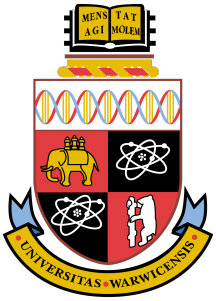
Measuring Parliament Polarization during the Covid-19 Pandemic - *A Comparison and Validation of Methods using Text-as-Data*

Lukas Birkenmaier

Master Thesis
Summer 2021
Centre for Interdisciplinary Methodologies (CIM) - Warwick University
Supervisor: Tessio Novack
Word Count: 10.326



Abstract

Political polarization constitutes a fundamental condition for the functioning of modern democracies. If levels of polarization are growing too high, however, political processes are harmed, and democratic principles can be violated. In order to study the causes and implications of this phenomenon, scholars of political science research have relied on different methods to operationalize and measure political polarization. Recently, a new branch of literature has emerged for doing so, which relies on methods from computational linguistics and the analysis of textual data. Hence, this dissertation aims to assess the measurement validity of the three most popular groups of methods in this field. In doing so, a validation framework is deployed to compare and validate the estimated levels of polarization for 15 parliamentary debates in the German Bundestag. Results show that only the method based on an ideological text scaling approach is able to pass all tests in the validation framework, whereas the other methods clearly fail to detect relevant aspects of political polarization. Nevertheless, it is suggested that future research should carefully consider applying these groups of methods in applied research, and should always include additional validation steps to justify the inferences derived from their research designs.

Key Words: Political Polarization, Parliamentary Debates, Natural Language Processing, Sentiment Analysis, German Bundestag, Covid-19

Contents

List of Tables

List of Figures

List of Abbreviations

CA	Classification Accuracy
IfSG	Infektionsschutzgesetz
IGO	Index of the Intensity of Government and Opposition
IS	Ideological Scaling
MP	Member of Parliament
NLP	Natural Language Processing
SL	Sentiment Lexicon

1. Introduction

> *Science cannot progress without reliable and accurate measurement of what it is you are trying to study. The key is measurement, simple as that.*
>
> - Robert D. Hare, *Professor emeritus of the University of British Columbia*

In recent years, scholars have observed a worldwide renaissance of politicization and political polarization. Fostered by the ongoing globalization, economic and ecologic crises as well as the disruption of the classical media industry, divisions among citizens across countries are growing sharper (Goldberg et al. 2020). In a recent survey from Pew Research Center (2019), for instance, 85% of US respondents stated that the tone and nature of political debates have become more negative in the last years. Likewise, data from the V-Dem Institute (2019) shows that nearly all EU societies have become more polarized in the twenty-first century compared to previous time periods. Ultimately, this trend does not only reflect itself in the growing support of populist parties and politicians (Arbatli and Rosenberg 2021; Jiang et al. 2020), but also more individual feelings of partisanship and emotional distance to arguments contrary to someone's own position (Galston 2018).

In order to better understand the causes and factors involved in this phenomenon, researchers rely on an extensive body of literature founded on theories of public discourse, policy-making, and intra-party conflicts (cf. Scollon (2008)). However, one fundamental methodological challenge has always been to validly measure the degree of political polarization in a given setting. Whereas previous research has mainly relied on survey data or voting behavior to operationalize political polarization, a new emerging strand of literature has shifted its attention to analyzing texts, such as party manifestos or speech transcripts. This new field of computational linguistics, based on methods of natural language processing (NLP) and the slogan of "text-as-date" (Grimmer and Stewart 2013, p. 267), has thus gained considerable popularity in the last few years.

However, even though these methods promise a lot of insight for political science research, a growing number of scholars also point to the methodological challenges associated with these methods, such as the inherently high-dimensional nature and semantic complexity of written texts (cf. Gentzkow, Kelly, et al. (2019) and Goet (2019)). Validating these methods is therefore especially relevant since "scholars routinely make claims that presuppose the validity of the observations and measurements that operationalize their concept" (Adcock and Collier 2001, 529).

Due to these reasons, this paper aims to further explore this relevant field of research. This is done by comparing and validating the most popular algorithms in this field on a specific use case of 326 parliamentary speeches, which are centered around the handling of the Covid-19 pandemic in the German parliament (Bundestag) between February and July 2020.

Throughout this paper, several interrelated research questions are explored; first, what are the most relevant groups of methods to measure political polarization using text-as-data approaches?

Second, do these groups of methods then come to similar estimates of political polarization for the speeches identified? And third, how can researchers validate the estimates and, thus, ensure academic inference for applied research projects?

Thereby, this paper relates most closely to recent work by Goet (2019). In his paper, he uses parliamentary transcripts from the UK Congressional Record to measure party polarization since the late nineteenth century. However, his analysis only covers two methods, and he analyses highly aggregated data in the extensive time period from 1811 to 2015. To the author's best knowledge, this paper is therewith the first approach to compare text-as-data methods measuring political polarization on such a short and defined time period of political discourse. To clarify terminology, political polarization is subsequently defined as "the extent to which opinions on an issue are opposed in relation to some theoretical maximum" (DiMaggio et al. 1996, p. 693).

The remainder is as follows: the following chapter 2 introduces the relevant research on political polarization and its relation to text-as-data approaches. Chapter 3 then presents the systematic literature review on the most relevant methods to measure parliament polarization using textual data. The research design is outlined in chapter 5, before chapter 6 presents the results of the analysis. Chapter 7, ultimately, completes the dissertation with a concluding discussion, the presentation of limitations, and an outlook on potential further research.

2. On the Relevance of Measuring Political Polarization

"Hatred, anger, and violence can destroy us: the politics of polarization is dangerous"

- Rahul Gandhi, *Indian Politican*

A large body of literature exists, covering various aspects of political contestation and polarization. Rather than trying to provide an extensive overview of this field, this chapter aims to shed light on the relevance of measuring political polarization for political science research. Doing so, it first provides a concise overview of the impact of political polarization on democratic political systems (Section 2.1). Afterward, it emphasizes the necessity to measure political polarization in parliament (Section 2.2) and highlights the potential of using text-as-data for doing so (Section 2.3).

2.1 Polarization and Political Contestation

According to Easton (1965), political systems constantly perceive input from all parts of society. Likewise, this input is then converted into outputs, such as decisions or actions, which in turn affect the future demands of society towards the political system. Dealing with political polarization is no exception. Generally, research acknowledges that political polarization is inextricably linked to the functioning of any political system. However, when evaluating the impact of polarization on the effectiveness of the political process in modern democracies, scholars have come to ambivalent assessments.

For the one part, moderate levels of political polarization have been acknowledged to be beneficial for democracies (Downs 1957; McCoy et al. 2018; Sartori 1976, 2005). For instance, case studies from North America (Campbell 2018), Africa (LeBas 2006) and Europe (Enyedi 2016) all provide evidence that a polarized environment helps voters to correctly differentiate between political alternatives. Furthermore, it enables parties to mobilize supporters and strengthen internal cohesion. Speaking more generally, political polarization thus constitutes a basic condition for social and political progress. This process, which has been described by Mudde and Kaltwasser (2013) as "inclusionary populism" (p. 147) thus functions as a positive force for democracy by "[giving] voice to groups that do not feel represented [...] and [changing] the political agenda to include these marginalized voices" (p. 168).

For the other part, however, several studies provide evidence that critical levels of political polarization can seriously hamper the functioning and integrity of contemporary democratic systems (Bergmann et al. 2021; Maoz and Somer-Topcu 2010; Petri and Biedenkopf 2021; Sani and Sartori 1983; Sakamoto and Takikawa 2017). The theoretical explanation for this mechanism can be found in social identity theory introduced by Tajfel et al. (1979). In their popular paper, they

sought to explain intergroup conflict with a set of explanation patterns, such as social categorization, ingroup favoritism, and outgroup rejection. Hence, these patterns have been acknowledged to be especially strong when the levels of conflict between groups are high (cf. McGarty et al. (1992)). Assessing the negative impact of political polarization on democratic political systems, McCoy et al. (2018) consequently argue that the negative effects of political polarization are especially severe when large parts of society can be assigned to a few mutually exclusive identities and interests. If high levels of polarization further cement cleavages between these groups, people will tend to abstain from a constructive exchange and further develop stereotypes, such as feelings of someone's own superiority (Eidelson and Eidelson 2003) or the de-individualization of members of the other group (Reicher and Levine 1994). In addition to that, political polarization can also foster the spread of "exclusionary populism" (Mudde and Kaltwasser 2013, p. 147). This form of populist behavior has recently gained popularity in formerly stable democracies and has been described by scholars as power-driven, antagonistic against minority groups, and generally harmful to the foundations of representative democracy (Bergmann et al. 2021; Caramani 2017).

Empirically, several studies have found indices of a negative impact of critical levels of political polarization on the functioning of the political system. On the individual level, the main findings constitute a negative impact on individuals' willingness to contribute to public goods (Cornelson and Miloucheva 2020) and greater acceptance of growing authoritarianism (Hetherington and Weiler 2009; Somer and McCoy 2018). On the political level, high polarization is also associated with lessened coalition stability (Bergmann et al. 2021; Maoz and Somer-Topcu 2010), increased party fragmentation (Stroschein 2011) and inefficient policymaking (Strøm et al. 2008; Burns 2019; Petri and Biedenkopf 2021).

2.2 On the Necessity to Measure Parliament Polarization

Considering the various positive and negative effects of political polarization on politics and society presented in the previous section, scholars have recognized the importance to systematically study the phenomenon of political polarization from an early stage on (Peterson and Spirling 2018). However, in order to develop and test theories in the social sciences, researchers always depend on valid and reliable methods to operationalize latent constructs such as political polarization (cf. Popper (1963) and Reiter (2017)). Therefore, several approaches have been developed to serve this purpose while focusing on different levels of observations.

On the level of society, one branch of literature relies on census and cross-country survey data to measure individual and group-aggregated conflict. Thereby, polarization is usually operationalized through respondents' self-placement on a left-right scale (DiMaggio et al. 1996; Grechyna 2016; Silva 2018) or their emotional attachments towards political actors (Boxell et al. 2017). Likewise, another branch of literature focuses on people's interactions on social networks like *Twitter* or *Facebook* (Morales et al. 2015; Yardi and Boyd 2010; Kiran and Weber 2017). Here, polarization is usually measured by examining user's network, their tweeting behavior, and the content they share, such as posts, retweets, or hashtags.

On the level of politics, the most common approach constitutes the analysis of parliamentary data. This is because the parliamentary arena constitutes the most central part of any democratic

discourse (Bayley 2004; Catt 2002). Thereby, its topics and conflicts stand representatively for the political and social cleavages in society. Analyzing the outputs of the political system thus allows scholars to justify their research by examining the most salient topics facing societies these days (Goet 2019).

Empirically, one proven method to operationalize political polarization has been to use party fragmentation as a proxy for political conflict (Tepe 2014; Corrales 2005). However, the scope for interpretation remains restricted to analyzing greater trends spanning long time intervals. This is because party composition in parliament usually changes at a relatively slow pace and might be influenced by the political system and its election procedure too (Golosov 2015).

On a more granular level of political interaction, a large body of work builds on legislative roll-call votes to assess the coherence of decisions from member of parliament (MP) (Laver et al. 2003; Poole and Rosenthal 1985; Wright 2007). However, from a methodological point of view, these approaches entail some important shortcomings as well. Primarily, unobserved factors like party discipline or strategic voting behavior represent a severe constrain to the validity of these approaches (Spirling and McLean 2007). Furthermore, the direction of causality might be misinterpreted, since voting behavior is "more a product of the political process [...] than causally prior to [its polarization process]" (Laver et al. 2003, p. 311)

Due to that, a growing group of scholars has instead shifted their attention to analyzing parliamentary texts and debate transcripts. The last section in this chapter will thus explain and outline the potential of measuring political polarization using textual data.

2.3 The Potential of Measuring Polarization using Text-as-Data

Measuring political polarization using parliamentary texts and debate transcripts has gained reasonable popularity in the last years (Abercrombie and Batista-Navarro 2020). From an academic point of view, a combination of different reasons and social and political developments have contributed to this trend.

First, scholars agree that textual data represent a valuable data resource that facilitates substantially valuable inferences about political discourse (Grimmer and Stewart 2013; Gentzkow, Shapiro, et al. 2019). Hence, the content and topics of what members of parliament speak about in plenary are usually revolving around the core conflict lines within society. This is particularly reflected by the fact that a parliamentary speech enables MPs to express dissent and polarized views in a more nuanced way when compared to simple roll-call votes. Therefore, analyzing political speeches enables researchers to explore and measure speakers' positions and patterns of argumentation in a high-dimensional setting (Goet 2019). Additionally, political speeches are typically held by one member of parliament (MP) on one specific topic. Thereby, they remain at a conceptually relevant level of analysis and can be aggregated up to any desired level, such as e. g. MPs position on a specific topic, MP sentiments towards a bill or, more generally, aggregated parliamentary polarization over a chosen time period (Proksch et al. 2019).

Another reason for the popularity of text-as-data approaches is connected to the immense liberating potential for political science research. This is because the digitization and provision of parliamentary transcripts as open data has enabled scholars to easily get access to vast amounts of

textual data previously hidden from research (Beelen et al. 2017). Often, several hundred speeches are being delivered by MPs in one parliamentary session alone, which results in hundreds of thousand textual data points over the year. Consequently, this enables the systematic assessment of large-scale text collections without massive funding support for researchers (Grimmer and Stewart 2013). Additionally, once the data is public, it can be "analyzed, reanalyzed again without becoming jaded or uncooperative [...] and others can replicate, modify, and improve the estimates involved or can produce completely new analyses using the same tools" (Laver et al. 2003, p. 311). Even though other areas of political science research have made advantages in providing data publicly as well (*Open Science*, cf. Vicente-Sáez and Martınez-Fuentes (2018)), analyzing parliamentary textual data is still one of the most promising approaches to provide easy and freely accessible data in the area of political discourse research (Hardwicke et al. 2018).

Lastly, technical and theoretical advantages in NLP methods promise an even greater potential for future research. As Gentzkow, Kelly, et al. (2019) note, the majority of popular NLP methods today are admittedly too simplistic and methodologically unsound to draw inferences about the *true* process of data generation. Nevertheless, even comparatively simplistic models can already provide researchers with highly relevant information. As Abercrombie and Batista-Navarro (2020) note this is true for the analysis of parliamentary debates as well. Here, several studies have relied on NLP methods to improve the process of information retrieval. Additionally, the future promises an even deeper understanding of language and textual data. With the ongoing rapid developments in deep learning and artificial intelligence, scholars will likely be able to further improve the semantic understanding of textual data to gain new insights into political discourse (cf. Devlin et al. (2018)).

Coming to the end of this chapter, it can be summarized that scholars emphasize the relevance of measuring political polarization using text-as-data approaches. But how should researchers actually evaluate the validity of studies performing this task? And how should they judge the implications drawn from their research designs? In order to systematically answer this question, the next chapter will provide a structured literature review on all studies identified to evaluate the current state of research on measuring parliament polarization from textual data.

3. Literature Review: Measuring Parliament Polarization using Text-as-Data

"All quantitative models of language are wrong - but some are useful."

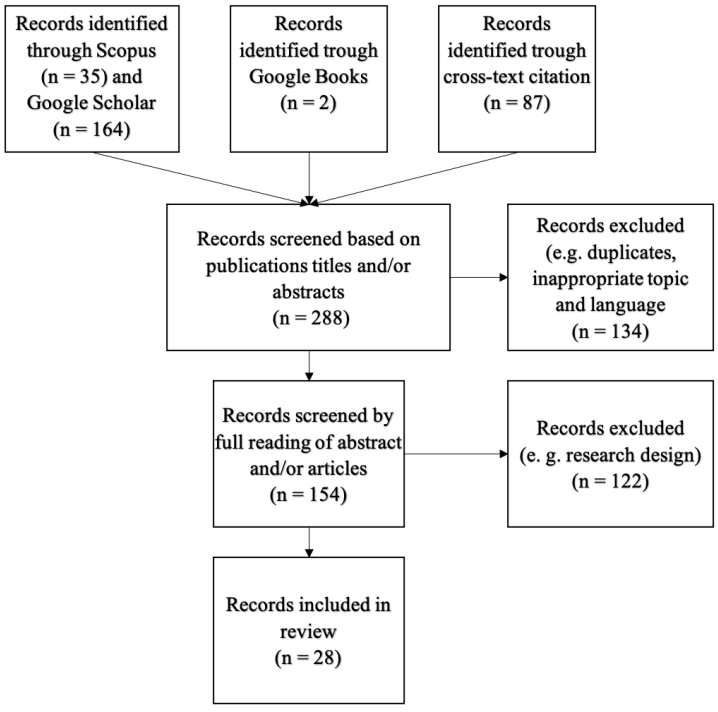
- Justin Grimmer and Brandon Stewart 2013,

At the beginning of this chapter, Section 3.1 provides a short summary of the structure of the systematic literature review. Afterwards, Section 3.2 presents the main overview of all methods identified including their main assumptions and limitations. Lastly, these methods are discussed in Section 3.3, which highlights similarities and differences between the methods and emphasizes possible pitfalls in applying these methods in a given study context.

3.1 PRISMA Method

In order to demonstrate the transparency of the literature search and to ensure the completeness of the studies under review, the literature review was conducted based on a slightly simplified variant of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Liberati et al. 2009). The adopted framework is graphically depicted in Figure 1.

Figure 1. PRISMA Method



Starting with the literature search, two central literature databases (*Google Scholar* and *Scopus*) were identified, which ensure access to a whole scientific range of relevant publications. Since the search terms ["political polarization" and “parliament”] generated a very high number of search results in the databases (>17,000), it was specified in a second round using the keywords [”political polarization” AND ”parliament” AND ”measurement” AND ”transcripts”], which limited the preliminary literature list to 1,140 results. Using the PRISMA scheme, the preliminary literature list was further restricted with the help of a number of selection criteria. On the one hand, only texts in English or German were selected that had been published online after 1990 and were freely accessible to users of the University of Warwick. Furthermore, only peer-reviewed articles were selected, and so-called "grey literature" was only considered in exceptional cases (e.g. direct reference in the literature). After removing duplicates, those articles in the selected list were then sorted out by checking the title and abstract, to ensure that they applied a quantitative method to measurement parliament polarization. Lastly, in a final step, cross-referencing in the most relevant publications for this work was considered in more detail (*Snowballing*, cf. Gentles et al. (2015)). This finally reduced the risk of overlooking important literature that was not found by the keywords mentioned. At the end of this selection procedure, 28 articles could be included in the final literature selection. Table 3.1 gives an overview over all studies covered in this literature review, which will be subsequently discussed in Section 3.2

3.2 Approaches to Measure Parliament Polarization

3.2.1 Sentiment Lexicons

From an early stage on, several authors have relied on lexical-based methods to measure polarization in parliamentary documents (Ahmadalinezhad and Makrehchi 2018; Balahur et al. 2009; Biessmann 2016; Haselmayer and Jenny 2017; Liu and Lei 2018; Proksch et al. 2019; Dzieciątko 2019; Onyimadu et al. 2014; Rauh 2018). These approaches, which are also known as *polarity classification*, *opinion mining*, or *sentiment analysis*, thus assume that higher levels of polarization reflect themselves in the use of more polarized vocabulary (Abercrombie and Batista-Navarro 2020). Thereby, these studies rely on lexicons and annotated datasets, where each word is associated with a polarity score on either a metrical (-1 to $+1$) or ordinal (“positive”, “neutral”, “negative”) scale. After assigning the dictionary values to all of the respective words in a given document, the scores are then summarized (e. g. by taking the average) and an overall polarity score is calculated for each document (Yadollahi et al. 2017). In the past, several studies have applied lexical-based methods to study the polarity in parliamentary debates. Balahur et al. (2009), for instance, combined three different lexicon sources to evaluate both the polarity of opinion and emotions expressed from the transcripts of U.S. Congressional floor debates. By comparing their results to the vote outcome of legislative bills, they are able to reach a classification accuracy of around 76% on the individual speaker segments based on their computed polarity score. Likewise, Proksch et al. (2019) apply their multilingual sentiment approach on the ParlSpeech data set (cf. Rauh (2018)) consisting of 3.9 million plenary speeches of seven European states. Their findings suggest that the calculated sentiment scores were able to capture important aspects of legislative conflict, regardless of the national background under study.

Table 3.1. *Overview Systematic Literature Review*

Method	Method Type	Publications
Sentiment Lexicons	Dictionary	Balahur, Kozareva,and Montoyo (2009)[a]; Onyimadu et al. (2013)[a]; Biessmann (2016)[b]; Haselmayer and Jenny (2017)[c]; Ahmadalinezhad and Makrehchi (2018)[a]; Dzieciatko (2018)[a]; Liu and Lei (2018)[d]; Proksch et al. (2019)[e];
Ideological Scaling	Wordscores	Laver, Benoit, and Garry (2003)[a]; Herzog and Benoit (2015)[a]; Baturo, Dasandi, and Mikhaylov (2017) [a]
	Wordfish	Slaping and Proksch (2008)[f]; Lowe and Benoit (2013)[g]; Lauderdale and Herzog (2016)[g]; Schwarz, Traber, and Benoit (2017)[h]; Frid-Nielsen (2018)[i]; Goet (2019)[g]; Curini,Hino, and Osaka (2020) [j]
	Others	Jensen et al. (2012)[k]; Gentzkow,Shapiro, and Taddy (2016)[l]; Spirling, Huang, and Patrick (2018)[m]; Kim, Londregan, and Ratkovic (2018)[g]; Gentzkow, Shapiro, et al. (2019)[n]; Yan et al. (2019)[a]
ML Accuracy	Linear	Peterson and Spirling (2018)[g]; Goet (2019)[g]; Idelberger (2020)[m]; Søyland (2020)[m]
	Non-Linear	Abercrombie and Batista-Navarro (2018)[a]

Note: Publications are ordered by publication date for each group of method.
Source of Publication: [a] *Conference Paper*; [b] *arXiv*; [c] *Quality & Quantity*; [d] *Discourse, Context & Media*; [e] *Legislative Studies Quarterly*; [f] *American Political Science Review*; [g] *Political Analysis*; [h] *Political Science Research and Methods*; [i] *European Union Politics*; [j] *Government and Opposition*; [k] *Brookings Papers on Economic Activity*; [l] *eSocialSciences*; [m] *Others*; [n] *Journal of Economic Literature*;

However, *lexical-based methods* are not without shortcomings. Even though researchers can implement them easily and resource-efficient to big corpora of textual data, their simplicity prevents them from detecting different semantic meanings of words in different contexts. Rheault et al. (2016) illustrate this point with the word "health", which is generally considered to bear a positive sentiment. In the context of political debates, however, the word health also often relates to formal definitions, such as e. g. a country's "Ministry of Health", which bears a more descriptive and neutral sentiment. Furthermore, empirical evidence points to the crucial role of domain specificity in sentiment analysis effectiveness (Haselmayer and Jenny 2017). Once a sentiment dictionary is trained on a specific purpose, researchers should be very cautious to apply it to other political contexts without thorough customization of the vocabulary (Loughran and McDonald 2011).

3.2.2 Ideological Text Scaling

Another popular approach in the measurement of political polarization can be summarized under the umbrella term *ideological text scaling* (Baturo et al. 2017; Gentzkow et al. 2016; Jensen et al. 2012; Kim et al. 2018; Laver et al. 2003; Schwarz et al. 2017; Spirling et al. 2018; Curini et al. 2020; Gentzkow, Shapiro, et al. 2019). These methods are generally founded on a spatial model of politics, underlying early research from Downs (1957) and Sartori (1976) in the field of party polarization and party competition. Hence, the basic idea behind these frameworks assumes that political parties are usually aligned along one conflict line, such as a left-right or a liberal-conservative continuum (Adams et al. 2005). In times of elections, the party positions on this policy space are then evaluated by the voters, which will typically select the party most proximate to their own position (Dalton 2008). As Dalton (2008) notes, this theoretical concept also inherently implies "a concern for the degree of polarization in a party system" (p. 901) given by the distance between the parties and, thus, their political proximity. Hence, researchers have begun to apply this spatial concept of politics to measure political polarization in parliamentary debate transcripts (Grimmer and Stewart 2013). Doing so, they apply several methods to derive ideological positions from textual data such as parliamentary transcripts. Thereby, they resort to measuring polarization by assessing the similarity between groups based on the sociological concept of intra-group homogeneity and inter-group heterogeneity (cf. Deutsch (1971)). This idea behind this concept is described by Esteban and Ray (1994) as such:

> Suppose that a population of individuals may be grouped according to some vector of characteristics into "clusters," such that each cluster is very "similar" in terms of the attributes of its members, but different clusters have members with very "dissimilar" attributes. In that case we say that the society is polarized. (p. 819)

In applied research, different measurement implementations based on this spatial concept of politics have been developed. Assessing the most relevant methods, Laver et al. (2003) implemented the *Wordscores*, which computes subject positions based on a set of selected reference texts known to represent the extremes of the political space. However, one important shortcoming of the *Wordscores* algorithm constitutes the selection of appropriate reference texts. In order to tackle this issue, Slapin and Proksch (2008) have developed the text scaling model *Wordfish*.

Like *Wordscores*, the model assumes that "the relative word usage of parties provides information about their placement in a policy space" (p. 708). However, the main difference lies in the assumption that word frequencies are drawn from a statistical Poisson distribution with:

$$\begin{aligned} y_{ijt} &\sim Poisson(\lambda_{ijt}) \\ \lambda_{ijt} &= exp(\alpha_{it} + \psi_j + \beta_j * \omega_{it}) \end{aligned} \tag{3.1}$$

where y_{ijt} is the count of word j in party i's text at time t, α_{it} is a set of document fixed effects (captures document lengths), Ψ_j is a set of word fixed effects (captures average word frequency), β_j is an estimate of a word specific weight capturing the importance of the word j in discriminating between parties, and ω_{it} is the estimate of party i's position in the year t (p. 709). Applying the *Wordfish* model to parliamentary debates, the model is thus not only able to estimate the ideological position of MP speeches ω_{it} based on the shared word usage across all texts but is also able to recover which words have more explanatory power to differentiate across political speeches as given by the word weight β_j (cf. Curini et al. (2020)).

Being able to apply this unsupervised technique for ideological scaling, several authors have relied on *Wordfish* to linearly order parties and politicians across political dimensions. Lowe and Benoit (2013), for instance, focus on validating the model estimates on an austerity debate in the Irish parliament in 2009 against the systematic ranking of around twenty human readers. Running their analysis, they find a "high degree of correspondence" between the model's estimates and human judgment, even though they claim that some assumptions of the statistical model are violated (p. 312). Examining asylum debates in the European parliament, Frid-Nielsen (2018) furthermore analyze 876 speeches held between 2004 and 2014. Developing their own *Wordshoal* model, Lauderdale and Herzog (2016) apply a two-step approach by adding Bayesian factor analysis on the *Wordfish* debate specific estimate ω_{it} to aggregate them into one general latent position for each legislator.

However, one crucial assumption of these approaches is that the use of language is primarily dominated by varying positions on one single underlying policy dimension. Grimmer and Stewart (2013) show that once the standard policy space is not dominated by ideology, "the model clearly fails to separate [US] senators". This might be especially likely if the respective texts are not only related to one topic but also contain a diverse set of topics and messages (Goet 2019).

3.2.3 Machine Learning Classification Accuracy

As a last method, the literature review also identified a third branch of literature to measure parliament polarization, namely the classification accuracy of a machine learning (ML) algorithm (Peterson and Spirling 2018; Frech et al. 2018; Søyland 2020; Goet 2019; Idelberger 2020). This approach has been introduced by Peterson and Spirling (2018) and has gained considerable attention in relevant journals like *Political Analysis* since then. Similar to the ideological scaling methods, this approach is based on a measure of similarity between texts. However, the computation of polarization scores at the document level follows a different approach. As Peterson and Spirling (2018) explain:

> "How distinguishable [MPs] are in practice is determined by a set of machine learning algorithms. Put very crudely, after being trained on a portion of the speeches, the models are then required to predict the most likely 'label'—that is, party identity [...] When the machine learning accuracy—in the technical sense—is low [..] we deduce then that we are in a world of relatively low polarization. By contrast, when accuracy is high, and the machine does well at discriminating between partisans based on their utterances, [...] we are in a more polarized era." (p. 2-3).

They illustrate their method by applying four different ML algorithms to measure varying levels of parliament polarization in the United Kingdom between 1935 and 2013. By validating their approach with qualitative and quantitative historical records, they claim that their estimates are able to trace back major trends in parliament polarization. Others have come to similar conclusions. Goet (2019), for instance, compares the results of an ideological scaling algorithm and an ML classifier against a validation framework. Hence, he finds that the ML approach showed a high degree of face-, construct-, and convergent validity" (p. 535). Most recently, other scholars have started to apply non-linear neural networks as well. Abercrombie and Batista-Navarro (2018), for instance, rely on a multi-layered perceptron neural network to classify the sentiment polarity of speakers in the *Hansard* corpus of UK parliamentary debate transcripts with government vs. opposition labels.

Generally, this group of techniques strongly builds on the assumption that ML classification accuracy is an adequate metric to measure political polarization. One fundamental and questionable hypothesis thus constitutes that the respective algorithms primarily measure dissimilarity in ideological speech, and not speech length, speaker variety, or dominant non-ideological features instead. On top of that, some assumptions are made which apply to machine learning classification tasks in general. Most importantly, these are a balanced dataset and a low bias for misclassifying texts of one specific group in a particular direction (Hopkins and King 2010).

3.3 Evaluation and Comparison

As presented in the previous section, several methods have been applied to measure political polarization from parliamentary speech data. Thereby, the literature review hints at the fact that all methods identified have their unique methodological approach, focusing on different aspects of polarization in texts while each relying on their own set of assumptions and limitations. Table 3.2 provides a taxonomy over some key characteristics of all methods under study.

Table 3.2. *Comparing Approaches to Measure Political Polarization using Text-as-Data*

Method	**Dimension**	**Reference**	**Measurement**	**Mode**	**Method**
Sentiment lexicons	Speech sentiment	Dictionary values	Sentiment score	Absolute	Unsupervised
Ideological scaling	Dissimilarity word usage	Spatial position	Actors distance	Relative	Unsupervised / Supervised
Machine learning	Dissimilarity word usage	Classification accuracy	Classification confidence	Relative	Supervised

However, taking a broader perspective on this field of research, scholars have also expressed concerns questioning the validity and reliability of current quantitative methods (Louwerse et al. 2021; Goet 2019; Grimmer and Stewart 2013). Even though the use of NLP techniques provides great potential for researchers, there might be some serious methodological challenges present which should be considered when analyzing text-as-data as well.

First and foremost, scholars have criticized the absence of a consistent validation framework (Grimmer and Stewart 2013). *Post hoc* statistical validation of measuring instruments, however, constitutes a crucial precondition when claiming the validity of results from academic research. In previous studies, the most popular approach to tackle this issue has been to compare computer-based estimates with hand-coded assessments from human coders. What Haselmayer and Jenny (2017) describes as the "gold standard [...] from human coding" (p. 2631), however, constitutes not only an expensive and time-consuming task, but is also inherently subjective as well (Petri and Biedenkopf 2021). This is because human coders will always possess prior knowledge and individual assessments on political issues, and these attitudes will be unevenly distributed across participants (Laver and Garry 2000). Admittedly, best practices are available to tackle this issue, such as the use of code-books or parallel coding from independent third-party coders (cf. Mayring (2015)). However, hand-coded estimates still represent an approximation to the *true* quantities of interest, which, ultimately, will remain unobservable (Lowe and Benoit 2013).

Furthermore, other scholars have pointed to the crucial and difficult process of data preprocessing. Goet (2019), for instance, points to the fact that political speeches are often not solely related to one specific topic. Instead, speakers often "go off-topic, speaking to different matters, or combining their statement on the discussion topic with several other messages" (p. 520). Moreover, the greater semantic context of language might also influence the outcome of the research methods. This is generally true for the majority of bag-of-words approaches (cf. Harris (1954) and Zhang et al. (2010)), which completely ignore grammar, sentence structure as well as word order altogether. Additionally, human language has always been the subject of constant evolution, often changing the semantic meaning of words altogether (Jatowt and Duh 2014). If researchers do not take all these possible biases and pitfalls in their research design into account, the academic findings of text-as-data approaches are likely to be biased and inconsistent.

Lastly, another shortcoming merely discussed in research constitutes the question of highly aggregated data (Proksch et al. 2019). Professedly, research assessing the long-term trends of political polarization spanning wide time intervals across legislative periods constitutes a relevant field of research (cf. Peterson and Spirling (2018)). However, scholars are often interested in measuring polarization on a much more granular level. However, little is know about the applicability of text-based methods to measure political polarization to more specific policy contexts, potentially spanning only a couple of parliamentary debates (Grimmer and Stewart 2013). Hence, missing guidelines on the limitations of text-based methods in narrow research contexts constitute another potential obstacle in the use of these methods.

Evaluating the major findings of this chapter ultimately draws an ambivalent assessment. On the one hand, this literature review has revealed that methods measuring political polarization using text-as-data can be characterized as highly promising and theoretically founded, even though they usually focus on one specific sub-dimension of speech characteristics to operationalize po-

litical polarization. On the other hand, however, the high dimensional character of textual data might also question the usefulness of common methods. Hence, doubts have been expressed on the methodological validity and reliability of text-based measures due to the variety of possible theoretical and practical pitfalls.

Therefore, the question arises: Should scholars ultimately rely on text-as-data methods for measuring political polarization in applied research projects? And which aspects do they have to consider when applying text-based measures to ensure the reliability and validity of their findings? Having the goal to answer these interrelated research questions, the next chapter will introduce the research design, which provides an approach to compare the measurements of the most important methods identified for a use case of parliamentary speeches in German parliament during the Covid-19 pandemic.

4. Research Design

This chapter presents all relevant steps within the research design. It starts with describing the process of case selection and data retrieval (Section 4.1). Subsequently, it introduces all groups of methods as well as their implementations to measure political polarization for the data sample (Section 4.2). Lastly, it discusses relevant performance metrics and presents the validation framework (Section 4.3).

4.1 Case Selection and Data Retrieval

4.1.1 German Politics and the Covid-19 Pandemic

Having the goal to compare the performance of techniques to measure political polarization from textual data, the ideal data for this research design should be both characteristic for political debates in general and contain at least some varying levels of polarization to be measured. Moreover, the recommendations of Slapin and Proksch (2008) are followed, who argue that scholars should restrict their data sample to only one specific policy dimension (cf. section 3.3).

Hence, MP speeches in the German Bundestag centered around the epidemic control of the Covid-19 pandemic in Germany are defined as the main source of data.

There are several reasons for that. First, parliamentary debates were chosen since they contain direct, unmediated interactions between MPs on political issues (Wendler 2014). Therefore, it is argued that they are better suited for comparing NLP methods in comparison to more mediated channels, such as for instance written communication or formal requests. Second, one should also expect varying levels of political polarization throughout the course of the pandemic. Generally, Germany experienced similar pandemic progressions as other European countries. Between March and May 2020, the first wave of the pandemic reached its peak, leading to school closures, strict contact restrictions, and curfews imposed by the federal government from March 22 on (Bosen 2021)[1]. After a decline in infection rates and increased public pressure from May 2020 onwards (Dostal 2020), Germany experienced relatively low infection rates and a relaxation of government regulations throughout the summer months (cf. Figure 9 in the Appendix).

While the Covid-19 pandemic had a highly disruptive impact on German society, first empirical results assume the same impact on political discourse as well (Balmford et al. 2020; Louwerse et al. 2021; Juhl et al. 2021).

Most prominent, a recent qualitative study by Louwerse et al. (2021) examines the degree of German parliament polarization between February and July 2020. Their findings suggest a

1. From a legal perspective, decisions were primarily coordinated between the federal and state level. Except for issues of border affairs, the federal government usually issued formal guidelines based on the German Infection Protection Act (IfSG) (*Infektionsschutzgesetz*). These guidelines were then implemented into legally binding regulations by the state governments, cf. Saurer (2020)

parabolic progression, with more negative sentiments expressed towards the government at the beginning and the end of their reference period, and more positive sentiments on the peak of the first wave in March and April 2020. Their hand-coded estimates are consistent with the ”rally around the flag” literature, which argues that, in times of crisis, public support for governments usually increases and opposition parties position themselves closer to the government side (Chowanietz 2011; Lee 1977). In another recent study, Naumann et al. (2020) observe increased levels of political contestation in German politics once the infection rates started to decrease in May 2020 as well. Even though their paper mainly focuses on public reactions towards politics in general, their findings also provide evidence that the issues surrounding epidemic containment policies are a highly emotional issue and, thus, will likely lead to varying levels of political polarization on the level of both society and party politics.

4.1.2 Data Collection

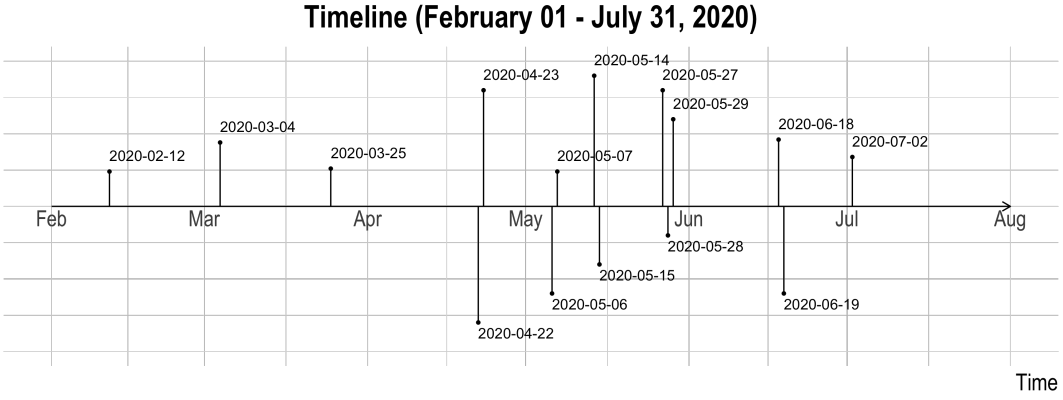
In order to test the hypotheses, an individually composed dataset of speech transcripts from the Bundestag is compiled. The dataset was collected using a gradual approach combining different publicly available data sources. As a starting point, data was retrieved from the non-profit project ”Open Discourse”, which covers 899.526 individual speeches and statements in the time period between 1949 and 2020 (Richter et al. 2020). While the administration of the Bundestag provides parliamentary protocols only in (non-machine readable or complex) PDF and XML format, Open Discourse provides a fully scraped database of every parliamentary speech held in the Bundestag together with some metadata, such as speaker characteristics or a direct link to the PDF protocol from the Bundestag Website.[2] However, in order to restrict the sample to only debates covering the handling of the Covid-19 pandemic, it is necessary to draw additional information from the Bundestag Information system (*Documentation and Information System for Parliamentary Materials of the German Bundestag*) (Bundestag 2021). On their website, the Bundestag provides an advanced search option to filter all operations, documents, and activities within the Bundestag. Making use of these filter options, the parliamentary materials were restricted to contain only individual speeches with the keywords [”Covid-19” AND ”Gesundheit” (*Health*)] held between February 01, 2020, and August 17, 2020 present in the Bundestag plenary records. Even though public officials have been discussing preventive measures before the first confirmed infection on January 27 (Rothe et al. 2020), the database lists no major debates before the time period selected. Ultimately, this restricted the search query to 344 individual speeches from 15 plenary debates meeting these conditions. In a second step, these search results were then exported as a word-document and scraped into tabular format.[3]. Finally, both data sets were merged on the variables *Date of Speech*, *MP Last Name*, and *MP Party*, thus keeping only those speeches in the dataset also present in the selection of the parliamentary materials. In four cases, a member of parliament would speak more than one time per debate, providing the algorithm with more than one unique option. In these cases, a manual approach was deployed and both data sources were merged manually by the author.

2. In order to access the dataset, please visit the official documentation at: https://open-discourse.github.io/open-discourse-documentation/1.0.0/index.html

3. For a detailed description, please visit the Code at the authors Github Profile online retrievable at: https://github.com/lukasbirki/Thesis

In a second screening round, descriptive statistics were inspected to identify data imbalances. Starting with the overall ratio of speeches, the data appears reasonably balanced, both on the aggregate level between government (49.4%) and opposition (50.6%) as well as on the level between parties per session (cf. Table 6.2 in the Appendix). Evaluating the within-debate ratio, no heavy imbalances were found as well, with MPs from the government speaking between 44% and 58% per debate. Considering the total count of speakers, all debates surpassed the threshold of fewer than 5 speeches per debate, which were defined as a lower limit (cf. Table 6.1 in the Appendix). Next, possible biases in speech length were assessed. Generally, the plotted histogram confirms a bell-shaped distribution of word counts per speech, with an average of 574 and a median of 607 words. However, 16 speeches were identified as outliers (less than 250 words or more than 1250 words) and subsequently removed (cf. Figure 10 in the Appendix). Figure 2 displays the temporal distribution of parliamentary debates throughout the time period. All speeches were held between

Figure 2. Timeline of Parliamentary Debates



February 12 and July 02, 2020, and are slightly clustered around April and July.[4].

Ultimately, a total of 326 speeches from 15 parliamentary debates were selected for further analysis. Figure 3 displays a summary of the total count of speeches per debate.

4.2 Measurement

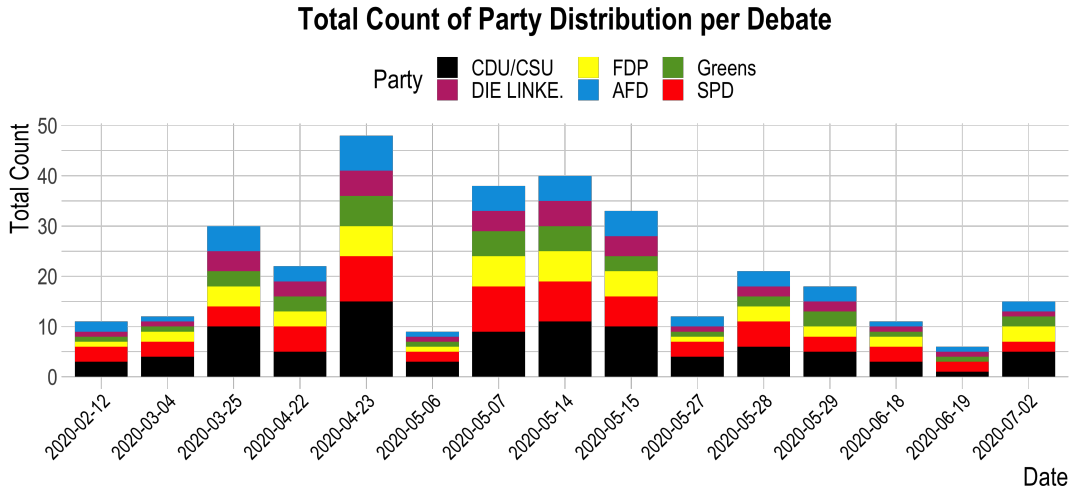
In principle, several aspects have to be considered when measuring latent constructs like political polarization. Hence, Section 4.2.1 first provides an overview of relevant preprocessing steps, followed by a detailed introduction of the respective measurement instruments in Section 4.2.2.

4.2.1 Data Preprocessing

In the field of NLP, *data preprocessing* describes a narrowly defined set of steps to reduce text complexity while preserving the substantive characteristics and information of the data (Denny

4. Between July 04 to September 06, no speeches were held due to the parliamentary summer recess in the Bundestag.

Figure 3. Total Count of Speeches per Bundestag Debate by Party



and Spirling 2018; Grimmer and Stewart 2013). Most commonly, these steps entail the tokenization of words, the removal of stopwords, the stemming of words to their most basic form, and (potentially) the inclusion of n-grams to capture word context (Zong et al. 2021). For this analysis, best practices are followed. First, all punctuation, numbers, and special characters were removed, since they are assumed to be uninformative for this analysis. Second, all words were lowercased and common greetings and procedural words at the beginning of the speeches were removed to avoid procedural phrases affecting the model results (cf Abercrombie and Batista-Navarro (2018)). Lastly, word lemmatization was applied to reduce the total amount of vocabulary used, which results in a decrease from 17,951 to 12,818 unique words in the speech sample.

4.2.2 Measuring Instruments

Sentiment Approach

To measure the polarity of parliamentary speeches based on a sentiment lexicon (SL), the publicly available German-language dictionary *SentiWS* was selected (Remus et al. 2010). The dictionary contains 1,650 negative and 1,818 positive words together with their inflections, which sums up to 32,734 words entries in the lexicon. For each word, the dictionary thus contains an assigned polarity weight ranging between -1 and $+1$ on a continuous scale.[5] Since the lexical entries all share "an affect-related meaning or connotation" (Remus et al. 2010, 1168), it is thus argued that the *SentiWS* lexicon constitutes a suitable resource to measure the sentiment polarity in parliamentary speeches. After assigning the sentiment scores to each word within the dataset, around 15.5% of words could be matched with a sentiment score. Lastly, the assigned values were then aggregated by taking the average for each individual speech. Table 6.3 in the Appendix displays the most frequent 50 words with and without an assigned sentiment score.

5. For the analysis, the algebraic signs were rearranged so that a low sentiment score represents low levels of polarization and vice versa.

Ideological Scaling Approach

In order to develop a measuring instrument based on the concept of ideological scaling (IS), the *Wordfish* model implemented by Slapin and Proksch (2008) was selected. Given the popularity of the model in academic research and its good accessibility, it is argued to be particularly well suited for being tested on the task of detecting political polarization from political speeches.

In order to calculate the polarisation scores, the index of the intensity of government and opposition (IGO) from Curini et al. (2020) is applied. The IGO index is based on the popular *Dalton* index of party system polarization (cf. Dalton (2008)) and provides a direct measurement of how far speakers and parties are apart on the latent ideological dimension for each parliamentary debate. The IGO index is calculated as such:

$$IGO_k = \sqrt{\sum_{j=1} VS_j * ([P_{jk} - \bar{P}_k]/6)^2} \tag{4.1}$$

where IGO_k is the value of the IGO index during the parliamentary debate k, VS_j is the share of party j in the legislative period, P_{jk} is the aggregated position of party j during debate k over the latent policy scale, and $\bar{P}_k$ is the average position of all speakers during debate k. Given that a debate k is usually centered around the same topics, large distances between parties might then indicate a more polarized setting of political discourse.

In order to feed the data into the model, the stemmed speeches were first transformed into a document-feature matrix, where each column represents one word, each row contains one speech and the values represent the total count of words in a given speech. Consistent with the literature, all terms which appear less than three times across all documents were removed in order to reduce the dimensionality of the matrix (cf. Lowe and Benoit (2013)). Ultimately, the model is implemented via the `quanteda` package by Benoit et al. (2018) for the R language.

Classification Accuracy Approach

To measure political polarization based on the classification accuracy (CA) of a machine learning approach, the aggregated classification accuracy between parties of the government and opposition is used as a proxy for political polarization for each debate. Hence, it is argued that the main conflict line of interest for the classifier is centered around a government-opposition dimension (cf. Curini and Zucchini (2012)). Theoretically, Wendler (2014) describes this phenomenon as characteristic for multiparty systems, since "institutional rules in parliamentary politics [leads MPs to be] either in favor of or against the coherence, effectiveness and goal attainment of national executives" (p. 554f).[6]

In order to compute the CA polarization scores, a simple linear stochastic gradient descent (SGD) classifier was implemented using the `Scikit-Learn` package (Pedregosa et al. 2011) in Python 3. Thereby, a log loss function and l2 regularization were applied, similar to the settings of Goet (2019) and Peterson and Spirling (2018). Following the advice of the `Scikit-Learn` manual for text classification, speeches were first transformed into a term-frequency times inverse

6. Furthermore, Wendler (2014) argues that this mechanism applies to heterogeneous ideological compositions of both the government and opposition side as well. This is because pragmatic disputes primarily based on a rationality principle dominate the plenary debates, instead of issues like normative beliefs (p.555)

document-frequency matrix, which diminishes the weight of terms that occur very frequently and are thus assumed to be less uninformative. [7]

In order to train the model, one challenge constitutes the small sample size (test set) of speeches on the Covid-19 debates ($n = 326$). Generally, machine learning theory states that the training and test sets are assumed to be drawn from the same probability distribution (Webb and Ting 2005; González and Abu-Mostafa 2015). A violation of this assumption, which is discussed in the literature under the term ”covariate shift” or ”dataset shift” (Quiñonero-Candela et al. 2009, 3), however, does not necessarily imply that researchers should abstain from using different datasets for testing and training in the presence of practical limitations. In some settings, a different dataset for training and testing might even improve model performance, especially if the training data is not fundamentally different from the test data (González and Abu-Mostafa 2015).

For this study, the SGD classifier was trained using all remaining speeches not selected for analysis (n = 18,510) in the 19th electoral term between October 24, 2017, and December 2017, 2020. Hence, it is assumed that the supervised character of the model will still enable the classifier to correctly identify characteristic interaction patterns between government and opposition. This might be supported by the fact that the ”training set” consists of speeches held in the same institutional setting, and the distributions between speeches from the government and opposition are assumed to be relatively stable across the electoral term.[8]

4.3 Method Validation

Validating the estimates of a scientific method covers the process ”to prove that [the method] consistently yields what it is expected [...] to do with adequate accuracy and precision” (Bruce et al. 1998, 93). In a broader sense, measurement validity thereby constitutes a fundamental methodological precondition to enable ”descriptive inference” in social science research (King et al. 1994, 34).

As outlined in section 3.3, validating the estimates of textual models is generally a difficult and multifaceted endeavor (Grimmer and Stewart 2013). This is especially relevant since even ”gold standard” estimates from human judgment are not immune to biases and subjective judgment, especially if the coders were not trained adequately on the respective task (Song et al. 2020). Hence, it is important to accept that any validation task of text-based measures will be inherently complex and contain some degree of uncertainty. In order to face these constraints, special attention has been put on a transparent and concise procedure to validate the methods selected. In doing so, this paper relies on an adapted validation framework by Goet (2019), which aims to test different dimensions of measurement validity, namely *face validity*, *convergent validity*, and *construct validity*. Table 4.1 depicts the validation framework visually.

Starting with the the first dimension, *face validity* describes the ”appropriateness, sensibility, or relevance of [a method]” (Holden 2010, 1). Thereby, the construct leaves room for interpre-

7. Cf. Scikit-Learn Manual, online retrievable at:https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

8. For a detailed discussion on the relevance of dataset shift when dealing with highly imbalanced distributions, cf. López et al. (2014)

Table 4.1. *Overview Validation Framework (modified from Goet (2019))*

Test	Key Question	Test
Face Validity		
1.1. General	Is there a reasonable level of stability of estimates from debate to debate?	Visual
1.2 Detailed	Do the estimates correspond with a priori expectations of levels of polarization?	Visual
Convergent Validity		
2.1 Debate level	Do the debate-level estimates of polarization correspond with a comparable exogenous measure?	Correlation
Construct Validity (conditional)		
3.1 - 3.3 Consistency	Does a method-specific assessment provides evidence of consistent estimates?	Analytic

tation and distinguishes itself from more technical and empirical assessment methods (Quinn et al. 2010). For this analysis, two tests of face validity are applied:

> **1.1 General Test**: The level of polarization should be at a reasonable level. It is expected, that there is no at-random pattern of switches between periods of high and low levels of polarization across debates.

As noted by Goet (2019), objectively assessing the stability of estimates on a general level might be difficult, especially since this paper aims to measure polarization within a comparatively short time period. Therefore, a more concrete detailed test is applied as well, which compares the method estimates against a priori expected levels of polarization based on the overall pandemic progression in Germany.

Based on Dostal (2020), two major turning points of public mood are identified, which are displayed in Table 4.2. The first turning point T_1 is centered around the beginning of March and the second turning point T_2 around the beginning of May. In terms of polarization, it is expected that political polarization should decrease after the first turning point T_1, given that rising infection rates and a national health crisis were foreseeable. After the second turning point T_2, however, it is expected that polarization would increase due to concerns over the escalating economic costs and infringement of fundamental basic rights. Therefore, it is said:

> **1.2 Detailed Test**: The level of polarization should correspond with a priori expectations derived from authoritative (secondary) sources.

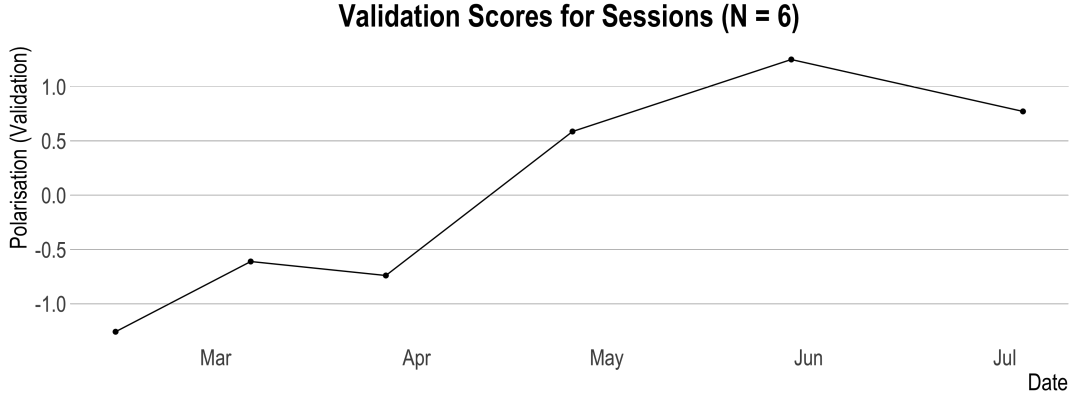
Next, *convergent validity* describes how much the estimates of a method correspond to another external, unrelated measure (Trochim and Donnelly 2001). For this study, the measured levels of polarization are compared against hand-coded assessments by Louwerse et al. (2021). In their paper, they examine expressed opposition in four countries across parliamentary debates that deal with the direct response to the Covid-19 crisis. For Germany, they hand-code 37 speeches across six plenary debates between February 12 and June 25, 2020, on the level of single paragraphs

Table 4.2. *Detailed Test: Expected Level of Polarization (cf. Dostal (2020))*

Turning Point	Months	Main Events	Polarization
$< T_1$	February	Public ignorance about the potential severity of the coronavirus threat; Occasional mentioning of the Virus in national media with a regional focus on China	Medium
$> T_1$ & $< T_2$	March, April	March 11 - Chancellor Merkel's press conference, stating that 60 to 70 per cent of Germans would at some point or other be infected with Covid-19; Lockdown from March 22 on; Rising infection rates (maximum value: 6992 new cases on April 6, cf. Figure 9 in the Appendix), Fear of the collapse of healthcare facilities; Economic bailouts for all sectors of society	Low
$> T_2$	May, June, July	Shift of public mood; Concerns over the escalating economic costs and ongoing restrictions to constitutional basic rights; Decreasing infection rates; Increased levels of party politics	High

using a five-point scale of opposition party expressed sentiment. Since the authors convincingly demonstrate a methodological consistent coding process, it is argued that their estimates provide a well-suited benchmark for comparison. Figure 4 depicts the validation scores from Louwerse et al. (2021) visually [9]. Therefore, convergent validity is determined as such:

Figure 4. Validation Scores (based on Louwerse et al. 2021)



9. As one can see, there are some minor differences between the expected levels of polarization based on Dostal (2020) and the hand-coded debate scores from Louwerse et al. (2021). In particular, the debate scores show a growing trend of polarization for April 2020. According to the argumentation of Dostal (2020), this rise of political polarization is observed only after April 2020. However, given that both studies focus on different aspects of political polarization (public vs. specific debate polarization), it can be acknowledged that both scores appear to be plausible and do not necessarily exclude each other

2.1 Debate Level: Do the debate-level estimates of polarization correspond with a comparable exogenous measure on the individual level?

Lastly, Goet (2019) also examines *construct validity* as a metric of method consistency. However, the specific tests conducted in his analysis cannot be replicated due to the diversity of approaches tested in this analysis. Therefore, all methods which successfully passed the test of *face* and *convergent validity* will be put to a final and method-specific consistency test. Given that construct validity "is a multifaceted process [to prove that a method is consistent] [...] that is never finished (Grimm and Widaman 2012, 621), it is hence argued that an individual analytical evaluation of the method's internal mechanisms can provide important insights into the internal consistency of the estimates. Therefore, it is determined:

3.1 - 3.3 Consistency Test (method-specific): Does an *individual* and method-specific assessment provides evidence of consistent estimates?

5. Results

This chapter presents and validates the estimated levels of polarization for all methods identified. These are: the sentiment scores derived from the sentiment dictionary (SD), the ideological distance of parties as implemented in the *Wordfish* model (IS), and the classification accuracy (CA) of the SGD classifier which is used as a proxy for political polarization.

Generally, results show that the estimated levels of polarization vary widely for the time period under study and do not correspond closely to each other ($\rho_{SL-IS} = -.09$, $\rho_{SL-CA} = .01$, $\rho_{IS-CA} = -.32$). These results provide strong evidence that a more careful validation of the methods and their estimates appears to be necessary.

In order to do so, Table 5.1 summarizes the test results for all methods inspected using the validation framework. Hence, the findings of the validation exercise are subsequently discussed in more detail.

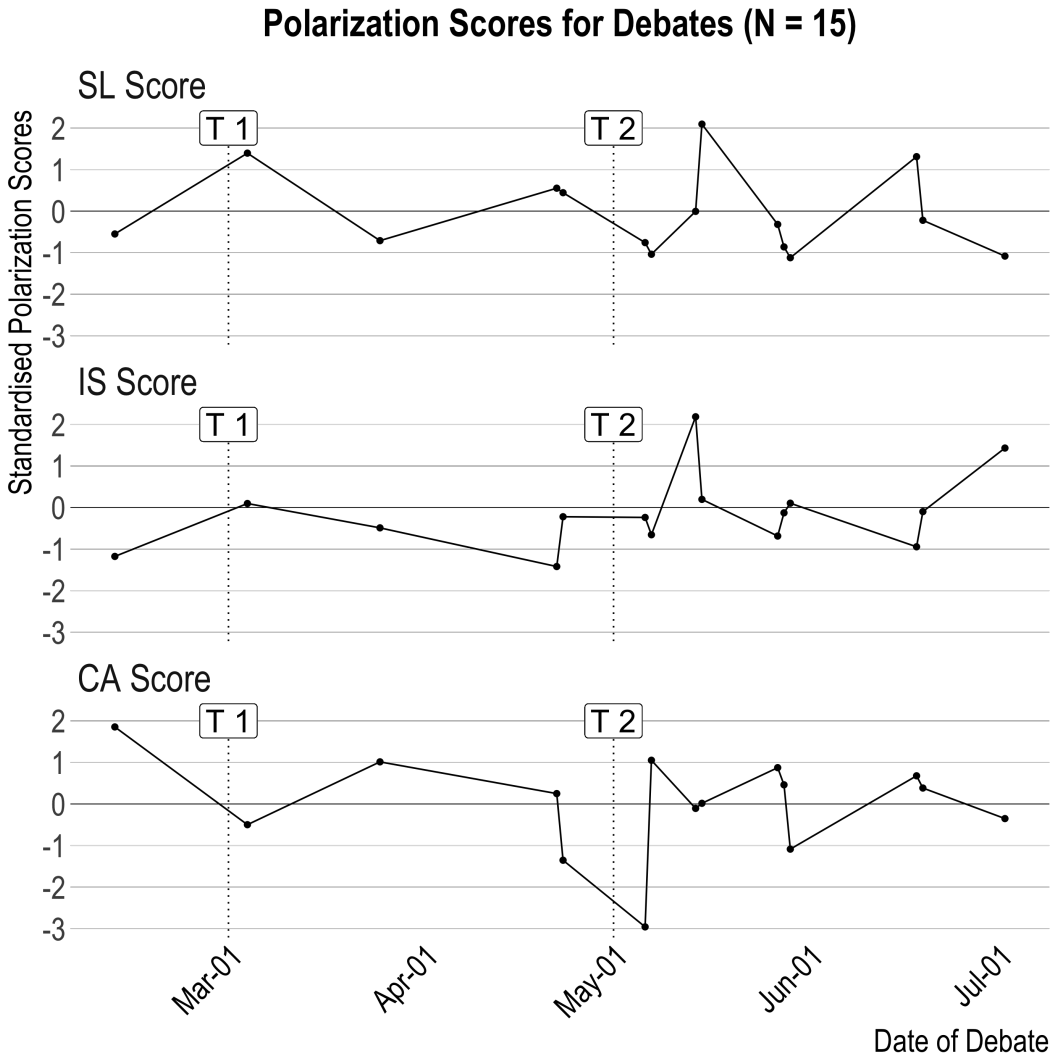
Table 5.1. *Results Validation Exercise*

Validation	**Test**	**SL**	**IS**	**CA**
Face Validity	1.1 General Test	✓	✓	✓/✗
	1.2 Detailed Test	✓/✗	✓/✗	✗
Convergent Validity	2.1 Debate Level Test	✗	✓	✗
Construct Validity	3.1 Consistency Test: SL	n.a.	-	-
(Method specific)	3.2 Consistency Test: IS	-	✓	-
	3.3 Consistency Test: CA	-	-	n.a.

Note: n.a. = not applicable (the methods did not pass the test of *face-* and *convergent validity*)

Starting with the first criterion of *face validity*, the **general test (1.1)** considers the stability of estimates from debate to debate. Figure 5 displays the z-standardized polarisation scores for all three methods. Generally, one can observe varying levels of polarization for all methods tested, with several peaks and lows across the inspected time period between February to July. For the SL Scores, the biggest variations can be located throughout May and June, with a maximum level of parliament polarization on May 15. The estimates for the IS score appear to be comparatively more stable, given that the estimates are less fluctuating. Similar to the SL estimates, they also predict maximum levels of parliament polarization around the middle of May. For the AC Scores, results seem to include some variation as well. In particular, the most remarkable change in parliament polarization can be observed from May 6 (-2.96) to May 7 (1.05). This drastic increase might question whether the method is capturing aspects of the textual data different than parliament polarization.

Figure 5. Comparing Method Estimates



Note: All method estimates were z-standardized to enable a visual comparison.

Based on this first impression, the **detailed test (1.2)** furthermore enables a more concrete evaluation of the method estimates. The expected turning points T_1 and T_2 are thereby represented by the dashed lines in Figure 5.

Starting with the first turning point, it is thus expected that parliament polarization should decrease after T_1 due to the national health crises caused by the threatening pandemic progression in Germany.

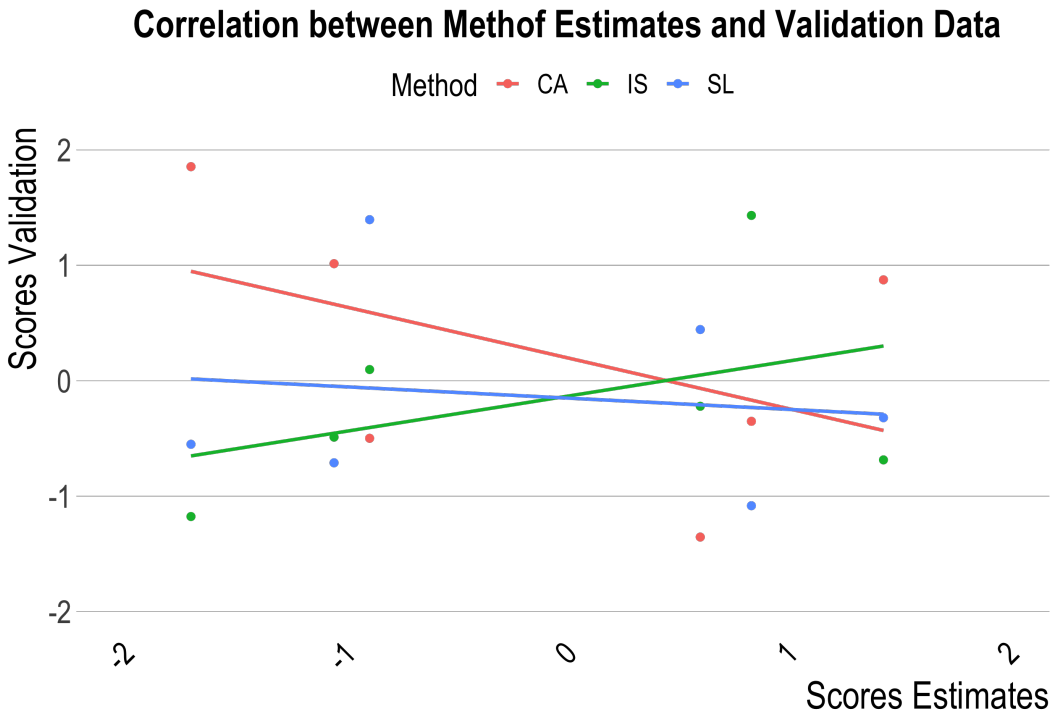
Empirically, this trend seems to be captured adequately by the relatively low IS Scores, and partially by the SL score, which again shows increasing levels of polarization for the parliamentary debates on April 22 and 23. The CA scores, on the other hand, appear to be contrary to the expected levels of polarization. Instead, they suggest increased levels of parliament polarization throughout March and April.

For the second turning point T_2, it is expected that parliament polarization would increase due to the continuing lockdown and increased public pressure. Assessing the method estimates

thereby draws a rather ambiguous picture. Generally, both SL and IS Scores measure a peak of polarization around mid of May, even though it is striking that the peaks predicted from the SL (May 14) and IS (May 15) are on different debates. Nevertheless, especially the IS seems to be able to capture increased levels of polarization, which might indicate that, after the second turning point T_2, MPs would speak more dissimilar and, thus, could be placed further apart on the latent ideological scale. For the AC Score, estimates remain ambiguous, since the method predicts a heavy drop of polarization around the beginning of May, which would be contrary to the a priori assumptions.

Moving on to the criterion of *convergent validity*, the **debate level test (2.1)** compares the model estimates against the hand-coded estimates from Louwerse et al. (2021). Results show that only the IS scores correlate positively with the validation data $\rho_{Validation-IS} = .42$, even though the results are not statistical significant ($p = 0.40$) (cf. Figure 11 in the Appendix for the complete correlation matrix). One obvious explanation for this might be connected to the small sample size ($n = 6$), which should generally prevent statistically significant correlation estimates. Nevertheless, the IS scores appear to share at least some variation with the hand-coded estimates, which provides further evidence that the method might be able to arrive at trustworthy estimates for the debates under study. Surprisingly, both the SL Scores $\rho_{Validation-SL} = -.13$ and CA Scores $\rho_{Validation-CA} = -.46$ correlate negatively with the validation data. This, in turn, provides strong evidence that both methods are capturing fundamentally different aspects of the textual data different from parliament polarization. Figure 6 displays the correlations visually.

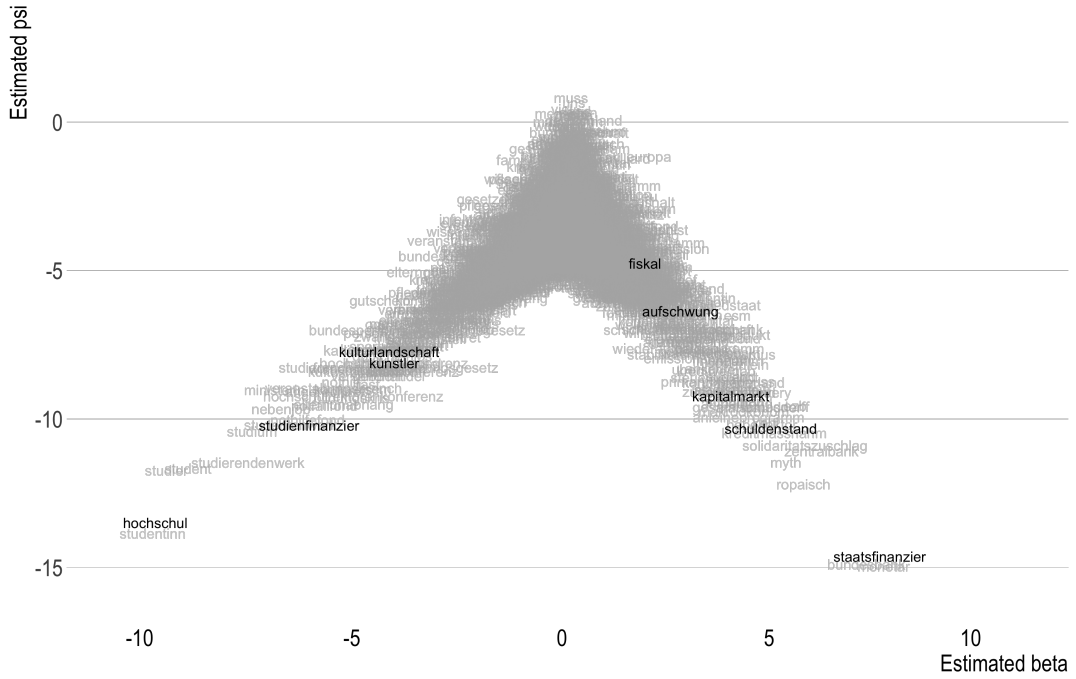
Figure 6. 2.1 Debate Level Test: Visual Interpretation



Note: All correlations are not significant at the < 0.05 level.

Given the questionable test results for both the SL and CA scores, only the IS estimates were put to a final **consistency test (3.2)** in order to explore the issue of *construct validity*. Thereby, the latent ideological scale derived from the Wordfish model is evaluated to examine whether the assumptions of a single underlying policy dimension is satisfied. More specifically, it is asked whether the individual speech positions show a clear separation between parties in the German Bundestag. Figure 7 displays the scatterplot for the estimated word positions from the Wordfish model. The x-axis represents the importance of a given word in determining the orientation of the

Figure 7. 3.2 Consistency Test - IS: Scatterplot

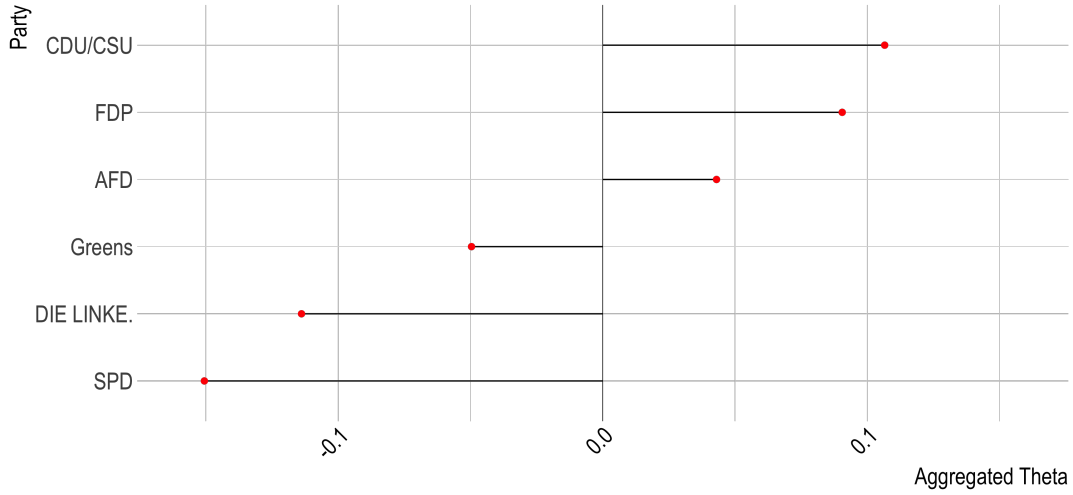


Note: The graph is generated using the `quanteda` package.

word on the latent scale (beta), whereas the y-axis represents the frequency of the word (psi). The two opposing sides provide evidence for a separation between more social and economic terms, with (stemmed) words like ”hochschul” (university), ”studienfinanzier” (student finance), ”kunstler” (artist) and ”kulturlandschaft” (cultural landscape) on the one side, and ”staatsfinanzier” (public finance), ”schuldenstand”(debt level), ”kapitalmarkt” (capital market), ”aufschwung” (upswing) and ”fiskal” (fiscal) on the other side.

Generally, this might support the assumption that the scale is indeed dominated by ideology, given that a more conservative position might highlight the importance of solid public finances and the negative side effects of government spending, whereas a more social position might stress the importance of helping out those people affected by the pandemic, such as e.g. artists, students and low-income families (cf. Grafton and Permaloff (2005)). This assumption is further strengthened when aggregating the individual speech positions for each party. Figure 8 depicts the aggregated scores visually. Hence, results show that the conservative CDU/CSU, the liberal FDP, and right-wing AFD are all placed on the right-hand side of the scale, whereas the Greens, the left LINKE, and the social-democrats SPD are placed on the left-hand side. Given these results, it is thus

Figure 8. 3.2 Consistency Test - IS: Mean Theta per Party



argued that the internal mechanisms associated with the IS Wordfish model are consistent with its model assumptions, and the estimated levels of polarization appear to show at least some degree of construct validity.

6. Conclusion

6.1 Discussion

Measuring the level of polarization between political actors is a fundamental condition for the analysis of politics and parliamentary dynamics in political science research (Curini and Zucchini 2012; Grimmer and Stewart 2013). In doing so, a growing body of literature has shifted its attention to the analysis of textual data to operationalize polarization while relying on methods from computational linguistics and statistics. Hence, advocates of these approaches emphasize the enormous liberating potential for political science research, such as the reduction of research expenses and the analysis of formerly impenetrable data. However, doubts can be expressed on the measurement validity of current approaches. Motivated by ambivalent empirical findings in this field (Goet 2019; Gentzkow, Shapiro, et al. 2019), this paper, thus, has taken a more critical perspective. Thereby, the goal of this analysis has been to compare and validate different methods on an applied use case of German parliamentary speeches during the Covid-19 pandemic. In doing so, measurements of political polarization from three different groups of methods were compared and validated using an adjusted validation framework (cf. Goet (2019)).

Results from the validation exercise draw an inconclusive picture. As summarized in Table 5.1, both SL and CA methods clearly failed relevant tests in the validation exercise, thereby raising serious concerns about the applicability of these methods in similar study contexts. The IS method, on the other hand, showed more promising results. Estimated levels of polarization passed not only the tests of *face validity* but also correlated positively with the external validation scores by Louwerse et al. (2021) (*convergent validity*). Hence, estimates based on the ideological scaling model Wordfish indeed appear to be able to capture relevant aspects of parliament polarization. Moreover, the consistency test (*construct validity*) provided additional arguments that the Wordfish model was able to place political speeches on a meaningful ideological scale, one important precondition for the calculation of the IGO index and the calculation of the IS scores. Nevertheless, it cannot be claimed that the method estimates can be trusted without more ado, given that construct validity is a multi-faceted construct (Grimm and Widaman 2012).

To summarize, it can be noted that the recent enthusiasm regarding approaches using text-as-date for political science research is not completely unfounded. The quantitative analysis of textual data does enable researchers to analyze large amount of data to explore and measure issues like political polarization. Thereby, this analysis has shown that these methods can indeed add value to applied research projects, even in relatively restricted research contexts. However, given the empirical evidence, a wide range of possible methodological pitfalls remain. Admittedly, even the IS estimates (viz. the only method which passed the criterion of *face-, convergent* and *construct validity*) provide rather vague tendencies of measurement validity. Instead, it appears that current methods in this field still rely on extensive validation work to ensure the appropriateness and measurement validity necessary for academic inference. Therefore, it is concluded that a

cautious and transparent procedure in connection with a convincing validation exercise constitutes a necessary precondition for researchers who want to apply text-as-data methods in future real-world research contexts.

6.2 Limitations and Outlook

Even though this analysis is based on a transparent procedure and strong methodological fundament, several limitations need to be addressed in this section.

Starting with the first limitation, it cannot be ruled out that preprocessing decisions significantly influenced the respective method estimates. As shown in a recent study by Denny and Spirling (2018), data preprocessing can have a profound impact on method estimates from textual data, such as for instance the Wordfish text scaling model. In order to tackle this issue, this analysis has relied on a transparent and established set of common preprocessing steps, such as the tokenization of words, the removal of stopwords and procedural phrases as well as the stemming of words to their most basic form (Zong et al. 2021). However, the sensitivity of model estimates to varying preprocessing regimes still constitutes an important shortcoming, especially for methods that are very sensitive to changes in the document-feature matrix.

Next, the interpretation and comparison of different measures of political polarization remains restricted to the specific study context. There are two interrelated reasons for that. First, this analysis is not interested in the absolute estimates of political polarization. Instead, the main metric of interest is rather the *relative* score changes of one method in relation to the change of another method for one specific time period (Peterson and Spirling 2018). On the other hand, both IS and CA methods measure polarization in relation to the whole document corpus. Consequently, they are likely to be sensitive to changes in the overall aggregated dataset caused by different time periods. Due to these reasons, the explanatory power of the analysis is limited, given that interpretations of the model estimates account only for the specific time period and in relation to the other methods.

Additionally, doubts can be expressed on the choice of validation data. Generally, one challenge constitutes the availability of high-quality (peer-reviewed) validation data for the debates under study. Therefore, only the hand-coded estimates from Louwerse et al. (2021) were used to determine *convergent validity*. Optimally, researchers can instead rely on a set of different validation sources to assess measurement validity from various perspectives. On top of that, one potential problem of the data by Louwerse et al. (2021) might constitute that the authors only code speeches from opposition parties. However, it is argued that the debate-level estimates should be consistent with the overall level of parliament polarization since the opposition usually takes an active role to influence conflict in times of crisis: Either "to cooperate with the government in order to overcome the crisis [or take advantage of] the weakened government to stress their adversarial position" (Moury and De Giorgi 2015, 1).

Besides, little is still know about the institutional dynamics and their impact on text-based estimates. Referring to earlier work from Eggers and Spirling (2014) and Benedetto and Hix (2007), Goet (2019) notes that how MPs engage in parliamentary debates is subject to a diverse set of factors— "dynamics which [models using text-as-data] should incorporate" (p. 538). For instance,

parliamentary speeches might be influenced by the roles of procedure, such as overlapping items on the agenda, direct reactions to preceding speeches, or daily developments. Furthermore, little is known about the institutional dynamics in parliament during a worldwide pandemic like the one caused by the SARS-CoV-2 virus (Rayment and VandenBeukel 2020). For instance, it could be conceivable that debated topics from one policy domain will be dominated by words related to the pandemic, and these words will be interpreted differently by the respective measurement method. Given that the methods applied in this study did not take these information into account, a large amount of variation in the data will consequently be explained by external factors, and will likely lead to biased estimates when these factors are imbalanced.

Despite these limitations, this study might stimulate future research beyond the scope of this analysis. As Goet (2019) notes, the use of speech data to measure parliament polarization is still in its infancy. Therefore, researchers are encouraged to test and validate these methods in different study contexts and settings, thereby contributing to the growing body of literature in this field. In particular, scholars are encouraged to challenge these methods in specific research contexts beyond highly aggregated data in order to examine whether text-based methods ultimately add value to applied political science research. Besides, scholars should also apply more advanced methods beyond simple bag-of-words approaches. Especially recent developments in deep learning (cf. Devlin et al. (2018) promise a better and more holistic semantic understanding of written texts.

Ultimately, time will show whether methods using text-as-date to measure political polarization will come up to the high expectations they promise. At this point, evidence from this analysis at least questions the measurement validity of the methods under study. However, it can only be hoped that more advanced methods in this field will contribute to a better understanding of political polarization in the future. Because one thing has become clear: Political polarization is likely to increase in the next decades, and political science has an obligation to study its causes and consequences to prevent harmful developments for societies worldwide.

Bibliography

Abercrombie, Gavin, and Riza Batista-Navarro. 2018. 'aye'or 'no'? speech-level sentiment analysis of hansard uk parliamentary debate transcripts. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).*

———. 2020. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science* 3:1–26.

Adams, James F., Samuel Merrill III, and Bernard Grofman. 2005. *A unified theory of party competition: a cross-national analysis integrating spatial and behavioral factors.* Cambridge University Press. ISBN: 113944400X.

Adcock, Robert, and David Collier. 2001. Measurement validity: a shared standard for qualitative and quantitative research. *American political science review* 95 (3): 529–546.

Ahmadalinezhad, Mahboubeh, and Masoud Makrehchi. 2018. Detecting agreement and disagreement in political debates. In *Social, cultural, and behavioral modeling,* edited by Robert Thomson, Christopher Dancy, Ayaz Hyder, and Halil Bisgin, pp. 54–60. Cham: Springer International Publishing. ISBN: 978-3-319-93372-6.

Arbatli, Ekim, and Dina Rosenberg. 2021. United we stand, divided we rule: how political polarization erodes democracy. *Democratization* 28 (2): 285–307.

Balahur, Alexandra, Zornitsa Kozareva, and Andrés Montoyo. 2009. Determining the polarity and source of opinions expressed in political debates. In *Computational linguistics and intelligent text processing,* edited by Alexander Gelbukh, pp. 468–480. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-00382-0.

Balmford, Ben, James D Annan, Julia C Hargreaves, Marina Altoè, and Ian J Bateman. 2020. Cross-country comparisons of covid-19: policy, politics and the price of life. *Environmental and Resource Economics* 76 (4): 525–551.

Baturo, Alexander, Niheer Dasandi, and Slava J. Mikhaylov. 2017. Understanding state preferences with text as data: introducing the un general debate corpus. *Research & Politics* 4 (2): 1–9.

Bayley, Paul. 2004. *Cross-cultural perspectives on parliamentary discourse.* Vol. 10. John Benjamins Publishing.

Beelen, Kaspar, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, and Ludovic Rheault. 2017. Digitization of the canadian parliamentary debates. *Canadian Journal of Political Science/Revue canadienne de science politique* 50 (3): 849–864.

Benedetto, Giacomo, and Simon Hix. 2007. The rejected, the ejected, and the dejected: explaining government rebels in the 2001-2005 british house of commons. *Comparative Political Studies* 40 (7): 755–781.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. Quanteda: an r package for the quantitative analysis of textual data. *Journal of Open Source Software* 3 (30): 774. https://quanteda.io.

Bergmann, Henning, Hanna Bäck, and Thomas Saalfeld. 2021. Party-system polarisation, legislative institutions and cabinet survival in 28 parliamentary democracies, 1945–2019. *West European Politics,* 1–19.

Biessmann, Felix. 2016. Automating political bias prediction. *arXiv preprint arXiv:1608.02195.*

Bosen, Jens, Ralf; Thurau. 2021. Chronology: how covid has spread in germany. *Deutsche Welle* (June 29, 2021). Accessed August 12, 2021. https://www.dw.com/en/chronology-how-covid-has-spread-in-germany/a-58026877.

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro. 2017. Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences* 114 (40): 10612–10617.

Bruce, P, P Minkkinen, and M-L Riekkola. 1998. Practical method validation: validation sufficient for an analysis method. *Microchimica Acta* 128 (1): 93–106.

Bundestag, German. 2021. Documentation and information system for parliamentary materials of the german bundestag, accessed August 12, 2021. https://dip.bundestag.de/erweiterte-suche?f.wahlperiode=19&rows=25.

Burns, Charlotte. 2019. In the eye of the storm? the european parliament, the environment and the eu's crises. *Journal of European Integration* 41 (3): 311–327.

Campbell, James E. 2018. *Polarized: making sense of a divided america.* Princeton University Press. ISBN: 0691180865.

Caramani, Daniele. 2017. Will vs. reason: the populist and technocratic forms of political representation and their critique to party government. *American political science review* 111 (1): 54–67.

Catt, Helena. 2002. *Democracy in practice.* Routledge. ISBN: 9780415168403.

Chowanietz, Christophe. 2011. Rallying around the flag or railing against the government? political parties' reactions to terrorist acts. *Party Politics* 17 (5): 673–698.

Cornelson, Kirsten, and Boriana Miloucheva. 2020. *Political polarization, social fragmentation, and cooperation during a pandemic.* University of Toronto, Department of Economics.

Corrales, Javier. 2005. In search of a theory of polarization: lessons from venezuela, 1999-2005. *European Review of Latin American and Caribbean Studies/Revista Europea de Estudios Latinoamericanos y del Caribe,* no. 79, 105–118.

Curini, Luigi, Airo Hino, and Atsushi Osaka. 2020. The intensity of government–opposition divide as measured through legislative speeches and what we can learn from it: analyses of japanese parliamentary debates, 1953–2013. *Government and Opposition* 55 (2): 184–201.

Curini, Luigi, and Francesco Zucchini. 2012. Government alternation and legislative party unity: the case of italy, 1988–2008. *West European Politics* 35 (4): 826–846. ISSN: 0140-2382.

Dalton, Russell J. 2008. The quantity and the quality of party systems: party system polarization, its measurement, and its consequences. *Comparative Political Studies* 41 (7): 899–920.

Denny, Matthew J, and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis* 26 (2): 168–189.

Deutsch, Morton. 1971. Conflict and its resolution. *Conflict Resolution: Contributions of the Behavioral Sciences.*

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

DiMaggio, Paul, John Evans, and Bethany Bryson. 1996. Have american's social attitudes become more polarized? *American journal of Sociology* 102 (3): 690–755.

Dostal, Jörg Michael. 2020. Governing under pressure: german policy making during the coronavirus crisis. *The Political Quarterly* 91 (3): 542–552.

Downs, Anthony. 1957. An economic theory of democracy.

Dzieciątko, Mariusz. 2019. Application of text analytics to analyze emotions in the speeches. In *Information technology in biomedicine,* edited by Ewa Pietka, Pawel Badura, Jacek Kawa, and Wojciech Wieclawek, 525–536. Cham: Springer International Publishing. ISBN: 978-3-319-91211-0.

Easton, David. 1965. A framework for political analysis. *Englewood Cliffs, HJ: Prentice-Hall.*

Eggers, Andrew C, and Arthur Spirling. 2014. Electoral security as a determinant of legislator activity, 1832–1918: new data and methods for analyzing british political development. *Legislative Studies Quarterly* 39 (4): 593–620.

Eidelson, Roy J., and Judy I. Eidelson. 2003. Dangerous ideas: five beliefs that propel groups toward conflict. *American psychologist* 58 (3): 182–192.

Enyedi, Zsolt. 2016. Populist polarization and party system institutionalization: the role of party politics in de-democratization. *Problems of Post-communism* 63 (4): 210–220.

Esteban, Joan-Maria, and Debraj Ray. 1994. On the measurement of polarization. *Econometrica: Journal of the Econometric Society,* 819–851.

Frech, Elena, Niels D. Goet, and Simon Hug. 2018. Shirking and slacking in parliament. *Legislative Studies Quarterly* 46 (2): 493–523. ISSN: 0362-9805.

Frid-Nielsen, Snorre Sylvester. 2018. Human rights or security? positions on asylum in european parliament speeches. *European union politics* 19 (2): 344–362. ISSN: 1465-1165.

Galston, William A. 2018. The populist challenge to liberal democracy. *Journal of Democracy* 29 (2): 5–19. ISSN: 1086-3214.

Gentles, Stephen J., Cathy Charles, Jenny Ploeg, and K. Ann McKibbon. 2015. Sampling in qualitative research: insights from an overview of the methods literature. *The qualitative report* 20 (11): 1772–1789. ISSN: 1052-0147.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature* 57 (3): 535–74.

Gentzkow, Matthew, Jesse Shapiro, and Matt Taddy. 2016. *Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech.* Working Papers id:11114. eSocialSciences, July. https://ideas.repec.org/p/ess/wpaper/id11114.html.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica* 87 (4): 1307–1340.

Goet, Niels D. 2019. Measuring polarization with text analysis: evidence from the uk house of commons, 1811–2015. *Political analysis* 27 (4): 518–539. ISSN: 1047-1987.

Goldberg, Andreas C., Erika J. van Elsas, and Claes H. de Vreese. 2020. Mismatch? comparing elite and citizen polarisation on eu issues across four countries. *Journal of European Public Policy* 27 (2): 310–328. ISSN: 1350-1763.

Golosov, Grigorii V. 2015. Factors of party system fragmentation: a cross-national study. *Australian journal of political science* 50 (1): 42–60.

González, Carlos R, and Yaser S Abu-Mostafa. 2015. Mismatched training and test distributions can outperform matched ones. *Neural computation* 27 (2): 365–387.

Grafton, Carl, and Anne Permaloff. 2005. Liberal and conservative dissensus in areas of domestic public policy other than business and economics. *Policy Sciences* 38 (1): 45–67. ISSN: 00322687, 15730891. http://www.jstor.org/stable/4532650.

Grechyna, Daryna. 2016. On the determinants of political polarization. *Economics Letters* 144 (April). https://doi.org/10.1016/j.econlet.2016.04.018.

Grimm, Kevin J, and Keith F Widaman. 2012. Construct validity.

Grimmer, Justin, and Brandon M. Stewart. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21 (3): 267–297. ISSN: 1047-1987.

Hardwicke, Tom E, Maya B Mathur, Kyle MacDonald, Gustav Nilsonne, George C Banks, Mallory C Kidwell, Alicia Hofelich Mohr, Elizabeth Clayton, Erica J Yoon, Michael Henry Tessler, et al. 2018. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal cognition. *Royal Society open science* 5 (8): 180448.

Harris, Zellig S. 1954. Distributional structure. *Word* 10 (2-3): 146–162. ISSN: 0043-7956.

Haselmayer, Martin, and Marcelo Jenny. 2017. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & Quantity* 51 (6): 2623–2646. ISSN: 1573-7845.

Hetherington, Marc J., and Jonathan D. Weiler. 2009. *Authoritarianism and polarization in american politics.* Cambridge University Press. ISBN: 0521884330.

Holden, Ronald R. 2010. Face validity. In *The corsini encyclopedia of psychology,* 1–2. American Cancer Society. ISBN: 9780470479216. https://doi.org/https://doi.org/10.1002/9780470479216.corpsy0341.

Hopkins, Daniel J., and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54 (1): 229–247. ISSN: 0092-5853.

Idelberger, Felix. 2020. Do the populists have a say? estimating their effect on topic prevalence and polarization in german state legislature. *SocArXiv* (September). https://doi.org/10.31235/osf.io/y5j2u.

Jatowt, Adam, and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Ieee/acm joint conference on digital libraries,* 229–238. https://doi.org/10.1109/JCDL.2014.6970173.

Jensen, Jacob, Suresh Naidu, Ethan Kaplan, Laurence Wilse-Samson, David Gergen, Michael Zuckerman, and Arthur Spirling. 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. ISSN: 0007-2303.

Jiang, Julie, Emily Chen, Shen Yan, Kristina Lerman, and Emilio Ferrara. 2020. Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies* 2 (3): 200–211.

Juhl, Sebastian, Roni Lehrer, Annelies G Blom, Alexander Wenz, Tobias Rettig, Ulrich Krieger, Marina Fikel, Carina Cornesse, Elias Naumann, Katja Möhring, et al. 2021. Preferences for centralized decision-making in times of crisis: the covid-19 pandemic in germany. In *Jahrbuch für handlungs- und entscheidungstheorie: band 11,* edited by Marc Debus, Markus Tepe, and Jan Sauermann. Springer Fachmedien Wiesbaden.

Kim, In Song, John Londregan, and Marc Ratkovic. 2018. Estimating spatial preferences from votes and text. *Political analysis,* ISSN: 1047-1987.

King, Gary, Robert O Keohane, and Sidney Verba. 1994. *Designing social inquiry.* Princeton university press.

Kiran, Garimella, and Ingmar Weber, eds. 2017. *A long-term analysis of polarization on twitter.* Vol. 1. ISBN: 2334-0770.

Lauderdale, Benjamin E., and Alexander Herzog. 2016. Measuring political positions from legislative speech. *Political analysis,* 374–394. ISSN: 1047-1987.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American political science review* 44 (3): 311–331. ISSN: 0003-0554.

Laver, Michael, and John Garry. 2000. Estimating policy positions from political texts. *American Journal of Political Science,* 619–634.

LeBas, Adrienne. 2006. Polarization as craft: party formation and state violence in zimbabwe. *Comparative Politics,* 419–438. ISSN: 0010-4159.

Lee, Jong R. 1977. Rallying around the flag: foreign policy events and presidential popularity. *Presidential Studies Quarterly* 7 (4): 252–256.

Liberati, Alessandro, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P. A. Ioannidis, Mike Clarke, Philip J. Devereaux, Jos Kleijnen, and David Moher. 2009. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology* 62 (10): e1–e34. ISSN: 0895-4356.

Liu, Dilin, and Lei Lei. 2018. The appeal to political sentiment: an analysis of donald trump's and hillary clinton's speech themes and discourse strategies in the 2016 us presidential election. *Discourse, Context & Media* 25:143–152. ISSN: 2211-6958.

López, Victoria, Alberto Fernández, and Francisco Herrera. 2014. On the importance of the validation technique for classification with imbalanced datasets: addressing covariate shift when data is skewed. *Information Sciences* 257:1–13.

Loughran, Tim, and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance* 66 (1): 35–65.

Louwerse, Tom, Ulrich Sieberer, Or Tuttnauer, and Rudy B Andeweg. 2021. Opposition in times of crisis: covid-19 in parliamentary debates. *West European Politics,* 1–27.

Lowe, Will, and Kenneth Benoit. 2013. Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political analysis* 21 (3): 298–313. ISSN: 1047-1987.

Maoz, Zeev, and Zeynep Somer-Topcu. 2010. Political polarization and cabinet stability in multiparty systems: a social networks analysis of european parliaments, 1945-98. *British Journal of Political Science,* 805–833. ISSN: 0007-1234.

Mayring, Philipp. 2015. Qualitative content analysis: theoretical background and procedures. In *Approaches to qualitative research in mathematics education: examples of methodology and methods,* edited by Angelika Bikner-Ahsbahs, Christine Knipping, and Norma Presmeg, 365–380. Dordrecht: Springer Netherlands. ISBN: 978-94-017-9181-6.

McCoy, Jennifer, Tahmina Rahman, and Murat Somer. 2018. Polarization and the global crisis of democracy: common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist* 62 (1): 16–42. ISSN: 0002-7642.

McGarty, Craig, John C. Turner, Michael A. Hogg, Barbara David, and Margaret S. Wetherell. 1992. Group polarization as conformity to the prototypical group member. *British journal of social psychology* 31 (1): 1–19. ISSN: 0144-6665.

Morales, Alfredo Jose, Javier Borondo, Juan Carlos Losada, and Rosa M. Benito. 2015. Measuring political polarization: twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25 (3): 033114. ISSN: 1054-1500.

Moury, Catherine, and Elisabetta De Giorgi. 2015. Introduction: conflict and consensus in parliament during the economic crisis. *The Journal of Legislative Studies* 21 (1): 1–13.

Mudde, Cas, and Cristóbal Rovira Kaltwasser. 2013. Exclusionary vs. inclusionary populism: comparing contemporary europe and latin america. *Government and Opposition* 48 (2): 147–174. ISSN: 0017-257X.

Naumann, Elias, Katja Möhring, Maximiliane Reifenscheid, Alexander Wenz, Tobias Rettig, Roni Lehrer, Ulrich Krieger, Sebastian Juhl, Sabine Friedel, Marina Fikel, et al. 2020. Covid-19 policies in germany and their social, political, and psychological consequences. *European Policy Analysis* 6 (2): 191–202.

Onyimadu, Obinna, Keiichi Nakata, Tony Wilson, David Macken, and Kecheng Liu. 2014. Towards sentiment analysis on parliamentary debates in hansard. In *Semantic technology,* edited by Wooju Kim, Ying Ding, and Hong-Gee Kim, 48–50. Cham: Springer International Publishing. ISBN: 978-3-319-06826-8.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Peterson, Andrew, and Arthur Spirling. 2018. Classification accuracy as a substantive quantity of interest: measuring polarization in westminster systems. *Political analysis* 26 (1): 120–128. ISSN: 1047-1987.

Petri, Franziska, and Katja Biedenkopf. 2021. Weathering growing polarization? the european parliament and eu foreign climate policy ambitions. *Journal of European Public Policy,* 1–19. ISSN: 1350-1763.

Pew Research Center. 2019. Public highly critical of state of political discourse in the us. *Pew Research Center,* accessed August 24, 2021. https://www.pewresearch.org/politics/2019/06/19/public-highly-critical-of-state-of-political-discourse-in-the-u-s/.

Poole, Keith T., and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science,* 357–384. ISSN: 0092-5853.

Popper, Karl R. 1963. Science as falsification. *Conjectures and refutations* 1 (1963): 33–39.

Proksch, Sven–Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. Multilingual sentiment analysis: a new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly* 44 (1): 97–131. ISSN: 0362-9805.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54 (1): 209–228. ISSN: 0092-5853.

Quiñonero-Candela, Joaquin, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. 2009. *Dataset shift in machine learning.* Mit Press.

Rauh, Christian. 2018. Validating a sentiment dictionary for german political language—a workbench note. *Journal of Information Technology & Politics* 15 (4): 319–343.

Rayment, Erica, and Jason VandenBeukel. 2020. Pandemic parliaments: canadian legislatures in a time of crisis. *Canadian Journal of Political Science/Revue canadienne de science politique* 53 (2): 379–384.

Reicher, Stephen, and Mark Levine. 1994. Deindividuation, power relations between groups and the expression of social identity: the effects of visibility to the out–group. *British journal of social psychology* 33 (2): 145–163. ISSN: 0144-6665.

Reiter, Bernd. 2017. Theory and methodology of exploratory social science research.

Remus, R., U. Quasthoff, and G. Heyer. 2010. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th international language resources and evaluation (lrec'10),* 1168–1171.

Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PloS one* 11 (12): e0168843. ISSN: 1932-6203.

Rothe, Camilla, Mirjam Schunk, Peter Sothmann, Gisela Bretzel, Guenter Froeschl, Claudia Wallrauch, Thorbjörn Zimmer, Verena Thiel, Christian Janke, Wolfgang Guggemos, et al. 2020. Transmission of 2019-ncov infection from an asymptomatic contact in germany. *New England journal of medicine* 382 (10): 970–971.

Sakamoto, Takuto, and Hiroki Takikawa. 2017. Cross-national measurement of polarization in political discourse: analyzing floor debate in the U.S. and the japanese legislatures. *CoRR* abs/1711.02977.

Sani, Giacomo, and Giovanni Sartori. 1983. Polarization, fragmentation and competition in western democracies. *Western European party systems,* 307–340.

Sartori, Giovanni. 1976. *Parties and party systems.* CUP Archive. ISBN: 0521291062.

———. 2005. *Parties and party systems: a framework for analysis.* ECPR press. ISBN: 0954796616.

Schwarz, Daniel, Denise Traber, and Kenneth Benoit. 2017. Estimating intra-party preferences: comparing speeches to votes. *Political Science Research and Methods* 5 (2): 379–396. ISSN: 2049-8470.

Scollon, Ronald. 2008. *Analyzing public discourse: discourse analysis in the making of public policy.* Routledge. ISBN: 9780415540872.

Silva, Bruno Castanho. 2018. Populist radical right parties and mass polarization in the netherlands. *European Political Science Review: EPSR* 10 (2): 1–26. ISSN: 1755-7739.

Slapin, Jonathan B., and Sven–Oliver Proksch. 2008. A scaling model for estimating time–series party positions from texts. *American Journal of Political Science* 52 (3): 705–722. ISSN: 0092-5853.

Somer, Murat, and Jennifer McCoy. 2018. *Déjà vu? polarization and endangered democracies in the 21st century.* SAGE Publications Sage CA: Los Angeles, CA. ISBN: 0002-7642.

Song, Hyunjin, Petro Tolochko, Jakob-Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich, Fabienne Lind, Sebastian Galyga, and Hajo G Boomgaarden. 2020. In validations we trust? the impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication* 37 (4): 550–572.

Søyland, Martin. 2020. Multi-party classification of parliamentary debates: intra-party cohesion and inter-party relations measured in text. *University of Oslo - Working paper.*

Spirling, Arthur, Leslie Huang, and Perry Patrick. 2018. Boring in a new way: estimation and inference for political style at westminster, 1935–2018. *Social Science Research Network,* 1–32.

Spirling, Arthur, and Iain McLean. 2007. Uk oc ok? interpreting optimal classification scores for the u.k. house of commons. *Political Analysis* 15 (1): 85–96.

Strøm, Kaare, Wolfgang C. Müller, and Torbjörn Bergman. 2008. *Cabinets and coalition bargaining: the democractic life cycle in western europe.* Oxford University Press. ISBN: 0199587493.

Stroschein, Sherrill. 2011. Demography in ethnic party fragmentation: hungarian local voting in romania. *Party Politics* 17 (2): 189–204. ISSN: 1354-0688.

Tajfel, Henri, John C. Turner, William G. Austin, and Stephen Worchel. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader* 56 (65): 9780203505984–16.

Tepe, Sultan. 2014. The perils of polarization and religious parties: the democratic challenges of political fragmentation in israel and turkey. In *Religiously oriented parties and democratization,* 43–68. Routledge.

Trochim, William MK, and James P Donnelly. 2001. *Research methods knowledge base.* Vol. 2. Atomic Dog Pub. ISBN: 1592602916.

V-Dem Institute. 2019. Polarization in europe. Accessed August 12, 2021. https://www.v-dem.net/en/news/polarization-europe/.

Vicente-Sáez, Rubén, and Clara Martınez-Fuentes. 2018. Open science now: a systematic literature review for an integrated definition. *Journal of business research* 88:428–436.

Webb, Geoffrey I, and Kai Ming Ting. 2005. On the application of roc analysis to predict classification performance under varying class distributions. *Machine learning* 58 (1): 25–32.

Wendler, Frank. 2014. Justification and political polarization in national parliamentary debates on eu treaty reform. *Journal of European Public Policy* 21 (4): 549–567. ISSN: 1350-1763.

Wright, Gerald C. 2007. Do term limits affect legislative roll call voting? representation, polarization, and participation. *State Politics & Policy Quarterly* 7 (3): 256–280. ISSN: 1532-4400.

Yadollahi, Ali, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)* 50 (2): 1–33. ISSN: 0360-0300.

Yan, Hao, Sanmay Das, Allen Lavoie, Sirui Li, and Betsy Sinclair. 2019. The congressional classification challenge: domain specificity and partisan intensity. In *Proceedings of the 2019 acm conference on economics and computation,* 71–89.

Yardi, Sarita, and Danah Boyd. 2010. Dynamic debates: an analysis of group polarization over time on twitter. *Bulletin of science, technology & society* 30 (5): 316–327. ISSN: 0270-4676.

Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* 1 (1-4): 43–52. ISSN: 1868-8071.

Zong, Chengqing, Rui Xia, and Jiajun Zhang. 2021. *Text data mining.* Springer. ISBN: 9789811601002.

Appendix

Figure 9. Covid-19 Cases in Germany (Daily)

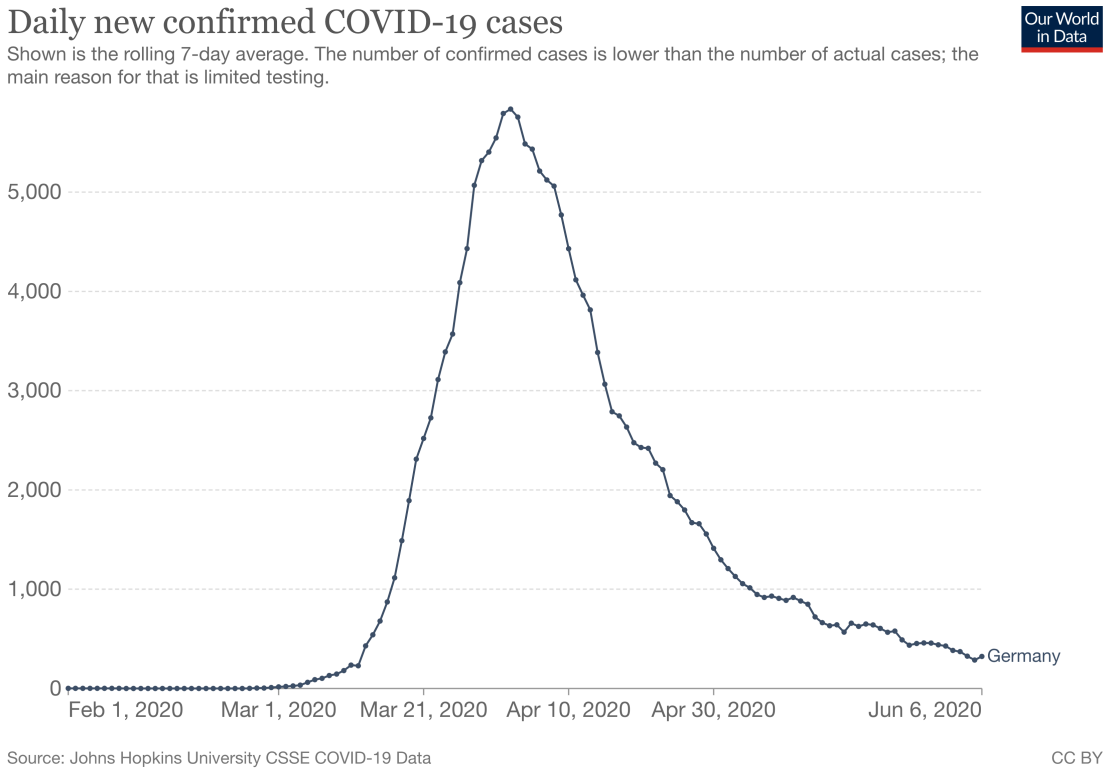


Figure 10. Count of Words per Speech (Length)

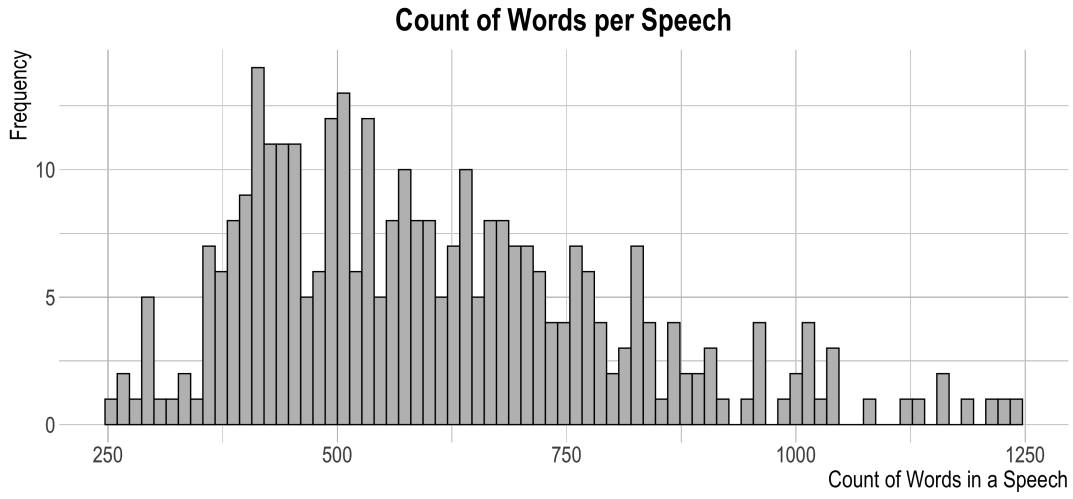


Table 6.1. *Total Ratio of Speeches between Government and Opposition*

No	Date Debate	# Opposition	# Government	# Total	Share Government
1	2020-02-12	5	6	11	0.55
2	2020-03-04	5	7	12	0.58
3	2020-03-25	16	14	30	0.47
4	2020-04-22	12	10	22	0.45
5	2020-04-23	24	24	48	0.50
6	2020-05-06	4	5	9	0.56
7	2020-05-07	20	18	38	0.47
8	2020-05-14	21	19	40	0.47
9	2020-05-15	17	16	33	0.48
10	2020-05-27	5	7	12	0.58
11	2020-05-28	10	11	21	0.52
12	2020-05-29	10	8	18	0.44
13	2020-06-18	5	6	11	0.55
14	2020-06-19	3	3	6	0.50
15	2020-07-02	8	7	15	0.47

Note: Speeches from parliamentary debates were removed when they include less then 5 speeches per debate or the share of government speeches was less than 40% or more than 60%.

Figure 11. Results: Debate Level Test

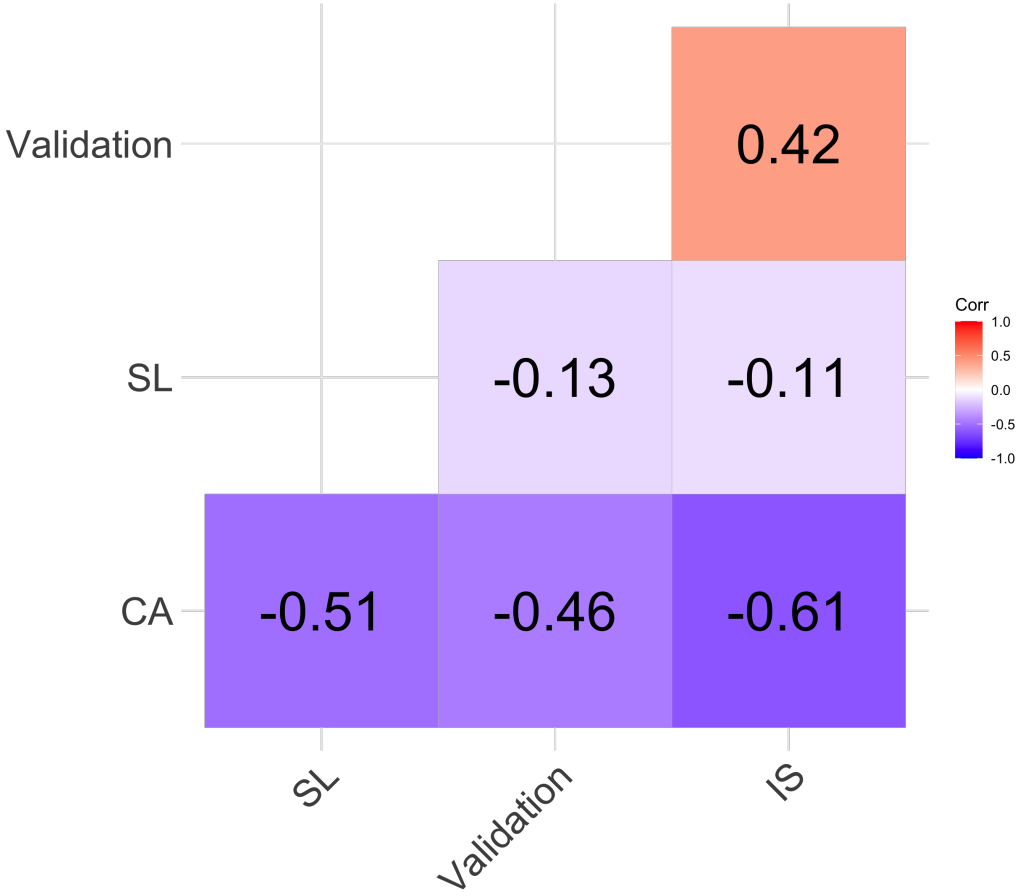


Table 6.2. *Total Count of Speeches by Party and Debate*

No	Date Debate	# AFD	# CDU/CSU	# Left	# FDP	# Greens	# SPD	# Total
1	2020-02-12	2	3	1	1	1	3	11
2	2020-03-04	1	4	1	2	1	3	12
3	2020-03-25	5	10	4	4	3	4	30
4	2020-04-22	3	5	3	3	3	5	22
5	2020-04-23	7	15	5	6	6	9	48
6	2020-05-06	1	3	1	1	1	2	9
7	2020-05-07	5	9	4	6	5	9	38
8	2020-05-14	5	11	5	6	5	8	40
9	2020-05-15	5	10	4	5	3	6	33
10	2020-05-27	2	4	1	1	1	3	12
11	2020-05-28	3	6	2	3	2	5	21
12	2020-05-29	3	5	2	2	3	3	18
13	2020-06-18	1	3	1	2	1	3	11
14	2020-06-19	1	1	1	0	1	2	6
15	2020-07-02	2	5	1	3	2	2	15

Table 6.3. *Count of top 50 Words (stemmed) with and without Sentiment Scores*

(a)

No	Word	Count	Score
1	schon	498	-0.01
2	gut	430	-0.37
3	dank	351	-0.19
4	wichtig	319	-0.38
5	wirtschaft	314	-0.00
6	imm	280	-0.00
7	richtig	278	-0.00
8	gross	271	-0.37
9	klar	218	-0.00
10	stark	215	-0.00
11	moglich	214	-0.00
12	neu	196	-0.00
13	lieb	195	-0.11
14	schnell	192	-0.12
15	genau	187	-0.00
16	unterstutz	180	-0.00
17	sich	179	-0.37
18	gemeinsam	170	-0.00
19	besond	168	-0.00
20	helf	165	-0.37
21	bereit	149	-0.00
22	sorg	141	0.00
23	bess	135	-0.00
24	verantwort	133	-0.00
25	bereich	132	-0.07
26	einfach	120	-0.00
27	aktuell	111	-0.00
28	pfleg	109	-0.23
29	leid	108	0.16
30	gesund	105	-0.16
31	fall	103	0.22
32	ford	99	-0.00
33	klein	97	0.27
34	herzlich	93	-0.00
35	bitt	89	0.34
36	kurz	88	0.00
37	schaff	86	-0.00
38	nachhalt	85	-0.00
39	falsch	75	0.76
40	erhalt	72	-0.00
41	schwierig	66	0.02
42	fuhr	66	-0.00
43	nutz	62	-0.00
44	ausbild	61	-0.00
45	selbststand	60	-0.00
46	gefahr	57	0.64
47	schuld	57	0.00
48	solidar	56	-0.00
49	genug	55	-0.10
50	beitrag	51	-0.00

(b)

No	word	n
1	muss	792
2	uns	682
3	viel	544
4	kris	522
5	europa	518
6	land	516
7	mensch	473
8	mehr	470
9	herr	463
10	ganz	460
11	sag	448
12	heut	435
13	euro	428
14	deutschland	398
15	ja	392
16	geht	384
17	brauch	367
18	massnahm	355
19	gibt	326
20	deshalb	309
21	erst	301
22	jahr	299
23	dafur	293
24	gerad	293
25	unternehm	286
26	mal	280
27	zeit	258
28	stell	257
29	desweg	257
30	milliard	256
31	antrag	247
32	bundesregier	245
33	frag	239
34	kolleg	229
35	natur	229
36	deutsch	220
37	dam	220
38	woch	214
39	weit	210
40	eben	208
41	wurd	207
42	letzt	207
43	seit	206
44	komm	204
45	situation	199
46	tun	199
47	glaub	198
48	geld	198
49	gesetz	193
50	arbeit	189

Note: The code and data for this analysis are available at https://github.com/lukasbirki/Thesis

```
# Part 1: Loading Data

rm(list = ls())

source("Scripts/helper_functions.R") # Load Packages and helper
    functions

as_tibble(read_csv("./data/speeches.csv")) %>%
  filter(electoralTerm == 19) %>%
  mutate(count_words = sapply(strsplit(.$speechContent, "␣"),
      length)) %>% #Counting words
  filter(count_words > 100 | count_words > 1413 )  %>%
  filter(factionId %in% c(0, 3, 4,6, 13,23)) %>%
  mutate(Party = case_when(
    factionId == 0 ~ "AFD",
    factionId == 3 ~ "Greens",
    factionId == 4 ~ "CDU/CSU",
    factionId == 6 ~ "DIE␣LINKE.",
    factionId == 13 ~ "FDP",
    factionId == 23~ "SPD",
    TRUE ~ NA_character_
  ))  %>% mutate(Ruling_Party = case_when(
    Party == "CDU/CSU" | Party == "SPD" ~ 1,
    TRUE ~ 0)) %>%
  mutate(date = as.Date(date))-> df_speeches

read_docx("data/DIP_Export.docx") -> DIP
```

```
# Part 2: Data Preprocessing
# DIP Data ----

# Filtering and Reshaping
DIP %>%
  docx_summary() %>%
  as_tibble()  %>%
  filter(style_name == "heading␣1" | grepl('Datum', text), style_
     name != "HeaderStandard") %>% # Filter Date and
  select(style_name, text, doc_index) %>%
  pivot_wider(names_from = style_name, values_from = text) %>%
  mutate(index_to_merge = rep(1:(nrow(.)/2), each = 2)) %>%
  group_by(index_to_merge) %>%
  summarise_all(list(~trimws(paste(., collapse = '')))) %>%
  select(-1:-2) -> r1

r1 %>%
  plyr::rename(c('heading␣1' = 'MP' , 'Werte' = 'date')) %>%
  mutate(date = str_remove_all(date, "^NADatum:␣")) %>%
  mutate(date = str_replace_all(date, '[.]','-')) %>%
  mutate(MP = str_sub(.$MP, 1, str_length(.$MP)-2)) %>%
  mutate(MP = str_remove_all(MP, "␣\\s*\\([^\\)]+\\)")) %>%
  splitstackshape::cSplit(., "MP", sep = ",", type.convert = F) %>%
  mutate(last_name = word(MP_1 , -1)) %>%
  mutate(last_name = case_when(
    last_name == "Ali" ~ "Mohamed␣Ali",
    T ~ last_name)) %>%
  as_tibble() -> r2

plyr::ldply(r2$date, reverse_words_helper) %>%
  cbind(.,r2) %>%
  as_tibble %>%
  select(-date) %>%
  plyr::rename(c('V1' = 'date')) %>%
  select(-MP_1,-MP_2, -MP_4, -MP_5) %>%
  plyr::rename(c( 'MP_3' = 'Party',  'last_name' = 'lastName')) %>%
  mutate(Party = case_when(
  Party == 'SPD' ~ 'SPD',
  Party == 'CDU/CSU' ~ 'CDU/CSU',
  Party == 'B NDNIS␣90/DIE␣GR NEN' ~ 'Greens',
  Party == 'DIE␣LINKE' ~ 'DIE␣LINKE.',
  Party == 'AfD' ~ 'AFD',
  Party == 'FDP' ~ 'FDP',
  T ~ NA_character_))  %>%
  mutate(lastName = case_when(
    lastName == 'Kotre' ~ 'Kotr ',
```

```
    lastName == 'Lambsdorff' ~ 'Graf Lambsdorff',
    lastName == 'Marschall' ~ 'von Marschall',
    lastName == 'Masi' ~ 'De Masi',
    lastName == 'Beeck' ~ 'in der Beek',
    T ~ lastName)) %>%  #Drops non-normal MPs
  drop_na() ->  df_docx_final

df_docx_final %>%
  filter(duplicated(.) == 'FALSE') %>%
  mutate(date = as.Date(date)) -> df_docx_final_without_duplicates

#Sessions where one speaker speaks more than one time
df_docx_final %>%
  filter(duplicated(.) == 'TRUE') %>%
  mutate(date = as.Date(date))-> df_docx_final_with_duplicates


df_speeches %>%
  select(lastName, date, Party, speechContent) -> data

data_no_duplicates <- data[!duplicated(data[,1:3]),]
data_duplicates <- data[duplicated(data[,1:3]),]

left_join(df_docx_final_without_duplicates, data_no_duplicates) ->
   x1
left_join(df_docx_final_with_duplicates, data_duplicates) -> x2

rbind(x1, x2) %>%
  mutate(date = as.Date(date))-> df_base

# Merging: Speeches Data Set + DIP Data Set ----


right_join(df_speeches, df_base, by = c('speechContent', 'Party', '
   date','lastName'))  %>%
  select(-positionShort, -positionLong, -electoralTerm, -factionId)
      %>%
  mutate(speechContent = str_remove_all(speechContent, greetings ))
      -> df_base

# Preprocessing II: Remove Stopwords and Punctuation; Apply
   Wordstemming
df_base %>%
  tidytext::unnest_tokens(word, speechContent, drop = F) %>%
  anti_join(.,sw, by = c("word" = "value")) %>%
  filter(str_detect(word, "[a-z]")) %>%
```

```
  mutate(token_stem = SnowballC::wordStem(.$word, language = "
     german")) %>%
  group_by(id) %>%
  summarize(speechContent_stemmed = str_c(token_stem, collapse = "␣
     ")) %>%
  ungroup() %>%
  right_join(df_base, y, by = 'id') %>%
  drop_na() %>%
  filter(count_words <= 1250 & count_words >= 250)-> df_base

# Filtering: Descriptive Statistics ----

# Total Count
df_base %>%
  count(Ruling_Party)

# Total Count Government/Oppossition per Session
df_base %>%
  count(session, Ruling_Party) %>%
  pivot_wider(id_cols = session,names_from = Ruling_Party, values_
     from =n ) %>%
  mutate(n = `0`+`1`) %>%
  mutate(share_government = `1`/n) %>%
  mutate(Removed = case_when(
    (session == 195 |  session == 186 |session == 187 |session ==
       202) ~ "Removed",
    T ~ "Not␣Removed"
  ))  %>%
  left_join(., df_base, by = 'session') %>%
  select(date, '0', '1', n, share_government) %>% unique() %>%
  mutate(date = as.character(date)) %>%
  xtable::xtable() %>%
  print(., file = "./Figures/Tables/ratio_sessions.tx")

# Total Count Government/Oppossition per Party
df_base %>%
  count(session, Party) %>%
  pivot_wider(id_cols = session,names_from = Party, values_from =n
     ) %>%
  mutate(n = AFD + `CDU/CSU` + `DIE LINKE.` + FDP + Greens + SPD)
     %>%
  left_join(., df_base, by = 'session') %>%
  select(date, AFD, 'CDU/CSU', 'DIE␣LINKE.', FDP, Greens, SPD, n)
     %>% unique() %>%
  mutate(date = as.character(date)) %>%
  xtable::xtable() %>%
```

```
  print(., file = "./Figures/Tables/ratio_party_session.tx")

#Speech Median & Mean

median(df_base$count_words)
mean(df_base$count_words)
sd(df_base$count_words)

# Speech Length

plot_speech_leght <- df_base %>%
  ggplot(aes(count_words)) +
  geom_histogram(color = "black", fill = "gray", bins = 75) +
  ylab("Frequency") + xlab("Count␣of␣Words␣in␣a␣Speech") +
  ggtitle('Count␣of␣Words␣per␣Speech') +
  hrbrthemes::theme_ipsum() +
  theme(
    plot.title = element_text(size=24, hjust = 0.5),
    axis.title.x = element_text(size = 18),
    axis.title.y = element_text(size = 18),
    axis.text.x = element_text(size = 17),
    axis.text.y = element_text(size = 17)
    )

ggsave("./Figures/Figures/Speech_lenght.png", plot = plot_speech_
    leght, height = 6)

# Temporal Distribution over time

#https://stackoverflow.com/questions/7492274/draw-a-chronological-
    timeline-with-ggplot2

df_base %>%
  count(session) %>%
  right_join(., df_base,  by = 'session') %>%
  select(date, session, n) %>%
  distinct() -> df_plot
df_plot$randoms <- c(0.12, 0.22, 0.13, -0.4, 0.40,
                     -0.3, 0.12, 0.45, -0.2, 0.4,
                     -0.1, 0.3, 0.23, -0.3, 0.17)

vjust = ifelse(df_plot$randoms > 0, -1, 1.5)
p2 <- df_plot %>%
  ggplot(aes(date, randoms)) +
  ggalt::geom_lollipop(point.size = 1) +
```

```
  geom_text(aes(x = date, y = randoms, label = as.character(date)),
      data = df_plot,
            hjust = 0, vjust = vjust, size = 4) +
  expand_limits(x = c(lubridate::ymd(20200201), lubridate::ymd
     (20200801)), y = c(-0.5, 0.5)) +
  scale_x_date(breaks = scales::pretty_breaks(n = 9)) +
  ggtitle("Timeline (February 01 - July 31, 2020)" ) +
  labs(x = "Time") +
  hrbrthemes::theme_ipsum() +
  theme(
    plot.title = element_text(size=24, hjust = 0.5),
    axis.title = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank(),
    axis.line = element_blank(),
    axis.title.x = element_text(size = 18),
    axis.title.y = element_blank(),
    axis.text.x = element_text(size = 17))
plot_timeline <- shift_axis(p2, lubridate::ymd(20200201), lubridate
   ::ymd(20200801))

ggsave("./Figures/Figures/Timeline.png", plot = plot_timeline,
   height = 5)

#Number of Speeches per Session

plot_speeches_per_session <- df_base %>%
  mutate(date = as.character(date)) %>%
  ggplot() + geom_bar(aes(date, fill = factor(Party,
                                              levels =  c("AFD","
                                                 DIE LINKE." , "
                                                 Greens","FDP" ,"
                                                 SPD"  ,"CDU/CSU"))
                                                 ), position  = '
                                                 stack', width =
                                                 0.8) +
  scale_fill_manual(breaks = c("CDU/CSU", "DIE LINKE.", "FDP", "AFD
     ", "Greens", "SPD"),
                    values=c("#000000", "#BE3075", "#FFFF00",  "
                       #009EE0",  "#64A12D","#FF0000"),
                    name = 'Party') +
  ggtitle("Total Count of Party Distribution per Debate" ) +
  labs(x = "Date", y = "Total Count") +
  hrbrthemes::theme_ipsum() +
  theme(
    plot.title = element_text(size=24, hjust = 0.5),
```

```
    axis.title.x = element_text(size = 20),
    axis.title.y = element_text(size = 20),
    axis.text.x = element_text(angle = 45, hjust = 1, color = "
       black", size = 15),
    axis.text.y = element_text(size = 17),
    legend.text=element_text(size=17),
    legend.title = element_text(size = 20),
    legend.position = "top") +
  guides(colour = guide_legend(nrow = 1))

ggsave("./Figures/Figures/Count_per_Session.png", plot = plot_
   speeches_per_session, height = 6)

# Number of Session (Latex)

df_base %>%
  count(session) %>%
  right_join(., df_base,  by = 'session') %>%
  select(date, session, n) %>%
  mutate(date = as.character(date)) %>%
  distinct() %>%
  xtable::xtable() %>%
  print(., file = "./Figures/Tables/table_sessions.tx")

#Covid-19 data

data <- read.csv("https://opendata.ecdc.europa.eu/covid19/
   nationalcasedeath_eueea_daily_ei/csv", na.strings = "",
   fileEncoding = "UTF-8-BOM")

data %>%
  select(countriesAndTerritories, cases, dateRep) %>% as_tibble()
     %>%
  filter(countriesAndTerritories == 'Germany') %>% tail()


#Data Preprocessing ----

df_base %>%
  tidytext::unnest_tokens(word, speechContent, drop = F) %>%
  select(word) %>%
  unique()

df_base %>%
  tidytext::unnest_tokens(word, speechContent_stemmed, drop = F)
     %>%
```

```
  select(word) %>%
  unique()

#Write test data
df_speeches %>%
  filter(electoralTerm == 19) -> x1

x1 %>%
  tidytext::unnest_tokens(word, speechContent, drop = F) %>%
  anti_join(.,sw, by = c("word" = "value")) %>%
  filter(str_detect(word, "[a-z]")) %>%
  mutate(token_stem = SnowballC::wordStem(.$word, language = "
      german")) %>%
  group_by(id) %>%
  summarize(speechContent_stemmed = str_c(token_stem, collapse = "␣
      ")) %>%
  ungroup() %>%
  right_join(., x1, by = 'id') %>%
  anti_join(., df_base, by = 'id') %>%
  write.csv(., file = "data/df_train_complete.csv")


rm(DIP,
  additional_stopwords,
  df_docx_final,
  df_docx_final_with_duplicates,
  df_docx_final_without_duplicates,
  x1,
  r1,
  r2,
  x2,
  sw,
  data_no_duplicates,
  data_duplicates,
  plot_speeches_per_session,
  df_plot,
  plot_timeline,
  plot_speech_leght,
  p2)
```

```r
# Part 3: Loading Validaton Data


df_handcoded  <- haven::read_dta("data/multilevel_nov20.dta")
df_validation <- readxl::read_excel("data/Coding_Germany_for_stata.
    xlsx")

df_validation %>%
  filter(chair != 1 & interruption == 0) %>%
  filter(score != 99) %>%
  group_by(date, speaker_lastname) %>%
  summarise(validation_speech = mean(score), .groups = 'keep' ) %>%
  drop_na()-> df_validation_scores

df_handcoded %>%
  splitstackshape::cSplit(., "speaker_name", sep = "␣", type.
      convert = F) %>%
  mutate(last_name = word(speaker_name_2 , -1)) %>%
  mutate(last_name = case_when(
    last_name == "Mohamed" ~ "Mohamed␣Ali",
    T ~ last_name)) %>%
  as_tibble() %>%
  left_join(., df_validation_scores, by = c("date" = "date", "last_
      name" = "speaker_lastname" )) %>%
  select(date, validation_speech,last_name,party_id) %>%
  distinct() -> df_handcoded_2

df_handcoded_2 %>%
  mutate(Party = case_when(
    party_id == 303 ~ "FDP",
    party_id == 304 ~ "Greens",
    party_id == 305 ~ "DIE␣LINKE.",
    party_id == 306 ~ "AFD",
    TRUE ~ NA_character_)) %>%
  rename('lastName' = 'last_name') %>%
  mutate(lastName = case_when(
    lastName ==  'Kotre' ~ 'Kotr ',
    T ~ lastName
  )) %>%
  right_join(., df_base, by = c('date', 'lastName', 'Party')) %>%
  distinct()-> df_base_running

df_base_running %>%
  aggregate(validation_speech ~ date, data = ., mean) %>%
  rename('validation_session' = 'validation_speech') %>%
  right_join(., df_base_running, by = 'date') %>%
```

```
  as_tibble() %>%
  mutate_at(vars(validation_session, validation_speech), list( ~. *
      -1))-> df_base # Rearrange algebraic sign

rm(df_base_running,
   df_handcoded_2,
   df_handcoded,
   df_validation,
   df_validation_scores)

# # Plots ----

## Plot: Validation Session over Time ----
rects <- data.frame(xstart = c("2020-02-01", '2020-03-01','
   2020-05-01'),
                    xend = c("2020-02-28", '2020-04-30','2020-07-30
                       '), col = letters[1:3])
rects$xstart <- as.Date(rects$xstart)
rects$xend <- as.Date(rects$xend)

p_validation <-
  df_base %>%
  select(date, validation_session) %>%
  distinct() %>% drop_na() %>%
  ggplot() +
  # geom_rect(aes(xmin = as.Date("2020-02-01", "%Y-%m-%d"),
  #                 xmax = as.Date("2020-03-01", "%Y-%m-%d"),
  #                 ymin = -Inf, ymax = Inf, fill = 'Medium'),
  #           alpha = .08) +
  #   geom_rect(aes(xmin = as.Date("2020-03-01", "%Y-%m-%d"),
  #                 xmax = as.Date("2020-05-01", "%Y-%m-%d"),
  #                 ymin = -Inf, ymax = Inf, fill = 'Low'),
  #             alpha = .08) +
  #   geom_rect(aes(xmin = as.Date("2020-05-01", "%Y-%m-%d"),
  #                 xmax = as.Date("2020-07-15", "%Y-%m-%d"),
  #                 ymin = -Inf, ymax = Inf, fill = 'High'),
  #             alpha = .08) +
  # scale_fill_brewer(palette = 'Dark2', name = 'Year')+
  geom_line(aes(x = as.Date(date), y = scale(validation_session)))
     +
  geom_point(aes(x = as.Date(date), y = scale(validation_session)))
      +
  ggtitle("Validation␣Scores␣for␣Sessions␣(N␣=␣6)") +
  labs(x = "Date", y = "Polarisation␣(Validation)") +
  hrbrthemes::theme_ipsum(grid = 'Y') +
  theme(
```

```
    plot.title = element_text(size=24, hjust = 0.5),
    axis.text.x = element_text(size = 17, hjust = 1),
    axis.title.x = element_text(size = 18),
    axis.title.y = element_text(size = 18),
    axis.text.y = element_text(size = 17),
    legend.position = "top")

ggsave("./Figures/Figures/validation.png", plot = p_validation,
   height = 5)
```

```
# Part 4: Sentiment Lexicons (SL)

# Loading Dictionary ----


read_tsv("./data/SentiWS_v2.0/SentiWS_v2.0_Negative.txt", col_names
    = FALSE) %>%
  rename("Wort_POS" = "X1", "Wert" = "X2", "Inflektionen" = "X3" )
     %>%
  mutate(Wort = str_sub(Wort_POS, 1, regexpr("\\|", .$Wort_POS)-1),
         POS = str_sub(Wort_POS, start = regexpr("\\|", .$Wort_POS)
            +1)) -> neg_df

read_tsv("./data/SentiWS_v2.0/SentiWS_v2.0_Positive.txt", col_names
    = FALSE) %>%
  rename("Wort_POS" = "X1", "Wert" = "X2", "Inflektionen" = "X3" )
     %>%
  mutate(Wort = str_sub(Wort_POS, 1, regexpr("\\|", .$Wort_POS)-1),
         POS = str_sub(Wort_POS, start = regexpr("\\|", .$Wort_POS)
            +1)) -> pos_df

bind_rows("neg" = neg_df, "pos" = pos_df, .id = "neg_pos") %>%
  select(neg_pos, Wort, Wert, Inflektionen, -Wort_POS) %>%
  mutate(token_stem = SnowballC::wordStem(.$Wort, language = "
     german")) %>%
  mutate(Polarity = .$Wert * -1)-> sentiment_df

rm(neg_df, pos_df)

# Analysis ----

## Matching Sentiment Values ----
df_base %>%
  tidytext::unnest_tokens(word, speechContent_stemmed, drop = F)
     %>%
  rowwise() %>%
  mutate(WordScore_token_stem = ifelse(word %in% sentiment_df$token
     _stem,
                               sentiment_df$Polarity[match(word,
                                  sentiment_df$token_stem)], NA)) ->
                                  sentiment_df_final


## Statistics ----

### Percent of words with an assigned Sentiment Score ----
```

```
length(sentiment_df_final$WordScore_token_stem[!is.na(sentiment_df_
    final$WordScore_token_stem)])/length(sentiment_df_final$
    WordScore_token_stem)

### Show most frequent 100 words with a sentiment Score ----
sentiment_df_final %>%
  filter(WordScore_token_stem != is.na(WordScore_token_stem)) %>%
  group_by(word) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  select(word, n, WordScore_token_stem) %>% unique() %>%
  arrange(desc(n)) %>%
  head(n = 50) %>%
  xtable::xtable() %>%
  print(., file = "./Figures/Tables/with_sentiment_Score.tx")

### Show most frequent 100 words without a sentiment Score ----
sentiment_df_final %>%
  filter(is.na(WordScore_token_stem)) %>%
  group_by(word) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  select(word, n) %>% unique() %>%
  arrange(desc(n)) %>%
  head(n = 50) %>%
  xtable::xtable() %>%
  print(., file = "./Figures/Tables/without_sentiment_Score.tx")


## Calculating Sentiment Scores for Speeches and Sessions ----
sentiment_df_final %>%
  group_by(id) %>%
  mutate(Score_id = mean(WordScore_token_stem, na.rm = T)) %>%
  ungroup() %>%
  group_by(date) %>%
  mutate(Score_date = mean(WordScore_token_stem, na.rm = T)) %>%
  ungroup() -> df_sentiment

#Merging with df_base
df_sentiment %>%
  group_by(speechContent) %>%
  summarise(sentiment_score_session = unique(Score_date),
            sentiment_Score_id = unique(Score_id)) %>%
  left_join(df_base, ., by = "speechContent")-> df_base

# Plots ----
```

```r
## Plot: Sentiment per Session ----

plot_sentiment <- df_base %>%
  group_by(date) %>%
  distinct(date, sentiment_score_session) %>%
  mutate(date = as.Date(date)) %>%
  ggplot(aes(x = date, y = sentiment_score_session)) +
  geom_line(stat="identity") +
  geom_point() +
  geom_hline(yintercept = 0)+
  ggtitle("Sentiment Score per Session" ) +
  labs(y = "Sentiment (Polarity) Score per Session", x = "Time") +
  hrbrthemes::theme_ipsum(grid = "X") +
  scale_x_date(date_labels="%b %y",date_breaks  ="1 month")

ggsave("./Figures/Figures/sentiment_session.png", plot = plot_
    sentiment, height = 9)

## Plot: Sentiment per Speech ----

plot_sentiment_box <- df_base %>%
  filter(sentiment_Score_id <=0.5) %>%
  ggplot(aes(x = as.character(date), y = sentiment_Score_id, group
      = as.character(date))) +
  geom_boxplot() +
  viridis::scale_fill_viridis(discrete = T, alpha=0.6) +
  hrbrthemes::theme_ipsum() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("Basic boxplot") +
  xlab("")

ggsave("./Figures/Figures/sentiment_session.png", plot = plot_
    sentiment_box, height = 9)
```

```
# Part 5: Ideological Scaling (IS)

# Selecting Speeches and Loading into Quanteda Corpus Class ----

corpus_wordfish <- quanteda::corpus(df_base,
                                    text_field = "speechContent_
                                       stemmed")

docnames(corpus_wordfish) <- df_base$lastName

summary(corpus_wordfish, n = 1000) %>% as_tibble() -> df_wordfish_
    running

#Creating Document-Feature Matrix and removing all words used less
    than 3 times

corpus_wordfish %>%
  tokens() %>%
  dfm() %>%
  dfm_trim(.,min_termfreq = 3) %>%
  textmodel_wordfish(., sparse = T) -> tmod_wf

# Summarising findings in dataframe

tmod_wf[c('docs','theta', 'alpha')] %>%
  as_tibble() %>%
  right_join(., df_wordfish_running, by = c('docs' = 'Text')) %>%
  mutate(date = as.Date(date)) -> df_wordfish_speech

# Calculating IGO Index

df_wordfish_speech %>%
  group_by(session) %>%
  mutate(session_mean = mean(theta)) %>%
  ungroup() %>%
  group_by(session, Party) %>%
  mutate(party_mean_session = mean(theta)) %>%
  ungroup() %>%
  mutate(Party_share = case_when(
    Party == 'CDU/CSU' ~ 0.34,
    Party == 'Greens' ~ 0.094,
    Party == 'DIE␣LINKE.' ~ 0.097,
    Party == 'FDP' ~ 0.11,
    Party == 'AFD' ~ 0.12,
    Party == 'SPD' ~ 0.21,
  )) %>%
```

```
  group_by(session) %>%
  mutate(IGO_session = sqrt(sum(((party_mean_session - session_mean
    )/6)**2))) %>%
  mutate(IGO_session_share = sqrt(sum(Party_share * ((party_mean_
    session - session_mean)/6)**2))) %>%
  ungroup() -> df_IGO

# Merging with df_base

df_IGO %>%
  select(speechContent, IGO_session, IGO_session_share) %>%
  right_join(., df_base, by = c("speechContent")) -> df_base

# Plots ----

# Polarisation over Time (Session)
df_IGO %>%
ggplot() +
  geom_line(aes(x = date, y = IGO_session)) +
  geom_point(aes(x = date, y = IGO_session)) +
  geom_line(aes(x = date, y = IGO_session_share)) +
  geom_point(aes(x = date, y = IGO_session_share))

df_IGO %>%
  group_by(date, Ruling_Party) %>%
  summarise(party_position = mean(theta)) %>%
  ggplot() +
  geom_line(aes(x = date, y = party_position, color = as.character(
    Ruling_Party)))

df_IGO %>%
  select(Ruling_Party, theta) %>%
  group_by(Ruling_Party) %>%
  summarize(mean = mean(theta))
```

```
# Part 6a: Python Scripts

Data Wrangling Plenar Protokols
https://medium.com/@bedigunjit/simple-guide-to-text-classification-
    nlp-using-svm-and-naive-bayes-with-python-421db3a72d34
import pandas as pd
import nltk
import numpy as np
import sklearn.datasets
from nltk.tokenize import word_tokenize
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.preprocessing import LabelEncoder
from collections import defaultdict
from nltk.corpus import wordnet as wn
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import model_selection, naive_bayes, svm
from sklearn.metrics import accuracy_score
pd.set_option('display.max_columns', None)
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from nltk.corpus import stopwords
nltk.download('stopwords')
german_stop_words = stopwords.words('german')
[nltk_data] Downloading package stopwords to /Users/lukas/nltk_data
    ...
[nltk_data]   Package stopwords is already up-to-date!
df_test = pd.read_csv('/Users/lukas/OneDrive␣-␣University␣of␣
    Warwick/Dokumente/Uni/Thesis/Thesis/data/df_base.csv')
df_train = pd.read_csv('/Users/lukas/OneDrive␣-␣University␣of␣
    Warwick/Dokumente/Uni/Thesis/Thesis/data/df_train.csv')

df_test = df_test.sample(frac = 1)
df_train = df_train.sample(frac = 1)

print(len(df_test))
print(len(df_train))

X_train = df_train['speechContent']
y_train = df_train['Ruling_Party']
X_test = df_test['speechContent']
Y_test = df_test['Ruling_Party']
326
18240
count_vect = CountVectorizer(stop_words = german_stop_words)
```

```
# Fit and tranform with X_train
count_vector = sklearn.feature_extraction.text.CountVectorizer(stop
    _words = german_stop_words)
word_counts = count_vector.fit_transform(X_train)
tf_transformer = sklearn.feature_extraction.text.TfidfTransformer(
    use_idf=True)
X_train = tf_transformer.fit_transform(word_counts)

SMV = svm.SVC(C=1.0, kernel='linear', degree=3, gamma='auto',
    probability=True)
SGD = sklearn.linear_model.SGDClassifier(loss = 'log', penalty='l2'
    )

SMV.fit(X_train, y_train)
SGD.fit(X_train, y_train)

# Transform X_test
test_word_counts = count_vector.transform(X_test)
ready_to_be_predicted = tf_transformer.transform(test_word_counts)
Transforming Term Frequency matrix

# predict the labels on validation dataset
predictions_SVM = SMV.predict(ready_to_be_predicted)
predictions_SGD = SGD.predict(ready_to_be_predicted)
# Use accuracy_score function to get tghe accuracy
print("SVM Accuracy Score -> ",accuracy_score(predictions_SVM, Y_
    test)*100)
print("SGD Accuracy Score -> ",accuracy_score(predictions_SGD, Y_
    test)*100)
SVM Accuracy Score ->  90.1840490797546
SGD Accuracy Score ->  86.80981595092024
class_probabilities_SMV = SMV.predict_proba(ready_to_be_predicted)
class_probabilities_SGD = SGD.predict_proba(ready_to_be_predicted)
confidence_0_SMV = class_probabilities_SMV.transpose()[0]
confidence_1_SMV = class_probabilities_SMV.transpose()[1]
confidence_0_SGD = class_probabilities_SGD.transpose()[0]
confidence_1_SGD = class_probabilities_SGD.transpose()[1]
df_test['SMV_predictions'] = predictions_SVM
df_test['SGD_predictions'] = predictions_SGD
df_test['Confidence_0_SMV'] = confidence_0_SMV
df_test['Confidence_1_SMV'] = confidence_1_SMV
df_test.to_csv('/Users/lukas/OneDrive - University of Warwick/
    Dokumente/Uni/Thesis/Thesis/data/df_base_ML.csv')
```

```
# Part 6b: Loading ML estimates

df_base_ML <- as_tibble(read_csv("./data/df_base_ML.csv" ))

#Create Confusion Matrix for each Session

#Confusion Matrixes
j <- 1
confusion_martices <- list()
for (i in sort(unique(as.character(df_base_ML$session)))){
  q <- caret::confusionMatrix(data      = factor(
     df_base_ML$Ruling_Party[df_base_ML$session == i]),
                         reference = factor(
                            df_base_ML$SMV_predictions[
                            df_base_ML$session == i]),
                         positive  = "1")$table
  confusion_martices[[j]] <- q
  j <- j + 1
}

#Accuracy Scores
j <- 1
accuracy_scores <- list()
for (i in sort(unique(as.character(df_base_ML$session)))){
  q <- caret::confusionMatrix(data      = factor(
     df_base_ML$Ruling_Party[df_base_ML$session == i]),
                         reference = factor(
                            df_base_ML$SMV_predictions[
                            df_base_ML$session == i]),
                         positive  = "1")$overall
  accuracy_scores[[j]] <- q
  j <- j + 1
}

lapply(accuracy_scores,'[[',1) %>%
  set_names(unique(df_base_ML$session)) %>%
  as_tibble(.name_repair = "unique") %>% t() %>%
  dplyr::as_data_frame(., rownames = "session") %>%
  plyr::rename(c( "V1" = "Accuracy_Score_SGD" )) %>%
  mutate(session = as.numeric(session)) %>%
  left_join(df_base, ., by = 'session') -> df_base
```

```
# Part 7: Method Comparison

df_base %>% select(-politicianId, -documentUrl, -count_words) %>%
      mutate(date = as.Date(date)) -> df_analysis

# Z-Standardizing ----

df_analysis %>%
  mutate(across(c(sentiment_score_session, IGO_session, validation_
      session, Accuracy_Score_SGD), scale)) %>%
  mutate(across(c(sentiment_score_session, IGO_session, validation_
      session, Accuracy_Score_SGD), as.vector)) -> df_analysis

df_analysis %>%
  select(date,
         session,
         sentiment_score_session,
         IGO_session,
         Accuracy_Score_SGD) %>%
  plyr::rename(c('sentiment_score_session' = 'SL␣Score','IGO_
      session' = 'IS␣Score', 'Accuracy_Score_SGD' = 'CA␣Score')) %>%
  pivot_longer(!c(date, session), names_to = "Metric", values_to =
      "values") %>% drop_na() %>%
  mutate(Metric = factor(Metric, levels=c('SL␣Score', 'IS␣Score', '
      CA␣Score'))) %>% unique() -> df_temp

# Plots ----

## Facet Wrap Line ----

df_temp %>%
  ggplot() +
  geom_line( aes(x = as.Date(date), y = values)) +
  geom_point(aes(x = as.Date(date), y = values)) +
  geom_hline(yintercept = 0,size=0.25 )+
  geom_vline(xintercept = as.Date('2020-03-01'), linetype="dotted")
      +
  geom_vline(xintercept = as.Date('2020-05-01'), linetype="dotted")
      +
  geom_label(aes(x =as.Date('2020-03-01'), y = 2, label = 'T␣1'),
      size = 7) +
  geom_label(aes(x =as.Date('2020-05-01'), y = 2, label = 'T␣2' ),
      size = 7) +
  ggtitle("Polarization␣Scores␣for␣Debates␣(N␣=␣15)") +
```

```
  labs(x = "Date␣of␣Debate", y = "Standardised␣Polarization␣Scores"
     ) +
  scale_x_date(date_labels = "%b-%d") +
  facet_wrap(~ Metric, nrow = 3) +
  hrbrthemes::theme_ipsum(grid = 'Y') +
  theme(
    plot.title = element_text(size=24, hjust = 0.5),
    axis.title.x = element_text(size = 18),
    axis.title.y = element_text(size = 18),
    axis.text.x = element_text(size = 18, angle = 45, hjust = 1,
       color = "black"),
    axis.text.y = element_text(size = 18),
    legend.text=element_text(size=17),
    legend.title = element_text(size = 20),
    strip.text.x = element_text(size = 22),
    panel.spacing = unit(0.5, "lines"))+
  guides(colour = guide_legend(nrow = 1)) -> comparison_line
ggsave("./Figures/Figures/comparison_line.png", plot = comparison_
   line, height = 10)

## Correlations----

df_analysis %>%
  select(date,
         validation_session,
         sentiment_score_session,
         IGO_session,
         Accuracy_Score_SGD) %>%
  distinct() %>%
  slice(1:6) %>%
  plyr::rename(c("sentiment_score_session"= "SL",
                 "IGO_session" = "IS" ,
                 "Accuracy_Score_SGD" = "CA",
                 "validation_session" = "Validation")) %>%
  pivot_longer(!c(date, Validation), names_to = "Metric", values_to
      = "values") %>%
  plyr::rename(c("Metric"= "Method")) %>%
ggplot(aes(x = Validation, y = values, color=Method)) +
  geom_point(size = 2) +
  geom_smooth(method=lm, se = F) +
  ggtitle("␣Correlation␣between␣Methof␣Estimates␣and␣Validation␣
     Data") +
  labs(x = "Scores␣Estimates", y = "Scores␣Validation") +
  hrbrthemes::theme_ipsum(grid = 'Y') +
  xlim(-2,2) +
  ylim(-2,2)+
```

```
  theme(
    plot.title = element_text(size=24, hjust = 0.5),
    axis.title.x = element_text(size = 24),
    axis.title.y = element_text(size = 24),
    axis.text.x = element_text(size = 20, angle = 45, hjust = 1,
       color = "black"),
    axis.text.y = element_text(size = 20),
    legend.text=element_text(size=17),
    legend.position="top",
    legend.title = element_text(size = 20),
    strip.text.x = element_text(size = 24),
    panel.spacing = unit(0.5, "lines"))+
  guides(colour = guide_legend(nrow = 1)) -> scatterplot
ggsave("./Figures/Figures/scatterplot.png", plot = scatterplot,
   height = 7)


## Correlation complete

df_analysis %>%
  select(sentiment_score_session,
         IGO_session,
         Accuracy_Score_SGD) %>%
  unique() %>%
  cor() %>% as_tibble()

## Correlations Validation Scores

df_analysis %>%
  select(validation_session,
         sentiment_score_session,
         IGO_session,
         Accuracy_Score_SGD) %>%
  distinct() %>%
  slice(1:6) %>%
  plyr::rename(c("sentiment_score_session"= "SL",
                 "IGO_session" = "IS" ,
                 "Accuracy_Score_SGD" = "CA",
                 "validation_session" = "Validation")) %>%
  cor() %>%
  ggcorrplot::ggcorrplot(hc.order = TRUE, type = "lower",
                         lab = TRUE,
                         lab_size = 12,
                         tl.cex = 24)-> p1

ggsave("./Figures/Figures/correlation.png", plot = p1, height = 8)
```

```
## Consistency Test

# Inspecting 'extreme' debates for each method

require(data.table) ## 1.9.2
df_temp <- as.data.table(df_temp)


df_max <- df_temp[df_temp[, .I[values == max(values)], by=Metric]$
    V1]
df_min <- df_temp[df_temp[, .I[values == min(values)], by=Metric]$
    V1]

df_extreme <- rbind(df_max, df_min)

df_base %>%
  write.csv(., file = "data/validation_df_base.csv")

#Analysing Wordfish Estimates: IS

vocabs <- tmod_wf$features %>% as_tibble()

tmod_wf[c('beta', 'features')] %>% as_tibble() -> scallles

corpus_wordfish %>%
  tokens() %>%
  dfm() %>%
  dfm_trim(.,min_termfreq = 3) %>%
  textmodel_wordfish(., sparse = T) -> tmod_wf

tmod_wf$features[tmod_wf$features == 'bafog'] <- NA
tmod_wf$features[tmod_wf$features == 'pepp'] <- NA



textplot_scale1d(tmod_wf, margin = 'features',
                 highlighted = c('staatsfinanzier', 'schuldenstand'
                    ,'kapitalmarkt','aufschwung','fiskal',
                                 'hochschul','studienfinanzier','
                                    kulturlandschaft','kunstler'))
                                    +
  ylim(-16, 3)+
  xlim(-11,11) +
hrbrthemes::theme_ipsum(grid = 'Y')  +
  theme(
```

```
    plot.title = element_text(size=24, hjust = 0.5),
    axis.title.x = element_text(size = 18),
    axis.title.y = element_text(size = 18),
    axis.text.x = element_text(size = 18, hjust = 1, color = "black
        "),
    axis.text.y = element_text(size = 18)) -> p2

ggsave("./Figures/Figures/scaling2.png", plot = p2, height = 8)

corpus_wordfish %>%
  tokens() %>%
  dfm() %>%
  dfm_trim(.,min_termfreq = 3) -> dfm_analysis


tmod_wf[c('docs','theta', 'alpha')] %>%
  as_tibble() %>%
  right_join(., df_wordfish_running, by = c('docs' = 'Text')) %>%
  mutate(date = as.Date(date)) -> df_wordfish_speech


df_wordfish_speech %>%
  group_by(Party) %>%
  summarise(Mean_party_position = mean(theta)) %>%
  ggplot(aes(x = Mean_party_position,y = fct_reorder(Party,Mean_
      party_position ))) +
  ggalt::geom_lollipop(horizontal=TRUE, point.size = 2, point.
      colour = 'red') +
  geom_vline(xintercept = 0,size=0.25 )+
  labs(x = "Aggregated␣Theta", y = "Party") +
  xlim(-0.16,0.16) +
  hrbrthemes::theme_ipsum(grid = 'Y')  +
  theme(
    plot.title = element_text(size=24, hjust = 0.5),
    axis.title.x = element_text(size = 18),
    axis.title.y = element_text(size = 18),
    axis.text.x = element_text(size = 18, angle = 45, hjust = 1,
        color = "black"),
    axis.text.y = element_text(size = 18)) -> p3
ggsave("./Figures/Figures/lollipop_graph.png", plot = p3, height =
    6)
```

```
# Part 8: Helper Functions

library(tidyverse)
library(tidytext)
library(SnowballC)
library(gridExtra)
library(grid)
library(ggplot2)
library(haven)
library(officer)
library(stringr)
library(splitstackshape)
require(devtools)
library(glmnet)
library(quanteda)
library(quanteda.textmodels)
library(quanteda.textplots)
library(readtext)


# Defining Functions ----

# Reverse words
reverse_words_helper <- function(string)
{
  # split string by blank spaces
  string_split = strsplit(as.character(string), split = "-")
  # how many split terms?
  string_length = length(string_split[[1]])
  # decide what to do
  if (string_length == 1) {
    # one word (do nothing)
    reversed_string = string_split[[1]]
  } else {
    # more than one word (collapse them)
    reversed_split = string_split[[1]][string_length:1]
    reversed_string = paste(reversed_split, collapse = "-")
  }
  # output
  return(reversed_string)
}



IsDate <- function(mydate, date.format = "%d/%m/%y") {
```

```
  tryCatch(!is.na(as.Date(mydate, date.format)),
           error = function(err) {FALSE})
}

shift_axis <- function(p, xmin, xmax, y=0){
  g <- ggplotGrob(p)
  dummy <- data.frame(y=y)
  ax <- g[["grobs"]][g$layout$name == "axis-b"][[1]]
  p + annotation_custom(grid::grobTree(ax, vp = grid::viewport(y=1,
      height=sum(ax$height))),
                        ymax=y, ymin=y) +
    annotate("segment", y = 0, yend = 0, x = xmin, xend = xmax,
             arrow = arrow(length = unit(0.1, "inches"))) +
    theme(axis.text.x = element_blank(),
          axis.ticks.x=element_blank())

}

sw <- tibble(stopwords::stopwords("de"))
sw$sw <- sw$`stopwords::stopwords("de")`
sw <- as_tibble(sw$sw)
additional_stopwords <- c('dass') %>% as_tibble()
sw <- rbind(sw,additional_stopwords )


c('Herr␣Pr sident', 'Frau␣Pr sidentin',
  'Herr␣Pr sident', 'Frau␣Pr sidentin',
  'Sehr␣geehrter␣Herr␣Pr sident', 'Sehr␣geehrte␣Frau␣Pr sidentin'
     ,
  'Sehr␣geehrte␣Damen␣und␣Herren', 'Meine␣Damen!␣Meine␣Herren',
  'Liebe␣Kolleginnen␣und␣Kollegen', 'Verehrte␣Kolleginnen␣und␣
     Kollegen',
  'Meine␣sehr␣verehrten␣Damen␣und␣Herren', 'Meine␣Damen␣und␣Herren'
     ,
  'Meine␣sehr␣geehrten␣Damen␣und␣Herren', 'Meine␣Damen␣und␣Herren',
  'Sehr␣geehrte␣Kolleginnen␣und␣Kollegen', 'Sehr␣geehrte␣Kollegen',
  'Werte␣Kollegen', '␣Meine␣Kolleginnen␣und␣Kollegen', 'Sehr␣
     geehrte␣Pr sidentin',
  'Liebe␣Kolleginnen!␣Liebe␣Kollegen', 'Sehr␣verehrte␣Kolleginnen␣
     und␣Kollegen',
  'Verehrte␣Kolleginnen', 'Verehrte␣Kollegen', 'Frau␣Pr sident',
  '␣Werte␣Kolleginnen␣und␣Kollegen', 'Liebe␣Zuh rerinnen␣und␣
     Zuh rer',
  'Kolleginnen␣und␣Kollegen', 'Liebe␣Landsleute', 'Verehrte␣
     Kollegen␣und␣Zuschauer',
```

```
'Liebe␣Damen␣und␣Herren', 'Meine␣sehr␣geehrten␣Kolleginnen␣und␣
   Kollegen', 'Kollegen',
'Meine␣lieben␣Kolleginnen␣und␣Kollegen', '␣Liebe␣Kollegen␣und␣
   Kolleginnen', 'Sehr␣geehrte␣Herren␣Pr sidenten',
'Liebe␣Zuschauerinnen␣und␣Zuschauer␣an␣den␣Bildschirmen', 'Meine␣
   sehr␣verehrten',
'Frau␣Bundeskanzlerin', 'Frau␣Bundesministerin', 'Herr␣
   Bundesminister', 'Herr␣Minister') %>%
paste(., "[,|.|!.?]␣", sep = "") %>%
str_c(.,collapse="|") -> greetings
```