be handled with reasonable success using traditional NLP methods such as entity extractors and classifiers respectively, but we also saw promising first results in using large language models to perform these two tasks jointly. However, owing to the inherent challenge in controlling the output generation of LLMs, the most effective strategy combines their capability to identify the correct actor and subsequently perform the canonicalization step within the traditional pipeline. For *claim classification* (Task 4), few-shot models show high potential, but they need human curation and re-calibration especially for infrequent claim categories.

While unable to fully automate annotation, current NLP tools go beyond just speeding up manual annotation processes. Topic agnostic claim detection models, few-shot learning, accounting for category hierarchies and models for

actor mapping have the potential to restructure qualitative social sciences text analysis workflows. Instead of starting from zero with a small set of completely manually annotated texts, the current tools allow researchers to immediately focus on relevant text sections and potential claim sentences. With this a traditionally sequential annotation process can be replaced by a parallel and focused approach in which human interaction is mainly focusing on curation tasks.

When we turn to the coarse-grained approach, which aims at identifying the positioning of political parties based on manifestos, the verdict is even more optimistic. It shows performances similar to human annotators when identifying the positioning of parties in the well-established left-right scale (RILE index) or regarding their similarities according to Wahl-o-Mat. These results carry over, to an extent, to the level of individual policy domains – results for the annotated policy-domains correspond well to human expert judgements – but the task becomes considerably more difficult for the models. There is a clear need for further research on assessing the limits of the coarse-grained approach, and specifically on improving the performance of the classifier across policy-domains.

Thus, both approaches have advantages and disadvantages. The fine-grained discourse network analysis offers greater insights into what is being articulated in the public sphere and identifies the key political actors influencing or engaging in those discussions. However, even though we have shown that the annotation load can be alleviated with NLP tools, the task still requires extensive labelling, and it is very domain focused – i.e., each domain demands a new codebook and round of annotations. Besides that, the generalizations derived from the networks are dependent on what is reported by the media, where the focused claims and actors are selected by the news outlets. The coarse-grained approach based on manifestos, on the other hand, gives direct access to parties' policies and higher-level ideological positioning, reaching high quality with little to no annotations. That being said, the coarse-grained approach cannot provide detailed information about individual actors or claims in the political discourse, instead focusing on the relation among parties either at a policy-domain or at an ideological level.

Ultimately, we contend that the two approaches complement each other by offering distinct perspectives onto the political process: One illuminates the precise agreements and disagreements among actors, whereas the other offers an overview of party relatedness at a level of ideology or policy domains. Both offer insights and challenges that can be traded off according to the type of data, resources and analysis requirements at hand.

Limitations. The studies we presented in this paper were carried out primarily on newspaper text and party manifestos. While these are arguably two important text types for political discourse, they are by far not the only ones. Future work is necessary to determine the extent to which our findings carry over to other text types, notably oral modes such as (parliamentary) debates or intermediate modes such as social media communication. Similarly, the bulk of our work was concerned with German language texts. On the methodological perspective, it could take advantage of the relatively good NLP resource situation for German,

leaving open the question of how to deal with similar situations in lower-resource languages. Our pilot studies [31,45] indicate that Machine Translation into a higher-resource language such as English appears a simple but effective solution for almost all languages at this point. At the data level instead, the annotations of German manifestos are recognized for their high quality due to the evaluation of inter-annotator agreement [25] – which may not be the case with manifestos from other countries. A crucial aspect to keep in mind are bias issues which could affect the models and thus result in unfair representations of the political discourse, i.e., overlooking actors from specific groups and/or their claims. While in [14] we have addressed frequency bias for claim detection (higher accuracy for claims by high frequency actors) a broader spectrum of unfairness sources is yet to be explored, in particular in the light of the use of LLMs.

Acknowledgments. The studies reported in this paper were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the projects MARDY-1 and MARDY-2 (grant number 375875969) within RATIO.

References

1. Barić, A., Padó, S., Papay, S.: Actor identification in discourse: a challenge for LLMs? In: Proceedings of the EACL CoDi Workshop, pp. 64–70. St. Julians, Malta (2024)
2. Barnes, J., Klinger, R.: Embedding projection for targeted cross-lingual sentiment: model comparisons and a real-world study. JAIR **66**, 691–742 (2019)
3. Benoit, K., Laver, M.: Party Policy in Modern Democracies. Routledge (2006)
4. Blokker, N., Blessing, A., Dayanik, E., Kuhn, J., Padó, S., Lapesa, G.: Between welcome culture and border fence: the European refugee crisis in German newspaper reports. LRE **57**, 121–153 (2023)
5. Brown, T., et al.: Language models are few-shot learners. In: Proceedings of NeurIPS, pp. 1877–1901 (2020)
6. Budge, I.: Validating the manifesto research group approach: theoretical assumptions and empirical confirmations. In: Laver, M. (ed.) Estimating the Policy Position of Political Actors, pp. 70–85. Routledge (2001)
7. Budge, I.: The standard Right–Left scale. Technical report, Comparative Manifesto Project (2013). https://manifesto-project.wzb.eu/down/papers/budge_right-left-scale.pdf
8. Budge, I., Klingemann, H.D., Volkens, A., Bara, J., Tanenbaum, E. (eds.): Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998. OUP, Oxford, New York (2001)
9. Ceron, T., Blokker, N., Padó, S.: Optimizing text representations to capture (DIS)similarity between political parties. In: Proceedings of CoNLL. Abu Dhabi, UAE (2022)
10. Ceron, T., Nikolaev, D., Padó, S.: Additive manifesto decomposition: a policy domain aware method for understanding party positioning. In: Findings of ACL. Toronto, Canada (2023)
11. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Proceedings of NeurIPS, Vancouver, Canada (2019)

12. Däubler, T., Benoit, K.: Scaling hand-coded political texts to learn more about left-right policy content. Party Politics 13540688211026076 (2021)
13. Dayanik, E., et al.: Improving neural political statement classification with class hierarchical information. In: Findings of ACL, Dublin, Ireland (2022)
14. Dayanik, E., Padó, S.: Masking actor information leads to fairer political claims detection. In: Proceedings of ACL, pp. 4385–4391. Online (2020)
15. Druckman, J.N., Martin, L.W., Thies, M.F.: Influence without confidence: upper chambers and government formation. LSQ **30**(4), 529–548 (2005)
16. Eger, S., Daxenberger, J., Stab, C., Gurevych, I.: Cross-lingual argumentation mining: machine translation (and a bit of projection) is all you need! In: Proceedings of COLING. Santa Fe, New Mexico, USA (2018)
17. Haunss, S., Blessing, A.: Revisiting the exit from nuclear energy in Germany with NLP. Zeitschrift für Diskursforschung (2023, under review)
18. Haunss, S., Blokker, N., Blessing, A., Dayanik, E., Lapesa, G., Kuhn, J., Padó, S.: Integrating manual and automatic annotation for the creation of discourse network data sets. Politics Govern. **8**(2), 326–339 (2020)
19. Haunss, S., Dietz, M., Nullmeier, F.: Der Ausstieg aus der Atomenergie. Diskursnetzwerkanalyse als Beitrag zur Erklärung einer radikalen Politikwende. Zeitschrift für Diskursforschung **1**(3), 288–316 (2013)
20. Heywood, A.: Political Ideologies: An Introduction. Bloomsbury Publishing (2021)
21. Iversen, T.: Political leadership and representation in West European democracies: a test of three models of voting. AJPS **38**(1), 45–74 (1994)
22. Kammerer, M., Ingold, K.: Actors and issues in climate change policy: the maturation of a policy discourse in the national and international context. Soc. Netw. **75**, 65–77 (2023)
23. Koopmans, R., Statham, P.: Political claims analysis: integrating protest event and political discourse approaches. Mobilization **4**(2), 203–221 (1999)
24. Kumar, S., Talukdar, P.: Reordering examples helps during priming-based few-shot learning. In: Findings of ACL-IJCNLP, pp. 4507–4518. Online (2021)
25. Lacewell, O.P., Werner, A.: Coder training: Key to enhancing coding reliability and estimate validity. Mapping Policy Preferences from Texts, Statistical Solutions for Manifesto Analysts (2013)
26. Leifeld, P.: Discourse network analysis: policy debates as dynamic networks. In: Victor, J.N., Montgomery, A.H., Lubell, M. (eds.) The Oxford Handbook of Political Networks. OUP (2016)
27. Leifeld, P., Haunss, S.: Political discourse networks and the conflict over software patents in Europe. Eur. J. Polit. Res. **51**(3), 382–409 (2012)
28. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. In: Proceedings of ACL, pp. 8086–8098. Dublin, Ireland (2022)
29. Margatina, K., Schick, T., Aletras, N., Dwivedi-Yu, J.: Active learning principles for in-context learning with large language models. In: Findings of EMNLP, pp. 5011–5034. Singapore (2023)
30. McGregor, R.M.: Measuring "correct voting" using comparative manifestos project data. J. Elect. Publ. Opinion Parties **23**(1), 1–26 (2013)
31. Nikolaev, D., Ceron, T., Padó, S.: Multilingual estimation of political party positioning: from label aggregation to long-input transformers. In: Proceedings of EMNLP. Singapore (2023)
32. Padó, S., Blessing, A., Blokker, N., Dayanik, E., Haunss, S., Kuhn, J.: Who sides with whom? Towards computational construction of discourse networks for political debates. In: Proceedings of ACL. Florence, Italy (2019)

33. Papay, S., Klinger, R., Padó, S.: Constraining linear-chain CRFs to regular languages. In: Proceedings of ICLR. Online (2022)
34. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual BERT? In: Proceedings of ACL. Florence, Italy (2019)
35. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (November 2019). https://doi.org/10.18653/v1/D19-1410, https://aclanthology.org/D19-1410
36. Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., Biemann, C.: Neural entity linking: a survey of models based on deep learning. Semant. Web **13**(3), 527–570 (2022)
37. Shimaoka, S., Stenetorp, P., Inui, K., Riedel, S.: Neural architectures for fine-grained entity type classification. In: Proceedings EACL. Valencia, Spain (2017)
38. Slapin, J.B., Proksch, S.O.: A scaling model for estimating time-series party positions from texts. Am. J. Polit. Sci. **52**(3), 705–722 (2008)
39. Su, J., Cao, J., Liu, W., Ou, Y.: Whitening sentence representations for better semantics and faster retrieval. arXiv:abs/2103.15316 (2021)
40. Touvron, H., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
41. Volkens, A., et al.: The Manifesto data collection, version 2019b (2019)
42. Vu, H.T., Liu, Y., Tran, D.V.: Nationalizing a global phenomenon: a study of how the press in 45 countries and territories portrays climate change. Glob. Environ. Chang. **58**, 101942 (2019)
43. Wagner, M., Ruusuvirta, O.: Matching voters to parties: voting advice applications and models of party choice. Acta Politica **47**(4), 400–422 (2012)
44. de Wilde, P.: No polity for old politics? A framework for analyzing the politicization of European integration. J. Eur. Integr. **33**(5), 559–575 (2011)
45. Zaberer, U., Padó, S., Lapesa, G.: Political claim identification and categorization in a multilingual setting: first experiments. In: Proceedings of KONVENS. Ingolstadt, Germany (2023)
46. Zürn, M.: The politicization of world politics and its effects: eight propositions. EPSR **6**(01), 47–71 (2014)