Datenbank Spektrum (2020) 20:143–153
https://doi.org/10.1007/s13222-020-00344-w

SCHWERPUNKTBEITRAG

Analysis of Political Debates through Newspaper Reports: Methods and Outcomes

Gabriella Lapesa[1] · Andre Blessing[1] · Nico Blokker[2] · Erenay Dayanik[1] · Sebastian Haunss[2] · Jonas Kuhn[1] · Sebastian Padó[1]

Received: 14 February 2020 / Accepted: 30 May 2020 / Published online: 16 June 2020
© The Author(s) 2020

Abstract
Discourse network analysis is an aspiring development in political science which analyzes political debates in terms of bipartite actor/claim networks. It aims at understanding the structure and temporal dynamics of major political debates as instances of politicized democratic decision making. We discuss how such networks can be constructed on the basis of large collections of unstructured text, namely newspaper reports. We sketch a hybrid methodology of manual analysis by domain experts complemented by machine learning and exemplify it on the case study of the German public debate on immigration in the year 2015. The first half of our article sketches the conceptual building blocks of discourse network analysis and demonstrates its application. The second half discusses the potential of the application of NLP methods to support the creation of discourse network datasets.

Keywords Computational Social Science · Discourse Network Analysis · Machine Learning

1 Introduction

Political decision making in democratic societies on all but the most technical issues builds on a prior public debate. One element of political debate is the exchange in parliaments, which is often ritualized. Another important part of public debate takes place in the news and especially in quality newspapers. The debate depicted there does not represent the positions and ideas of the general population, but it is a central source of information used by decision makers [26]. Indeed, most political decisions which are bound to affect large portions of the population attract public attention and thus happen in a politicized mode, in which more or less intense public debates accompany decision making [15, 39, 40]. To better understand democratic decision making, we need a fine-grained picture of such debates and of their dynamics that captures specific aspects of the domain at issue and represents how the structure of support/disagreement evolves around such aspects.

Political debates have a very complex structure, and their representation in the public sphere (approximated in this paper by the news coverage) is accordingly complex. First, they result from the interaction among different types of *actors*: politicians, parties, governments, but also groups of citizens (i.e., protesters). Second, the public debate does not target the topic as a whole (e.g., pro or against immigration?), but very specific aspects of it which are tightly connected to specific policy measures (e.g., Should a quota for refugees be established? Should empty flats be assigned to refugees?). Third, crucially, the dynamics of a public debate evolve over time, as a result of the shift in the opinion of leading figures (e.g., party leaders) or as result of external shocks (e.g., the Fukushima disaster reshaped the position of the public opinion).

A framework which has shown to be very helpful for capturing the abovementioned structural aspects of political debates is *discourse network analysis* [23], which combines recent innovations from political claims analysis [18] with network science. Crucially, this framework relies on the annotation of (large) newspaper corpora which is noto-

We acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG) through MARDY (Modeling Argumentation Dynamics) within SPP RATIO and by the Bundesministerium für Bildung und Forschung (BMBF) through Center for Reflected Text Analytics (CRETA).

✉ Gabriella Lapesa
gabriella.lapesa@ims.uni-stuttgart.de

Nico Blokker
blokker@uni-bremen.de

[1] Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany

[2] Research Center on Inequality and Social Policy, University of Bremen, Bremen, Germany

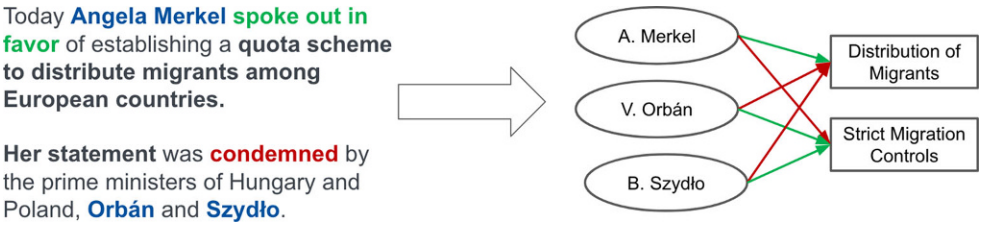

Fig. 1 From text to discourse networks through claim and actor annotation

Table 1 Textual spans, Actors, & Claims: annotation examples

Text	Actors	Claims	Full annotation
SPD and the Greens demanded, apart from the withdrawal of term extension for nuclear power plants, that the seven oldest and most insecure power plants should be switched off	SPD Green	term extension, closing older plants	TERM_EXTENSION+REJECT+SPD TERM_EXTENSION+REJECT+GREEN CLOSING_OLDER_PLANTS+SUPPORT+SPD CLOSING_OLDER_PLANTS+SUPPORT+GREEN
One could hear slogans from the demonstrators: *'No walls around Europe. Right to stay for everyone and for long!'*	demonstrators	border installations, residency right	RESIDENCY_RIGHT+SUPPORT+DEMONSTRATORS BORDER_INSTALLATIONS+REJECT+DEMONSTRATORS
The Government intends to halve pension entitlements for the unemployed in 2007.	Government	pension cutbacks	PENSION_CUTBACKS+SUPPORT+GOVERNMENT

riously tedious and time-consuming. So far, discourse network studies have been carried out manually, limiting their scope to the amount of data that can be considered within the breadth of a research project.

This is where Natural Language Processing becomes relevant: It opens up the use of large-scale newspaper archives as basis for network construction, enabling us to substantially broaden the empirical basis and applicability of discourse network studies. In this paper, we outline a methodological framework in Computational Social Science that integrates NLP with discourse network studies. We exemplify it on a case study investigating the relation between media coverage and policy making in a specific topic: the migration debate in Germany in the year 2015. Consider the example in Fig. 1. The text snippet on the left contains mentions to three actors: Angela Merkel, Viktor Orban, and Beata Szydlo. The actors take opposite positions with respect to two claims (distribution of migrants, strict migration controls): this allows us to cluster Orban and Szydlo together, and to capture their opposition to Merkel.

This paper makes two contributions: first, we outline the main features of the discourse network analysis framework and demonstrate its application on a manually annotated corpus of the German migration debate. Second, we demonstrate the potential of the application of NLP methods to scale-up the annotation of public debates from large corpora, thus allowing the network analysis to be conducted on a larger data sample.

2 Discourse Network Analysis (DNA)

Discourse Network Analysis (DNA) is a framework for the representation and the analysis of policy debates (i.e., political discourses centered on a given policy, e.g., immigration or pension). By modeling policy debates as dynamic networks, DNA effectively brings together political science and network analysis [23].

One conceptual building block of the DNA [24] framework is the notion of *political claim* [18]. Claims are statements concerning specific actions to be taken with respect to a specific aspect of a domain of interest. In more detail, they take the form of demands, proposals, criticisms, or collective actions. Crucially, each claim should be attributable to *actors* (individuals or groups) in a specific *polarity* (support or opposition).

Table 1 shows the example of three textual snippets containing claims (highlighted in italics), together with their annotation under a DNA-framework (claim category, polarity, actor) from three different debates: the nuclear phase-out debate in Germany after the Fukushima disaster (example 1), the domestic migration debate — the focus of this article (example 2) — and the pension debate (example 3).[1]

DNA represents actors and claims as the two types of nodes in a bipartite *affiliation network*, as shown in Fig. 2. We show actors as circles and claims as squares, which are linked by edges that indicate support (green) or opposition (orange). The projection of the affiliation network, the concept side, yields argumentative clusters present in the debate. The projection of the affiliation network on the ac-

[1] In this article, we use English translations of the original German texts.

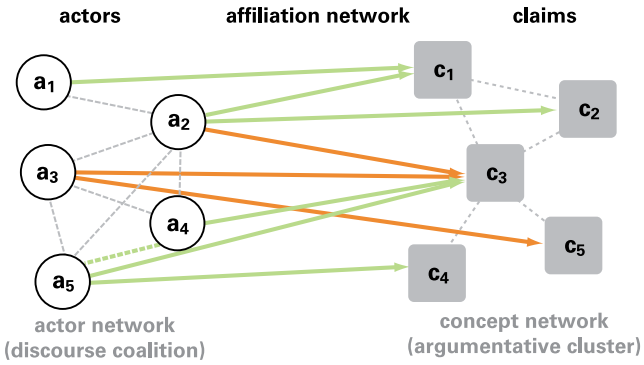

Fig. 2 Actor, affiliation, and concept networks, from [12, p. 131]

tor side (dotted edges) yields *discourse coalitions*, defined as groups of actors who share a social construct (e.g., in Table 1, example 1, SPD and Green party act as a coalition with respect to the two claims). Within a political debate, discourse coalitions are not static: their emergence and evolution (new actors, new claims) is considered of key importance as far as political decisions are concerned [10, 29]. For example, one highly influential political actor may change her/his mind about a specific aspect because of an external event, initiating a change in the opinion of other discourse participants.

The natural question is how to populate such networks for a debate of interest. The DNA approach targets text sources such as newspaper articles (most notably), parliamentary testimonies and other types of documents, depending on the debate to be analysed [23]. Claims are manually identified by trained/expert annotators, and assigned to the corresponding actors. Carrying out corpus annotation for DNA requires a number of design decisions, most prominently on the concept side of the bipartite networks (the claims), but also on the actor side.

From the conceptual side of the network, political claim annotation (or *coding*) requires an *ontologization* of the domain of interest. More concretely, this requires that experts employ their knowledge to establish a set of theory-neutral *categories* to be employed in the annotation [19] (e.g., in Table 1, "term extension" and "closing older plants" in example 1; "border control" and "right to residency" in example 2). The granularity of the claim ontology (in technical terms, *codebook*) is crucial because it targets the different issues of the phenomenon under debate, which are in turn the object of the policy making (a policy does not target immigration as a whole, but a specific practical aspect of it, i.e., border control, accommodation for refugees, etc.). The generation of a codebook is a core element of category-based content analysis approaches in the social sciences [38].

Populating the actor side of the network can be implemented by simply marking of the textual span corresponding to the actor, but quite often implies additional annotation, too. The most obvious one is the distinction between person and organization actors (e.g., "Angela Merkel" vs. "SPD", with interesting intermediate cases such as "the Federal Government"), and, potentially, the assignment of actors to political parties.

Finding the key players: DNA and SNA Once the bipartite network has been populated, the next step is to ask who are the most influential actors and the most popular claims, and how the discourse evolves over time.

This is the point at which the synergy with (social) network analysis (SNA) comes in handy. The notion of *centrality* has been devised to detect influential nodes (the "key players") in a social network, and it reflects the position of a node (in our case an actor, or a claim) within the network. Centrality can be quantified according to different metrics: degree, betweenness, closeness, eigenvector centrality (for a comprehensive review, refer to [5, 21, 25]). Degree of a node v is calculated as the number of edges incident to v or as the number of nodes directly connected to v. The computation of betweenness and closeness is based on the impact of v on the connectivity of the network: how many shortest paths among other nodes pass through v (betweenness)? How close is v (average length of the shortest path) to the other nodes in the graph (closeness)? Eigenvector centrality of v, on the other hand, takes into account the importance of neighboring nodes for v. A classical example of its application is Google's PageRank.

Crucially, the SNA measures for centrality defined above apply to DNA networks. For a discussion of suitable methods and developments for inferential statistics for (temporal) DNA see [23] and [33].

3 Case study: the domestic immigration debate in Germany (2015) and *DEbateNet-mig15*

The year 2015 was a crucial one in the debate concerning migration in Germany. The extremely high number of people seeking to enter Europe from Africa and the Middle East challenged European societies, sparking plenty of discussions concerning how to react to what was perceived as a crisis. This has revealed cleavages along and within party lines addressing both moral as well as economic and pan-European obligations. Amidst such heated debate, governments have reacted to publicly voiced demands by adapting their migration policies (i.e., operational policies and the body of laws regulating actions in this specific domain).

The case study presented in this section targets the German side of the pan-European (in fact, world-wide) debate on the refugee "crisis" in 2015.

3.1 Annotation workflow

The dataset and analysis we present here are the output of a large annotation project which took place at the University of Bremen and took roughly a year, involving six political science students and two domain experts as annotators. Our textual source was *Die Tageszeitung (taz)*, a major national German quality newspaper. From the full 2015 taz issue, we selected and annotated 959 articles, which all together constitute *DEbateNet-mig15*, the annotated corpus which is documented in [22] and available as a CLARIN resource.[2] *DEbateNet-mig15* has been annotated using *MARDY*, an environment developed for and shaped by the needs of the Political Science workflow, described in detail in [1].

As anticipated in Sect. 2, corpus annotation for the purposes of DNA targets multiple levels, with different degrees of abstraction and complexity for the annotator. Fig. 3 illustrates the different levels of the annotation carried out in this project, and the corresponding tasks in the workflow:[3]

1. **Claim detection**: identification of the textual spans containing claims. Claim-bearing textual spans do not necessarily coincide with a sentence: they can be a subpart of a sentence, or span beyond the sentence boundary.
2. **Actor detection**: identification of the strings corresponding to actor mentions (e.g., "Merkel", "Die Kanzlerin", "Frau Merkel").
3. **Actor mapping**: different actor mentions need to be mapped to the same referent (e.g., "Merkel", "Die Kanzlerin", "Frau Merkel" → ANGELA MERKEL).
4. **Claim classification**: assignment of theoretically motivated claim categories to the textual spans. Note that a textual span can be assigned more than one claim category. The claim ontology, our annotation codebook, comprises 8 high-level categories (controlling migration, residency, integration, domestic security, foreign policy, economy/labor market, society, procedures) and 97 subcategories (e.g., "asylum right" and "border control" are among the sub-categories of "controlling migration").[4]
5. **Claim attribution**: textual spans identified as claims are assigned to the relevant actor. Note that a single claim can be attributed to more than one actor, and actors can be mentioned inside or outside the textual span. At this step, the annotator also annotates **polarity** (does the actor support or reject the categorized claim?) and **date** (by default the day preceding the publication of the article; but it may need to be further reconstructed based on textual information).
6. **Claim aggregation**: strictly speaking, is not part of the annotation workflow but takes place before the quantitative interpretation of the annotation that will be performed by DNA.

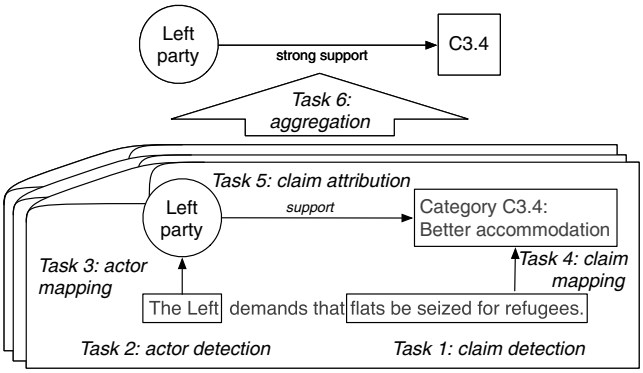

Fig. 3 From text to discourse network: workflow from [13]

DebateNet-migr15 contains 1815 textual spans, corresponding to 2274 distinct claims (recall that one textual span may contain more than one claim). Table 2 displays the 10 most frequently annotated claim subcategories, grouped by polarity (positive vs. negative). It shows category/polarity frequency (*Freq*), along with the overall frequency of the category in the entire dataset (positive plus negative, *Glob*). A comparison of the two lists provides the basis for a number of observations, whose common denominator is the clear separation between discourse coalitions on the left and right sides of the political spectrum. First, the claims "EU solution" and "Safe country of origin" dominate both rankings (first and third in the positive ranking, second and fourth in the negative ranking), albeit with a largest share for the positive polarity; this indicates that these two claims have been at the center of the political conflict and, thus, can serve as an indicator for the identification of discourse coalitions. Second, the societal claims appearing in both lists show a clear left-wing nuance: "Xenophobia" and "Right-wing extremism" are always criticized or attacked (negative polarity) and never supported. The claim "Refugees Welcome" is almost always supported. Third, still on the left vs. right divide, we observe that the claims "Deportation", "Ceiling/Upper Limit", as well as "Border Control" are mostly reported with a positive polarity, thus being dominated by more right-wing actors.

A full-fledged quantitative analysis of our resource falls out of the scope of this paper. In what follows, we fill focus on two (tightly interrelated) aspects already pointed out as particularly relevant in Sect. 1, namely the identification of

[2] http://hdl.handle.net/11022/1007-0000-0007-DB07-B.

[3] Note that this representation is in principle agnostic to who the annotator is: be it a human or a NLP classifier, these are the necessary steps and the corresponding tasks. In this section, we refer to data that are annotated manually, and Sect. 4 will tackle the the integration of ML modules in the workflow.

[4] The full codebook and the annotation guidelines are avaliable at the address https://github.com/mardy-spp/mardy_acl2019/blob/master/codebook.pdf.

Table 2 Top 10 claim categories by polarity, from [22]: category identifier (*Code*), category frequency in the targeted polarity (*Freq*), global category frequency in the whole dataset (*Glob*), and claim label (*Claim Category*)

Code	Freq	Glob	Claim Category
		POSITIVE	
501	152	193	EU solution (quotas for refugees)
812	97	104	Fast/Accelerated procedure
504	93	124	Safe country of origin
805	78	82	Additional financing
207	70	82	Deportation
102	69	80	Ceiling/Upper limit
105	59	73	Border controls
309	55	76	Care (medical, financial)
705	54	60	Refugees welcome
108	46	59	Immigration law
		NEGATIVE	
703	45	54	Xenophobia
501	41	193	EU solution (quotas for refugees)
190	36	51	Current migration policy
504	31	124	Safe country of origin
709	24	25	Right-wing radicalism
203	24	64	Centralized accommodation
104	21	52	Walls-up policy
309	21	76	Care (medical, financial)
110	17	41	Asylum right
202	17	34	Refugee accommodation

cues to find *discourse coalitions* and the use of discourse networks to characterize the evolution of the debate *over time*.

3.2 Network Analysis

We now analyze the discourse network of *DEbateNet-mig15*. Our annotation enables aggregation at several levels, most straightforwardly the levels of actors, claims, and time. We focus here on the dimension of time.

Table 3 displays aggregated statistics of the whole *DEbateNet-mig15* network, on a monthly basis. For each month, we report number of observations (*Obs.*), number of claims (*C-token*), number of distinct claim categories (*C-type*), number of distinct actors (*Actors*), as well as the average degree centrality in the network (*Degree*), measured as the number of incident connections (edges) to each node [37, p. 100]. Recall from Sect. 2, that degree can be interpreted as a measure of centrality and thus popularity/prestige of a node.

We observe that September is the month with the highest average degree: this is not surprising as the heated debate is the natural consequence of the peak of refugee arrivals over the summer and of the high number of casualties in the Mediterranean. Fig. 4 depicts the discourse network for March and September (the months with the lowest and

Table 3 *De*bateNet-mig15 aggregated statistics over months, from [22]: number of textual spans (*Obs*ervations), absolute number of claims (*C_token*), number of distinct claims (*C_type*) and of distinct actors (*Actors*), average degree centrality of the network (*Degree*)

Month	Obs	C_token	C_type	Actors	Degree
Jan	141	189	47	77	2.44
Feb	66	83	30	57	1.77
Mar	41	60	27	31	1.72
Apr	79	89	37	49	1.84
May	78	106	32	41	2.41
Jun	89	100	41	56	1.86
Jul	140	177	46	90	2.13
Aug	207	264	52	109	2.73
Sep	411	541	66	168	**3.38**
Oct	211	285	55	116	2.69
Nov	186	267	58	82	2.91
Dec	166	213	54	78	2.39

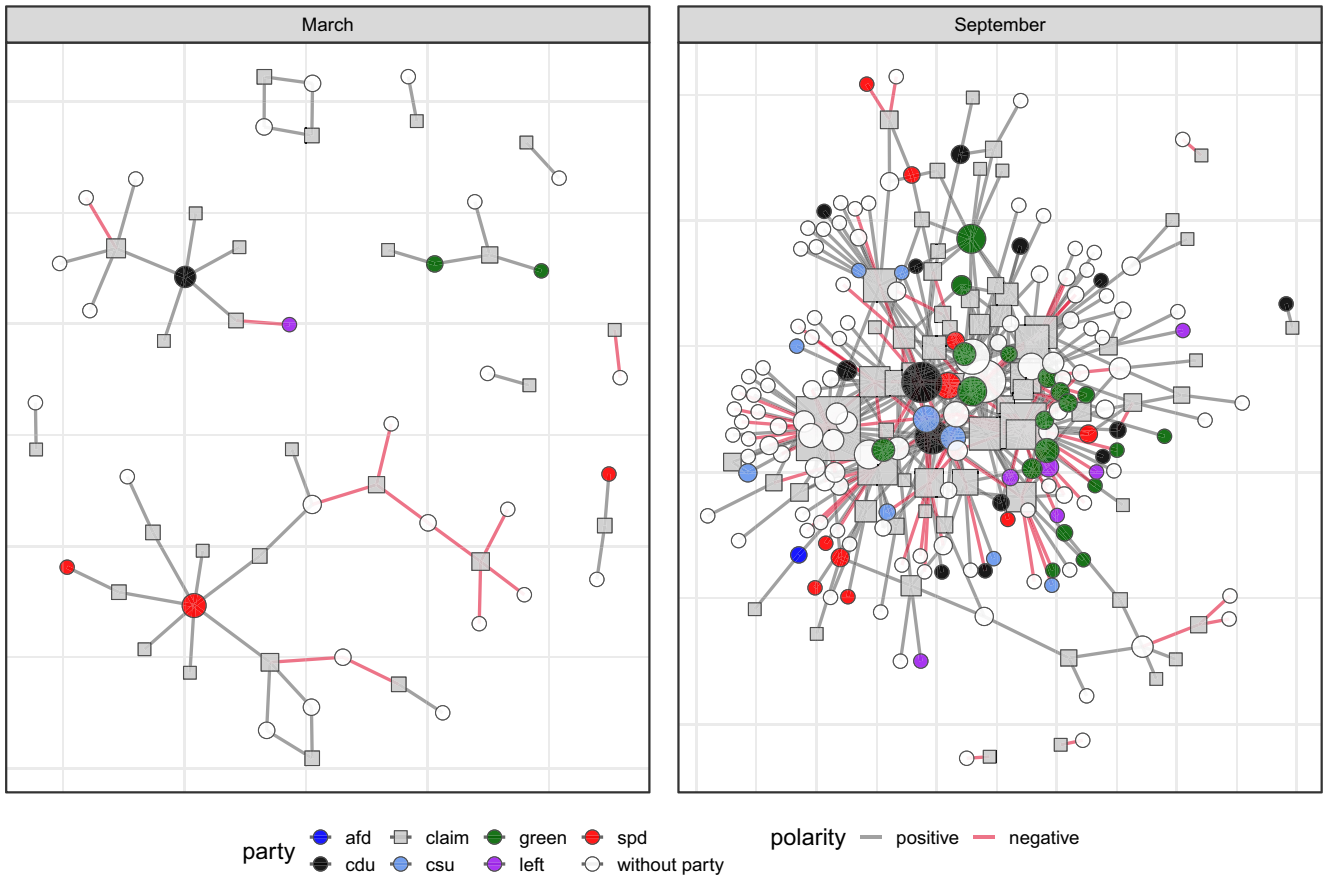

Fig. 4 Discourse network: March vs. September

highest average degree, respectively), showing how the increase in the density of the discourse is reflected in the state of the network. The discourse network for March shows several unconnected components, indicating a fragmented discourse. This impression is reinforced by the fact that members of distinct parties almost never address the same claims. For instance, the Christian Democratic Party (CDU) focuses on reducing the number of asylum-seekers, while the Social Democratic Party (SPD) aims to integrate migrants into the labor-market. This has changed in September, which shows much more coherence and connectedness. At the same time, the number of actors involved and proposed claims increases substantially. This is also reflected on the party-level. The CDU, its sister party CSU (Christian Social Union), and the oppositional Green Party are now occupying the center of the debate, while the Left party and the far-right leaning Alternative für Deutschland (AfD) are less salient.

Aggregation can also be carried out over multiple months to capture more stable discourse patterns in the debate. In this case, it is common in DNA to focus on the core structure of the discourse by focussing on repetitive actor-claim pairs and discarding those that occur only once as marginal opinions. More concretely, this amounts to counting the total number of times a specific actor A makes a certain claim C, add these values as edge weights into the network, and keep these pairs occurring at least twice per day in the designated time frame. The remaining network will only contain edges and adjacent nodes with edge weights greater than one. Technically, this view of the original network configuration is referred to as its 2-slice representation [27, p. 98]. Intuitively, the 2-slice, core network displays actors and claims which occur more than once in a specified time-frame and are more likely to have an influence on the political debate.

Fig. 5 displays actor-claim-pairs that have been reported at least twice on different days between September and December, the most lively phase of the debate (cf. Table 3). Node size corresponds to the prominence of actors and claims in terms of degree centrality. The centrality of the nodes in the network is a correlate of their prominence in the discourse, and only the most prominent nodes are labelled. In this case, we observe that Angela Merkel is, unsurprisingly, the leading figure of the debate, with claim 501 (the need for a EU-wide solution) being simultaneously the most central claim, and the most controversial one: note the high share of supporting (10) vs. rejecting edges (6). Other prominent figures are Thomas de Maizière, minister of the

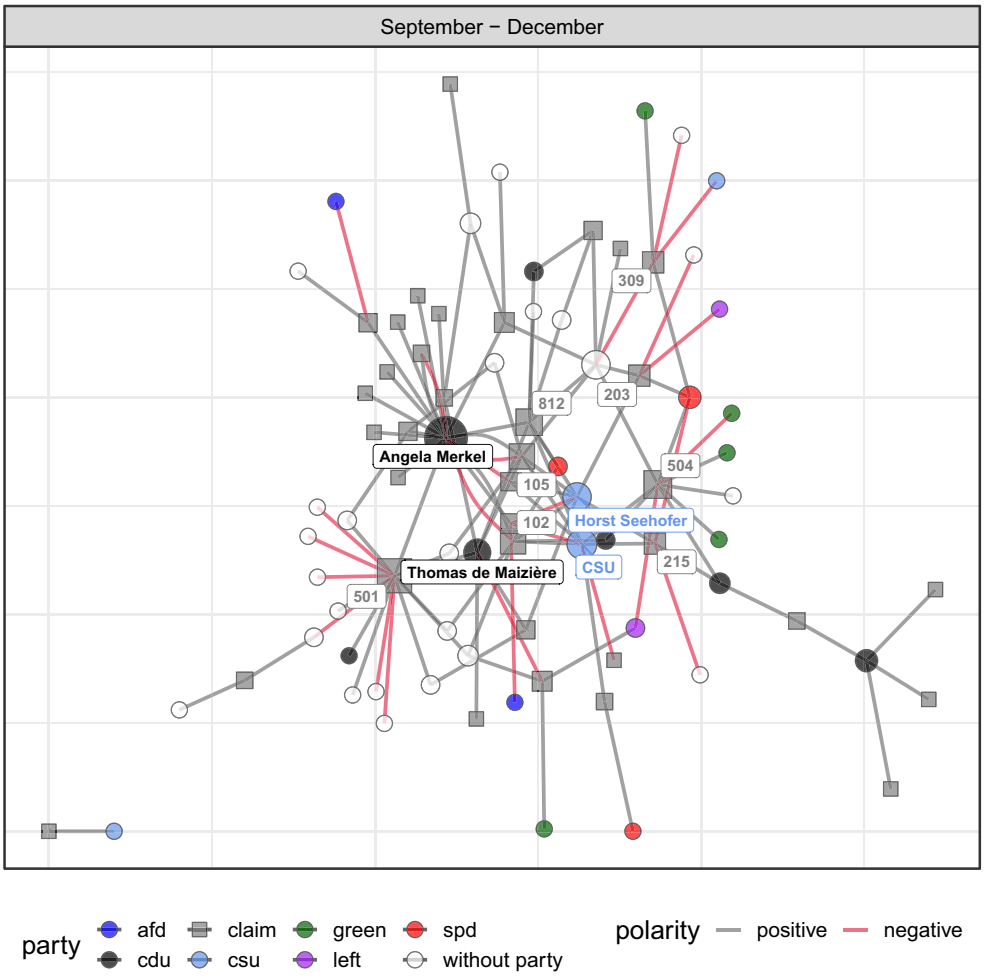

Fig. 5 Core discourse network (2-slice): Sept–Dec

interior and Horst Seehofer, prime minister of Bavaria at the time.

4 Network Analysis with Machine Learning: first steps

In the previous section, we have demonstrated the application of DNA in the (manual) analysis of the domestic debate on migration in Germany. So far, we have assumed human annotation: high quality, but costly. The next step is the exploitation of Natural Language Processing methods for the extraction of comparable discourse networks from large textual sources. NLP can support the need to scale-up the annotation in two (often overlapping) ways. The first one is the design of efficient annotation tools which integrate linguistic preprocessing and efficient support for annotators to deal with language data, leading to faster and more reliable annotation. We do not cover this level in the current article; see [1] for a discussion. The second one, in focus here, targets the development of machine learning models which can be applied to novel data to automate (or, more realistically, semi-automate) the workflow.

4.1 Related work: NLP for Political Science

Our work situates itself at the intersection of a number of well-explored tasks in NLP, but it is substantially different from all of them in the complexity of the phenomenon investigated (development of discourse network over time) and in the granularity of semantic analysis (following the detailed domain ontology which is needed to characterize policies and debates).

The natural point of comparison is the body of work on the manipulation of media coverage for *framing purposes* and *agenda setting* (e.g. [9, 35]). While such work does target the temporal dynamics of the monitored debates, it does not target fine-grained categories, and typically takes a single article as the object of annotation. Besides, it is in a sense complementary to our approach: framing/agenda setting targets the control of governments over media, our work targets the impact of media coverage on the actions taken by governments.

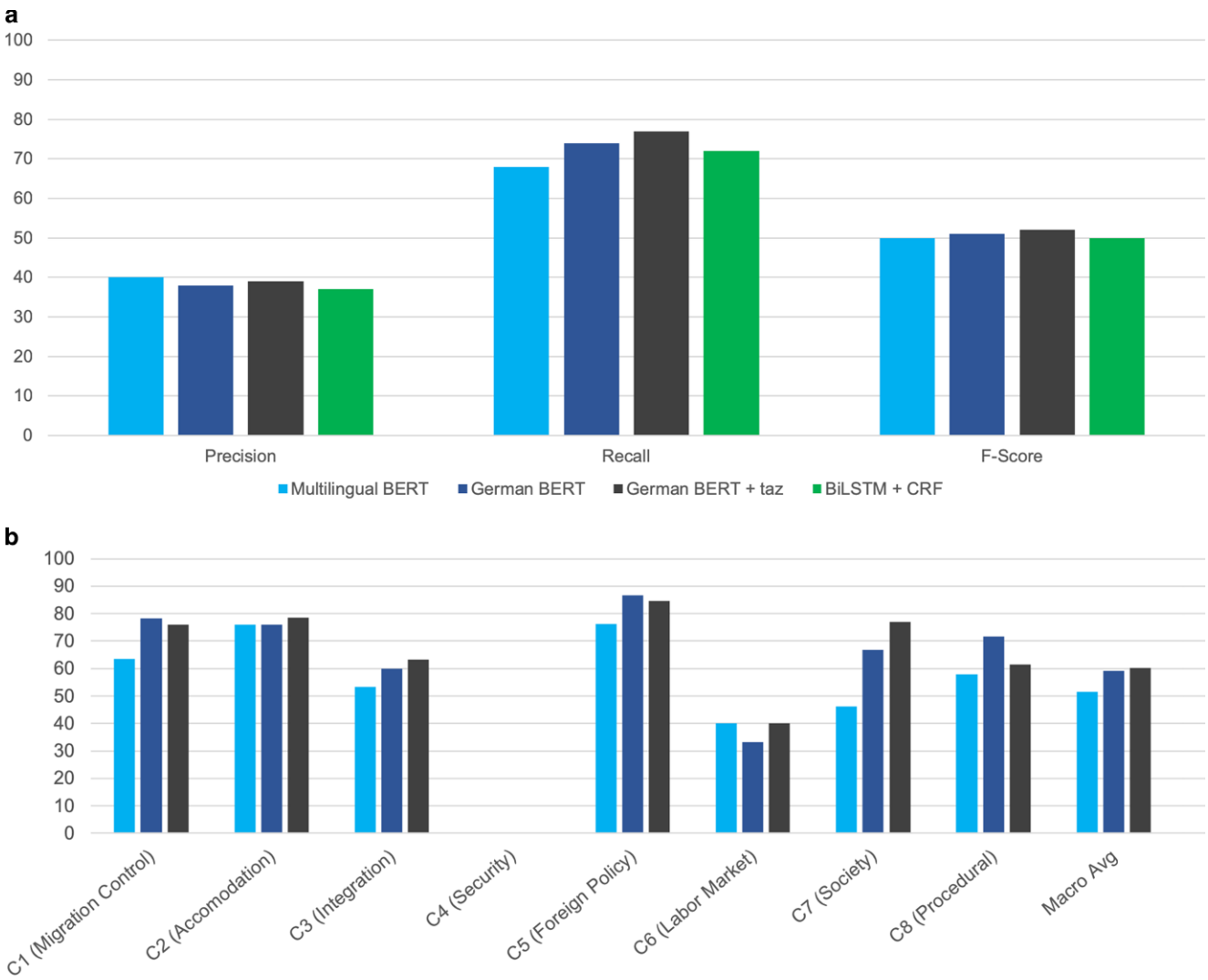

Fig. 6 Evaluation of claim detection (*top*, P/R/F1) and claim classification (*bottom*, F1 for major categories)

Our work also relates to research in Social Media Analysis using NLP – in particular sentiment analysis (e.g. [3]), but also going into fine-grained analysis of groups of users/actors [4]. Such analyses typically target relatively broad categories, such as party preferences [16], stance classification (e.g. [36]), or the measuring of ideology in speeches [30]. Also related is the growing field of argumentation analysis (e.g. [28, 32, 34]). Our approach differs in adapting a political science-based definition of claim ("a statement concerning specific actions to be taken"), which only covers a subset of what is often considered a claim in argument mining. Also, a core interest there is analyzing the argument structure of longer texts (i.e., claims and their justifications), whereas we focus on recognizing actors' core claims as reflected in news coverage.

It is within the Digital Humanities that we find the closest relative to our focus on the dynamics of the interaction among actors, namely in the extraction of actor/character networks from literary texts [8, 11, 17].

4.2 Automatic claim detection and classification

In the long run, all the annotation tasks shown in Fig. 3 should be automated. Therefore, we have started by automating two subtasks, namely claim detection and claim categorization, both framed as (supervised) sequence classification tasks.

Tasks. Claim detection takes a sequence of words as an input, and it answers a yes-no question: is the sequence a claim? If the answer is yes, we proceed to claim classification, where each claim is assigned one or more of the eight top level categories (multi-label classification) of the MARDY claim codebook. The restriction to top-level categories is a practical one, since for many of the more specific categories we do not have a sufficient number of examples

to learn reliable classifiers in the challenging multi-label situation.

Model architectures. We adopt the approach that has become standard for semantic tasks in NLP, namely neural network models. We consider two standard architectures for sequence classification, one based on recurrent networks (**BiLSTMs with a CRF layer** for structural prediction) and one based on transformers, more specifically BERT [7]. The main difference between those methods consists in their treatment of word meaning: The recurrent network has a simple input layer representing word meaning by way of word embeddings, while transformers *contextualize* the meaning of words based on the sentences in which they occur. For the BiLSTM model, we use in-domain word embeddings trained on the taz newspaper corpus by using FastText [2]. These are concatenated with character-based word embeddings obtained by learning character embeddings and feeding them through a CNN and max-pooling the output. As for BERT, we assess the effect of language-specific adaptations to the transformer model. We experiment with three variants: **Multilingual BERT**, trained in a language-agnostic fashion on the Wikipedia text of 104 languages; (2) **German BERT**, a German-specific BERT model[5] trained only on German corpora (including Wikipedia); (3) **German BERT + taz**, a variation of German BERT which we fine-tuned on the complete taz corpus by using BERT's 'next sentence prediction' objective (see [7] for details).

Optimization. We optimize and evaluate all models in the *DEbateNet-mig15* dataset. We use 75% for training, 10% for development and 15% for testing. We optimize hyperparameters on the development set. With a view to using the models as proposals to be reviewed by human annotators, we prioritize recall over precision: we select the model with the highest development set recall value among the subset of saved models where the recall/precision ratio is equal to or smaller than two.

Results. Fig. 6 shows evaluation results for the two tasks. The top panel shows that we obtain reasonable results for claim detection, in particular a high recall, albeit a somewhat lower precision, as a consequence of our model selection strategy. We believe this trade-off is reasonable for semi-automatic annotation support. Since human annotators review the ML predictions, they can easily rule out false positives, while due to the high recall the model has a chance of finding instances which may be missed by human annotators.

For both tasks, language specific pre-processing on BERT is important to achieve good results: The German BERT + taz model obtains the best Recall (0.77) and F-Score (0.52) for claim detection and the best F-Score for five of eight major groups and best macro-averaged F-Score (0.60) for claim classification. The move from multilingual BERT to German BERT appears to be more important than the additional fine tuning on the domain specific corpus. Our interpretation is that for the multilingual BERT model, a major limiting factor is subword tokenization [31] because the model has to learn cross-lingually valid patterns. Consequently, it struggles with the long German nouns, often compounds, that are indicative of domain-specific claims and categories [22]. Examples include Asylbewerber (Asylum seeker), Flüchtling (refugee), Einschränkungen (limitations), or Forderungen (requirements). German BERT, with its language-specific subword tokenization, handles this aspect better.

Even if using language adaptation improves overall performance, there are still problems in the data which notably affect claim classification. The first is data imbalance. We observe that none of the claim classifiers is able to perform well on the category C4 (Security) – arguably, this is because we have less than 60 instances in the training set. Another reason for the high variance among major categories lies in the idiosyncrasies of the categories: Categories with a specific technical jargon (e.g. Dublin Procedure) are generally easy to learn from a few examples, while other categories, expressed more colloquially, may require more examples (e.g. limiting migration, social commitment, or integration offers).

4.3 Further research directions

Our evaluation of the ML claim detection and classification has shown that, at the current state of the art, we cannot automate full, end-to-end discourse network creation at a level of accuracy that is useful for Political Science analysis. There is still an open question, however, whether the ML predictions can be used to support the annotation process, in a semi-automatic fashion. Note that this requires a shift in the way evaluation is designed: a classifier does not need to be *perfect* to be an efficient support to a human annotator. Newspaper articles are naturally redundant, because both political actors and journalists tend to repeat themselves, and such (near)-repetitions are often ignored in discourse network analysis because they do not provide substantial new information: for the purpose of the network, only one instance (more often two, in the 2-slice case discussed before) of each actor-claim pair needs to be detected in a specified time-frame. As a consequence, network construction is more forgiving than a text-based evaluation perspective might suggest at first glance: it can proceed even based on a (somewhat) incomplete manual or automatic annotation. This makes optimizing for recall not only good for (semi)automatic annotation, but also advantageous for (semi)automatic network extraction.

[5] https://deepset.ai/german-bert.

We have conducted an experiment investigating the impact of redundancy/recall both on network extraction and on the annotation workflow [13]. We applied the recall-optimized classifier to 40 randomly selected articles (not in the training data) and found out that (a) while the claim detection only reached a 77% of recall on the full network, it did score a 100% recall on the core network (2-slice): in other words, it was able to recover all relevant claims; and (b) interestingly, while the integration of ML suggestions in the annotation environment did not uniformly affect speed (some annotators became faster, some even became slower), it significantly improved inter-annotator agreement.

That being said, the evaluation discussed so far has still been in-domain, leaving open the empirical question of the transferability of our ML approach to other debates. Another challenge for ML systems which is currently very much in the focus of the NLP community is the presence of *bias*, particularly relevant for us because a biased classifier would result in an "unfair" representation of the debate at issue. Transferability and bias surely relate to one another, as we would like the quality of automatic analyses to be as independent as possible of textual properties of the debate at issue. They are, however, not fully overlapping: ensuring transferability does not necessarily guarantee that models will not be biased, and the other way round.

Concretely, in [6] we have targeted the issue of actor frequency bias in claim identification. Actor frequency is one of the textual properties which we would not want our ML to rely on. This would amount to learning that every span containing the mention of a salient actor is a guaranteed to also contain a claim and it would generate a bias towards a subset of the political sphere, effectively "silencing" less prominent actors. Several debiasing methods on claim identification have been tested in [6]: among them, simply masking the actor mention in the training stage turned out to be the best one. It reduced frequency bias (i.e., improving performance on actors from a low frequency band) while keeping the overall performance stable. Interestingly, no bias for political party was found. Next, [6] also tackled the transferability issue by evaluating on the debate on the future of nuclear energy use in Germany in the four months after the Fukushima disaster, in 2011 [14]. Once again, actor masking resulted in improved performance on a debate different in topic, year, and thus with a non overlapping set of actors.

Summing up, the recall/redundancy interaction uncovered in [13] and the debiasing and transferability potential tested in [6] indicate a promising avenue for (semi)automatic discourse network analysis. In this connection, current work targets the evaluation of the classifiers on new topics (e.g., the pension debate) and text types (e.g., party manifestos).

5 Conclusion

In this paper, we have outlined a hybrid methodology for the analysis of political claims on text corpora. Sect. 2 has defined the building blocks of Discourse Network Analysis and Sect. 3 translated them into concrete steps for the analysis of the domestic debate on migration in Germany in 2015. Sect. 4 reported the results of ML experiments on the first modules of the workflow (claim detection and classification) and discussed the potential of its application in a (semi)automatic annotation setting.

Current developments target the representation of the argumentative structure in the discourse network. As pointed out in [23], besides claims, a very important cue to identify coalitions in a discourse network are the justifications actors use to support their claims. We annotated a new dataset on the German debate on pension, in the same year (2015) and on the same source (taz), keeping track of both claims (e.g., the trade unions oppose the pension cutbacks) and justifications (e.g., social justice). This shift raises new questions and increases the complexity of the workflow because (a) it adds a level of annotation and it requires a separate codebook for frames as well as the update of the ML architecture and (b) it turns the bipartite network into a tripartite one, opening a set of methodological issues for DNA. It will also, however, bring our approach closer to current research on Argument Mining, enabling us to benefit from existing datasets, guidelines, and insights.

The research framework summarised in this paper originates from the synergy of two examples of mixed-method approaches to research: DNA on the Political Science/SNA side [23], and our hybrid methodology on the CL/DH side [20]. Bringing together questions, methods, and concrete requirements from different fields has shaped our research program and brought it further for all involved sides, providing valuable criteria for the interpretation of the ML results. Such synergy has also shown that in a real world, application-oriented research scenario the quantification of performance needs to be rethought in a pragmatic way.

Funding Open Access funding provided by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Blessing A, Blokker N, Haunss S, Kuhn J, Lapesa G, Padó S (2019) An environment for the relational annotation of political debates. In: Proceedings of ACL Florence (system demonstrations), pp. 105–110
2. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. TACL 5:135–146
3. Ceron A, Curini L, Iacus SM, Porro G (2014) Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media Soc 16(2):340–358
4. Cesare N, Grant C, Nsoesie EO (2017) Detection of user demographics on social media: A review of methods and recommendations for best practices. CoRR abs/1702.01807
5. Cordeiro M, Sarmento R, Brazdil P, Gama J (2018) Evolving networks and social network analysis methods and techniques. In: Social Media and Journalism – Trends, Connections, Implications
6. Dayanik E, Padó S (2020) Masking actor information leads to fairer political claims detection. In: Proceedings of ACL Seattle (To appear)
7. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT Minneapolis, pp 4171–4186
8. Elson DK, Dames N, McKeown KR (2010) Extracting social networks from literary fiction. In: Proceedings of ACL Uppsala, pp 138–147
9. Field A, Kliger D, Wintner S, Pan J, Jurafsky D, Tsvetkov Y (2018) Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In: Proceedings of EMNLP Brussels, pp 3570–3580
10. Hajer MA (1993) Discourse coalitions and the institutionalization of practice: the case of acid rain in Britain. In: The argumentative turn in policy analysis and planning. Duke University Press, Durham, NC, pp 43–76
11. Hassan A, Abu-Jbara A, Radev D (2012) Extracting signed social networks from text. In: Proceedings of textgraphs-7/graph-based methods for NLP Jeju, pp 6–14
12. Haunss S (2013) Conflicts in the knowledge society. The contentious politics of intellectual property. Cambridge University Press, Cambridge
13. Haunss S, Blokker N, Blessing A, Dayanik E, Lapesa G, Kuhn J, Pado S (2020) Integrating manual and automatic annotation for the creation of discourse network data sets. Polit Gov 8(2):326–339
14. Haunss S, Dietz M, Nullmeier F (2013) Der Ausstieg aus der Atomenergie. Diskursnetzwerkanalyse als Beitrag zur Erklärung einer radikalen Politikwende. Z Diskursforsch 1(3):288–316
15. Haunss S, Hofmann J (2015) Entstehung von Politikfeldern – Bedingungen einer Anomalie. Dms – Mod Staat 8(1):29–49
16. Hong L, Yang W, Resnik P, Frías-Martínez V (2016) Uncovering topic dynamics of social media and news: the case of Ferguson. In: Proceedings of social informatics Bellevue, pp 240–256
17. Iyyer M, Guha A, Chaturvedi S, Boyd-Graber J, Daumé H III (2016) Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In: Proceedings of NAACL-HLT San Diego, pp 1534–1544
18. Koopmans R, Statham P (1999) Political claims analysis: integrating protest event and political discourse approaches. Mobilization 4(2):203–221
19. Kuckartz U (2010) Einführung in die computergestützte Analyse qualitativer Daten. VS, Wiesbaden
20. Kuhn J (2019) Computational text analysis within the humanities: How to combine working practices from the contributing fields? Lang Resources & Evaluation 53:565–602
21. Landherr A, Friedl B, Heidemann J (2010) A critical review of centrality measures in social networks. Bus Inf Syst Eng 2(6):371–385
22. Lapesa G, Blessing A, Blokker N, Dayanik E, Haunss S, Kuhn J, Padó S (2020) DEbateNet-mig15: tracing the 2015 immigration debate in Germany over time. In: Proceedings of LREC Marseille, pp 919–927
23. Leifeld P (2017) Discourse network analysis. policy debates as dynamic networks. In: Victor JN, Lubell MN, Montgomery AH (eds) The Oxford handbook of political networks, chap. 12. Oxford University Press, Oxford, pp 301–325
24. Leifeld P, Haunss S (2012) Political discourse networks and the conflict over software patents in Europe. Eur J Political Res 51(3):382–409
25. Lü L, Chen D, Ren XL, Zhang QM, Zhang YC, Zhou T (2016) Vital nodes identification in complex networks. Phys Rep 650:1–63
26. Magin M (2019) Elite versus popular press. In: The international encyclopedia of journalism studies. American Cancer Society, Hoboken, NJ, pp 1–7
27. de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge
28. Peldszus A, Stede M (2013) From argument diagrams to argumentation mining in texts: a survey. Int J Cogn Informatics Nat Intell 7(1):1–31
29. Sabatier PA, Weible CM (2007) The advocacy coalition framework: innovations and clarifications. In: Theories of the policy process. Westview Press, Boulder, CO, pp 189–220
30. Sim Y, Acree BDL, Gross JH, Smith NA (2013) Measuring ideological proportions in political speeches. In: Proceedings of EMNLP Seattle, pp 91–101
31. Singh J, McCann B, Socher R, Xiong C (2019) BERT is not an interlingua and the bias of tokenization. In: Proceedings of DeepLo Hong Kong, pp 47–55
32. Stab C, Gurevych I (2017) Parsing argumentation structures in persuasive essays. Comput Linguist 43(3):619–659
33. Stadtfeld C, Hollway J, Block P (2017) Dynamic network actor models: Investigating coordination ties through time. Sociol Methodol 47(1):1–40
34. Swanson R, Ecker B, Walker M (2015) Argument mining: Extracting arguments from online dialogue. In: Proceedings of SIGDIAL Prague, pp 217–226
35. Tsur O, Calacci D, Lazer D (2015) A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In: Proceedings of ACL Beijing, pp 1629–1638
36. Vilares D, He Y (2017) Detecting perspectives in political debates. In: Proceedings of EMNLP Copenhagen, pp 1573–1582
37. Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
38. Wiedemann G (2016) Text mining for qualitative data analysis in the social sciences. Springer, Berlin Heidelberg
39. de Wilde P (2011) No polity for old politics? A framework for analyzing the politicization of european integration. J Eur Integr 33(5):559–575
40. Zürn M (2014) The politicization of world politics and its effects: eight propositions. Eur Polit Sci Rev 6(1):47–71