Matching agendas

Franziska Löw [*]

Department of Industrial Economics,
Helmut Schmidt University,
Hamburg, Germany

March 8, 2019

Abstract

To measure the political slant of german online newspaper the topics addressed in newspapers are compared with topics addressed in press releases of political parties. To find the latent topics in the corpus a structural topic model is conducted.

[*] Electronic address: `loewf@hsu-hh.de`

Contents

1 Introduction

In recent years, the media and their role in the perception and decision of individuals in the political context have been increasingly subject to criticism. Terms such as "fake news" or "quality journalism" are currently part of almost every debate regarding the role of the media. Critics accuse the media of reporting biased on certain parties or political events and thus influencing the political consciousness of voters. This raises the unavoidable question of what biased reporting actually means or, on the contrary, what objective reporting is and if this is even possible. A journalist who writes an article about a certain topic puts rough facts (e.g. figures on economic indicators) into a context, such that each article is shaped by the subjectivity of this journalist. Similarly, an editor of a media outlet has to select the topics to be discussed in the medium from a large pool of reports. Thus, to a certain extent, media is always filtered by journalists' perceptions and editorial decisions.

A legitimate question, however, could be which factors or incentives lead to the selection or deselection of certain topics. On the one hand, one could assume that editors select the topics and articles that correspond to their own political views. A profit-maximizing editor, on the other hand, would tend to adapt the selection to readers' preferences. (Some populist voices would even claim that - at least the so called "mainstream media" - is controlled by the government.)

In order to answer these and other media-related questions in the political context, quantifying the content of media is essential. In other words, the key challenge is to determine the features that are used to describe text content. Studies that rely on quantifying media content for their analyses use, for example, visibility (how often political actors appear in the media) or tonality (how they are evaluated). In contrast to these actor-based approaches, issue-based methods exist where the topics discussed in the media are analyzed in order to identify whether political actors are able to place their own policy positions in the media. Leading studies from economic literature, for example, examine how often a newspaper quotes the same think tanks (Groseclose and Milyo, 2005; Lott and Hassett, 2014) or uses the same language (M. A. Gentzkow and Shapiro, 2004) as members of Congress.

To measure the ideological content of several online news services in Germany, in the present paper the topics discussed in these media outlets are compared with the press releases of the parties in the german "Bundestag". The dataset contains nearly 12.000 online news articles from seven major news provider dated from June 1, 2017 to March 1, 2018 as well as over 1.900 press releases of the parties in the german "Bundestag". As the German federal elections took place on 24th of September 2017 and the formation of the government has taken up a period of about five months, the articles considered inform their readers about both the election promises of the parties (before the election) and the coalition talks (after the election). To discover the latent topics in the corpus of text data, the structural topic model (STM) developed by M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016) is applied. The STM is an unsupervised machine learning approach that models topics as multinomial distributions of words and documents (as a synonym for news articles) as multinomial distributions of topics, allowing the incorporation of external variables that affect both, topical content and topical prevalence. The results of the generative process of the STM are two posterior distributions: One for the topic prevalence in a document (what is the article or press release about?) and one for the content of a topic (what is the topic about?). The topic prevalence is used to estimate a slant-index by comparing press releases (i.e., the party agenda) with the news articles (i.e., the mediated party agenda). The results of this analysis are then compared with data on the political orientation of consumers in order to evaluate whether there may be a link between consumer preferences and the slant-index of a medium.

The use of online news as data input accurately reflects the changed conditions in the news market as the importance of the internet as a source of information for political topics has grown strongly in recent years.[1] This trend has as strong effect on the market for media content as neither supply nor demand is tied to specific times and can adapt to events in real time. Users can consume their preferred news sources at any time and providers can adapt their offerings to events in real time without waiting for the next issue or TV show.

The research contribution of this paper is twofold: First, a new method for the calculation of the slant-index is presented that allows an extensive content analyses of newspaper coverage and party press releases and at the same time reduces human induced bias and makes research more traceable and comparable. In addition, a new dataset of online news is used, which has a significant relation to the current discussion of the media in the political context.

The remaining course of the paper is as follows: The following section provides an overview of the related literature. Section 3 gives an introduction to the political trends within the considered time (June 2017 to March 2018). The data used to conduct the model is described in section 4.1. Section **??** explains the generative process of the structural topic model as well as the selected parameters to run the model. The empirical analysis is conducted in section **??**.

2 Literature

In unbiased media reporting all political sides should be equally represented according to some kind of benchmark for balance or neutrality (Hopmann et al., 2012). Bias is then defined as the extent to which media reporting deviates from this benchmark.

However, the discussion about media bias has not only been studied since the growing importance of the internet. Different research disciplines investigate the concept of media bias, albeit with different focus. Regardless of the discipline, media bias research can be roughly divided into three research questions: (1) What types of bias exist, (2) how and why does it emerge and (3) what influence does it have on the political outcomes. While 1 and 3 are dealt with in economic, political and communication research, question 2 is mainly addressed in economic literature. It investigates which forces on the supply and demand side drives to media bias.

The general hypothesis within economic literature is that (1) different media outlets tend to report "biased" or "slanted" (Groseclose and Milyo, 2005; Lott and Hassett, 2014) and (2) that media reporting about political news may have a profound influence on political outcomes (DellaVigna and Kaplan, 2006; M. Gentzkow, 2006; M. A. Gentzkow and Shapiro, 2004; Snyder and Strömberg, 2010; Strömberg, 2004).

In the economic literature the market dynamics that lead to a possible bias are analysed. In principle, media bias can come from the supply side, and reflect the preferences of journalists (Baron, 2006), editors, or owners (Besley and Prat, 2004). Alternatively, media bias can come from the demand side, and reflect the news providers' profit-maximizing choice to cater to the preferences of the consumers.

[1] Even though television remains the most widely used source of news in Germany (2018: 74%), numbers watching continue to decline while use of the internet for news has grown significantly in the last year (+5%, 2018: 65%) (Hölig and Hasebrink, 2018).

While economic research investigates the market dynamics that lead to media bias, communication science is more concerned with the question of what different types of bias exist. According to a D'Alessio and Allen (2000) the concept of media bias actually encompasses different subtypes: (1) Coverage bias, (2) tonality bias und (3) agenda bias. These three concepts measure how often political actors appear in the media (coverage bias), how they are evaluated (tonality bias) and whether they are able to present their own political positions and talk about their issues in the media (agenda bias).

There is visibility bias when a party is the subject of an undue amount of coverage compared to the benchmark of that party at a given point in time. Media that reports biased in that sense influences voters behaviour in such way, that voters tend to prefer parties that are more visible in their media repertoire (Eberl et al., 2017). Studies combining media content data with voter surveys have indeed found that the mere visibility of parties and candidates is an important factor influencing vote choice (Oegema Kleinnijenhuis, 2000). The amount of a parties campaign communication or their standing in polls are commonly used as a reference point (Hopmann et al., 2012; Junqué de Fortuny et al., 2012).

Eberl et al. (2017) use the average visibility of all parties in each media outlet during the period of their analysis as a key benchmark to capture whether party visibility is biased in comparison to what is typical for that outlet and are therefore able to compare party visibility between outlets. Applying a similar logic to tonality and agenda bias, they measure the effect of the different bias on user voting behaviour using an online panel survey from the Austrian parliamentary election campaign of 2013.

- Druckman and Parkin (2005) argue that the audience's conclusions about parties are automatically drawn from positive or negative descriptions in texts about the parties.

- Similarly, valence framing suggests that public awareness of parties is affected depending on whether they are highlighted with positive or negative aspects in the media (Hurtíková, 2017; Vreese and Boomgaarden, 2006).

To measure tonality in a text, studies differ between manually coded data (e.g. Eberl et al. (2017)) and dictionary-based analysis (e.g. (Junqué de Fortuny et al., 2012)). The latter approach is widely used outside the area of media content analysis. To conduct such an analysis, a lists of words (dictionary) associated with a given emotion, such as negativity is pre-defined by the analyst. The document is then deconstructed into individual words and the frequencies of words contained in a given dictionary are calculated. Such lexical or "bag-of-words" approaches are widely presented in the finance literature to determine the effect of central banks' monetary policy communications on asset prices and real variables (Nyman et al. (2018) Tetlock (2007), Tetlock et al. (2008)). Hansen and McMahon (2016) use a similar approach to explore the effects of FOMC (Federal Open Market Committee) statements on both market and real economic variables. To calculate their score, they subtract the negative words from the positive words und divide this by the number of total words of the statement. A similar score is used by Nyman et al. (2018), who measure the effect of narratives and sentiment of financial market text-based data on developments in the financial system.

In the domain of media content analysis Junqué de Fortuny et al. (2012) count the sentiment words in a window of two sentences before and after the mention of a political party and assuming uniformity of sentiment distribution among parties to measure the bias.[2] They use a text-mining approach to automate the analysis of a large text corpus showing techniques to measure both visibility and tonality bias. The former is benchmarked by the amount of preference votes for that party. Similar to Junqué de Fortuny et al. (ibid.) the present study uses techniques that allow the computational

[2] A similar approach for target identification with a 10-word window is used in Balahur et al. (2013)

analysis of a large dataset of text-data. However, a different reference point is used to allow for a comparison between media outlets. Additionally, agenda bias is measured based on the comparison between the content of online news and parties press releases Eberl et al. (2017).

Overall, most research tends to disregard agenda bias as the operationalization is more challenging. In order to know which news stories have been held out by journalists, the true universe of news stories at a given point in time has to be known D'Alessio and Allen (2000). However, a greater dissemination of a party's political content may have a positive impact on attitudes towards that party (Benewick et al., 1969; Eberl et al., 2017). Brandenburg (2006) measure partisan tendencies in reporting in terms of all three biases. Utilizing content analysis data from the 2005 General Election campaign they show that increasingly ambiguous endorsements translate into an absence of open support for political parties. Similarly, Eberl et al. (2017) find that voters evaluate parties more favorably if those parties addressed their own favored topics more prominently in media coverage. In their analysis media content was analyzed using manual content analysis of political claims on a sentence level.

What drives media bias? In principle, media bias can come from the supply side, and reflect the preferences of journalists (Baron, 2006), editors, or owners (Besley and Prat, 2004). Alternatively, media bias can come from the demand side, and reflect the news providers' profit-maximizing choice to cater to the preferences of the consumers.

The STM developed by M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016) is a recent extension of the standard topic modelling technique, labeled as latent Dirichlet allocation (LDA), which refers to the Bayesian model in Blei et al. (2003) that treats each word in a topic and each topic in a document as generated from a Dirichlet - distributed prior.[3] Since its introduction into text analysis, LDA has become hugely popular and especially useful in political science.[4] Wiedmann (2016) uses topic model methods on large amounts of news articles from two german newspapers published between 1959 and 2011, to reveal how democratic demarcation was performed in Germany over the past six decades. Paul (2009) compares editorial differences between media sources, using cross-collection latent Dirichlet allocation (ccLDA), an LDA-based approach that incorporates differences in document metadata. They use a dataset of 623 news articles from August 2008 from two American media outlets - msnbc.com and foxnews.com - to compare how they discuss topics. Reviewing the top words of the word-topic distribution, they find some content differences between the two media sources under review.

The difference between the widely used LDA and the STM approaches lies in how θ and ϕ are determined. LDA assumes that θ Dirichlet(α) and ϕ Dirichlet(β), where α and β are fitted with the model. While for STM, the prior distributions for θ and ϕ depend on document-level covariates (e.g. the author or date of a document). For this purpose, the the STM specifies two design matrices of covariates, where each line defines a vector of covariates for a specific document. In X, the covariates for topic prevalence are given, so that the probability of a topic for each document varies according to X, rather than resulting from a single common prior. The same applies to Z, in which the covariates for the word distribution within a topic are specified.

The model has been applied to multiple academic fields: M. E. Roberts, B. M. Stewart, Tingley, et al. (2014) uses STM to analyze open-ended responses from sur-

[3] See also Griffiths and Steyvers (2002), Griffiths and Steyvers (2004) and Hofmann (1999). Pritchard et al. (2000) introduced the same model in genetics for factorizing gene expression as a function of latent populations.

[4] see Blei (2012), Grimmer and B. Stewart (2013) and Wiedmann (2016) for an overview in social science and M. Gentzkow et al. (2017) give an overview of text mining applications in economics.

veys and experiments, Farrell (2016) applies the model to scientific texts on climate change, revealing links between corporate funding and the framing of scientific studies. Mishler et al. (2015) show that "STM can be used to detect significant events such as the downing of Malaysia Air Flight 17" when applied to twitter data. Another study shows how STM can be used to explore the main international development topics of countries' annual statements in the UN General Debate and examine the country-specific drivers of international development rhetoric (Baturo et al., 2017). Mueller and Rauh (2016) use newspaper text to predict armed conflicts in different regions. They use the estimated topic shares in linear fixed effects regression to forecast conflict out-of-sample. M. Roberts, B. Stewart, and Tingley (2016a) use STM to examine the role of partisanship in topical coverage using a corpus of 13,246 posts that were written for 6 political blogs during the course of the 2008 U.S. presidential election. With the aim of revealing the effect of partisan membership on topic prevalence, each blog is assigned to be either liberal or conservative. To explore the differences between the two, they look at the expected proportion of topics and examine the posts most associated with a respective topic. This approach is similar to M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016).

The present analysis differs from earlier approaches to measure agenda bias in that a machine learning technique is used to identify the underlying topics in the text corpus applying a structural topic model (ibid.). Furthermore, I use text-mining techniques to measure coverage and tonality bias. However, I shall refrain as far as possible from interpreting the results at political level. Rather, it is my goal to show how text mining techniques enable an efficient and objective analysis of today's online media landscape and simultaneously allow the analysis to be reproduced.

Much of the research on online content and political trends have focused on traditional weblogs and social media websites, such as Twitter, Facebook, MySpace, and YouTube. These studies have shown that social media is used to spread political opinions and that these considerations reflect the political landscape of the offline world. Tumasjan et al. (2010) investigate Tweets between August 13th and September 19th, 2009, prior to the German national elections to examine whether Twitter messages reflect the current offline political sentiment and whether it can be used to predict the popularity of parties or coalitions in the real world. With regard to the later question, they compare the share of attention the political parties receive on Twitter with the election result to examine whether the activity on Twitter can serve as a predictor of the election outcome. They found that the number of tweets reflects the election result and even comes close to traditional election polls.

Fu and Chan (2013) use a corpus of online posts from discussion forums and blogs to examine the extent to which online sentiment reflected in social media content can predict phone survey results in Hong Kong. They build a sentiment classifier conducting a support vector machine analysis on a training set of 2,000 manually labeled posts. In order to evaluate the temporal relationship between the time series of the online sentiment score and the results of the telephone survey, a cross correlation analysis was conducted, using the Box and Jenkins autoregressive integrated moving average (ARIMA) method (Box et al., 2008). Estimating the cross-correlation functions of the residuals, they find that online sentiment scores can lead phone survey results by about 8â15 days.

In a more recent conference paper, Padmaja et al. (2014) identify the scope of negation in news articles for two political parties in India (BJP and UPA) to analyze how the choice of certain words used in these texts influence the sentiments of public in polls. Comparing three different sentiment analysis methods (two machine learning and one dictionary method), they observe that the choice of certain words used in political text was influencing the sentiments in favor of BJP. They conclude that this sentiment bias might be one of the causes for the election results in 2014.

Dewenter et al. (2018) use human-coded data from leading media in Germany together with the German Politbarometer survey to investigate how media coverage affects short- and long-term political preferences between February 1998 and December 2012. They find a positive correlation between the media coverage and the short-term voting intention for a political party. In the long-term, however, voting preferences are stable.

3 Background on the federal election in Germany (2017)

The articles analyzed in this paper cover a period from June 1, 2017 to March 1, 2018 and thus cover both the most important election campaign topics for the Bundestag elections on September 24, 2017 and the process of forming a government that lasted until February 2018. After four years in a grand coalition with the Social Democrats (SPD), German Chancellor Angela Merkel, member of the conservative party CDU/CSU (also known as Union), ran for re-election. The SPD nominated Martin Schulz as their candidate.

On the right side of the political spectrum, AfD (alternative for Germany) managed to be elected to the German Bundestag for the first time in 2017. The political debate about the high refugee numbers of the past years brought a political upswing to the AfD, which used the dissatisfaction of parts of the population to raise its own profile. In the course of the reporting on the federal elections, leading party members of the AfD as well as party supporters repeatedly accused the mass media of reporting unilaterally and intentionally presenting the AfD badly.

After the election, the formation of a government was difficult due to the large number of parties elected to the Bundestag and the considerable loss of votes by the major parties CDU/CSU and SPD. Since all parties rejected a coalition with the AfD, numerically only two coalitions with an absolute parliamentary majority were possible: a grand coalition ("GroKo" - from the German word Große Koalition) of CDU/CSU and SPD, and a Jamaica coalition (coalition of CDU/CSU, FDP (economic liberal party) and B90/Die Grünen (Bündnis 90/Die Grünen, green party)). The grand coalition was initially rejected by the SPD. The four-week exploratory talks on the possible formation of a Jamaica coalition officially failed on November 19, 2017 after the FDP announced its withdrawal from the negotiations. FDP party leader Christian Lindner said that there had been no trust between the parties during the negotiations. The main points of contention were climate and refugee policy. CDU and CSU regretted this result, while B90/Die Grünen sharply criticized the liberals' withdrawal. The then Green leader Cem Özdemir accused the FDP of lacking the will to reach an agreement.

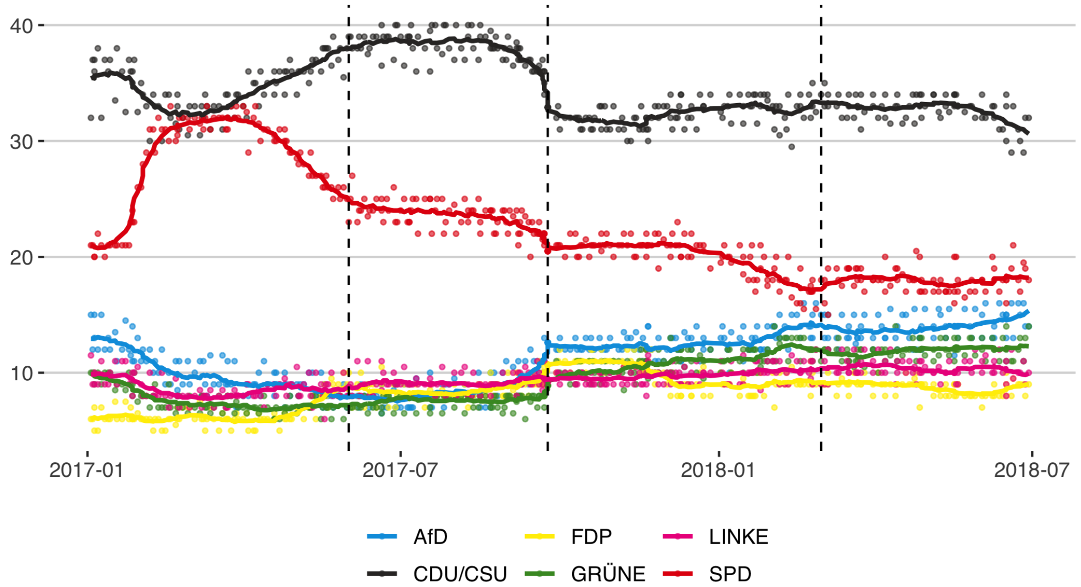
After the failure of the Jamaica coalition talks, a possible re-election or a minority government as alternatives were discussed in the media before the SPD decided to hold coalition talks with the CDU/CSU. This led to great resistance from the party base, which called for a party-internal referendum on a grand coalition. After the party members voted in favor of the grand coalition, a government was formed 171 days after the federal elections.

Figure 1 shows that support for the two major popular parties has been declining in recent months since August 2017, with the CDU/CSU again showing positive survey results since November 2017.[5] However, the poll results of the SPD have been falling since

[5] For each party the survey results of the seven major institutes are considered. To calculate a smooth line for each party on each day, the moving average within 15 days (7 before the day, 7 after the day, and the day itself) is estimated. The data source is https://www.wahlrecht.de/.

March 2017. At the same time, the AfD in particular has been recording increasingly positive survey results since June 2017.

Figure 1: Election Polls



4 Measuring slant-index

My approach to measuring the slant of a newspaper is to compare the topics in the newspaper with topics in parties' press releases to identify which parties the topics in the newspaper are most similar to. Parties want the media agenda to be congruent with their own agenda to define the issue-based criteria on which they will be evaluated by voters (Eberl et al., 2017). Thus, I assume that parties instrumentalize their press releases in order to highlight issues that they are perceived to be competent on, that they "own" and that are important to their voters (Kepplinger and Maurer, 2004). Editors can select from this universe and decide which topics will be discussed in the news. In that sense the ideological content of a newspaper refers to the extent to which the topics promoted by the parties correlate with the topics discussed in the news articles.

To discover the latent topics in the corpus of press releases (1.942) and news articles (11.880), a structural topic model (STM) developed by M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016) is applied. The STM is an unsupervised machine learning approach that models topics as multinomial distributions of words (topical content) and documents as multinomial distributions of topics (topical prevalence of a document), allowing to incorporate external variables that effect both, topical content and topical prevalence. I estimate a model in which the source[6] of the document is included as a control for the topical prevalence, e.g. I assume that the possibility that a topic appears in a document depends on the source. Additionally, the type of source[7] is included as a control for the topical content, e.g. I assume that words used to describe the same topics differ between press releases and news articles. The results of the generative process of the STM are two posterior distributions: One for the topic prevalence in a document (what

[6] Bild.de, DIE WELT, FOCUS ONLINE, SPIEGEL ONLINE, stern.de, ZEIT ONLINE, Tagesschau.de, CDU/CSU, SPD, AfD, FDP, B90/Die Grünen, DIE LINKE

[7] press release or news article

is the article or press release about?) and one for the content of a topic (what is the topic about?). The topic prevalence is used to calculate the agenda of each source as the mean over all documents that belong to that source. Subsequently the bivariate correlations between party agendas and the mediated party agendas in the online news are estimated. These correlations represent the agenda selectivity each party experiences in each media outlet. The higher the correlation, the more congruent both agendas are. In the last step I compare the agenda selectivity of each medium with the user preferences expressed in the Reuters Institute Digital News Survey. A more formale description of this process is described in Section 4.3 and 4.4.

The following section describes the data on news articles and press releases and how they have processed in order to use them as input for the model.

4.1 Data sources

4.1.1 Press releases

The press releases were scraped from the publicly accessible press portals of the parties[8] and parliamentary groups[9].

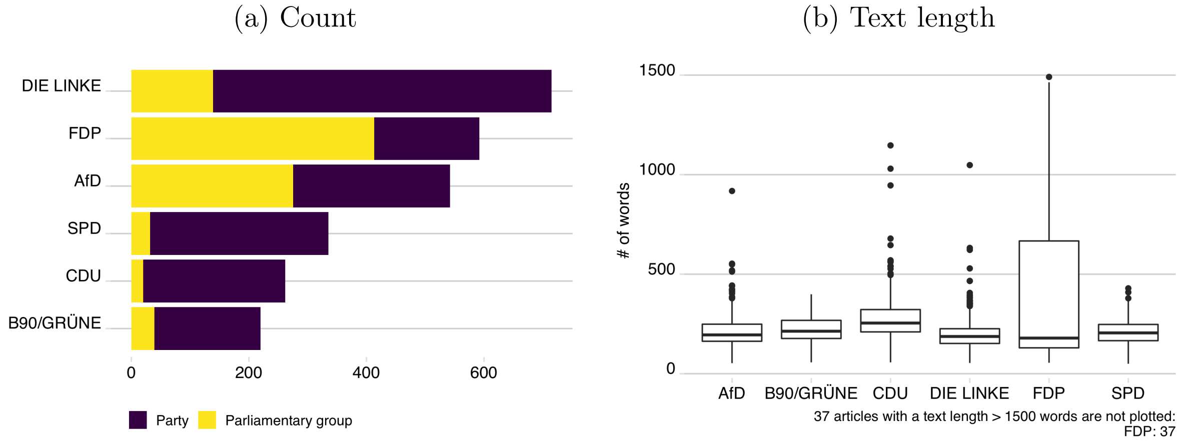
It should be noted that there is a legal distinction between press releases from political parties and parliamentary groups in the Bundestag, as parties are financed by membership dues, donations and campaign expenses, while parliamentary groups are financed by state funds. According to Parteigesetzt §25 (2) state funded parliamentary groups may not support parties, as there would be a disadvantaged for parties that are not in the Bundestag. However, since it is difficult to draw the line between the activity of parliamentary groups and election campaign assistance (Kepplinger and Maurer, 2004), I assume that parliamentary groups intervene in the public perception of this party with their press releases, which is why both the press releases of the federal party and the parliamentary groups are included.

Articles with less than 50 words, as well as articles about party internal elections, were filtered out resulting in a record of 2666 press releases. Figure 2a displays the number of press releases for each party and Figure 2b shows the text length of those articles. Overall, DIE LINKE has published the most press releases, B90/Die Grünen the fewest. Looking at the number of words, it is noticeable that the FDP publishes press releases that are significantly longer compared to the other parties.

[8] afd.de, spd.de, die-linke.de, fdp.de, gruene.de, cdu.de

[9] afdbundestag.de, spdfraktion.de, die-linke.de/start/presse/aus-dem-bundestag, fdpbt.de, gruene-bundestag.de/, presseportal.de/nr/7846

Figure 2: Press releases


(a) Count

(b) Text length

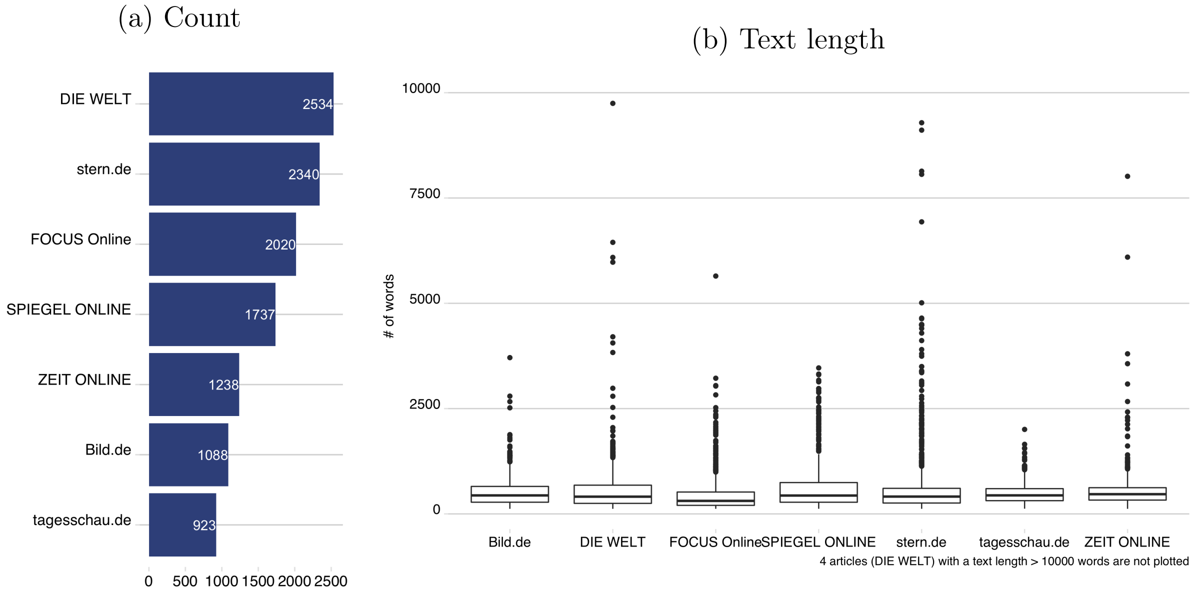
4.1.2 News articles

The data on news articles includes online articles from seven German news providers about domestic politics[10] dated from June 1, 2017 to March 1, 2018. I first extract all online articles using the Webhose.io API.[11] In order to limit the analysis to articles on domestic politics, the articles are filtered using the url of the article. For example, the URL *https://www.welt.de/politik/deutschland/article166629242/Tegel-ist-gefuehlt-das-Einzige-in-Berlin-was-funktioniert.html* indicates that this article belongs thematically to ’politics’ and ’deutschland’ (Germany). After further cleaning processes (e.g. articles that only contain video content) the total sample size was 11,880. As shown in Figure 3a most articles were published by "Die Welt" and "tagesschau.de" has the smallest number of articles. Also in terms of text length is DIE WELT on top and tagesschau.de at the bottom.

[10] Bild.de, DIE WELT, FOCUS ONLINE, SPIEGEL ONLINE, stern.de, ZEIT ONLINE, Tagesschau.de

[11] For more information see https://docs.webhose.io/v1.0/docs/getting-started. The scraping code was written in Python and can be made available on request.

Figure 3: News articles


(a) Count

(b) Text length

As shown in Figure 4a, the selected news providers are - in terms of site visits - among the top German online news providers in the period under review, with only Tagesschau.de belonging to the public media. The Figure displays the total visits of the top 30 websites in Germany in January 2018.[12] Figure 4b shows the results of a survey on digital news by the Reuters Institute (**newman_reuters_2018**). The online survey for German data was undertaken between 19th - 22nd January 2018 by the Hans Bredow Institute[13] with a total sample size of 2038 adults (aged 18+) who access news once a month or more. Among other questions, participants were asked which news sources they use to access news online.[14] The results indicate that the media used for the analysis play a relevant role.

[12] The term visit is used to describe the call to a website by a visitor. The visit begins as soon as a user generates a page impression (PI) within an offer and each additional PI, which the user generates within the offer, belongs to this visit.

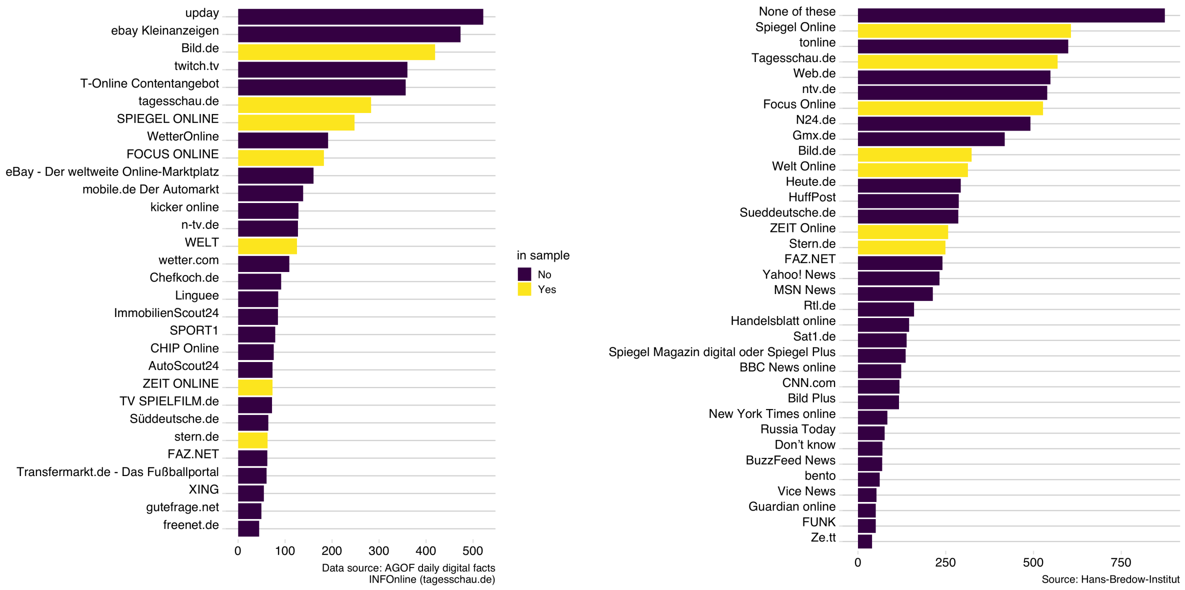
[13] https://www.hans-bredow-institut.de/de/projekte/reuters-institute-digital-news-survey

[14] The exact question was: "Which of the following brands have you used to access news online in the last week (via websites, apps, social media, and other forms of Internet access)? Please select all that apply"

Figure 4: Selected news brands

(a) Total visits in million (Jan 2018)

(b) Use of a brand to access news online



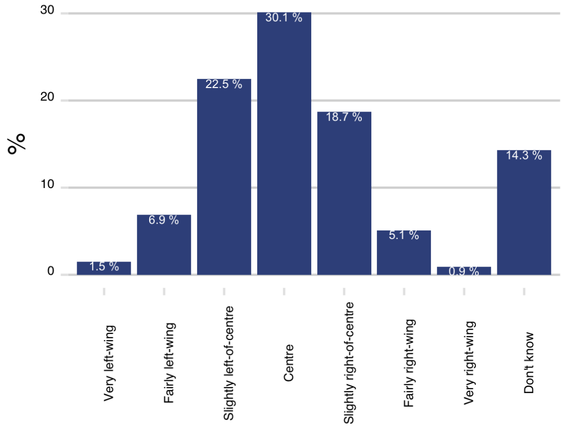
Data source: AGOF daily digital facts
INFOnline (tagesschau.de)

Source: Hans-Bredow-Institut

4.1.3 Reader preferences

In addition to media consumption, the participants of the survey on digital news were asked where they would rank themselves on a political scale defined by "left", "center" and "right".[15] Figure 5 shows that a large proportion of the respondents place themselves in the political centre (30%) or slightly to the right (18.7%) or the left (22.5%) of the center. Only a small part would consider themselves to be fairly or very right (6.1%) or left (8.3%) of the political spectrum.

Figure 5: Political orientation

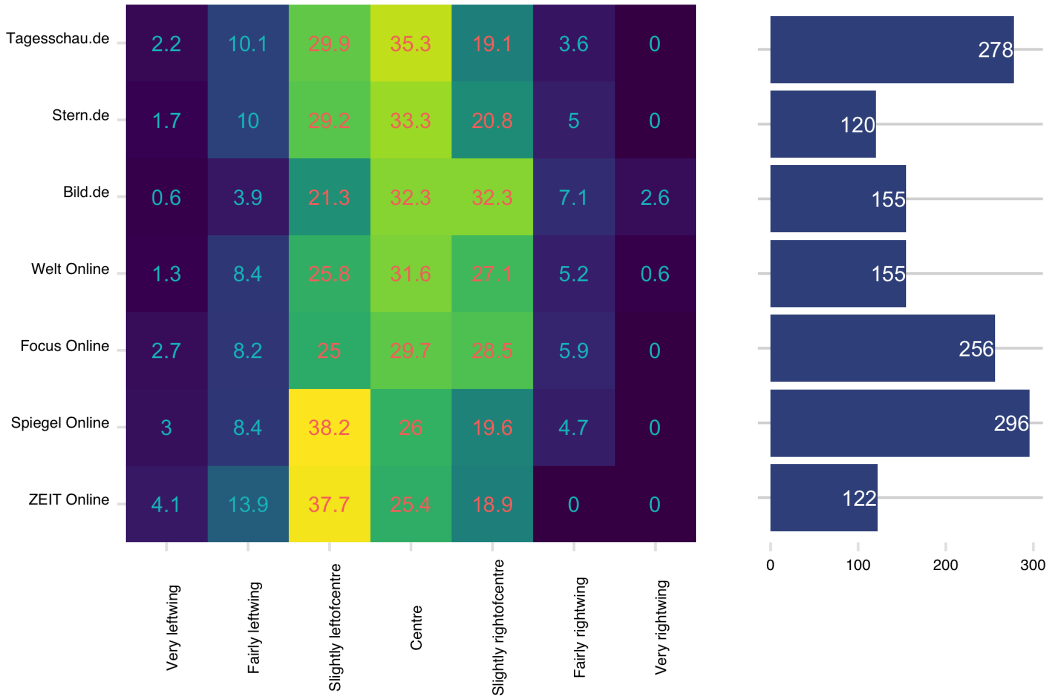


Combining both the political orientation and the brand preferences can give an indication about the political orientation of the consumers of a specific medium. Thus, for

[15] The exact question was: Some people talk about 'left', 'right' and 'centre' to describe parties and politicians. (Generally socialist parties would be considered 'left wing' whilst conservative parties would be considered 'right wing'). With this in mind, where would you place yourself on the following scale?

example, it is possible to answer the question of how many percent of the respondents who stated that they consume stern.de classify themselves as rather left of the political spectrum. To calculate this value, the absolute values for each medium-political orientation combination in Figure 6 was divided by the total number of respondents, who consume a specific medium. As the results show, the majority of consumers of ZEIT Online and Spiegel Online place themselves rather on the left, while the opposite is true for Focus Online, Welt Online and Bild.de. Consumers of Tagesschau.de and stern.de would rather place themselves in the center of the political spectrum. In Section 5.3, these data is used to compare the political positioning of readers with the index calculated on the basis of text data (press releases and news articles).

Figure 6: Political orientation



4.2 Data preparation

To use text as data for statistical analysis, different pre-processing steps have to be conducted. In fact, in order to use text as data and reduce the dimensionality to avoid unnecessary computational complexity and overfitting, pre-processsing the text is a central task in text mining (Bholat et al., 2015; M. Gentzkow et al., 2017). Intuitively the term frequency (tf) of a word is a measure of how important that word may be for the understanding of the text. To visualize these terms, word clouds are a commonly used technique in text mining as they translate the tf into the size of the term in the cloud. As can be seen in Figure 7a, problems arise with words that are highly frequent. For example "die", or "der (eng. "the"), "und" (eng. "and"), and "ist" (eng. "is") are extremely common but unrelated to the quantity of interest. These terms, often called stop words (M. Gentzkow et al., 2017), are important to the grammatical structure of a text, but typically don't add any additional meaning and can therefore be neglected.

To remove distorting words, the pre-defined stop word list from the Snowball project[16]

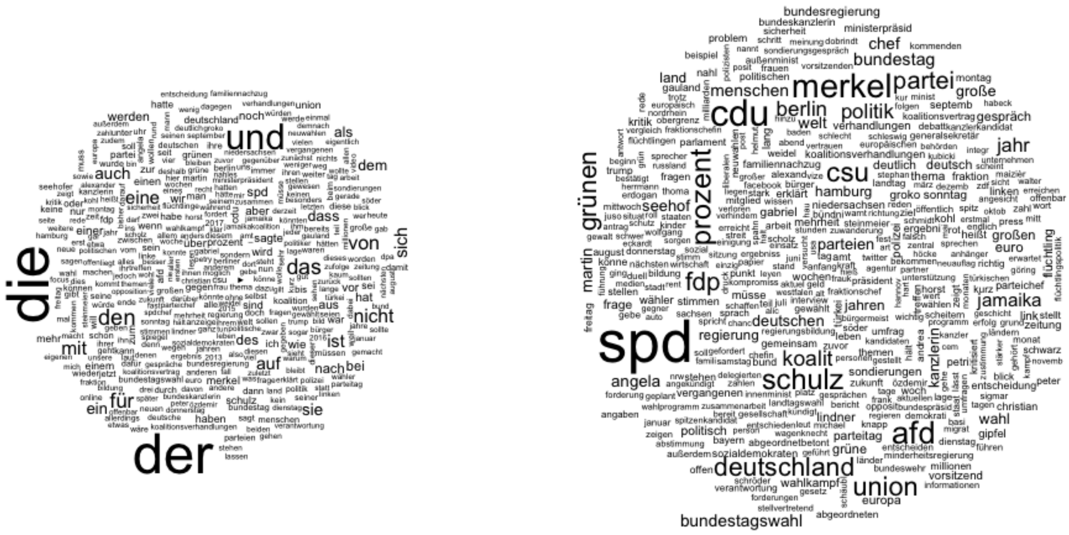
[16] http://snowball.tartarus.org/algorithms/german/stop.txt

is used together with a customized, domain-specific list of stop-words. Additionally punctuation character (e.g. ., „ !, ?, etc.) and all numbers are removed from the data. A next step to reduce the dimensionality of text data is to apply an adequate stemming technique. Stemming is a process by which different morphological variants of a word are traced back to their common root. For example, "voting" and "vote" would be treated as two instances of the same token after the stemming process. There are many different techniques for the stemming process. I apply the widely used Porter-Stemmer algorithm, which is based on a set of shortening rules that are applied to a word until it has a minimum number of syllables.[17] After completing these steps 63,360 unique terms were left in the vocabulary. The word clouds in Figure 7b represent the most frequent words after pre-processing the text data. It becomes evident that the words "SPD", "CDU" and "AfD" among others seems to be highly frequent.

Figure 7: Wordclouds

(a) before pre-processing

(b) after pre-processing



The next step is to divide the entire dataset into individual documents and to represent these documents as a finite list of unique terms. In this setting, each news article and each press release represents a document d, whereby each of these documents can be assigned to a news website or a party. The sum of all documents forms what is called the corpus. For each document $d \in \{1, ..., D\}$ the number of occurrences of term v in document d is computed, in order to obtain the count $x_{d,v}$, where each unique term in the corpus is indexed by some $v \in \{1, ..., V\}$ and where V is the number of unique terms. The D x V matrix $\boldsymbol{X}$ of all such counts is called the document-term matrix. Each row in this matrix represents a document, where each entry in this row counts the occurrences of a unique term in that document. This representation is often referred to as the bag of words model (M. Gentzkow et al., 2017), since the order in which words are used within a document is disregarded.

[17] https://tartarus.org/martin/PorterStemmer/

4.3 Structural topic model

To find out the latent topics of each document, a structural topic model (STM) is estimated. In general, topic models formalize the idea that documents are formed by hidden variables (topics) that generate correlations among observed terms. They belong to the group of unsupervised generative models, meaning that the true attributes (topics) cannot be observed. The STM developed by M. E. ROBERTS, B. M. STEWART, and E. M. AIROLDI (2016) is a recent extension of the standard topic modelling technique, labeled as latent Dirichlet allocation (LDA), which refers to the Bayesian model in BLEI et al. (2003) that treats each word in a topic and each topic in a document as generated from a Dirichlet - distributed prior.[18].

One crucial assumption to be made for such models is the number of topics (K) that occur over the entire corpus. The underlying idea for these models suggests that each individual topic k potentially contains all of the unique terms within the vocabulary V with different probability. Therefore, each topic k can be represented as a probability vector ϕ_k over all unique terms V. Simultaneously, each individual document d in the corpus can be represented as a probability distribution θ_d over the K topics.

The difference between the widely used LDA and the STM approaches lies in how the posterior distributions (θ and ϕ) are determined. LDA assumes that θ Dirichlet(α) and ϕ Dirichlet(β), where α and β are fitted with the model. While for STM, the prior distributions for θ and ϕ depend on document-level covariates (e.g. the author or date of a document). For this purpose, the the STM specifies two design matrices of covariates (X and Z), where each line defines a vector of covariates for a specific document. In X, the covariates for topic prevalence are given, so that the probability of a topic for each document varies according to X, rather than resulting from a single common prior. The same applies to Z, in which the covariates for the word distribution within a topic are specified. The underlying data generating process to generate each individual word $w_{d,n}$ in a document d for the n^{th} word-position can be described as follows:[19]

- for each document i, draw its distribution of topics θ_d depending on the metadata included in the model defined in X;
- for each topic k, draw its distribution of words ϕ_k depending on the metadata included in the model defined in Z;
- for each word n, draw its topic z_n based on θ_i;
- for each word word n, draw the term distribution for the selected topic $\phi_{z_{d,n}}$.

This process generates two posterior distribution parameters:

1. ϕ is a K-by-V matrix (where K = number of topics and V = vocabulary or unique terms), where the entry ϕ_{kvc} can be interpreted as the probability of observing the v-th word in topic k for the covariate level c (the news website).
2. θ is a D-by-V matrix (where D = number of documents and V = vocabulary or unique terms) of the document-topic distributions, where the entry θ_{dk} can be interpreted as the proportion of words in document d which arise from topic k, or rather as the probability that document d deals about topic k.

[18] See also GRIFFITHS and STEYVERS (2002), GRIFFITHS and STEYVERS (2004) and HOFMANN (1999)

[19] A more detailed description of the generative process of the STM can be found in section A.1

Inference of mixed-membership models, such as the one applied in this paper, has been a thread of research in applied statistics (BLEI et al., 2003; BRAUN and MCAULIFFE, 2010; EROSHEVA et al., 2004). Topic models are usually imprecise as the function to be optimized has multiple modes, such that the model results can be sensitive to the starting values (e.g. the number of topics). Since an ex ante valuation of a model is hardly possible, I compute a variety of different models and compare their posterior probability. This enables me to check how results vary for different model solution (M. ROBERTS, B. STEWART, and TINGLEY, 2016a). I then cross-checked some subset of assigned topic distributions to evaluate whether the estimates align with the concept of interest (M. GENTZKOW et al., 2017). These manual audits are applied together with numeric optimization based on the topic coherence measure suggested by MIMNO et al. (2011).

This process revealed that a model with 60 topics best reflects the structure in the corpus. Furthermore, the source (news website or party) of a document is used as covariate in the topic prevalence. In other words, I assume that the probability distribution of topics for a specific document is influenced by the source of that document. Additionally the type of that source is used as a covariate for the term frequency as I assume that the words used for the same topic differ between news articles or press releases.

4.4 Agenda Correlation

5 Results

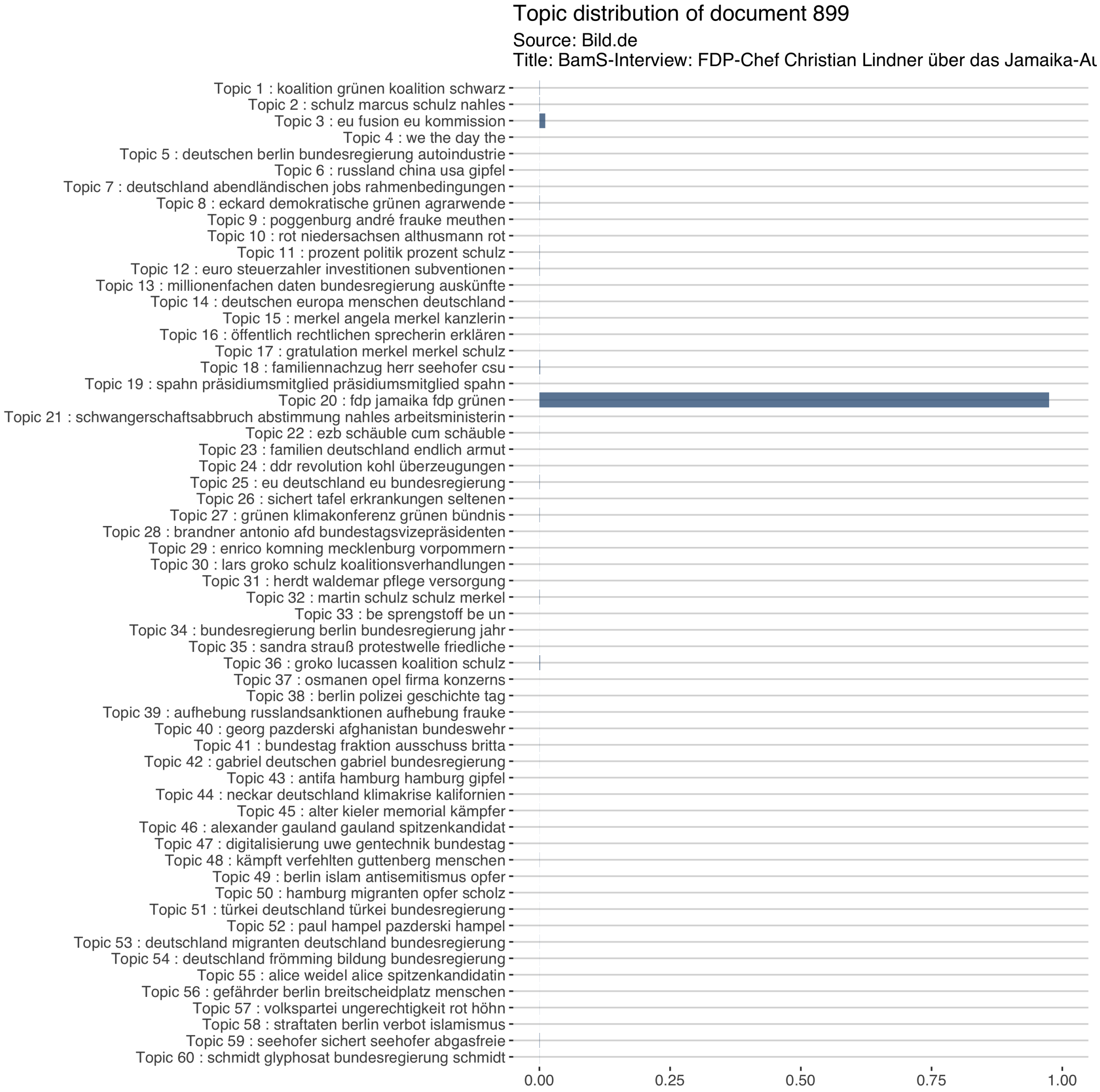
5.1 Topic distribution

As stated in section 4.3, the generative process of the STM results in two posterior distributions: (1) a distribution over all topics for each document and (2) a distribution over all words for each topic. Since the type of source was assumed as a covariates in the model, the result consists of two different term distributions ϕ for each topic. These distributions are used to understand what each topic is about and to label it.[20].

For each document d, we have a distribution θ_d over all topics k. An example of this distribution is shown in Figure 8.

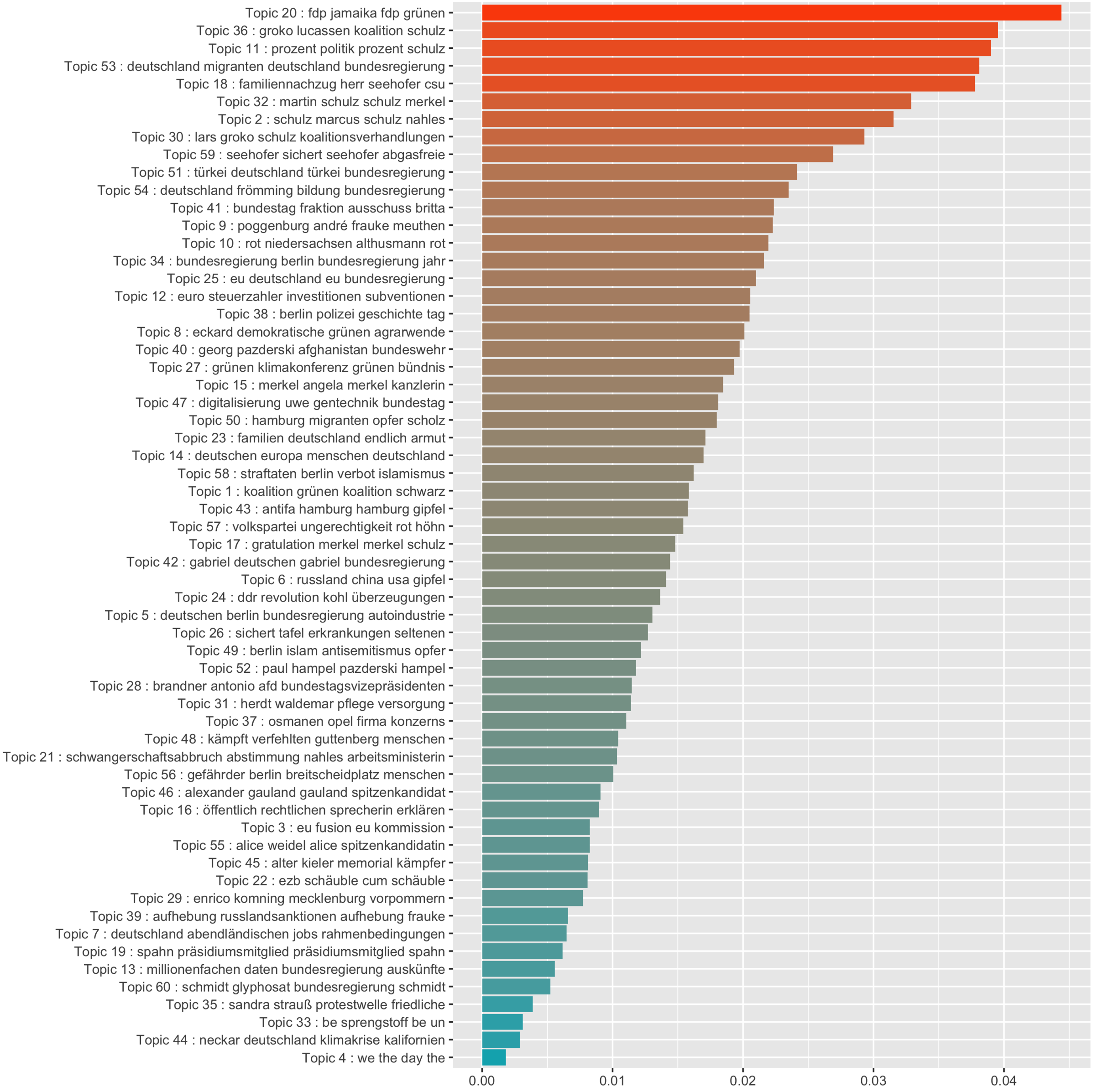
[20] See A.2 for a table of the most frequent words in each topic for the different covariates.

Figure 8: Document-Topic distribution



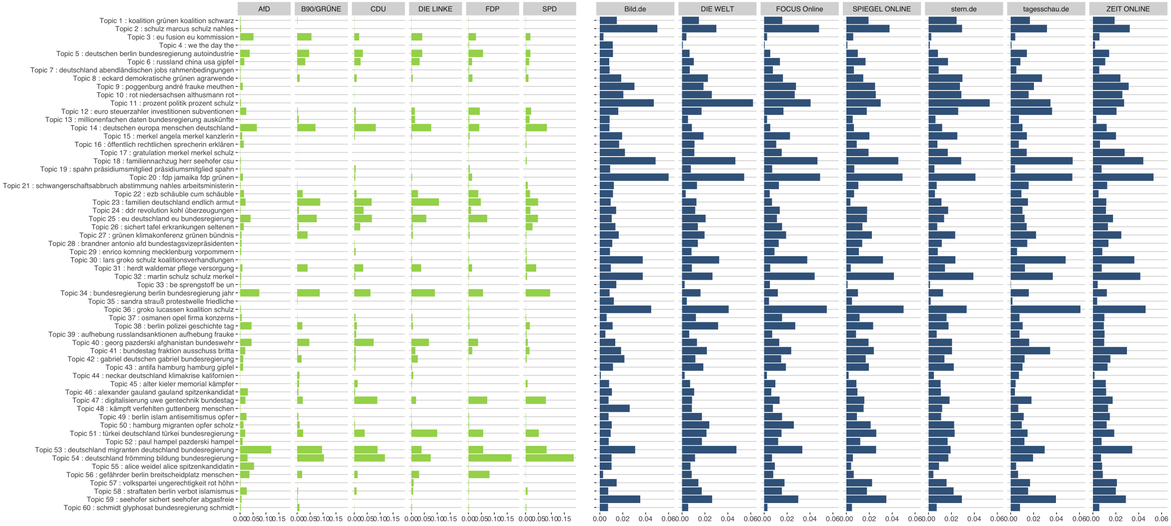
The average of each topic across all documents in the corpus results in the expected probability of a topic across the corpus. As shown in Figure 11 topic 20 about the so-called Jamaica parties (CDU/CSU, FDP and B90/Die Grünen) is the topic with the highest expected frequency in the whole corpus, followed by topic 36 about the coalition talks between CDU/CSU and SPD - the "Grand coalition" or "GroKo".

Figure 9: Document-Topic distribution



However, the model was calculated under the assumption that each document source has a different distribution across all topics. Looking at the expected frequency of a topic for each source separately, some differences become apparent.

Figure 10: Expected frequency



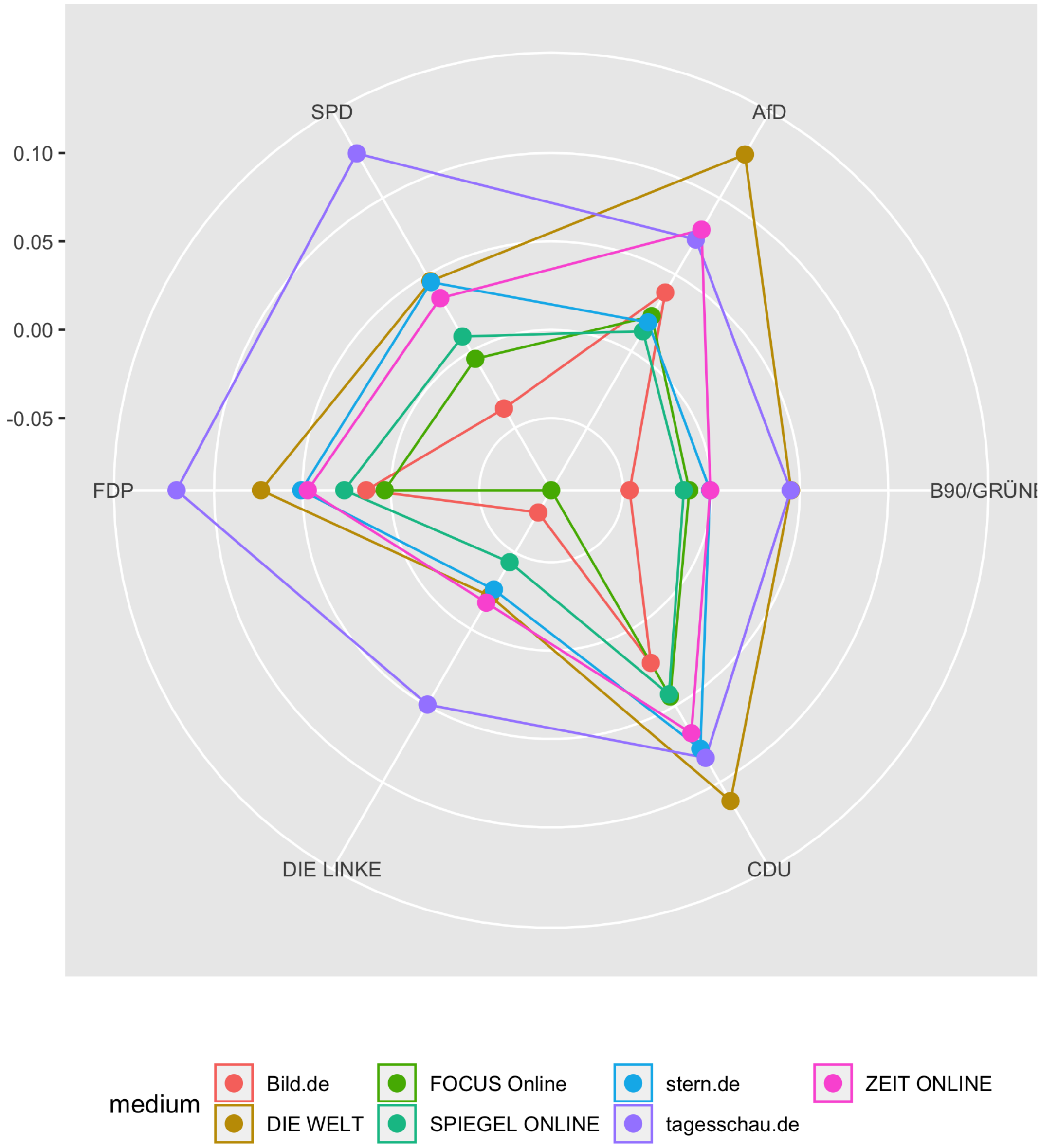
(a) Online news

(b) Press releases

5.2 Agenda correlation

Agendas were measured in terms of percentage distributions across the 80 topics. For each source the average distribution of each topic is calculated for each month. The following pictures show the overall topic distribution. Then, we estimated bivariate correlations between party agendas and the mediated party agendas in the online news. These correlations represent the agenda selectivity each party experiences in each media outlet. The higher the correlation, the more congruent both agendas are.

Figure 11: Document-Topic distribution



5.3 Reader Preferences

6 Conclusion

The ongoing discussion about the influence of digital media on the political opinion-forming process addresses the question whether there are convergence tendencies within the mass media and whether the reporting in the media correlates with the voting preferences. To analyze this question, this paper examines ....

Using text data of 14,937 online news articles from seven German news providers about domestic politics, I first estimate a Structural Topic Model to find the latent topics in the news articles ...

References

Balahur, Alexandra et al. (Sept. 24, 2013). “Sentiment Analysis in the News”. In: *arXiv:1309.6202 [cs]*. arXiv: 1309.6202. URL: http://arxiv.org/abs/1309.6202 (visited on 11/15/2018).

Baron, David P. (Jan. 1, 2006). “Persistent media bias”. In: *Journal of Public Economics* 90.1, pp. 1–36. URL: http://www.sciencedirect.com/science/article/pii/S0047272705000216 (visited on 01/19/2019).

Baturo, Alexander, Niheer Dasandi, and Slava J. Mikhaylov (Aug. 19, 2017). “What Drives the International Development Agenda? An NLP Analysis of the United Nations General Debate 1970-2016”. In: *arXiv:1708.05873 [cs]*. arXiv: 1708.05873. URL: http://arxiv.org/abs/1708.05873.

Benewick, R. J. et al. (June 1, 1969). “THE FLOATING VOTER AND THE LIBERAL VIEW OF REPRESENTATIONa”. In: *Political Studies* 17.2, pp. 177–195. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9248.1969.tb00634.x (visited on 08/14/2018).

Bholat, David M. et al. (June 29, 2015). “Text Mining for Central Banks”. In: *SSRN Electronic Journal*. URL: http://www.academia.edu/13430482/Text_mining_for_central_banks (visited on 11/06/2017).

Blei, David M. (Apr. 2012). “Probabilistic Topic Models”. In: *Commun. ACM* 55.4, pp. 77–84. URL: http://doi.acm.org/10.1145/2133806.2133826.

Blei, David M., Andrew Y Ng, and Michael I Jordan (Jan. 2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3, pp. 993–1022.

Box, George E. P., Gwilym M. Jenkins, and Gregory C. Reinsel (2008). *Time series analysis: forecasting and control*. 4th ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley. xxiv+746.

Brandenburg, Heinz (July 1, 2006). “Party Strategy and Media Bias: A Quantitative Analysis of the 2005 UK Election Campaign”. In: *Journal of Elections, Public Opinion and Parties* 16.2, pp. 157–178. URL: https://doi.org/10.1080/13689880600716027 (visited on 09/25/2018).

Braun, Michael and Jon McAuliffe (Mar. 2010). “Variational inference for large-scale models of discrete choice”. In: *Journal of the American Statistical Association* 105.489, pp. 324–335. arXiv: 0712.2526. URL: http://arxiv.org/abs/0712.2526 (visited on 01/19/2018).

D’Alessio, Dave and Mike Allen (Dec. 1, 2000). “Media Bias in Presidential Elections: A Meta-Analysis”. In: *Journal of Communication* 50.4, pp. 133–156. URL: https://academic.oup.com/joc/article/50/4/133/4110147 (visited on 08/14/2018).

DellaVigna, Stefano and Ethan Kaplan (Apr. 2006). *The Fox News Effect: Media Bias and Voting*. Working Paper 12169. National Bureau of Economic Research. URL: http://www.nber.org/papers/w12169 (visited on 08/22/2018).

Dewenter, Ralf, Melissa Linder, and Tobias Thomas (Apr. 2018). “Can Media Drive the Electorate? The Impact of Media Coverage on Party Affiliation and Voting Intentions”. In: *Working Paper Series, Helmut Schmidt University Hamburg, Department of Economics* 179.

Druckman, James N. and Michael Parkin (Nov. 1, 2005). “The Impact of Media Bias: How Editorial Slant Affects Voters”. In: *The Journal of Politics* 67.4, pp. 1030–1049. URL: https://www.journals.uchicago.edu/doi/full/10.1111/j.1468-2508.2005.00349.x (visited on 09/25/2018).

Eberl, Jakob-Moritz, Hajo G. Boomgaarden, and Markus Wagner (Dec. 1, 2017). “One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences”.

In: *Communication Research* 44.8, pp. 1125–1148. URL: https://doi.org/10.1177/0093650215614364 (visited on 10/20/2018).

Erosheva, Elena, Stephen Fienberg, and John Lafferty (Apr. 6, 2004). "Mixed-membership models of scientific publications". In: *Proceedings of the National Academy of Sciences* 101 (suppl 1), pp. 5220–5227. URL: http://www.pnas.org/content/101/suppl_1/5220 (visited on 01/19/2018).

Farrell, Justin (Jan. 5, 2016). "Corporate funding and ideological polarization about climate change". In: *Proceedings of the National Academy of Sciences* 113.1, pp. 92–97. URL: http://www.pnas.org/content/113/1/92 (visited on 11/09/2017).

Fu, King-wa and Chee-hon Chan (Sept. 2013). "Analyzing Online Sentiment to Predict Telephone Poll Results". In: *Cyberpsychology, Behavior, and Social Networking* 16.9, pp. 702–707. URL: http://online.liebertpub.com/doi/abs/10.1089/cyber.2012.0375 (visited on 03/19/2018).

Gentzkow, Matthew (Aug. 1, 2006). "Television and Voter Turnout". In: *The Quarterly Journal of Economics* 121.3, pp. 931–972. URL: https://academic.oup.com/qje/article/121/3/931/1917885 (visited on 01/19/2019).

Gentzkow, Matthew A. and Jesse M. Shapiro (Sept. 2004). "Media, Education and Anti-Americanism in the Muslim World". In: *Journal of Economic Perspectives* 18.3, pp. 117–133. URL: https://www.aeaweb.org/articles?id=10.1257/0895330042162313 (visited on 01/11/2019).

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy (Mar. 2017). *Text as Data*. Working Paper 23276. National Bureau of Economic Research. URL: http://www.nber.org/papers/w23276.

Griffiths, Thomas L. and Mark Steyvers (Jan. 1, 2002). "A probabilistic approach to semantic representation". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 24.24. URL: https://escholarship.org/uc/item/44x9v7m7 (visited on 11/16/2017).

— (Apr. 6, 2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101 (suppl 1), pp. 5228–5235. URL: http://www.pnas.org/content/101/suppl_1/5228 (visited on 10/12/2017).

Grimmer, Justin and Brandon Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21, pp. 267–297.

Groseclose, Tim and Jeffrey Milyo (2005). "A Measure of Media Bias". In: *The Quarterly Journal of Economics* 120.4, pp. 1191–1237. URL: https://www.jstor.org/stable/25098770 (visited on 01/07/2019).

Hansen, Stephen and Michael McMahon (Mar. 1, 2016). "Shocking language: Understanding the macroeconomic effects of central bank communication". In: *Journal of International Economics*. 38th Annual NBER International Seminar on Macroeconomics 99, S114–S133. URL: http://www.sciencedirect.com/science/article/pii/S0022199615001828 (visited on 03/07/2018).

Hofmann, Thomas (1999). "Probabilistic Latent Semantic Indexing". In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '99. New York, NY, USA: ACM, pp. 50–57. URL: http://doi.acm.org/10.1145/312624.312649.

Hölig, Sascha and Uwe Hasebrink (June 2018). "Reuters Institute Digital News Report 2018 – Ergebnisse für Deutschland". In: *Arbeitspapiere des Hans-Bredow- Instituts* 44.

Hopmann, David Nicolas, Peter Van Aelst, and Guido Legnante (Feb. 1, 2012). "Political balance in the news: A review of concepts, operationalizations and key findings". In: *Journalism* 13.2, pp. 240–257. URL: https://doi.org/10.1177/1464884911427804 (visited on 11/10/2018).

Hurtíková, Hanna (Dec. 21, 2017). "The Importance of Valence-Framing in the Process of Political Communicati on: Effects on the Formation of Political Attitudes among Viewers of Television News in the Czech Republic | Media Studies". In: 8.15. URL: https://hrcak.srce.hr/ojs/index.php/medijske-studije/article/view/6200 (visited on 11/16/2018).

Junqué de Fortuny, Enric et al. (Oct. 15, 2012). "Media coverage in times of political crisis: A text mining approach". In: *Expert Systems with Applications* 39.14, pp. 11616–11622. URL: http://www.sciencedirect.com/science/article/pii/S0957417412006100 (visited on 08/14/2018).

Kepplinger, Hans Mathias and Marcus Maurer (2004). "Der Einfluss der Pressemitteilungen der Bundesparteien auf die Berichterstattung im Bundestagswahlkampf 2002". In: *Quo vadis Public Relations? Auf dem Weg zum Kommunikationsmanagement: Bestandsaufnahmen und Entwicklungen*. Ed. by Juliana Raupp and Joachim Klewes. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 113–124. URL: https://doi.org/10.1007/978-3-322-83381-5_9.

Lott, John R. and Kevin A. Hassett (July 1, 2014). "Is newspaper coverage of economic events politically biased?" In: *Public Choice* 160.1, pp. 65–108. URL: https://doi.org/10.1007/s11127-014-0171-5 (visited on 01/07/2019).

Mimno, David et al. (2011). "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 262–272. URL: http://dl.acm.org/citation.cfm?id=2145432.2145462.

Mishler, Alan et al. (Aug. 2, 2015). "Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis". In: *HCI International 2015 - Posters' Extended Abstracts*. International Conference on Human-Computer Interaction. Communications in Computer and Information Science. Springer, Cham, pp. 639–644. URL: https://link.springer.com/chapter/10.1007/978-3-319-21380-4_108 (visited on 10/12/2017).

Mueller, Hannes Felix and Christopher Rauh (Sept. 1, 2016). *Reading between the Lines: Prediction of Political Violence Using Newspaper Text*. SSRN Scholarly Paper ID 2843535. Rochester, NY: Social Science Research Network. URL: https://papers.ssrn.com/abstract=2843535 (visited on 11/09/2017).

Nyman, Rickard et al. (May 1, 2018). "News and narratives in financial systems: exploiting big data for systemic risk assessment | Bank of England". In: *Bank of England Working Paper* 704. URL: https://www.bankofengland.co.uk/working-paper/2018/news-and-narratives-in-financial-systems (visited on 02/21/2018).

Padmaja, S., Prof S. Sameen Fatima, and Sasidhar Bandu (2014). "Evaluating Sentiment Analysis Methods and Identifying Scope of Negation in Newspaper Articles". In: *International Journal of Advanced Research in Artificial Intelligence (IJARAI)* 3.11. URL: http://thesai.org/Publications/ViewPaper?Volume=3&Issue=11&Code=IJARAI&SerialNo=1 (visited on 03/19/2018).

Paul, Michael (2009). "Cross-Collection Topic Models: Automatically Comparing and Contrasting Text". Master Thesis. University of Illinois at Urbana-Champaign.

Pritchard, J. K., M. Stephens, and P. Donnelly (June 2000). "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2, pp. 945–959.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi (July 2, 2016). "A Model of Text for Experimentation in the Social Sciences". In: *Journal of the*

American Statistical Association 111.515, pp. 988–1003. URL: http://dx.doi.org/10.1080/01621459.2016.1141684.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, et al. (Oct. 1, 2014). “Structural Topic Models for Open-Ended Survey Responses”. In: *American Journal of Political Science* 58.4, pp. 1064–1082. URL: http://onlinelibrary.wiley.com/doi/10.1111/ajps.12103/abstract.

Roberts, Margaret, Brandon Stewart, and Dustin Tingley (2016a). “Navigating the Local Modes of Big Data: The Case of Topic Models.” In: *Computational Social Science: Discovery and Prediction.* New York: Cambridge University Press.

— (Dec. 1, 2016b). “stm: R Package for Structural Topic Models”. In: *Journal of Statistical Software* forthcoming.

Roberts, Margaret, Brandon Stewart, Dustin Tingley, and Edoardo Airoldi (2013). “The Structural Topic Model and Applied Social Science”. In: *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.*

Snyder, James M. and David Strömberg (2010). “Press Coverage and Political Accountability”. In: *Journal of Political Economy* 118.2, pp. 355–408. URL: https://www.jstor.org/stable/10.1086/652903 (visited on 08/22/2018).

Strömberg, David (Feb. 1, 2004). “Radio’s Impact on Public Spending”. In: *The Quarterly Journal of Economics* 119.1, pp. 189–221. URL: https://academic.oup.com/qje/article/119/1/189/1876059 (visited on 01/11/2019).

Tetlock, Paul C. (June 1, 2007). “Giving Content to Investor Sentiment: The Role of Media in the Stock Market”. In: *The Journal of Finance* 62.3, pp. 1139–1168. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2007.01232.x/abstract.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy (June 1, 2008). “More Than Words: Quantifying Language to Measure Firms’ Fundamentals”. In: *The Journal of Finance* 63.3, pp. 1437–1467. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2008.01362.x/abstract (visited on 03/07/2018).

Tumasjan, Andranik et al. (2010). “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”. In: *Proceedings of the Fourth International Conference on Weblogs and Social Media.* INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA. Washington. URL: https://www.researchgate.net/publication/215776042_Predicting_Elections_with_Twitter_What_140_Characters_Reveal_about_Political_Sentiment (visited on 03/17/2018).

Vreese, Claes de and Hajo G. Boomgaarden (2006). “Valenced news frames and public support for the EU”. In: *Communications* 28.4, pp. 361–381. URL: https://www.degruyter.com/view/j/comm.2003.28.issue-4/comm.2003.024/comm.2003.024.xml (visited on 10/20/2018).

Wiedmann, Gregor (2016). *Text Mining for Qualitative Data Analysis in the Social Sciences.* 1st ed. Wiesbaden: VS Verlag für Sozialwissenschaften. URL: //www.springer.com/de/book/9783658153083 (visited on 11/26/2017).

A Appendices

A.1 Generative Process of STM

The following describes the generative process for filling the n^{th} word-position in document d in the case of the STM (M. Roberts, B. Stewart, Tingley, and E. Airoldi, 2013): As in the case of conventional models, first a specific topic z_{dn} is assigned, according to the topic distribution for that document θ through the process:

$$z_{dn}|\theta_d \sim \text{Multinomial}(\theta_d). \tag{1}$$

To incorporate the covariate values for that document, a topic-prevalence vector θ_d is drawn from a logistic-normal distribution:

$$\theta_d|y_{d\gamma}, \Sigma \sim \text{LogisticNormal}(\mu = y_{d\gamma}\Sigma), \tag{2}$$

where $y_d\gamma$ lists the values of the metadata covariates for document d and γ relates these covariate values to the topic-prevalence.

Conditional in the topic chosen (z_{dn}), a specific word w_{dn}, is selected from the overall corpus vocabulary V, using the following process:

$$w_{dn}|z_{dn}, \phi_{dkv} \sim \text{Multinomial}(\phi_{dk1}, ..., \phi_{dkV}), \tag{3}$$

where the word probability ϕ_{dkv} is parameterized in terms of log-transformed rate deviations from the rates of a corpus-wide background distribution m_v (M. Roberts, B. Stewart, Tingley, and E. Airoldi, 2013). The log-transformed rate deviations can then be specified by a collection of parameters $\{\boldsymbol{\kappa}\}$, where $\kappa^{(t)}$ is a K-by-V matrix containing the log-transformed rate deviations for each topic k and term v, over the baseline log-transformed rate for term v. This matrix is the same for all A levels of covariates. To put it differently, $\kappa^{(t)}$ indicates the importance of the term v given topic k regardless of the covariates. Similarly, $\kappa^{(c)}$ is a A-by-V matrix, indicating the importance of the term v given the covariate level c regardless of the topic. Finally, $\kappa^{(i)}$ is a A-by-K-by-V matrix, collecting the covariate-topic effects:

$$\phi_{dkv}|z_{dn} = \frac{\exp(m_v + \kappa^{(t)}_{kv}, \kappa^{(c)}_{y_d v} + \kappa^{(i)}_{y_d kv})}{\sum_v \exp(m_v + \kappa^{(t)}_{kv}, \kappa^{(c)}_{y_d v} + \kappa^{(i)}_{y_d kv})}. \tag{4}$$

The STM maximizes the posterior likelihood that the observed data were generated by the above data-generating process using an iterative approximation-based variational expectation-maximization algorithm[21] available in R's stm package (M. Roberts, B. Stewart, and Tingley, 2016b).

A.2 Topic labels

[21] A technical description of this maximization process can be found in M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016)

Joint label	News articles
Topic 1 : koalition grünen koalition schwarz	koalition grünen rot gr
Topic 2 : schulz marcus schulz nahles	schulz marcus bühl nal
Topic 3 : eu fusion eu kommission	eu fusion meuthen jörg
Topic 4 : we the day the	we the at is
Topic 5 : deutschen berlin bundesregierung autoindustrie	deutschen berlin diesel
Topic 6 : russland china usa gipfel	russland china us dialo
Topic 7 : deutschland abendländischen jobs rahmenbedingungen	deutschland abendländ
Topic 8 : eckard demokratische grünen agrarwende	eckard demokratische v
Topic 9 : poggenburg andré frauke meuthen	poggenburg andré sach
Topic 10 : rot niedersachsen althusmann rot	rot niedersachsen landt
Topic 11 : prozent politik prozent schulz	prozent politik flüchtlin
Topic 12 : euro steuerzahler investitionen subventionen	euro steuerzahler millia
Topic 13 : millionenfachen daten bundesregierung auskünfte	millionenfachen daten
Topic 14 : deutschen europa menschen deutschland	deutschen europa berli
Topic 15 : merkel angela merkel kanzlerin	merkel angela kanzlerin
Topic 16 : öffentlich rechtlichen sprecherin erklären	öffentlich rechtlichen a
Topic 17 : gratulation merkel merkel schulz	gratulation merkel berl
Topic 18 : familiennachzug herr seehofer csu	familiennachzug herr k
Topic 19 : spahn präsidiumsmitglied präsidiumsmitglied spahn	spahn präsidiumsmitgl
Topic 20 : fdp jamaika fdp grünen	fdp jamaika christian g
Topic 21 : schwangerschaftsabbruch abstimmung nahles arbeitsministerin	schwangerschaftsabbru
Topic 22 : ezb schäuble cum schäuble	ezb schäuble schaden b
Topic 23 : familien deutschland endlich armut	familien deutschland k
Topic 24 : ddr revolution kohl überzeugungen	ddr revolution friedlich
Topic 25 : eu deutschland eu bundesregierung	eu deutschland europa
Topic 26 : sichert tafel erkrankungen seltenen	sichert tafel menschen
Topic 27 : grünen klimakonferenz grünen bündnis	grünen klimakonferenz
Topic 28 : brandner antonio afd bundestagsvizepräsidenten	brandner antonio amac
Topic 29 : enrico komning mecklenburg vorpommern	enrico komning meckle
Topic 30 : lars groko schulz koalitionsverhandlungen	lars groko koalitionsver
Topic 31 : herdt waldemar pflege versorgung	herdt waldemar reform
Topic 32 : martin schulz schulz merkel	martin schulz merkel s
Topic 33 : be sprengstoff be un	be sprengstoff brandne
Topic 34 : bundesregierung berlin bundesregierung jahr	bundesregierung berlin
Topic 35 : sandra strauß protestwelle friedliche	sandra strauß palmer b
Topic 36 : groko lucassen koalition schulz	groko lucassen wales st
Topic 37 : osmanen opel firma konzerns	osmanen opel berlin po
Topic 38 : berlin polizei geschichte tag	berlin polizei prozent g
Topic 39 : aufhebung russlandsanktionen aufhebung frauke	aufhebung russlandsan
Topic 40 : georg pazderski afghanistan bundeswehr	georg pazderski bundes
Topic 41 : bundestag fraktion ausschuss britta	bundestag fraktion ant
Topic 42 : gabriel deutschen gabriel bundesregierung	gabriel deutschen deuts
Topic 43 : antifa hamburg hamburg gipfel	antifa hamburg storch
Topic 44 : neckar deutschland klimakrise kalifornien	neckar deutschland ber
Topic 45 : alter kieler memorial kämpfer	alter kieler günther sch
Topic 46 : alexander gauland gauland spitzenkandidat	alexander gauland spit
Topic 47 : digitalisierung uwe gentechnik bundestag	digitalisierung uwe digi
Topic 48 : kämpft verfehlten guttenberg menschen	kämpft verfehlten bete
Topic 49 : berlin islam antisemitismus opfer	berlin islam driesang d
Topic 50 : hamburg migranten opfer scholz	hamburg migranten ab
Topic 51 : türkei deutschland türkei bundesregierung	türkei deutschland deu
Topic 52 : paul hampel pazderski hampel	paul hampel bundesvor
Topic 53 : deutschland migranten deutschland bundesregierung	deutschland migranten