arXiv:1907.04126v1 [cs.CL] 9 Jul 2019

Sentiment and position-taking analysis of parliamentary debates: A systematic literature review

Gavin Abercrombie and Riza Batista-Navarro

School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, UK

CONTACT Gavin Abercrombie. Email: gavin.abercrombie@postgrad.manchester.ac.uk

ABSTRACT

Parliamentary and legislative debate transcripts provide access to information concerning the opinions, positions and policy preferences of elected politicians. They attract attention from researchers from a wide variety of backgrounds, from political and social sciences to computer science. As a result, the problem of automatic sentiment and position-taking analysis has been tackled from different perspectives, using varying approaches and methods, and with relatively little collaboration or cross-pollination of ideas. The existing research is scattered across publications from various fields and venues. In this article we present the results of a systematic literature review of 61 studies, all of which address the automatic analysis of the sentiment and opinions expressed and positions taken by speakers in parliamentary (and other legislative) debates. In this review, we discuss the available research with regard to the aims and objectives of the researchers who work on these problems, the automatic analysis tasks they undertake, and the approaches and methods they use. We conclude by summarizing their findings, discussing the challenges of applying computational analysis to parliamentary debates, and suggesting possible avenues for further research.

KEYWORDS

Sentiment analysis; opinion mining; text as data; parliamentary debates; legislative debates

1. Sentiment and position analysis of parliamentary debates

Debate transcripts from legislatures such as the United Kingdom (UK) and European Union (EU) parliaments and the United States (US) Congress, among others, provide access to a wealth of information concerning the opinions and attitudes of politicians and their parties towards arguably the most important topics facing societies and their citizens, as well as potential insights into the democratic processes that take place in the world's legislative assemblies.

In recent years, these debates have attracted the attention of researchers from diverse fields and research backgrounds. This includes computer scientists working in the field of natural language processing (NLP), who have investigated the application and adaptation to the political sphere of methods developed for sentiment analysis of product reviews and blogs, as well as tackling other tasks in this domiain, such as topic detection. In addition, political and social scientists, traditionally relying on

expert coding for the analysis of such transcripts, have increasingly explored the idea of viewing "text as data" (Grimmer & Stewart, 2013), and using computational methods to investigate the positions taken by debate participants.

As a result, a wide range of approaches to the problem of automatic debate analysis have been adopted, with research on this problem differing widely in its aims and methods. Within this work, there exist many inconsistencies in the use of terminology, with studies in some cases referring to very similar tasks by different names, while in others the same term may mean quite different things. For example, while Chen, Zhang, Wang, Yang & Li (2017) and Kapočiūtė-Dzikienė & Krupavičius (2014) both attempt to classify speakers according to party affiliation, the former refer to this as "political ideology detection", and the latter as "party group prediction". Conversely, a single term like "sentiment analysis" may be used to refer to, among other things, support/opposition detection (Thomas, Pang & Lee, 2006), a form of opinion-topic modeling (Nguyen, Boyd-Graber & Resnik, 2013), and psychological analysis (Honkela, Korhonen, Lagus & Saarinen, 2014). The methods adopted range from statistical analyses to predictive methods, including both supervised classification and unsupervised topic modeling. There are also contrasting approaches to modeling the textual data, the level of granuality of the analyses, and, for both supervised learning methods and the evaluation of other approaches, the acquisition and application of labels used to represent the ground-truth.

While Kaal, Maks & van Elfrinkhof (2014) assembled researchers from diverse fields to investigate the problem of text analysis in political texts, and both Hopkins & King (2009) and Monroe, Colaresi & Quinn (2017) discuss the general differences in the aims and objectives of social scientists and computer scientists when working on such problems, as far as we are aware there exists no comprehensive overview, systematic or otherwise, of research in this area. The aim, therefore, of this review is to bring together work from different research backgrounds, locating and appraising literature concerning the automatic analysis of sentiment and position-taking analysis that has been undertaken to date on the domain of parliamentary and legislative debate transcripts. We assess the research objectives, the types of task undertaken, and the approaches taken to this problem by scholars in different fields, and present suggested

directions for future work in this area.

A note on terminology

In the NLP literature, the terms *opinion mining* and *sentiment analysis* are used more or less interchangeably (for a discussion of this, see Pang & Lee (2008)), and are employed to describe both the specific task of determining a document's sentiment polarity (that is, *positive* or *negative*, or sometimes *neutral*), as well as the more general problem area of automatically identifying a range of emotional and attitudinal "private states" (that is, non-observable, subjective states)[1] such as "opinion, sentiment, evaluation, appraisal, attitude, and emotion"' (Liu, 2012). In a recent survey, Yadollahi, Shahraki & Zaiane (2017) list nine such different sentiment analysis subtasks.

Political and social scientists, meanwhile, seem to lack a single term to describe the act of determining from text the positions taken by legislators, and in the literature, such tasks are variously referred to as "political scaling", "position scaling", "ideal point estimation" and a range of other task- and datset-specific terms. One phrase that appears throughout such work is "text as data" (e.g., Grimmer & Stewart, 2013; Laver, Benoit & Garry, 2003; Proksch, Lowe, Wäckerle & Soroka, 2018). We therefore include this term in our systematic search to capture work from the field that utilises computational methods for speaker position analysis.

For the purposes of this article, we use "sentiment analysis" as a general umbrella term, encompassing any tasks concerned with the extraction of information relating to speakers' opinions and expressed positions, and "sentiment polarity classification" for the more specific, binary or ternary classification task.

Review scope and method

For this review, we followed the established systematic review guidelines of Petticrew & Roberts (2006) and Boland, Cherry & Dickson (2017). The use of systematic review methodology, while uncovering a substantial body of relevant work, excludes some potentially interesting studies. We are therefore unable to include a number of known

[1] See Quirk, Greenbaum, Leech & Svartvik (1985).

relevant results uncovered by an initial scoping search of the Google Scholar platform, which, due to the lack of transparency of its search algorithm, does not facilitate replication. While somewhat limiting in this sense, the decision to adhere to a systematic methodology provides a replicable and transparent method of synthesizing and summarizing the literature and identifying future research priorities.

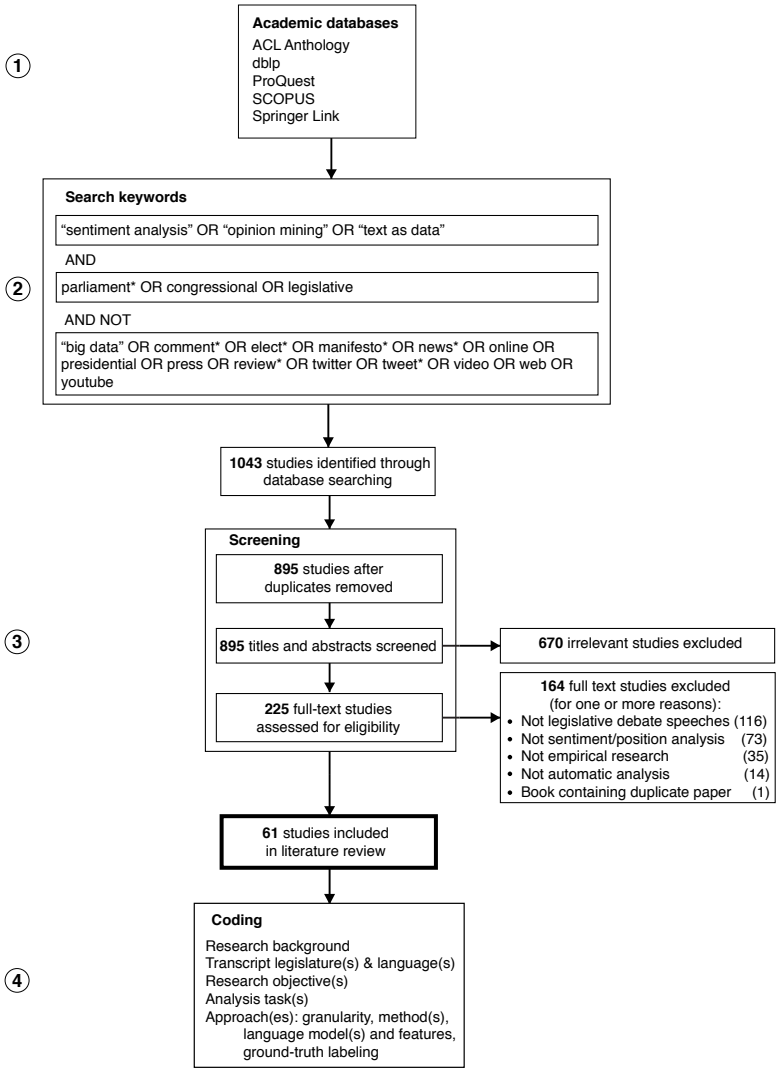
Although there exists interesting work on related domains such as political campaign speeches (e.g., Menini & Tonelli, 2016; Nanni, Zirn, Glavaš, Eichorst & Ponzetto, 2016; Sim, Acree, Gross & Smith, 2013) and electoral manifestos (e.g., Menini, Nanni, Ponzetto & Tonelli, 2017), we limited our search to studies concerning the automatic analysis of the sentiment, opinions, and positions expressed by participants in the transcripts of debates in parliaments and other legislatures, and also excluded any studies that do not report the results of empirical experiments.

The review covers all literature retrieved by systematic search of four digital library databases and repositories that provided high coverage of the results obtained by the prior scoping search. We conducted all searches on January 31st, 2019. Following deduplication, screening and eligibility assessment, 61 studies are included in the review. We coded these according to (1) their research backgrounds, (2) the legislature and language of the debate transcripts analysed, (3) their stated research objectives, (4) the sentiment and position analysis tasks undertaken, and (5) the approaches taken and methods used. The full review protocol pipeline in shown in Figure 1.

Research backgrounds

We categorize the research background of each study according to the institutional affiliation(s) of its author(s) and the nature of its venue of publication, coding them as either *computer science*, *political/social science*, or *multi-disciplinary*. We consider a study to be multi-disciplinary if it (a) is written by authors from two or more research backgrounds, or (b) the paper is published at a venue associated with a different research background than that of its author(s)' affiliations(s). While, it is of course possible that the work we class as being from computer science or the social sciences, actually involves some level of inter-disciplinary collaboration that does not fit within

Figure 1. Flow diagram of the phases of the systematic review process: 1. database selection; 2. keyword search; 3. screening and eligibility assessment; 4. manual coding.



our definition (for example, we do not investigate the authors' academic histories), this is a straightforward way of obtaining a general overview of the research community working in this area.

We find that the majority of studies are written from a computer science perspective (n = 35). Within this are researchers working on two kinds on problems. Firstly, there are those who approach the transcripts from a computational linguistics perspective, and whose work relates to properties of the language used such as argumentation structures and dialogue (Duthie & Budzynska, 2018; Naderi & Hirst, 2016). The second, larger group consists of work that can be characterized as belonging to the field

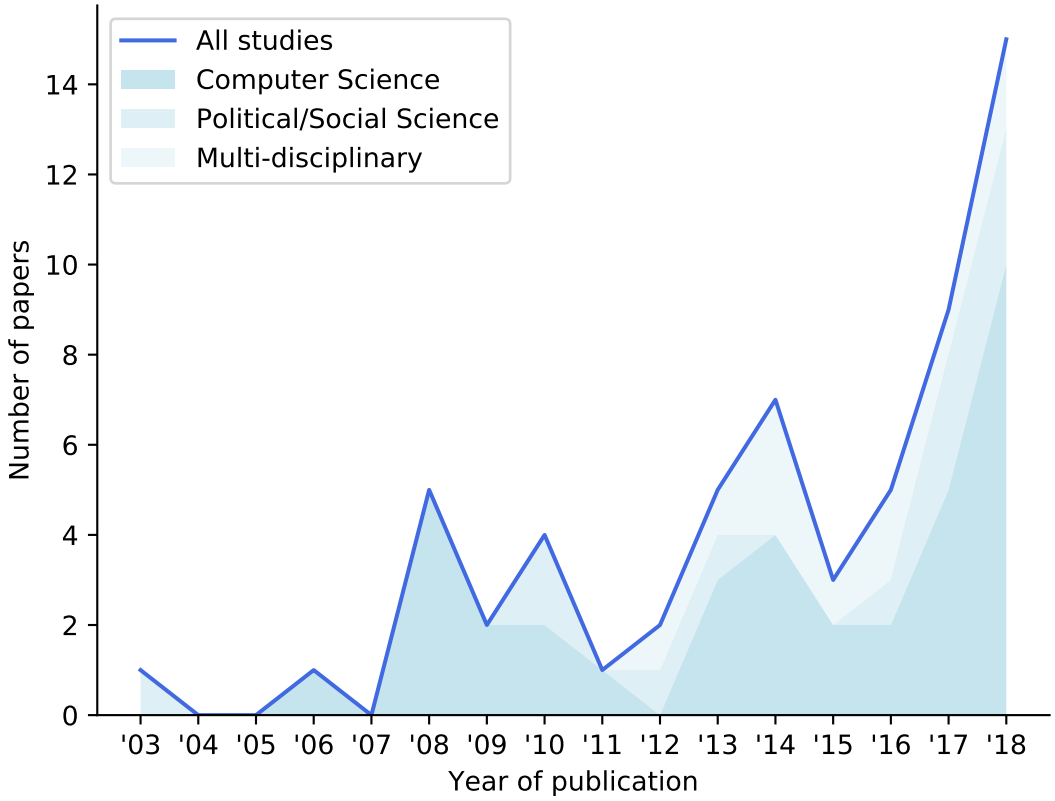
of NLP, and whose work is more focused on tools and applications (e.g., Ji & Smith, 2017; Thomas et al., 2006).

Political and social scientists are responsible for less than half the number of included studies as computer scientists (n = 14), and just 12 studies involve multi-disciplinary research. Of these, seven involve both computer scientists and political or social scientists (Kapočiūtė-Dzikienė & Krupavičius, 2014; Lapponi, Søyland, Velldal & Oepen, 2018; Rheault, 2016; Rheault, Beelen, Cochrane & Hirst, 2016; Rudkovsky et al., 2018; Sakamoto & Takikawa, 2017; Van der Zwaan, Marx & Kamps, 2016), three collaboration between linguists and computer scientists (Honkela et al., 2014; Iyyer, Enns, Boyd-Graber & Resnik, 2014; Nguyen et al., 2013), and two that include researchers from three different fields (Diermeier, Godbout, Yu & Kaufmann, 2012; Nguyen, Boyd-Graber, Resnik & Miler, 2015). According to the number of studies published on this subject annually, interest in this area has been increasing over time, particularly in recent years (see Figure 2.).

Parliaments and legislatures

Nearly all the included studies focus on one single legislature for analysis, with only Sakamoto & Takikawa (2017) and Proksch et al. (2018) comparing their approaches (to the analysis of the level of polarization (ideological division) in parliaments) on transcripts from two or more different chambers. The US Congress is by far the most popular legislature for analysis, attracting the attention of 31 of the studies. This can partly be attributed to the global power and influence of the US and of the English language, but is also explained by the widespread use by NLP researchers of the *ConVote* corpus (Thomas et al., 2006) as a benchmark dataset for the evaluation of sentiment analysis systems. Indeed, including its original authors, 17 of the included studies use this dataset, 15 of which are written from a computer science background, with Hopkins & King (2009) (social science) and Iyyer et al. (2014) (multi-disciplinary) the exceptions. In some cases, *ConVote* is used alongside one or more other non-legislative datasets (such as product reviews) for the evaluation of text classification methods (Allison, 2018; Burford, Bird & Baldwin, 2015; Chen et al., 2017; Iyyer et al., 2014; Ji

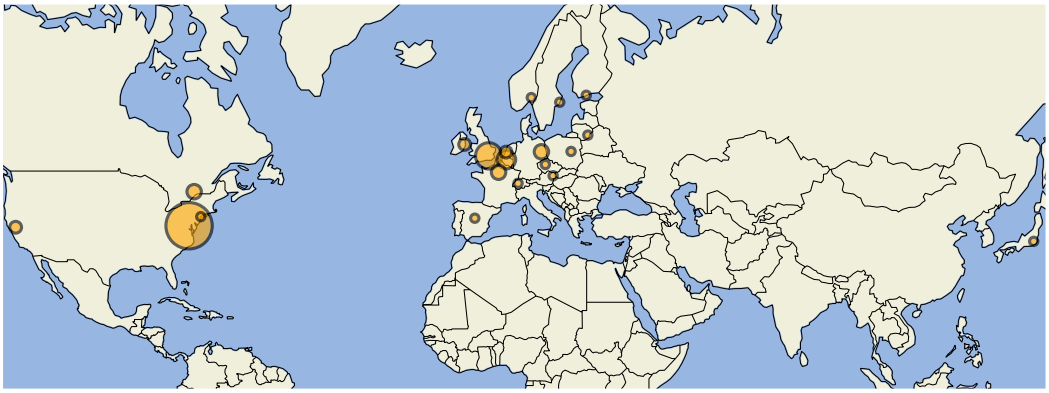
Figure 2. Rate of publication of papers featured in the review from 2003, the year of publication of the first paper, to 2018. The review additionally features one paper published in February 2019.



& Smith, 2017; Li, Chen, Wang & Huang, 2017; Martineau, Finin, Joshi & Patel, 2009; Yogatama & Smith, 2014a,b; Yogatama, Kong & Smith, 2015). In fact, only a little over half (37) of the studies are exclusively concerned with the analysis of legislative debates. On the whole, political and social scientists seemingly prefer to construct their own datasets from the congressional record to suit their research aims, while Sakamoto & Takikawa (2017) (multi-disciplinary computational social science) also do so.

Following Congress, the next most analysed legislatures are the UK Parliament (n = 10) and the EU Parliament (n = 5). The German Bundestag and the French and Canadian parliaments all appear in 3 studies, while the California State Legislature and the Irish Dáil are both analysed in two papers. The Austrian, Czech, Dutch, Finnish, Lithuanian, Norwegian, Spanish and Swiss parliaments, Polish Sejm, Japenese Diet, California State Legislature, and the UN General Assembly are all utilised in only one study each (see Figure 3). It is notable that, thus far, research in this area appears

Figure 3. Global distribution of parliaments and legislatures from which debate transcripts are analysed in the studies included in this review. The size of each marker is proportional to the number of studies carried out on the legislature in question.



to have been restricted to data from North America, Europe, and Japan. Sources of transcript data reported in the included studies are shown in Table 1.

Table 1. Sources of publicly available parliamentary and legislative debate data used in studies.

Legislature	Dataset name	url
Dutch parliament	ODE	http://ode.politicalmashup.nl/data/summarise/folia
Norwegian parliament	Talk of Norway	http://clarino.uib.no/korpuskel/corpus-list
Polish Sejm	*Kronika Sejmowa* (Sejm Chronicle)	www.sejm.gov.pl
UK Parliament	Hansard	https://hansard.parliament.uk
	Historic Hansard	http://hansard.millbanksystems.com
	TheyWorkForYou	https://www.theyworkforyou.com/
US Congress	Capitol Words	http://www.capitolwords.org/
	Congressional Record	https://www.congress.gov/congressional-record
	Govtrack	https://www.govtrack.us/
	ConVote	http://www.cs.cornell.edu/home/llee/data/convote.html
	SLE ConVote	http://www.cs.cornell.edu/ainur/sle-data.html

Nearly all the included studies consist of analysis of parliamentary or legislative data in a single language. Exceptions include Sakamoto & Takikawa (2017) who use two corpora in different languages (English and Japanese), and Glavaš, Nanni & Ponzetto (2017) who use a multilingual dataset (German, French, English, Italian and Spanish), as do Proksch & Slapin (2010) (English, French, and German translations). In the latter case, while the original data are multilingual (23 official languages of the European Parliament), the transcripts have been translated to English, French, and German. Similarly, for speeches not originally in English, Baturo, Dasandi & Mikhaylov (2017) use official translations from the other official languages of the UN (Arabic, Chinese, French, Russian, and Spanish). By far the most prominent language is English, analysed in 55 studies. This is followed by French and German, used in six studies each, and a long tail of languages that only appear in one study each (Czech, Dutch, Finnish,

Italian, Japanese, Lithuanian, Norwegian, Polish, Spanish and Swedish).

Research objectives, tasks, and approaches

We examine three aspects of the studies under review: the authors' stated research objectives; the task types undertaken in order to achieve those aims; and the approaches taken to tackle those tasks. For the latter, we report the granularity at which analysis is undertaken, the methods used, and the labels used to represent ground-truth sentiment or position (where applicable).

Objectives

We examine the principal stated objectives of each study in relation to the backgrounds of the researchers (see Table 2). While the aims of computer scientists are almost always to test a method or system against some baseline and/or state-of-the art system, the work of political and social scientists generally falls into two categories: (1) method performance evaluation, in which novel methods are presented and assessed (e.g., Bonica, 2016; Frid-Nielsen, 2018); and (2) political analysis, in which an existing method is used in order to answer a research question (e.g., Diermeier et al., 2012; Owen, 2017). For political science papers, even where the primary aim appears to be the former, a common approach is to combine these objectives, first presenting a text analysis method, and then illustrating its potential by employing it to answer one or more research questions or test hypotheses, as in Hopkins & King (2009) and Proksch & Slapin (2010).

Although computer scientists generally focus on system evaluation, they often state secondary application objectives which encompass motivations relating to contributions to civic technology or the development of tools for social scientists. For example, Burfoot (2008) suggests that "tools could assist researchers in understanding the nuances of contentious issues on the web by highlighting areas in which different sites or pages agree and disagree", while Budhwar et al. (2018) hope that their work will "give ordinary citizens a powerful tool to better organize and hold accountable legislators without the costs of physical presence and full-time lobbying representation".

Table 2. Included studies by background and stated research objectives. Individual studies may have more than one objective.

Objective	Computer science	Political and social sciences	Multi-disciplinary
Datasets/resources		Baturo et al. (2017); Rauh (2018)	Lapponi et al. (2018)
Linguistic analysis	Sokolova & Lapalme (2008)		
Political analysis		Baturo et al. (2017); Diermeier et al. (2012); Frid-Nielsen (2018) Jensen et al. (2012); Owen (2017); Proksch & Slapin (2010); Rheault et al. (2016); Schwarz, Traber & Benoit (2017)	
System/method performance evaluation	Abercrombie & Batista-Navarro (2018a); Abercrombie & Batista-Navarro (2018b); Akhmedova, Semenkin, & Stanovov (2018); Ahmadalinezhad & Makrehchi (2018); Allison (2018); Balahur,Kozareva & Montoyo (2009); Bansal, Cardie & Lee (2008); Bhatia & P (2018); Bonica (2016); Budhwar, Kuboi, Dekhtyar & Khosmood (2018); Burfoot (2008); Burfoot, Bird & Baldwin (2011); Burfood et al. (2015); Chen et al. (2017); Duthie & Budzynska (2018); Dzieciątko (2018); Glavaš et al. (2017); Ji & Smith (2017); Lefait & Kechadi (2010); Li et al. (2017); Martineau et al. (2009); Naderi & Hirst (2016); Onyimadu, Nakata, Wilson, Macken & Liu (2013); Plantié, Roche, Dray & Poncelet (2008); Salah (2014); Salah, Coenen & Grossi (2013a); Salah, Coenen & Grossi (2013b); Thomas et al. (2006); Vilares & He (2017); Yessenalina, Yue & Cardie (2010); Yogatama & Smith (2014b); Yogatama & Smith (2014a); Yogatama et al. (2015)	Frid-Nielsen (2018); Hopkins & King (2009); Kim, Londregan & Ratkovic (2018); Kapočiūtė-Dzikienė & Krupavičius (2014); Laver et al. (2003); Lowe & Benoit (2013); Monroe et al. (2017); Proksch et al. (2018); Rauh (2018); Rheault (2016); Schwarz et al. (2017); Taddy (2013)	Honkela et al. (2014); Iyyer et al. (2014); Kauffman, Khosmood, Kuboi & Dekhtyar (2018); Nguyen et al. (2013); Rheault et al. (2016); Rudkovsky et al. (2018); Rudkovsky et al. (2018); Sakamoto & Takikawa (2017); Van der Zwaan et al. (2016)

While the production of corpora and datasets are among the secondary contributions of many of the featured papers (e.g., Abercrombie & Batista-Navarro, 2018b; Salah, 2014; Thomas et al., 2006), in the cases of Lapponi et al. (2018) (linguistically annotated corpus) and Rauh (2018) (sentiment lexicon), this is their principal objective.

Hopkins & King (2009) claim that a fundamental difference between the objectives of computer scientists and political or social scientists is that, while the former are interested in making predictions about latent characteristics of individuals documents (such as sentiment polarity), the latter are more concerned with characterizing corpora or collections of documents as a whole, for example by the *proportion* of positive or negative examples it contains. This point is supported by Monroe et al. (2017), who agree that individual document classification is "inappropriate to the task", because, they suggest, under this model, representation of the whole data generation process is backwards—where classification presumes that class labels are manifestations of the underlying latent phenomena of interest, the reality is the other way around: people first hold opinions or positions, and subsequently express them in speech or writing.

Despite this dichotomy, as we see in the next section, we do find cases of political scientists tackling classification (Proksch & Slapin, 2010) or computer scientists undertaking the scaling task from political science (Glavaš et al., 2017).

Tasks

We find that the following eight categories of sentiment and position analysis task are performed in the studies:

- Agreement and alignment detection: analysis of the similarity of the position taken by a speaker and another entity (another speaker, or a person or organisation outside of the debate in question) (n = 3).
- Argument mining: tasks concerned with the identification of the speaker's differing positions towards different targets such as arguments or other speakers, including frame and ethos analysis: (n = 2).
- Emotion analysis: including emotion, anxiety and wellbeing analysis (n = 3).
- Ideology and party affiliation detection: in the literature, a speaker's party affiliation is often used as a proxy for their ideological position. This may be performed as either topic modeling or classification (n = 14).
- Opinion-topic analysis: attempts to simultaneously extract topics and the speakers' positions towards them. (n = 5)

- Polarization analysis: analysis of the extent to which debate in a legislature is polarized and speakers are ideologically divided, according to the positions expressed (n = 2).
- Position scaling: positioning of speakers or parties on a scale of one or more dimensions (n = 11).
- Sentiment/opinion polarity classification: binary or ternary analysis. As votes are frequently used as opinion polarity labels, this includes the task of vote prediction (n = 28).

By far the most frequently occuring task undertaken is sentiment or opinion polarity classification (although this is not always named as such). In the majority of cases this takes the form of learning from speeches the predictive features of speakers' votes (e.g., Salah, 2014) or manually annotated ground-truth labels (e.g., Onyimadu et al., 2013). Polarity classification is particularly prevalent in the computer science studies (24 out of 29), but despite the previously discussed claims of Hopkins & King (2009) and Monroe et al. (2017) that the task is incompatible with the aims of social scientists, some political scientists and multi-disciplinary teams also tackle this task (Hopkins & King, 2009; Proksch et al., 2018; Rudkovsky et al., 2018).

As all the tasks undertaken concern the analysis of the positions taken by debate participants, there is considerable overlap between them. Furthermore, there is sometimes some discrepancy between the name given to a task and the actual task performed. For example, Onyimadu et al. (2013) refer to the task they perform as both "sentiment analysis" and "stance detection", although they actually carry out only sentiment polarity classification as they do not specify a pre-chosen target, a requirement of stance detection (as defined by Mohammad, Sobhani & Kiritchenko, 2017). Meanwhile, Thomas et al. (2006) refer to this task as "predicting support", and Allison (2018), working on the same problem and the same dataset, call it variously sentiment "detection" and "classification". Although Rheault et al. (2016) consider their work to be a form of emotion detection, they actually perform a form of sentiment polarity analysis at the whole legislature level, while Akhmedovaet et al. (2018) simply refer to the task as an "opinion mining problem". Other terms used to refer to this include "attitude

detection" (Salah et al., 2013a), "vote prediction" (Budhwar et al., 2018), "emotional polarity" measurement, "predicting the polarity of a piece of text" (Yogatama & Smith, 2014b), "sentiment classification" (Yessalina et al., 2010; Yogatama et al., 2015), and simply "sentiment analysis" (Proksch & Slapin, 2010; Rauh, 2018; Rudkovsky et al., 2018; Salah, 2014; Yogatama & Smith, 2014a).

In some cases, more than one task is investigated. For example, by switching party labels for vote labels, Burfoot (2008) use the same method to perform both sentiment polarity and party affiliation (or ideology) detection. Sentiment polarity analysis is often used as part of an NLP pipeline as a subtask of a different opinion mining task, such as agreement detection (Salah et al., 2013a). Similarly, Kauffman et al. (2018) use sentiment analysis as a sub-task and the output scores as features for allignment detection, while Duthie & Budzynska (2018) do similar for ethos detection, and Budhwar et al. (2018) aim to predict vote outcome using the results of sentiment polarity analysis as features for the task. Conversely, Burfoot (2008) applies the results of classification by party affiliation to predict speaker sentiment.

Balahur et al. (2009) combine polarity with party classification, a task that we consider to be a form of ideology detection, but which they name "source classification". Indeed, this is another task that suffers from a lack of clarity over terminology, with some studies considering party affiliation to be a proxy for ideology (Diermeier et al., 2012; Jensen et al., 2012; Kapočiūtė-Dzikienė & Krupavičius, 2014; Taddy, 2013), while others do not make this connection, extracting information about speakers' ideologies from their sentiment towards different topics (Bhatia & P, 2018; Chen et al., 2017; Nguyen et al., 2013), or training a model on examples that have been explicitly labelled by ideology, and not party membership (Iyyer et al., 2014). Yet others perform party classification, making no mention of the relationship between party membership and ideology (Balahur et al., 2009; Burfoot, 2008; Lapponi et al., 2018; Lefait & Kechadi, 2010). Alternatively, Abercrombie & Batista-Navarro (2018a) explicitly assume that membership of the same party does not guarantee homogeneity of ideologies, investigating intra-party differences of opinion and positions. Also concerned with ideology, position scaling, which we code as a separate task, can be performed on different dimensions, one of the most common being the left- to right-wing scale.

The literature contains several efforts to simultaneously extract topics and speakers' attitudes towards them (opinion-topic analysis). A common approach here is to combine topic-modeling with forms of stance detection. Nguyen et al. (2015) use a supervised form of hierarchical Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003) to extract topics and polarity variables. Van der Zwaan (2016) generate separate topic models for different grammatical categories of words in efforts to obtain this information. And Nguyen et al. (2013) perform supervised topic modeling to capture ideological perspectives on issues to produce coarse-grained speaker ideology analysis. Topic modeling as also undertaken by Sakamoto & Takikawa (2017), who use it to analyze polarization, a task also tackled by Jensen et al. (2012). Meanwhile, Vilares & He (2017) also perform opinion-topic modeling to extract speakers' perspectives—"the arguments behind the person's position"—on different topics.

A number of other tasks which fit under the broader definition of sentiment analysis are also tackled. Polarization analysis is undertaken in both Jensen et al. (2012) and Sakamoto & Takikawa (2017), who investigate changes in the extent to which language in Congress is polarized over time. Meanwhile, emotion detection is the subject of three studies. Dzieciątko (2018) classifies speakers as expressing *happiness*, *anger*, *sadness*, *fear*, or *disgust*, while Rheault (2016) attempts to identify the level of anxiety exhibited by speakers, and Honkela et al. (2014) analyse a corpus of congressional speeches under the PERMA (*Positive emotion*, *Engagement*, *Relationships*, *Meaning*, and *Achievement*) model.[2] Agreement detection, an end in itself for Ahmadalinezhad & Makrehchi (2018) and Kauffman et al. (2018), is used by Burfoot (2008) Burfoot et al. (2011) and Burfood et al. (2015) to predict speaker sentiment, while Salah et al. (2013a) use agreement information to construct debate graphs. Finally, Naderi & Hirst (2016) automatically compare speeches with another type of labelled text (statements from online debates) to identify positive and negative framing of arguments.

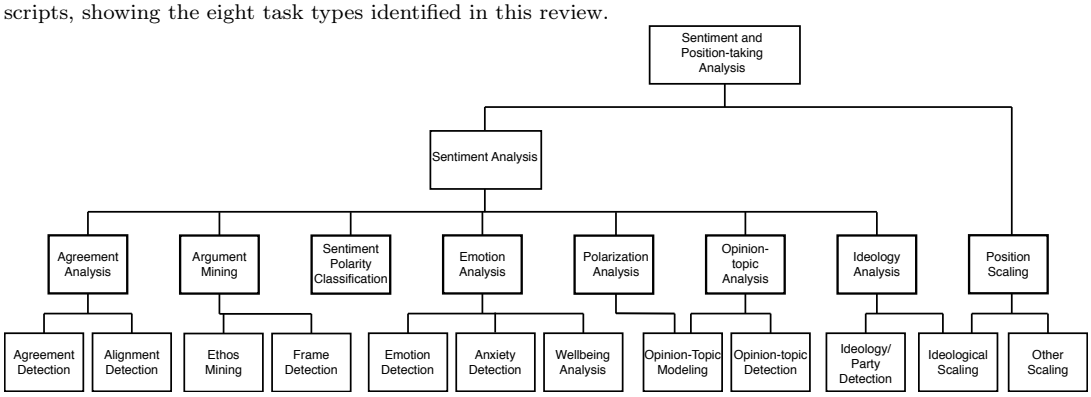
Although there exist exceptions (see above), a notable difference between the focus of tasks undertaken by NLP researchers and social scientists is that the former tend to perform analysis with regard to the target of expressed sentiment (a specific proposal (e.g., Allison, 2018), piece of legislation (e.g., Thomas et al., 2006), topic

[2]Seligman (2012).

(e.g., Van der Zwaan et al., 2016), or other entity), while the latter generally analyse speakers' aggregated speeches, ignoring the targets of individual contributions, and instead attempting to project actors onto a scale (such as *left-right*) (Iliev, Huang & Gel, 2019; Kim et al., 2018; Laver et al., 2003; Lowe & Benoit, 2013; Proksch & Slapin, 2010; Schwarz et al., 2017). Grimmer & Stewart (2013) note that this can be problematic as manual "validation is needed to confirm that the intended space has been identified", and suggest automatic detection of relevant "ideological statements" (or opinion-targets) as an important challenge.

For a typology of identified tasks in this domain, see Figure 3.

Figure 4. Typology of sentiment and position-taking analysis tasks performed on legislative debate transcripts, showing the eight task types identified in this review.



Approaches

We consider the granularity (level of analysis), methods, features, and ground truth labels used (in the cases of both supervised learning methods and evaluation of other methods) for each study.

Granularity

There are a number of approaches to segmenting the transcripts for analysis, ranging from breaking them down to the sentence or sub-sentence level to aggregating sentiment over entire corpora.

The vast majority of studies conduct analysis at the speech level (n = 39). However "speech" appears to mean different things to different researchers, and in some studies

it is not immediately clear just what the unit of analysis actually is.[3] In some, a speech is considered to be the concatenated utterances of each individual speaker in each debate (n = 16). For others (n = 24), analysis is conducted at the utterance or "speech segment" level (that is, an unbroken passage of speech by the same speaker), although Akhmedovaet et al. (2018) refer to these as "interventions", and Bansal et al. (2008) as "individual conversational turns". While several researchers who use Thomas et al. (2006)'s *ConVote* corpus claim to analyse "speeches", the dataset (usually used unaltered) is in fact labelled at the speech segment level. Similar use of terminology can be found in other work, such as Vilares & He (2017).

A further eight papers report analysis at the speaker level. That is, they consider a document to be the concatenation of all speeches given by the same representative (Bonica, 2016; Diermeier et al., 2012; Kauffman et al., 2018; Kim et al., 2018; Owen, 2017; Schwarz et al., 2017; Taddy, 2013).

Other approaches are to analyse speeches at the coarser political party (or bloc/coalition) level (Frid-Nielsen, 2018; Glavaš et al., 2017; Proksch & Slapin, 2010; Proksch et al., 2018; Sakamoto & Takikawa, 2017; Van der Zwaan et al., 2016), or the finer sentence (Duthie & Budzynska, 2018; Naderi & Hirst, 2016; Onyimadu et al., 2013; Rauh, 2018) or phrase (Jensen et al., 2012) levels. Although Rudkovsky et al. (2018) detect sentiment in sentences, they aggragate these scores to provide speech-level results. Iyyer et al. (2014) break speeches down to both these levels, while Naderi & Hirst (2016) do so for sentences and paragraphs.

At the highest possible level of granularity, four studies consider sentiment over entire corpora. Dzieciątko (2018) and Rheault et al. (2016) aggregate sentiment scores for all speeches, presenting analysis of the Polish and UK parliaments respectively. Meanwhile, Honkela et al. (2014) compare the overall sentiment of the European Parliament transcripts with other corpora at the whole dataset level, and, in addition to party-level analysis, Sakamoto & Takikawa (2017) compare the polarity of Japanese and US datasets.

Finally, Ahmadalinezhad & Makrehchi (2018) consider each document to be a 'conversation between two individuals'—that is, their combined utterances—in order

[3]In several cases, it has been necessary to contact the authors for clarification or to manually examine the datasets used to obtain this information.

to classify these as being either in agreement or disagreement.

Overall, computer scientists tend to work at finer-grained levels (speech, speech segment, paragraph, sentence, or phrase), while in political science, the preferred unit of analysis would seem to be the party, which is the target of most work on position scaling, a task very much associated with that field. This confirms to some extent the assertation of Hopkins & King (2009) that, while computer scientists are "interested in finding the needle in the haystack, ... social scientists are more commonly interested in characterizing the haystack". Exceptions, from the political and social sciences, are Iliev et al. (2019), and Hopkins & King (2009)—who actually propose a method of optimizing speech-level classification for social science goals, and from computer science, Glavaš et al. (2017), who also tackle the position scaling problem.

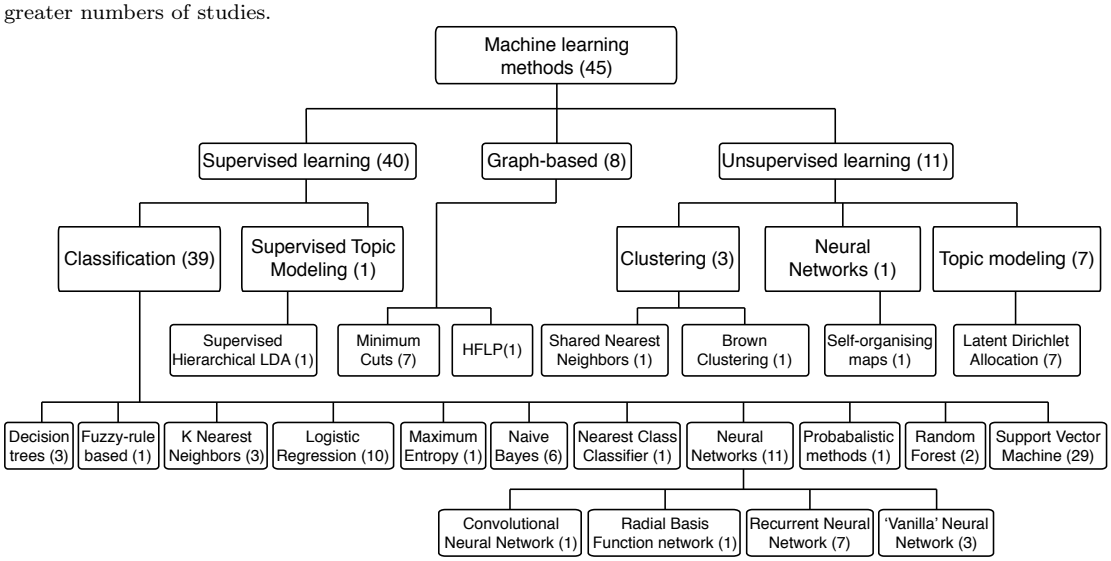
Methods

A wide range of methods are used, but these can be grouped into five main approaches: dictionary-based (n = 16), machine learning (n = 45), rule-based (n = 2), similarity measurement (n = 9), and word frequency analysis methods (n = 10).

Overall, political and social scientists tend to use statistical scaling approaches based on word counts, although a few make use of machine learning approaches (e.g., Diermeier et al., 2012; Hopkins & King, 2009; Rheault, 2016). The scaling task comes in two varieties: supervised, in which target speeches ("virgin texts") are compared to reference texts, and unsupervised, such as the Wordfish package introduced by Proksch & Slapin (2010), and used by Schwarz et al. (2017). Glavaš et al. (2017), in the only study conducted from an NLP perspective that takes on the position scaling problem so favoured by political scientists, use a combination of semantic similarity measurement and harmonic function label propagation, a semi-supervised graph-based machine learning algorithm.

In total, roughly three quarters of included studies (n = 45) make some use of machine learning, and within this area there are a multitude of different approaches (see Figure 4). These can be broadly categorised as supervised learning (n = 40), semi-supervised (n = 1), or semi-supervised (n = 11) methods. Supervised methods are the preferred approach for text classification, and a wide variety of algorithms are

Figure 5. Machine learning methods used for sentiment and position analysis. Darker colours signify use in greater numbers of studies.



used, including logistic regression (n = 10), naive Bayes (n = 6), decision trees (n = 3), nearest neigbor (n = 3), boosting (n = 1), a fuzzy rule-based classifier (n = 1), maximum entropy (n = 1), nearest class classification (n = 1). A further eleven studies make use of neural networks, which range in complexity from "vanilla" feed-forward networks such as the multi-layer perceptron to convolutional and recurrent neural networks, including the use of long short-term memory units (LSTMs) (see Figure 5). Of these, six are concerned with sentiment polarity analysis and four with ideology detection, which as previously discussed, are highly similar classification tasks, performed with different class labels. The exception is Duthie & Budzynska (2018), who use recurrent neural networks modules in their ethos mining task.

Rather than testing different classification algorithms, some work focuses on the use of different regularization methods (for logistic regression) (Yogatama & Smith, 2014a,b; Yogatama et al., 2015), while other approaches to improving performance of classifiers include Boosting (Budhwar et al., 2018), and Collective Classification (Burfoot et al., 2011).

Much of the work on unsupervised learning focuses on use of topic modeling methods, the majority of which are variations of LDA. The cross-perspective topic-model of Van der Zwaan et al. (2016), for example generates two topic models from

the data: one over nouns to derive topics, and the other over adjectives, verbs and adverbs, which is intended to produce opinion information.

Finally, several studies model debates as networks of connected speakers and employ graph-based methods (n = 7). Bansal et al. (2008), Burfoot (2008), Burfoot et al. (2011), Burfood et al. (2015), and Thomas et al. (2006) all approach sentiment polarity classification as a *minimum cuts* graph partition problem. Chen et al. (2017) construct an "opinion-aware knowledge graph", propagating known ideology information through the graph to infer opinions held by actor nodes towards entity nodes. For position scaling, Glavaš et al. (2017) use similarity measurements as edges between documents, and propogate the scores of known "pivot texts" to the other nodes.

Language models and feature selection

Although analysis of debate transcripts necessarily utilises textual features derived from the speeches, there are a variety of approaches to how this is modelled, and which types of features are selected. In terms of language modeling, the majority of studies represent text as bags of words.[4] However, some add contextual information with use of word embeddings. While their use is generally restricted to studies from computer science, Rudkovsky et al. (2018) (multi-disciplinary) and Glavaš et al. (2017) (computer science) explore their use for aims normally associated with the social sciences. Relatively little linguistic or structural analysis is undertaken. Exceptions re Ji & Smith (2017), who find that the application of Rhetorical Structure Theory does not produce good results in this domain because its "discourse structure diverges from that of news", and Balahur et al. (2009), who extract parse trees to determine the target of speech sentiment.

In addition to textual information, many studies also make use of metadata features of the speakers and the debates themselves, as well as other features such as those derived from the structure of the debates. The following six categories of features are used:

- Discourse features: including citations (Burfoot, 2008; Burfoot et al., 2011; Lefait & Kechadi, 2010), interruptions and speech length (Budhwar et al., 2018), n-grams from other nearby sentences (Yessalina et al., 2010), and utterance statis-

[4]That is, unstructured, unordered arrays of n-gram (term) counts.

tics (number and duration of speaker's utterances) (Kauffman et al., 2018).

- Metadata features: including bill authorship (Budhwar et al., 2018), debate and speaker IDs (Abercrombie & Batista-Navarro, 2018a; Salah, 2014), debate type (Lapponi et al., 2018), donations (Bonica, 2016; Kauffman et al., 2018), geographic provenance (Lapponi et al., 2018), party affiliation (Abercrombie & Batista-Navarro, 2018a,b; Burfoot, 2008; Kauffman et al., 2018; Rudkovsky et al., 2018; Sakamoto & Takikawa, 2017; Salah, 2014), and speaker gender (Lapponi et al., 2018).
- Polarity scores: the output of opinion polarity classification used as a feature for prediction of another phenomena (such as sentiment polarity (Bhatia & P, 2018), voting intention (Budhwar et al., 2018), argument frame Naderi & Hirst (2016) or alignment with another entity (Kauffman et al., 2018)).
- Relational and knowledge features: features based on the relationships between different speeches or speakers, speakers and other entitities (such as the targets of expressed opinion), or the measured similarity between speeches (Burfoot, 2008; Burfoot et al., 2011; Burfood et al., 2015; Chen et al., 2017; Thomas et al., 2006).
- Textual: including word (all studies) or character (Kapočiūtė-Dzikienė & Krupavičius, 2014) n-grams, punctuation, embeddings, custom dictionary keyword features (Budhwar et al., 2018), words from particular grammatical categories (Iyyer et al., 2014; Kapočiūtė-Dzikienė & Krupavičius, 2014; Lapponi et al., 2018; Monroe et al., 2017; Naderi & Hirst, 2016; Onyimadu et al., 2013; Sokolova & Lapalme, 2008; Van der Zwaan et al., 2016), presence of questions (Budhwar et al., 2018), word embeddings (Bhatia & P, 2018; Glavaš et al., 2017; Iyyer et al., 2014; Ji & Smith, 2017; Li et al., 2017; Naderi & Hirst, 2016; Rheault, 2016; Rheault et al., 2016) and sentence (Rudkovsky et al., 2018) embeddings, and parse trees (Balahur et al., 2009; Iyyer et al., 2014; Ji & Smith, 2017).
- Speaker vote: used as a feature to identify speaker ideology (Kim et al., 2018; Schwarz et al., 2017).

A key decision for researchers is that of whether or not to include non-textual, metadata features, and the answer to this is usually driven by their research objectives.

In some studies, particularly those from political science focused on position scaling, the object may be to examine intraparty differences or to compare speech and vote behaviour, in which cases features such as party affiliation or vote are the dependent variable under observation, and cannot be used as features for analysis. For classification, while some researchers compare performance with and without this additional information (e.g., Abercrombie & Batista-Navarro, 2018a,b; Salah et al., 2013a), others prefer to exclude them entirely in order to make their methods more generalizable to debates from other domains such as online debates, which do not have access to such information (Burfoot, 2008; Thomas et al., 2006).

Ground-truth labels

Depending on the nature of the task being tackled, for supervised classification methods, and in some cases, validation of unsupervised methods, several different data sources are used to represent the ground-truth sentiment or opinion. In most cases, researchers opt to make use of some pre-existing data, with the most common being the speakers' roll-call or division[5] votes (n = 21). Here, the most common approach is to consider each speaker's vote on a given debate to represent the ground truth position taken in their speeches, which may be analysed at the whole (concatenated) speech level (as in Salah (2014)) or broken down into smaller units (as in Thomas et al. (2006)), whereby each vote label is attached to multiple examples. A general difference in approach is that, while computer science studies use these as ground truth, political scientists tends to view speech and vote as unconnected, and explicitly compare the two on this basis (Schwarz et al., 2017). Lowe & Benoit (2013) use human annotations for validation of the output of their scaling method. Whether or not votes are actually reliable as ground-truth is a matter of contention. Although some computer science studies assume this to be the case (e.g., Salah et al., 2013b; Salah, 2014; Thomas et al., 2006), Schwarz et al. (2017) compare speeches to votes for scaling, producing quite different results, and Abercrombie & Batista-Navarro (2018a) who compare votes with human produced labels, conclude that it is the latter which more closely reflect sentiments expressed by speakers.

[5]Terms used in the US Congress and Westminster system parliaments, respectively.

An alternative approach, used in six studies, is to use manually annotated labels. While some researchers make use of already existing expert annotations, such as the Chapel Hill expert surveys[6] (Glavaš et al., 2017; Proksch & Slapin, 2010), others produce labelled datasets specifically for their purposes. Onyimadu et al. (2013) and Rauh (2018) had in-house annotators label speech sentences as being *positive*, *negative*, or *neutral*, while Rheault (2016) used crowd-sourced coders to label sentences as *anxious* or *non-anxious*. Rudkovsky et al. (2018) also use crowd-sourced labels, but for evaluation rather that training purposes, in their case to assess negativity detection. To create labels for the validation of their scaling of speakers' positions towards a given topic, Frid-Nielsen (2018) have experts follow the coding scheme of the Comparative Manifestos Project[7] to produce policy postion labels, although the reliability of these is also controversial (Mikhaylov, Laver & Benoit, 2008).

Other data used as ground-truth labels are the speakers' DW-nominate scores (scores derived from congressional legislators' voting records[8]) (Diermeier et al., 2012; Nguyen et al., 2013), their constituency vote shares (Taddy, 2013), "issue" labels from the Library of Congress's Congressional Research Service[9] (Bonica, 2016), word perplexity (Van der Zwaan et al., 2016), sentiment analysis scores obtained from prior experiments on the same data (Sokolova & Lapalme, 2008), and party affiliations (n = 13). While the latter are widely used as a proxy for speaker ideology (Bhatia & P, 2018; Jensen et al., 2012; Li et al., 2017), Hirst, Riabinin & Graham (2010) suggest that party membership is actually a confounding factor for this task.

Performance and outcomes

With the research reviewed here having such varied objectives and undertaking many different analysis tasks, it is not possible to directly compare the reported performances of the methods proposed. Nevertheless, in this section we attempt to summarize some conclusions of the included studies that are potentially relevant to future work in this area.

[6]https://www.chesdata.eu/
[7]https://manifesto-project.wzb.eu/
[8]https://legacy.voteview.com/dwnomin.htm
[9]https://www.loc.gov/crsinfo/

For classification, machine learning methods, and particularly neural networks, seem to outperform other approaches. Here, just as in other domains, such as product reviews, dictionary-based sentiment analysis methods appear to have been superceded by machine learning approaches. In a direct comparison, Salah (2014) found that machine learning classification methods outperform those utilising both generic and parliament-specific lexica, while Balahur et al. (2009) improved lexicon-based performance with the addition of a support vector machine classifier. Given this, and also considering the conclusion of Allison (2018) that "classifier choice plays at least as important a role as feature choice", which learning algorithms should be selected for classification in this domain? In the work reviewed here, support vector machines, used in 29 of the studies, are the most popular option—both as a common baseline, and as a default algorithm choice. Although, in NLP in general, the last decade has seen an explosion in interest in deep learning methods, here we see relatively little use of neural network-based machine learning. Those studies that directly compare the performance of such methods with other classifiers, tend to report better performance using neural networks (Abercrombie & Batista-Navarro, 2018a; Budhwar et al., 2018; Iyyer et al., 2014; Li et al., 2017).

For position scaling, political and social scientists do not tend to place the same emphasis on performance metrics such as accuracy, preferring to make comparisons between output manual analyses in order to investigate theory-based hypotheses. Indeed, discussion of technical performance in these papers often focuses on whether or not computational text analysis is valid at all when compared with expert examination. In this respect, Diermeier et al. (2012); Frid-Nielsen (2018); Laver et al. (2003) conclude that, with some caveats, it is (Lowe & Benoit (2013) note that their method appears to position some speakers on a different dimension to that of their expert analysis). In the one computer science paper to tackle this problem, Glavaš et al. (2017) report equally promising results on mono- and multilingual data, as well as superior performance using word embeddings over a bag of words model.

In the reviewed studies, a large range of feature types are extracted from the transcripts. Most studies rely primarily on the bag-of-words model, and for textual features, the benefits of adding higher-order n-grams (bi-, tri-grams, etc.) appear in-

conclusive. While Plantié et al. (2008) report improved performance with the addition of bigrams to their feature set, Abercrombie & Batista-Navarro (2018a) do not see significant improvement with the use of bi- and tri-grams. With the most common method of n-gram feature selection being TF-IDF weighting, Martineau et al. (2009), noting that IDF favours rare features, find that, for the relatively homogenous domain of a particular parliament's transcripts, their alternative Delta TF-IDF leads to better classification performance.

As we have seen, the appropriateness of using metadata features depends on the objectives of the research. However, if optimal classification performance is the goal and information regarding the speakers' party affiliations is available, this has been found to be highly predictive of expressed sentiment (Abercrombie & Batista-Navarro, 2018a; Salah, 2014). Inter-document relationship information regarding agreement between speakers also assists in sentiment polarity classification, and has been applied successfully by Bansal et al. (2008), Burfoot (2008), and Thomas et al. (2006), as has network information (Burfoot et al., 2011; Burfood et al., 2015). The latter show that is possible to model these relationships for any dataset using n-gram overlap. In another approach to modeling debate structure, Balahur et al. (2009) use dependency parsing to find targets, which seems to improve classification and helps to balance results obtained in the positive and negative classes. While Iyyer et al. (2014) also report success in using parse trees as features for classification with a recurring neural network, Ji & Smith (2017) do not find improvement in the parliamentary domain (although they do in news articles).

When it comes to representing ground-truth, votes are not necessarily indicative of the opinions expressed in speeches, but for speech-level polarity analysis they can be a convenient option. The results of computational analysis by Schwarz et al. (2017) supports manual analysis in political science (Proksch & Slapin, 2015) to indicate that representatives position themselves differently in their speeches than in their voting behaviour. However, the relatively small difference between votes and manual annotations (less than four per cent of their corpus) found by Abercrombie & Batista-Navarro (2018a), suggests that relatively small gains are to be had by investing in human labeling where other forms of class label are available.

A number of observations arise about the use of language in this domain. For the UK Parliament, Onyimadu et al. (2013) find that "compound opinions", sarcasm, and comparative structures are all confounding elements for classifiers. In German, Rauh (2018) notes that "positive language is easier to detect than negative language", while Salah et al. (2013a,b) make a similar observation for the UK Hansard. The latter study explains this phenomena as an artifact of the "polite parliamentary jargon" used in Parliament. This point is also backed up by Abercrombie & Batista-Navarro (2018a), who observe that, the most indicitive features, even of negative polarity, are words not typically thought of as conveying negativity. Where negative adjectives and verbs *are* present, Sokolova & Lapalme (2008) find that these are highly discriminative features.

Discussion and conclusion

Considering the nature of the problem at hand—computational methods for the analysis of political text—it is somewhat surprising how little crossover can be found in this domain between ideas from computer science and political science, and how seldom the methods used by researchers from these different fields are adopted by researchers from the other disciplines. As an explanation for this, Hopkins & King (2009), Monroe et al. (2017), Lowe & Benoit (2013) provide insights into the differing aims of the two fields. However, despite these differences, researchers in computer science may well be able to benefit from the theoretical expertise of political and social scientists, such as the rigourous labelling schema and expertly coded corpora already existing in the field. Similarly, more political and social scientists could consider going beyond the simple bag-of-words n-gram language models they currently rely on to investigate the use of more advanced NLP methods of representing text and handling feature sparsity in natural language, such as word embeddings.

A problematic issue that arises from surveying the work included in this review is the wildly inconsistent use of terminology, even just from within each of the fields represented here. There is a clear need for greater agreement on which terms to use to refer to the affective targets of interest and the names of the tasks designed to analyse them, as well as the varying levels at which analysis is performed. These inconsistencies

often mean that it is difficult—or even, without further investigation, impossible—for the reader to understand just what is done in a given study.

Studies included in this review approach analysis of legislative transcripts at a wide variety of granularities, from the phrase-level to comparisons aggregating sentiment over entire corpora. However, for the sake of convenience, and in order to make use of existing labels such as votes, the majority conduct analysis at the speech level, or even if they do so at a more fine-grained sentence or phrase level, they tend not to consider the discourse structure of the debates. As Burfoot (2008) points out, parliamentary and legislative debates are complex, with many topics discussed and sentiment directed towards varying targets in ways that a document level classifier can struggle to identify. There is therfore room to develop more complex analyses, capable of recognising the relationships between entities and targets in fine-grained sections of the transcripts, perhaps using argument mining methods that harnesses theories from fields such as communication theory (e.g., Naderi & Hirst, 2016) or even philosophy (e.g., Duthie & Budzynska, 2018) in order to explore the relationships between actors, opinion, targets and other entities in debates.

There have also been few attempts to link expressed opinion with topic information. While there are some efforts to do so from debate motions (Abercrombie & Batista-Navarro, 2018b), at the political party level (Van der Zwaan et al., 2016), and as as a form of perspective analysis (Vilares & He, 2017), as well as by scaling on pre-defined topic dimensions (Owen, 2017), the majority of studies simply conduct analysis of sentiment towards a target, such as a Bill or motion, the topic of which is unknown. In order to provide truly useful information, it may make sense to focus efforts on the extraction of topic-centric opinions and to conduct analysis at the level at which different topics are found in the data.

While the majority of studies that focus on supervised classification rely on votes as ground-truth labels, it is debatable whether these actually represent the target phenomena—the opinion or position taken by the speaker. Manual analysis in political science (Proksch & Slapin, 2015) certainly suggests that, in many legislatures, representatives express different positions in speech than in their votes, a point supported by Schwarz et al. (2017), who compare the scaling of speeches and votes. However,

as Abercrombie & Batista-Navarro (2018a) point out, gains made from seeking more reliable ground-truth may be small and not worth the associated costs in a practical setting. An alternative approach, for which there is still plenty of scope for further research, is to develop semi-supervised or unsupervised approaches, which require few or no labels.

There also exist other possible directions for future work. With an increasing quantity of transcript data becoming available, in the case of some parliaments stretching back over hundreds of years, one such possibility is the analysis of language change over time. While Diermeier et al. (2012) suggest using changes in classification performance to infer changes in agenda control, and Kapočiūtė-Dzikienė & Krupavičius (2014) found that performance worsened when transcripts from different sessions were used for training and testing, language drift in parliamentary debates remains relatively unexplored. Although Ahmadalinezhad & Makrehchi (2018), note a performance drop when training and testing on different debate data (Canadian Hansard and US election debates), domain adaptation and inter-legislature transfer learning also remains under-explored. Additionally, given the successes achieved with neural networks and deep learning in other domains, as well as the results reported by studies that use such methods, there would appear to be considerable scope for further investigation of their application to legislative debates. Finally, while some of the included studies mention potential applications of their work for civic technology (e.g., Budhwar et al., 2018; Burfoot, 2008), with the exception of Bonica (2016)'s CrowdPac Voter Guide, as far as we know, the methods used are not currently being applied to any real-world systems. There is therefore room to explore the application of the approaches used to the area of civic technology, and provide tools that could genuinely assist people in processing information about their elected representatives.

The computational analysis of sentiment and position-taking in parliamentary debate transcripts is an area of growing interest. While the researchers working on this problem have varied backgrounds and objectives, in this review we have identified some of the common challenges they face. With the majority of work emanating from computer science focusing on unknown targets (Bills or debate motions, the topic of which is not assessed), and political scaling being conducted on very coarse grained

scales (*left-right*, *pro/anti-EU*), there has thus far been little effort to direct efforts towards examining the targets of the opinions expressed. For the aims of both political scholarship and civic technology, what is required in many cases is identification of these targets, namely the policies and policy preferences that are discussed in the legislative chambers. It is our belief, therefore, that future work should be directed towards such target-specific analyses.

Acknowledgements

The authors would like to thank Federico Nanni for his helpful comments and suggestions.

References

Abercrombie, G. & Batista-Navarro, R. (2018a). 'Aye' or 'no'? Speech-level sentiment analysis of Hansard UK parliamentary debate transcripts. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Abercrombie, G., & Batista-Navarro, R. (2018b). Identifying opinion-topics and polarity of parliamentary debate motions. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 280–285.

Ahmadalinezhad, M., & and Makrehchi, M. (2018). Detecting agreement and disagreement in political debates. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 54–60.

Akhmedova, S., Semenkin, E. & Stanovov, V. (2018). Co-operation of biology related algorithms for solving opinion mining problems by using different term weighting schemes. *Informatics in Control, Automation and Robotics*, 73–90.

Allison, B. (2018). Sentiment detection using lexically-based classifiers. *International Conference on Text, Speech and Dialogue*, 21–28.

Balahur, A., Kozareva, Z. & Montoyo, A. (2009). Determining the polarity and source of opinions expressed in political debates. *International Conference on Intelligent Text Processing and Computational Linguistics*, 468–480.

Bansal, M., Cardie, C. & Lee, L. (2008). The power of negative thinking: Exploiting label

disagreement in the min-cut classification framework. *COLING 2008: Companion Volume: Posters*, 15–18.

Baturo, A., Dasandi, N. & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: introducing the UN General Debate Corpus. *Research & Politics*, *4*(2).

Bhatia, S. & P, D. (2018). Topic-specific sentiment analysis can help identify political ideology. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 79–84.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research*, *3*, Jan, 993–1022.

Boland, A., Cherry, G. & Dickson, R. (2017). *Doing a Systematic Review: A Student's Guide.* London: Sage.

Bonica, A. (2016). A data-driven voter guide for U.S. elections: adapting quantitative measures of the preferences and priorities of political elites to help voters learn about Candidates. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *7*(2), 11–32.

Budhwar, A., Kuboi, T., Dekhtyar, A. & Khosmood, F. (2018). Predicting the vote using legislative speech. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 35.

Burfoot, C. (2008). Using multiple sources of agreement information for sentiment classification of political transcripts. *Proceedings of the Australasian Language Technology Association Workshop 2008*, 11–18.

Burfoot, C., Bird, S. & Baldwin, T. (2011). Collective classification of congressional floor-debate transcripts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 1506–1515.

Burford, C., Bird, S. & Baldwin, T. (2015). Collective document classification with implicit inter-document semantic relationships. *Proceedings of the fourth joint conference on lexical and computational semantics*, 106–116.

Chen, W., Zhang, X., Wang, T., Yang, B. & Li, Y. (2017). Opinion-aware knowledge graph for political ideology detection. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3647–3653.

Diermeier, D., Godbout, J.-F., Yu, B. & Kaufmann, S. (2012). Language and ideology in Congress. *British Journal of Political Science*, *42*(1), 31–55.

Duthie, R. & Budzynska, K. (2018). A deep modular RNN approach for ethos mining. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence,*

(IJCAI-18), 4041–4047.

Dzieciątko, M. (2018). Application of text analytics to analyze emotions in the speeches. *Information Technology in Biomedicine*, 525–536.

Fang, X. & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1).

Frid-Nielsen, S. S. (2018). Human rights or security? Positions on asylum in European Parliament speeches. *European Union Politics*, *19*(2), 344–362.

Glavaš, G., Nanni, F. & Ponzetto, S. P. (2017). Unsupervised cross-lingual scaling of political texts. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, *2*, 688–693.

Grijzenhout, S., Jijkoun, V. & Marx, M. (2010). Opinion mining in dutch hansards. *Proceedings of the Workshop From Text to Political Positions, Free University of Amsterdam*.

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

Hirst, G., Riabinin, Y. & Graham, J. (2010). Party status as a confound in the automatic classification of political speech by ideology (2010). *Proceedings, 10th International Conference on Statistical Analysis of Textual Data/10es Journées internationales dâĂŹAnalyse statistique des Données Textuelles (JADT 2010), Rome*, 731–742.

Hirst, G., Feng, V. W., Cochrane, C. & Naderi, N. (2014). Argumentation, Ideology, and Issue Framing in Parliamentary Discourse. *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing (ArgNLP)*.

Honkela, T., Korhonen, J., Lagus, K. & Saarinen, E. (2014). Five-dimensional sentiment analysis of corpora, documents and words. *Advances in Self-Organizing Maps and Learning Vector Quantization*, 209–218.

Hopkins, D. J. & King. (2009). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229-247.

Iliev, I. R., Huang, X. & Gel, Y. R. (2019). Political rhetoric through the lens of non-parametric statistics: are our legislators that different? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 583-604.

Iyyer, M., Enns, P., Boyd-Graber, J. & Resnik, P. (2014). Political ideology detection using recursive neural networks. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1113–1122.

Jensen, J., Naidu, S., Kaplan, E., Wilse-Samson, L., Gergen, D., Zuckerman, M. & Spirling,

A. (2012). Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech [with comments and discussion]. *Brookings Papers on Economic Activity*, 1–81.

Ji, Y. & Smith, N. A. (2017). Neural discourse structure for text categorization. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 996–1005.

Kaal, B., Maks, I. & van Elfrinkhof, A. (2014). *From Text to Political Positions: Text Analysis Across Disciplines*. Philadelphia: John Benjamins Publishing Company.

Kapočiūtė-Dzikienė, J. & Krupavičius, A. (2014). Predicting party group from the Lithuanian parliamentary speeches. *Information Technology And Control*, *43*(3), 321–332.

Kauffman, D., Khosmood, F., Kuboi, T. & Dekhtyar, A. (2018). Learning Alignments from Legislative Discourse. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 119:1–119:2.

Kim, I. S., Londregan, J. & Ratkovic, M. (2018). Estimating spatial preferences from votes and text. *Political Analysis*, *26*(2), 210–229.

Lapponi, E., Søyland, M. G., Velldal, E. & Oepen, S. (2018). The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. *Language Resources and Evaluation*, *52*(3), 873–893.

Laver, M., Benoit, K. & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(2), 311–331.

Lefait, G. & Kechadi, T. (2010). Analysis of deputy and party similarities through hierarchical clustering. *2010 Fourth International Conference on Digital Society*, 264-268.

Li, X., Chen, W., Wang, T. & Huang, W. (2017). Target-specific convolutional bi-directional LSTM neural network for political ideology analysis. *Web and Big Data*, 64–72.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge, UK: Cambridge University Press.

Lowe, W. & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, *21*(3), 298–313.

Martineau, J., Finin, T., Joshi, A. & Patel, S. (2009). Improving binary classification on text problems using differential word features. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2019–2024.

Menini, S. & Tonelli, S. (2016). Agreement and disagreement: Comparison of points of view in the political domain. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2461–2470.

Menini, S., Nanni, F., Ponzetto, S.P. & Tonelli, S. (2017). Topic-based agreement and disagreement in US electoral manifestos. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2938–2944.

Mikhaylov, S., Laver, M. & Benoit, K. (2008). Coder reliability and misclassification in comparative manifesto project codings. *66th MPSA Annual National Conference*, *3*(6) 3.

Mohammad, S. M., Sobhani, P. & Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3), 26:1–26:23.

Monroe, B. L., Colaresi, M. P. & Quinn, K. M. (2017). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict, *Political Analysis*, *16*(4), 372–403.

Naderi, N. & Hirst, G. (2016). Argumentation Mining in Parliamentary Discourse. *Principles and Practice of Multi-Agent Systems*, 16–25.

Nanni, F., Zirn, C., Glavaš, G., Eichorst, J. & Ponzetto, S.P. (2016). TopFish: topic-based analysis of political position in US electoral campaigns. *PolText 2016 : The International Conference on the Advances in Computational Analysis of Political Text : proceedings of the conference*.

Nguyen, V.-A., Boyd-Graber, J.-L. & Resnik, P. (2013). Lexical and hierarchical topic regression. *Advances in neural information processing systems*, 1106–1114.

Nguyen, V.-A., Boyd-Graber, J., Resnik, P. & Miler, K. (2015). Tea party in the House: A hierarchical ideal point topic model and its application to Republican Legislators in the 112th Congress. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1438–1448.

Onyimadu, O., Nakata, K., Wilson, T., Macken, D. & Liu, K. (2013). Towards sentiment analysis on parliamentary debates in Hansard. *Joint International Semantic Technology Conference*, 48–50.

Owen, E. (2017). Exposure to offshoring and the politics of trade liberalization: Debate and votes on free trade agreements in the US House of Representatives, 2001–2006. *International Studies Quarterly*, *61*(2), 297–311.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends®*

in Information Retrieval, *2*,(1–2), 1–135.

Petticrew, M. & Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide*. Malden, MA: Blackwell.

Plantié, M., Roche, M., Dray, G. & Poncelet, P. (2008). Is a voting approach accurate for opinion mining? *Data Warehousing and Knowledge Discovery*, 413–422.

Proksch, S.-O. & Slapin, J.B. (2010). Position taking in European Parliament speeches. *British Journal of Political Science*, *40*(3), 587–611.

Proksch, S.-O. & Slapin, J.B. (2015). *The Politics of Parliamentary Debate*. Cambridge, UK: Cambridge University Press.

Proksch, S.-O., Lowe, W., Wäckerle, J. & Soroka, S. (2018). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, *44*(1), 97–131.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Rauh, C. (2018). Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics*, *15*(4), 319-343.

Rheault, L. (2016). Expressions of anxiety in political texts. *Proceedings of the First Workshop on NLP and Computational Social Science*, 92–101.

Rheault, L., Beelen, K., Cochrane, C. & Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLOS ONE*, *11*(12), 1-18.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š. & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), 140-157.

Sakamoto, T. & Takikawa, H. (2017). Cross-national measurement of polarization in political discourse: Analyzing floor debate in the US the Japanese legislatures. *2017 IEEE International Conference on Big Data (Big Data)*, 3104–3110.

Salah, Z., Coenen, F. & Grossi, D. (2013). Extracting debate graphs from parliamentary transcripts: A study directed at UK House of Commons debates. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, 121–130.

Salah, Z., Coenen, F. & Grossi, D. (2013). Generating domain-specific sentiment lexicons for opinion mining. *Advanced Data Mining and Applications*, 13–24.

Salah, Z. (2014). *Machine Learning and Sentiment Analysis Approaches for the Analysis of Parliamentary Debates* (PhD thesis). University of Liverpool, UK.

Schwarz, D., Traber, D. & Benoit, K. (2017). Estimating Intra-Party Preferences: Comparing Speeches to Votes. *Political Science Research and Methods*, *5*(2), 379–396.

Seligman, M. E. P. (2012). *Flourish: A Visionary New Understanding of Happiness and Well-being*. New York: Simon and Schuster.

Sim, Y., Acree, B. D., Gross, J. H. & Smith, N. A. (2013). Measuring ideological proportions in political speeches. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 91-101.

Sokolova, M. & Lapalme, G. (2008). Verbs speak loud: Verb categories in learning polarity and strength of opinions. *Advances in Artificial Intelligence*, 320–331.

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, *108*(503), 755-770.

Thomas, M., Pang, B. & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 327–335.

Van der Zwaan, J. M., Marx, M. & Kamps, J. (2016). Validating cross-perspective topic modeling for extracting political parties' positions from parliamentary proceedings. *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, 28–36.

Vilares, D. & He, Y. (2017). Detecting Perspectives in Political Debates. *Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing*, 1573–1582.

Yadollahi, A., Shahraki, A. G. & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, *50*(2), 25:1–25:33.

Yessenalina, A., Yue, Y. & Cardie, C. (2010). Multi-level structured models for document-level sentiment classification. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1046–1056.

Yogatama, D. & Smith, N. (2014a). Linguistic structured sparsity in text categorization. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 786–796.

Yogatama, D. & Smith, N. (2014b). Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. *International Conference on Machine Learning*, 656–664.

Yogatama, D., Kong, L. & Smith, N. A. (2015). Bayesian optimization of text representations. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2100–2105.

Zirn, C. (2016). *Fine-grained Position Analysis for Political Texts* (PhD thesis). Universität Mannheim.