Matching agendas

Franziska Löw *

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

April 29, 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. 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. However, in the case of public services, one would assume that the reporting reflects the mandate to contribute to the formation of individual and public opinion and to act politically and economically independently.

In order to answer these and other media-related questions in the political context, quantifying the content of media is a prerequisite. One of the key challenges is to determine the features that are used to describe media content (audio, video, text). 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). Other studies examine the topics discussed or the language used in the media, 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.

Following this approach, the present paper compares topics discussed in media outlets with topics addressed in the press releases of the parties in the german "Bundestag", to measure the ideological content of several online news services in Germany. The dataset contains over 15.000 online news articles from seven major news provider as well as 2.666 press releases of the parties in the german "Bundestag", both dated from June 1, 2017 to March 1, 2018. 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 estimating the correlation between the posterior distribution of topics in press releases with the posterior distributions of topics in news articles. The resulting slant index of a newspaper is used to identify possible drivers of the differences between the media outlets: 1) the political orientation

of consumers/consumer preferences, 2) current poll values, 3) differences between public or private media. 1) and 2) no evidence for causality!

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 **??**. 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

This paper combines research in the field of media bias with the literature of computer-aided analysis of text data in general and topic models in particular. To put this paper into context within both research areas, this section starts with an overview of the literature on media bias, mainly limited to the economic literature and referring to related disciplines where appropriate. Subsequently, a literature overview on topic models in the relevant research areas is given.

The concept of media bias has been investigated in different research fields in the social sciences (e.g. politics, communication research, economics), although based on different hypotheses and assumptions. The majority of this research attempts to answer one or more of the following questions: (1) Is media biased? (2) how and why does it emerge? (3) What influence does it have on (political/societal) outcomes? From an economic point of view, question (2) analyses the market dynamics that lead to a possible 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 that (2) 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). Basically, media bias can be influenced by the supply side and reflect the preferences of journalists (Baron, 2006), editors or owners (Besley and Prat, 2006). Alternatively, media distortion can also reflect the profit-maximising behaviour of news providers who are responsive to consumer preferences. In this case, the distortion would rather be initiated by the demand side (M. Gentzkow and Shapiro, 2006; Mullainathan and Shleifer, 2005; Suen, 2004).

Since private media outlets are almost exclusively platforms, that connect the market of advertising with the reader market, broadcasters coordinate the two sides to exploit the indirect network effects between them. Advertising-financed media generate their sales on the market side of the advertisers, who in turn are attracted by the readers as potential customers. Anderson and Gabszewicz (2006) argue, that competition for viewers of the demographics most desired by advertisers implies that programming choices will be biased towards the tastes of those with such demographics. However, M. Gentzkow and Shapiro (2010) find that media outlets respond strongly to consumer preferences for like-minded news, which account for 20 percent of the variation in measured slant in their sample of U.S. daily newspapers.

In order to answer the question whether the media is biased, the preliminary question of what media bias means has to be answered. In order to arrive at a better operational definition of bias, some scholars argue that an objective news report is a neutral or balanced report, thus one where all sides are equally represented according to some kind of benchmark for balance or neutrality (Eberl et al., 2017).

The major empirical challenge for the analysis of media content, and one I address head on in this paper, is to convert the raw text into meaningful quantities which can be systematically analysed.

Some approaches focus on the quantitative visibility from a political entity, e.g. how often this entity is mentioned in the article. The frequency of mentioning or mere visibility is then used as an indicator of a party's media coverage (Eberl et al., 2017; Junqué de Fortuny et al., 2014; Oegema and Kleinnijenhuis, 2009). Although this approach analyses the visibility of political actors, it says nothing about the way in which they are covered. The concept of 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 take into account this effect, sentiment analysis is a commonly used approach to measure how parties are discussed in media. To measure tonality in a text, studies differ between manually coded data (Dewenter et al., 2018; Eberl et al., 2017) and dictionary-based analysis (Junqué de Fortuny et al., 2012). To conduct the latter, 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. Junqué de Fortuny et al. (ibid.) 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.[1]

A further approach to quantifying media content in terms of political content is the evaluation of domain-specific language. Groseclose and Milyo (2005) count the times that particular media outlet cites various think tanks and policy groups, and then compare this with the times that members of Congress cite the same groups. Similarly M. Gentzkow and Shapiro (2010) define a set of all phrases used by members of Congress in the 2005 Congressional Record, and identify those that are used much more frequently by one party than another. In both cases, news articles are scanned for the previously defined language. News providers who quote think tanks more frequently or use sentences used by one party rather than another have a bias toward that party.

In addition to domain-specific language, topics can also be used as an indicator of political content, following the assumption that 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) compares topics in press releases of political parties with topics in news agencies in order to find out which media cover the topics addressed by the parties.

The main methodological contribution in this paper is to use computational linguistics, and particularly the combination of topic modelling and dictionary methods, in order to examine the content of what central banks are trying to communicate to the markets and the public.

The first obvious advantage of the use of automated techniques rather than a purely narrative approach to study the statements is scalability without concerns about consistency of the application of the method. With automated methods it is then easy to extend the sample to include more recent data, other sources of communication such as FOMC speeches, or to extend it to other central banks.

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

In terms of the computational approaches, I use a structural topic model (STM) and dictionary methods to extract the content of online news articles of seven major news provider and press releases of the parties in the German Bundestag.

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.[2] Since its introduction into text analysis, LDA has become hugely popular and especially useful in political science.[3] 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.

Furthermore, 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 surveys 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 main contribution of this paper relative to the above mentioned research work is that I combine the effect of topic correlation (Eberl et al., 2017) and sentiment analysis (Dewenter et al., 2018; Eberl et al., 2017; Junqué de Fortuny et al., 2014) to estimate the slant index of media. I also apply different tools from computational linguistics (both STM for topic modelling and dictionary methods to measure sentiment).

The remainder of the paper proceeds as follows...

[2]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.

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

3 Background information

3.1 The political situation in Germany (June 2017 - March 2018)

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.

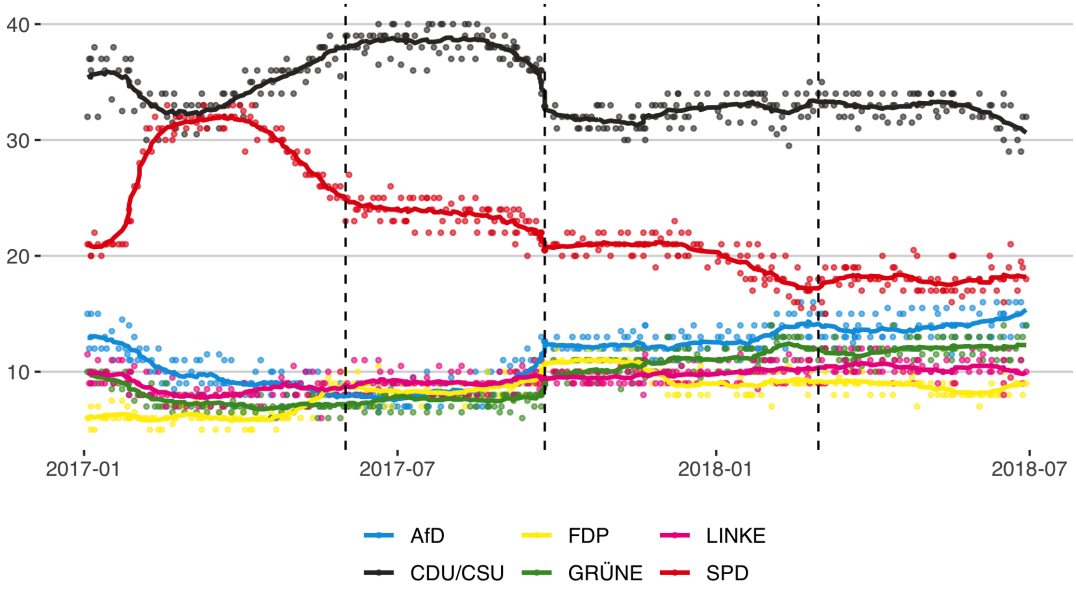
After the election, the formation of a government was difficult due to the large number of parties elected to the Bundestag and the considerable loss of votes by the major parties CDU/CSU and SPD. Since all parties rejected a coalition with the AfD, numerically only two coalitions with an absolute parliamentary 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.[4] However, the poll results of the SPD have been falling since March 2017. At the same time, the AfD in particular has been recording increasingly positive survey results since June 2017.

[4]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/.

Figure 1: Election Polls



3.2 German online news market

The analysis performed in this paper is based on the news articles of the following news websites: Bild.de, DIE WELT, FOCUS ONLINE, SPIEGEL ONLINE, stern.de, ZEIT ONLINE, Tagesschau.de. As can be seen from Figure 2a, these media outlets are among the top 30 German online news providers in the period under review in terms of visits.[5]. Of the selected websites only Tagesschau.de belongs to the public broadcasting and is financed thus by coercive fees - the remaining websites belong to privately managed media houses. The main source of income for private media is digital advertising, even though paid content is playing an increasingly important role. However, according to a survey on digital news by the Reuters Institute (Newman et al., 2018) only 8% of respondents pay for online news. The online survey for German data was undertaken between 19th - 22nd January 2018 by the Hans Bredow Institute[6] 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.[7] The results displayed in Figure 2b indicate that the media used for the analysis play a relevant role in their consumption.

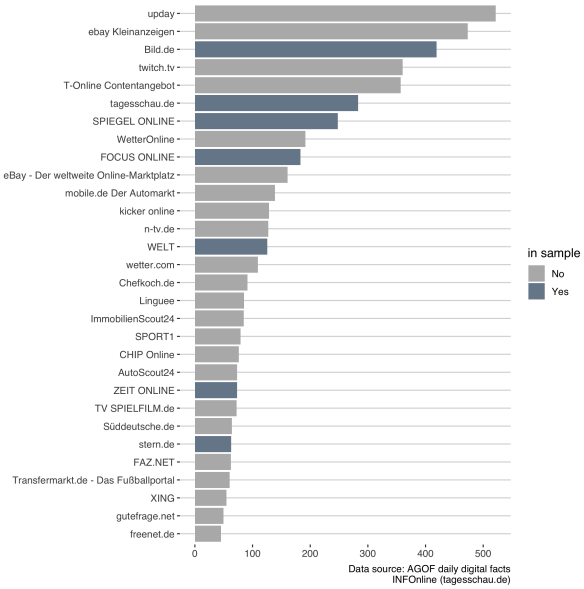
[5]The term visit is used to describe the call to a website by a visitor. The visit begins as soon as a user generates a page impression (PI) within an offer and each additional PI, which the user generates within the offer, belongs to this visit.

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

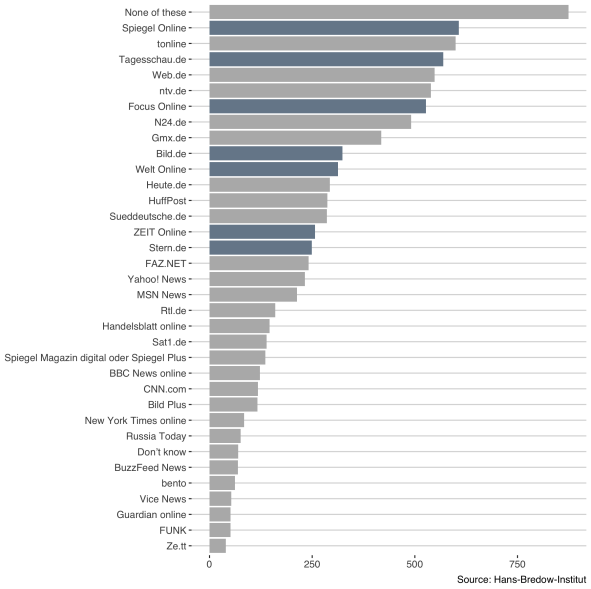
[7]The exact question was: "Which of the following brands have you used to access news online in the last week (via websites, apps, social media, and other forms of Internet access)? Please select all that apply"

Figure 2: Selected news brands

(a) Total visits in million (Jan 2018)

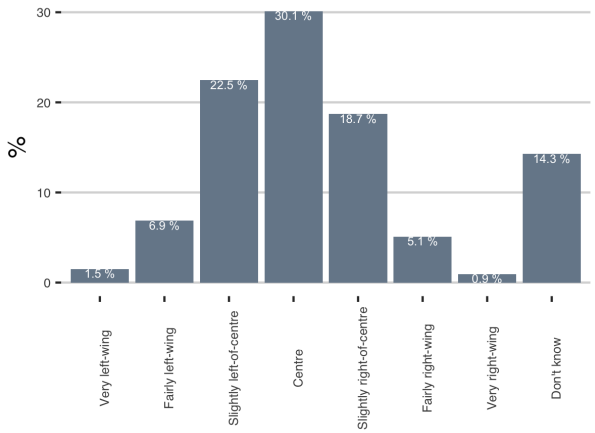


(b) Use of a brand to access news online



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".[8] Figure 3 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 3: Political orientation

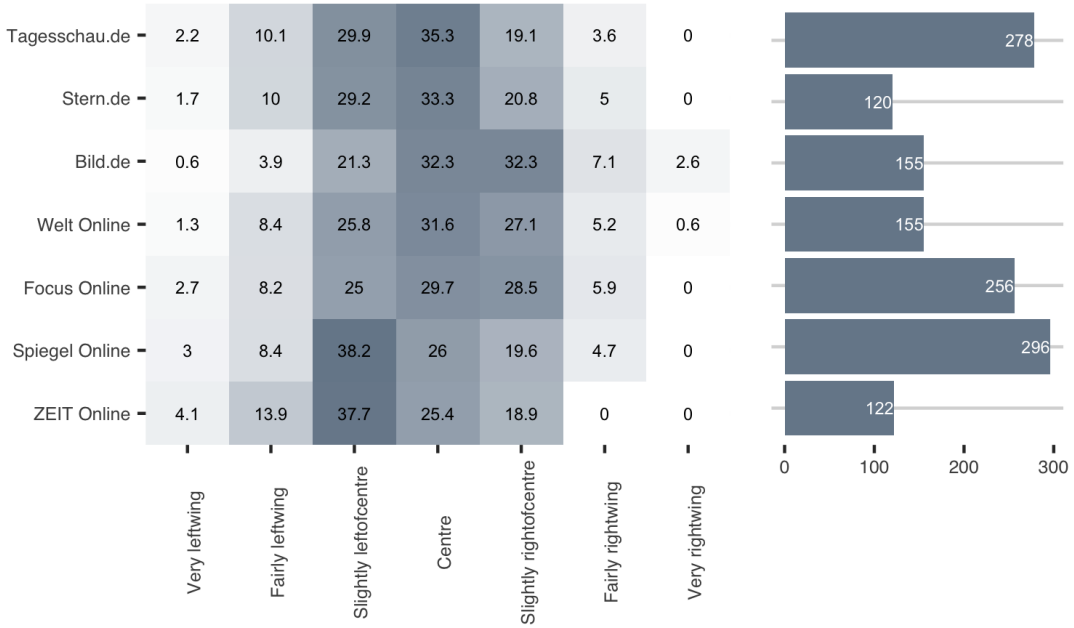


Combining both the political orientation and the brand preferences can give an indication about the political orientation of the consumers of a specific medium. Thus, for example, it is possible to answer the question of how many percent of the respondents, who

[8]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?

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 4 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.

Figure 4: Political orientation



Based on the finding from M. Gentzkow and Shapiro (2010) that readers prefer like-minded news and that firms respond strongly to consumer preferences, the hypothesis can be made that media reporting of ZEIT Online and Spiegel Online tends to be left biased, while Focus Online and Welt Online are expected to be more biased to the right. However, Anderson and Gabszewicz (2006) argue, that competition for viewers of the demographics most desired by advertisers implies that programming choices will be biased towards the tastes of those with such demographics. This effect would tend to lead news providers to report similarly as they compete for the same customers on the advertising market.

In contrast to the other providers, tagesschau.de is financed by compulsory fees and has the constitutional mandate to contribute to the formation of individual and public opinion and to secure basic provision. This implies a nationwide reception of broadcasting for the general public as well as the supply of a diversified content.[9] This leads to the assumption that the reporting of tagessschau.de is not aligned to the market conditions of the advertising and reader market, but rather reports balanced on all parties and possible topics.

[9]The mandate of public service broadcasting is based on Art. 5 Para. 1 Sentence 2 of the German Basic Law (Grundgesetz).

4 Data

To measure the slant index of a media outlet, the topics discussed in that media outlet are compared to topics discussed in the press releases of political parties.

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 (Brandenburg, 2005; 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.

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 Press releases

The press releases were scraped from the publicly accessible press portals of the parties[10] and parliamentary groups[11].

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 (ibid.), 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.

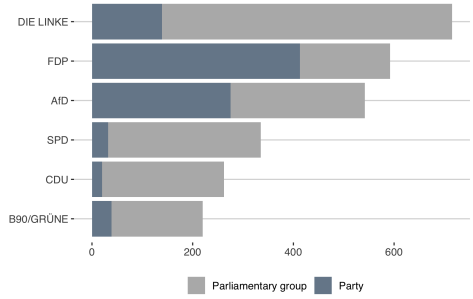
Articles with less than 50 words, as well as articles about party internal elections, were filtered out resulting in a record of 2666 press releases. Figure 5a displays the number of press releases for each party and Figure 5b shows the text length of those articles. Overall, DIE LINKE has published 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.

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

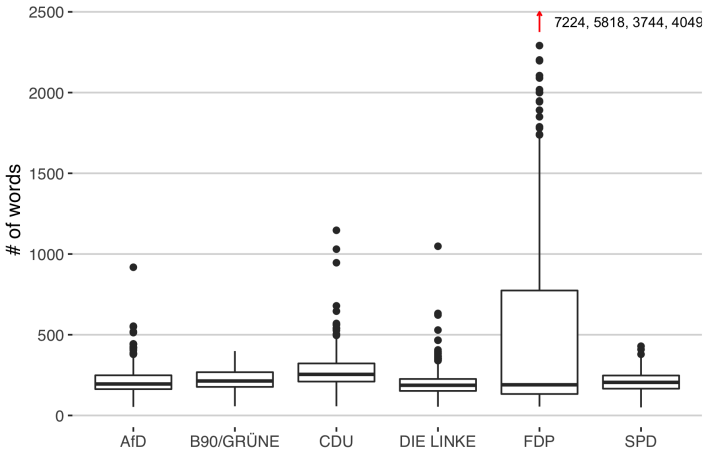
[11] afdbundestag.de, spdfraktion.de, die-linke.de/start/presse/aus-dem-bundestag, fdpbt.de, gruene-bundestag.de/, presseportal.de/nr/7846

Figure 5: Press releases

(a) Count



(b) Text length



4.2 News articles

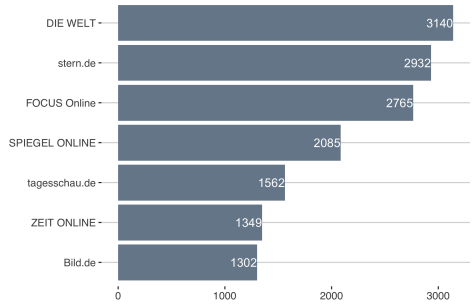
The data on news articles includes online articles from seven German news providers about domestic politics[12] dated from June 1, 2017 to March 1, 2018. First all online articles from the corresponding news websites are scraped using the Webhose.io API.[13] In order to limit the analysis to articles on domestic politics, the articles are filtered using the url of the article. For example, the URL *http://www.spiegel.de/politik/deutschland/christian-lindner-ukraine-kritisiert-fdp-chef-scharf-der-morgen-live-a-1161196.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 15,135. As shown in Figure 6a 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.

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

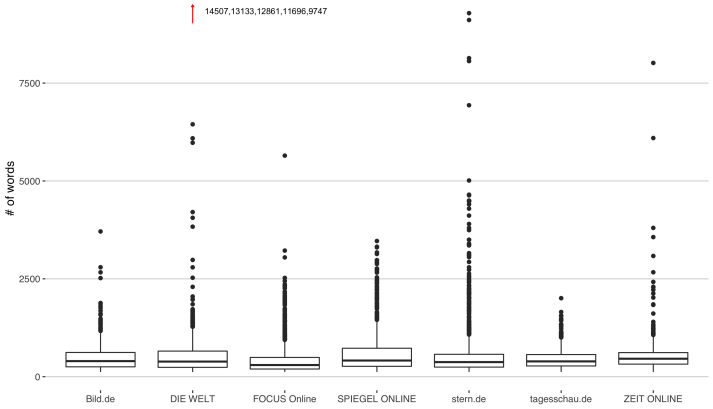
[13]For more information see https://docs.webhose.io/v1.0/docs/getting-started. The scraping code was written in Python and can be made available on request.

Figure 6: News articles

(a) Count



(b) Text length

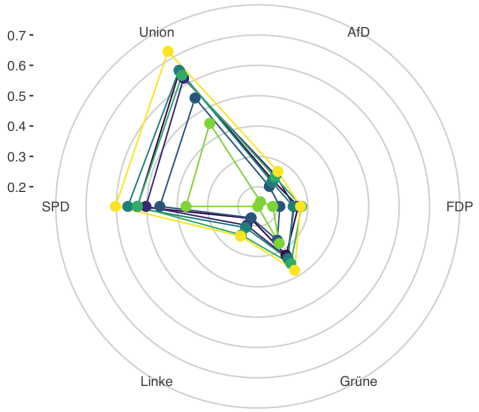


In order to get a deeper understanding of the extent to which the newspapers report about parties, the visibility of the individual parties in the news articles can be considered (D'Alessio and Allen, 2000; Eberl et al., 2017; Junqué de Fortuny et al., 2012). The visibility of a party in a newspaper is defined as the number of news articles published by the newspaper on that party (e.g. articles that mention the name of the party), normalised on the total amount of articles by that newspaper in the corpus. A party is treated as visible in an article if the party itself is mentioned in an article (if an article contains the word "SPD", "CDU"/"CSU/Union","FDP","Grüne","AfD" or "Linke"). Obviously, due to their political or social relevance, not all parties receive the same amount of attention in the media. Thus, Figure 7a shows that the major popular parties CDU and SPD are mentioned most frequently among all news providers, in contrast to the other parties. However, these external factors should influence all media outlets equally. To show the relative visibility of a party on a medium, the values are rescaled on an interval between 0 and 1 displayed in Figure 7b.

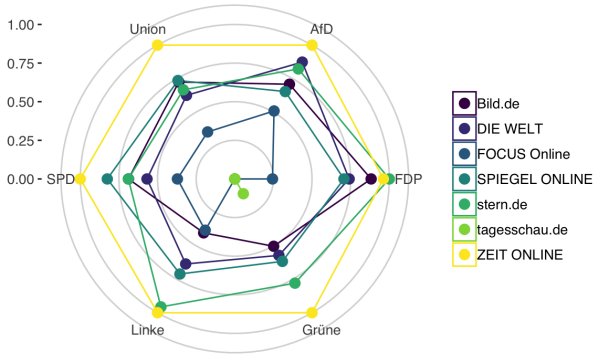
- Bei Zeit online haben alle Parteien am meisten Sichtbarkeit / bei tagesschau.de am wenigsten - Insgesamt scheint die Sichtbarkeit gleichmäßig verteilt zu sein, größere Unterschiede sind erkennbar bei Focus (wenig Grüne, mehr AfD) und Bild (weniger Linke, mehr FDP).

Figure 7: Visibility

(a) absolute values



(b) relative values



4.3 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, Kelly, 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 8a, 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, Kelly, 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[14] 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.[15] After completing these steps 63,360 unique terms were left in the vocabulary. The word clouds in Figure 8b 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.

[14]http://snowball.tartarus.org/algorithms/german/stop.txt
[15]https://tartarus.org/martin/PorterStemmer/

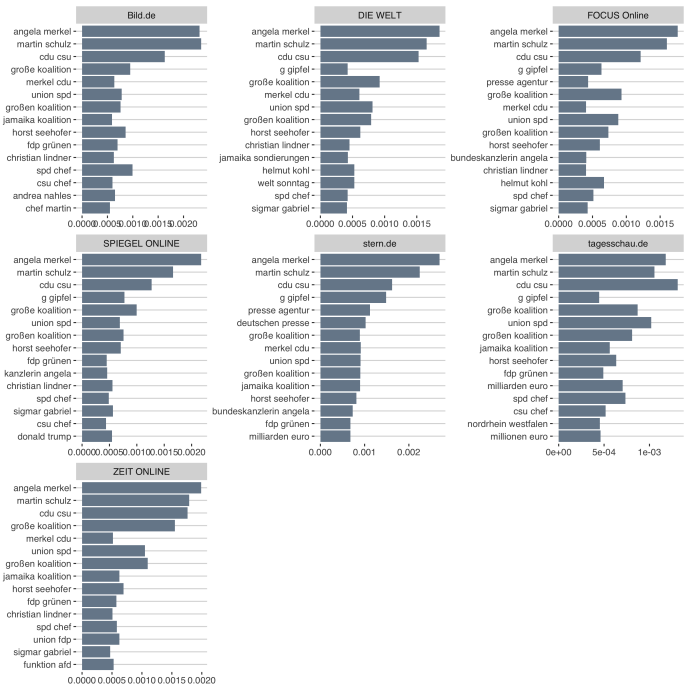
Figure 8: Wordclouds

(a) before pre-processing

(b) after pre-processing

Figure 9: Term frequency

(a) News articles



(b) Press releases

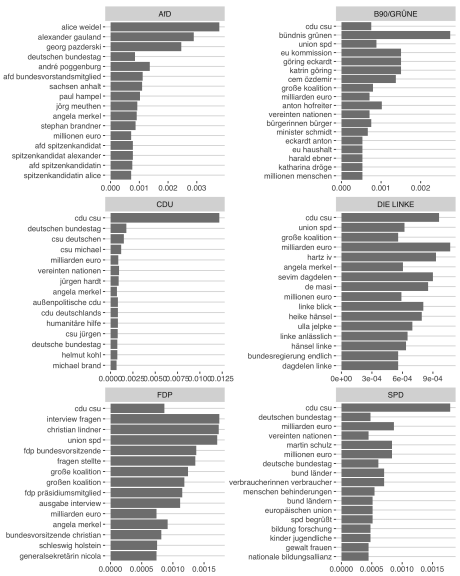


Figure 13b and 13a show the top 15 bigrams (pairs of two consecutive words) for each source in terms of frequency. While no obvious differences are visible for the news websites, the top words in the press releases make it easy to deduce the corresponding party.

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, Kelly, et al., 2017), since the order in which words are used within a document is disregarded.

5 Measuring slant-index

To discover the latent topics in the corpus of press releases (2.666) (see Section 4.1) and news articles (15.135) (see Section 4.2), 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[16] 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[17] 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 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. A more formale description of this process is described in Section 5.2 and 5.3. In the last step I compare the agenda selectivity of each medium with the user preferences expressed in the Reuters Institute Digital News Survey (see Section 6.4).

5.1 Sentiment score

To estimate the sentiment of each document a dictionary-based method is applied with the aim to measure the tone (or sentiment) of a document. The idea of a sentiment analysis is to determine the attitude of a writer toward the overall tonality of a document. 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.

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

[17] press release or news article

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 measure "the two Ts" (Topic and tone). They explore the effects of FOMC (Federal Open Market Committee) statements on both market and real economic variables. To understand the latent topic of a statement, they apply LDA on a corpus of 142 FOMC decision statements split into sentences. They then measure how the central bank is talking about that topic, using a dictionary approach. 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. They count the number of occurrences of excitement words and anxiety words and then scale these numbers by the total text size as measured by the number of characters.

The present paper uses a dictionary that lists words associated with positive and negative polarity weighted within the interval of $[-1; 1]$. SentimentWortschatz[18], is a publicly available German-language resource for sentiment analysis, opinion mining, etc.. The current version of SentiWS (v1.8b) contains 1,650 positive and 1,818 negative words, which sum up to 15,649 positive and 15,632 negative words including their inflections, respectively.

The sentiment score for each document d is calculated based on the weighted polarity values for a word, defined on an interval between -1 and 1. The score is then calculated from the sum of the words in a document (which can be assigned to a word from the dictionary) divided by the total number of words in that document:

$$\text{SentScore}_d = \frac{|\text{positive polarity score}_d| - |\text{negative polarity score}_d|}{|\text{TotalWords}_d|} \tag{1}$$

5.2 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.[19].

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

[18]SentiWS for short. available here: http://wortschatz.uni-leipzig.de/de/download

[19]See also Griffiths and Steyvers (2002), Griffiths and Steyvers (2004) and Hofmann (1999)

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:[20]

- 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 $\phi_{z_{d,n}}$.

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, Kelly, 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 50 topics best reflects the structure in the corpus. Furthermore, the source of a document[21] 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 (news website or party) 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.

5.3 Weighted topic correlation

The process described in 5.2 generates two different kinds of posterior distributions that can be represented as the following matrices:

1. Φ_c is a K-by-V matrix (where K = number of topics and V = vocabulary or unique terms), where the entry $\phi_{k,v,c}$ can be interpreted as the probability of observing the v-th word in topic k for the covariate level c (type of source: press release or news website).

[20] A more detailed description of the generative process of the STM can be found in section A.2

[21] Each party as well as each media outlet represent one source.

2. Θ is a D-by-K matrix (where D = number of documents and K = number of topics), where the entry $\theta_{d,k}$ 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.

To combine the sentiment value with the topic probability of each document, the sentiment score is multiplied with the kx1 vector for each document. Subsequently, to calculate the average topic sentiment for a source s, the mean value of all documents belonging to source s is calculated. This results in a kx1 vector representing the mean distributions $\bar{\theta_s}$ of this source, weighted by the sentiment scores:

$$\bar{\theta_s} = \begin{bmatrix} \bar{\theta_1} \\ . \\ . \\ \bar{\theta_k} \end{bmatrix} \tag{2}$$

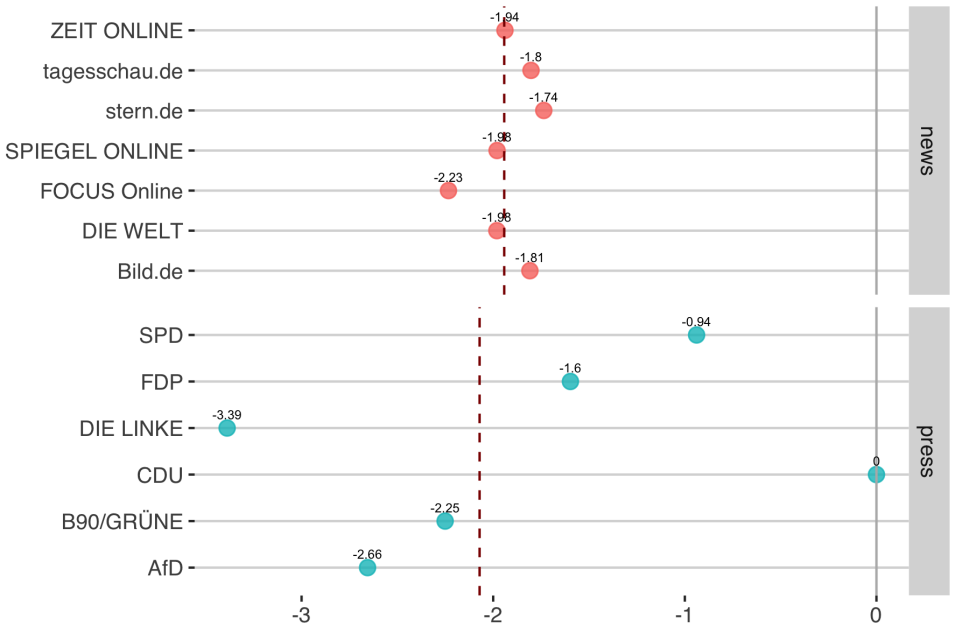
On the basis of these values, the bivariate Pearson correlation coefficients are calculated for each pair of media outlet and party on a monthly basis. The higher the correlation coefficient, the higher the slant index for a party in a media outlet.

6 Results

6.1 Sentiment

- Figure 10 shows the aggregated sentiment value for each source. Dashed line shows the median value. - overall: all values are negative or 0 (CDU). - parties: higher sentiment variance. DIE LINKE is most negative, CDU most positive - media: FOCUS Online most negative, stern.de most positive.

Figure 10: Sentiment score



6.2 Topics

As stated above, the generative process of the STM results in a matrix Φ_c representing the probability of words in each topic. The most frequent words in each topic help to understand what each topic is about. Since the type of source is included as a covariate in the model specification, the result consists of two matrices (one for each covariate level c). It is therefore important to check the different word-distributions for each topic. The following Figure shows the 20 most frequent words for the first 3 topics separated between the covariate levels "news article" and "press release".[22].

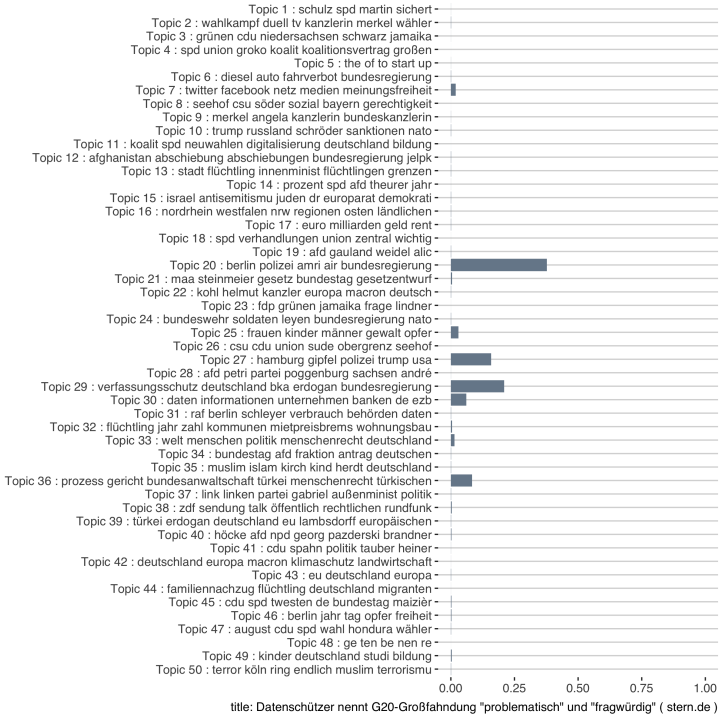
Figure 11: Most frequent terms

(a) Topic 1 (b) Topic 2 (c) Topic 3

The second posterior distribution from the STM results in a topic distribution θ_d for each document d over all topics k. An example of this distribution is shown in Figure 12.

[22]See A.3 for a table of the most frequent words for each topic.

Figure 12: Topic frequency



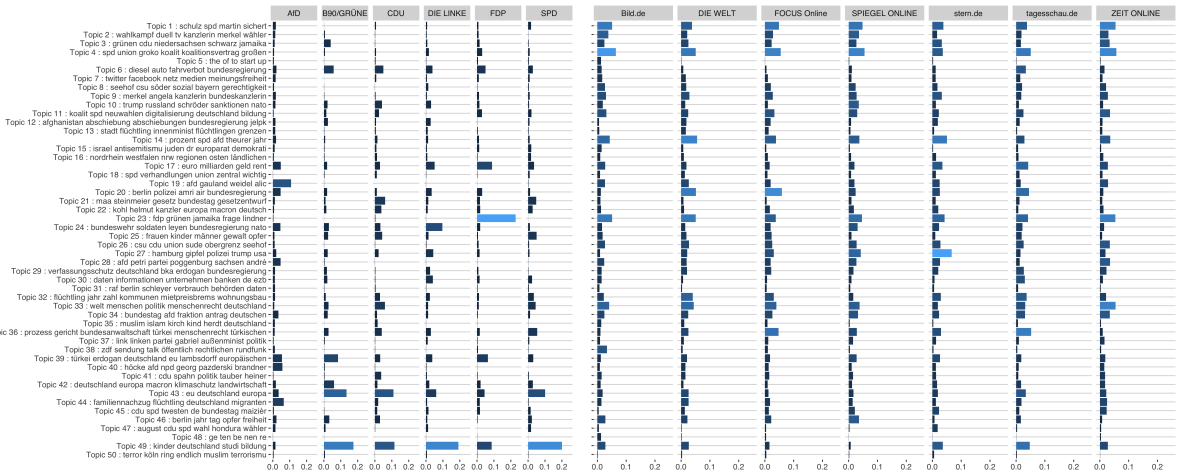
The average of each topic across all documents results in the expected probability of a topic across the whole corpus. As shown in Figure **??**,[23] topic 23 - which apparently deals with the issue of the Jamaica Coalition - is the most common topic. Topic 4, which deals with the coalition negotiations of the grand coalition ("GroKo") follows closely behind.

To calculate the mean topic probability for each source, the documents the average value of all documents belonging to that source is calculated. For each source the average distribution of each topic is calculated.

Figure 13: Topic probablity

(a) Press releases

(b) News articles



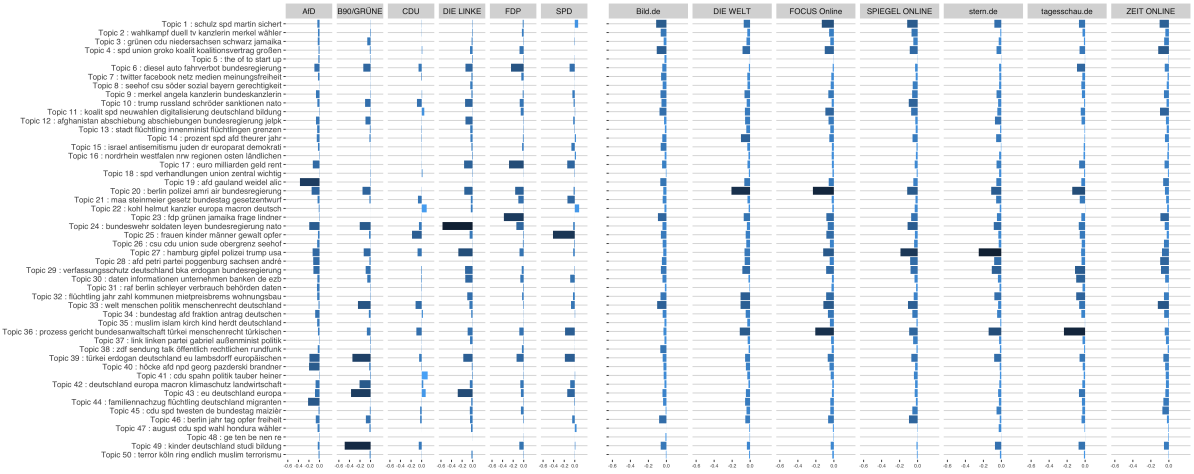
[23]Results a shown in Appendix A.4

6.3 Weighted topic correlation

First, the topic probability vector of each document is weighted by its sentiment value

Figure 14: Weighted topic probability

(a) Press releases (b) News articles



Subsequently, the bivariate correlations are estimated.

Figure 15: Monthly topic correlation

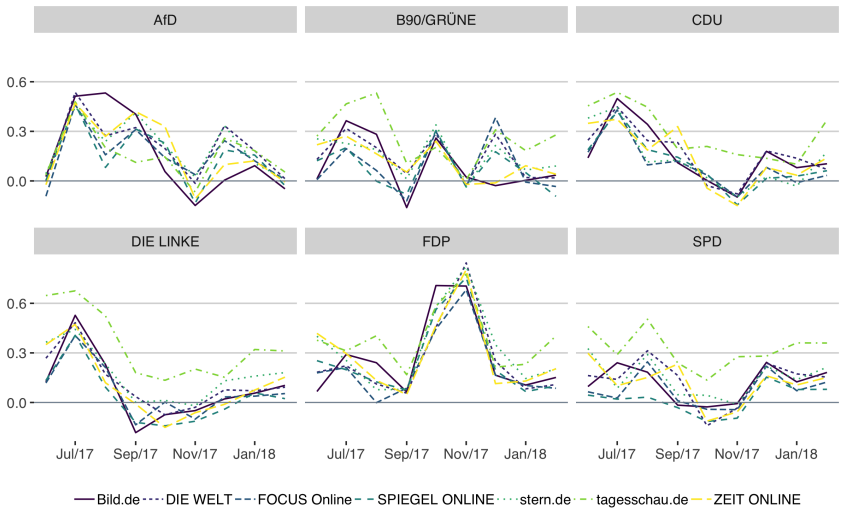
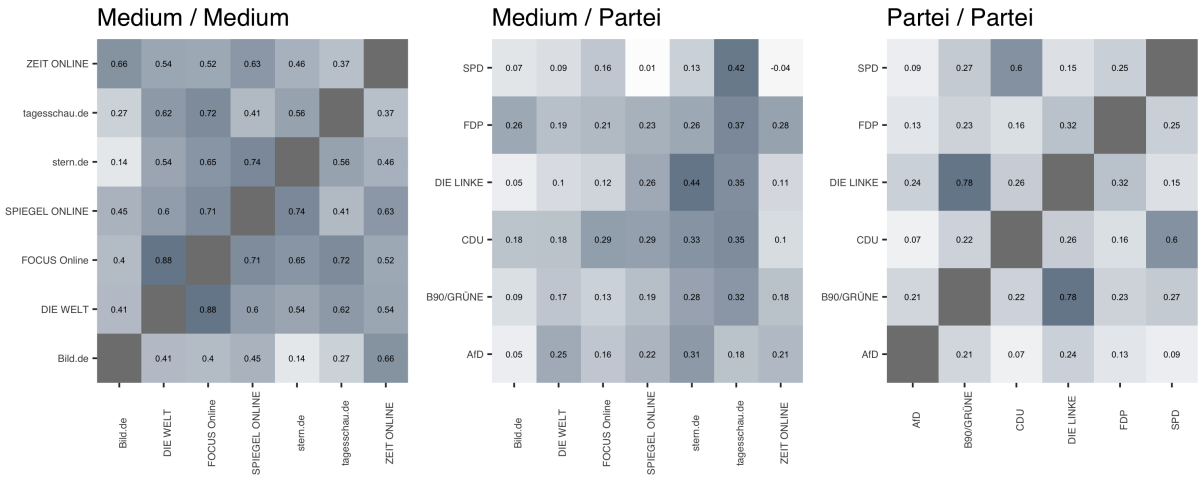
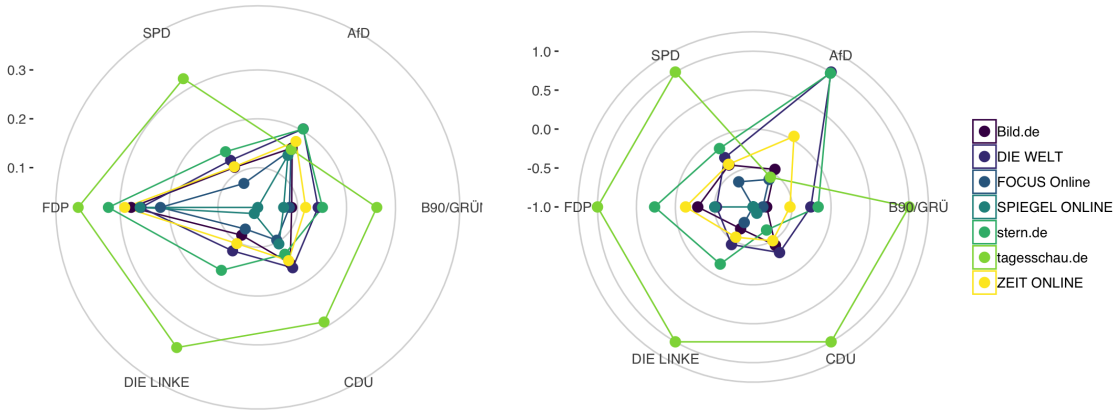


Figure 16: Topic correlation



Obviously, due to their political or social relevance, not all potential topics receive the same amount of attention in the media. However, these factors should influence all media outlets equally. To calculate the relative topic correlation, I estimate the deviation of the topic correlation of a party in one medium from the average topic correlation of that party over all news paper.

Figure 17: Topic correlation - Radarcharts



(a) absolute values

(b) relative values

6.4 Reader Preferences

7 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

Anderson, Simon P. and Jean J. Gabszewicz (Jan. 1, 2006). "The Media and Advertising: A Tale of Two-Sided Markets". In: *Handbook of the Economics of Art and Culture*. Ed. by Victor A. Ginsburg and David Throsby. Vol. 1. Elsevier, pp. 567–614. URL: http://www.sciencedirect.com/science/article/pii/S1574067606010180 (visited on 04/19/2019).

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.

Besley, Timothy and Andrea Prat (June 2006). "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability". In: *American Economic Review* 96.3, pp. 720–736. URL: https://www.aeaweb.org/articles?id=10.1257/aer.96.3.720 (visited on 03/30/2019).

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.

Brandenburg, Heinz (Sept. 1, 2005). "Political Bias in the Irish Media: A Quantitative Study of Campaign Coverage during the 2002 General Election". In: *Irish Political Studies* 20.3, pp. 297–322. URL: https://doi.org/10.1080/07907180500359350 (visited on 10/20/2018).

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.

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).

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.

Gentzkow, Matthew and Jesse M. Shapiro (Apr. 1, 2006). "Media Bias and Reputation". In: *Journal of Political Economy* 114.2, pp. 280–316. URL: https://www.journals.uchicago.edu/doi/10.1086/499414 (visited on 03/30/2019).

— (2010). "What Drives Media Slant? Evidence From U.S. Daily Newspapers". In: *Econometrica* 78.1, pp. 35–71. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7195 (visited on 01/07/2019).

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.

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).

— (Mar. 1, 2014). "Evaluating and understanding text-based stock price prediction models". In: *Information Processing & Management* 50.2, pp. 426–441. URL: http://www.sciencedirect.com/science/article/pii/S0306457313001143 (visited on 10/11/2018).

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).

Mullainathan, Sendhil and Andrei Shleifer (Sept. 2005). "The Market for News". In: *American Economic Review* 95.4, pp. 1031–1053. URL: https://www.aeaweb.org/articles?id=10.1257/0002828054825619 (visited on 01/19/2019).

Newman, Nic et al. (2018). *Reuters Institute Digital News Report 2018*. Reuters Institute for the Study of Journalism. URL: http://media.digitalnewsreport.org/wp-content/uploads/2018/06/digital-news-report-2018.pdf?x89475.

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).

Oegema, Dirk and Jan Kleinnijenhuis (2009). "Personalization in political Television News: A 13-Wave Survey Study to Assess Effects of Text and Footage". In: *Communications* 25.1, pp. 43–60. URL: https://www.degruyter.com/view/j/comm.2000.25.issue-1/comm.2000.25.1.43/comm.2000.25.1.43.xml (visited on 04/19/2019).

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).

Suen, Wing (2004). "The Self-Perpetuation of Biased Beliefs". In: *Economic Journal* 114.495, pp. 377–396. URL: https://ideas.repec.org/a/ecj/econjl/v114y2004i495p377-396.html (visited on 01/19/2019).

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).

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 Visibility

medium	AfD	FDP	Grüne	Linke	SPD	Union
Bild.de	0.23	0.27	0.32	0.18	0.53	0.65
DIE WELT	0.25	0.25	0.33	0.21	0.50	0.62
FOCUS Online	0.21	0.21	0.26	0.18	0.46	0.55
SPIEGEL ONLINE	0.23	0.25	0.33	0.22	0.56	0.65
stern.de	0.25	0.28	0.35	0.24	0.53	0.63
tagesschau.de	0.15	0.18	0.28	0.14	0.37	0.45
ZEIT ONLINE	0.27	0.28	0.38	0.25	0.60	0.73

Table 1: Visibility

A.2 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:

$$z_{dn}|\theta_d \sim \text{Multinomial}(\theta_d). \tag{3}$$

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

$$\theta_d|y_{d\gamma}, \Sigma \sim \text{LogisticNormal}(\mu = y_{d\gamma}\Sigma), \tag{4}$$

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:

$$w_{dn}|z_{dn}, \phi_{dkv} \sim \text{Multinomial}(\phi_{dk1}, ..., \phi_{dkV}), \tag{5}$$

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 (ibid.). 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:

$$\phi_{dkv}|z_{dn} = \frac{\exp(m_v + \kappa^{(t)}_{kv}, \kappa^{(c)}_{y_dv} + \kappa^{(i)}_{y_dkv})}{\sum_v \exp(m_v + \kappa^{(t)}_{kv}, \kappa^{(c)}_{y_dv} + \kappa^{(i)}_{y_dkv})}. \tag{6}$$

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[24] available in R's stm package (M. Roberts, B. Stewart, and Tingley, 2016b).

A.3 Most frequent words

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

Topic	News articles	Press releases
1	schulz spd martin gabriel partei	schulz martin sichert w
2	wahlkampf duell tv lang spricht	kanzlerin merkel wähle
3	grünen cdu niedersachsen grüne fdp	grünen schwarz jamaik
4	spd union groko koalit koalitionsverhandlungen	koalit koalitionsvertrag
5	the of to on it	start up the german it
6	diesel auto fahrverbot autoindustri deutschen	diesel bundesregierung
7	twitter facebook netz tweet medien	medien facebook meinu
8	seehof csu söder bayern horst	sozial bayern gerechtigk
9	merkel angela kanzlerin cdu bundeskanzlerin	merkel angela bundeska
10	trump russland schröder usa putin	russland sanktionen na
11	koalit spd neuwahlen jamaika große	digitalisierung deutschl
12	afghanistan abschiebung abschiebungen abgeschoben deutschland	afghanistan bundesregi
13	stadt flüchtling innenminist herrmann flüchtlingen	flüchtling flüchtlingen g
14	prozent spd afd bundestagswahl umfrag	prozent theurer jahr za
15	israel antisemitismu juden deutschland israelisch	dr europarat demokrati
16	nordrhein westfalen nrw sachsen schwesig	regionen osten ländlich
17	euro milliarden geld millionen jahr	euro milliarden rent ge
18	spd verhandlungen union einigung cdu	zentral wichtig verhand
19	afd gauland weidel partei alexand	weidel alic gauland alex
20	berlin polizei amri anschlag polizisten	berlin air bundesregieru
21	maa steinmeier gesetz heiko bundespräsid	gesetz bundestag gesetz
22	kohl helmut kanzler einheit altkanzl	europa macron deutsch
23	fdp grünen jamaika lindner grüne	frage lindner grünen ku
24	bundeswehr soldaten leyen nato einsatz	bundeswehr bundesregi
25	frauen kinder männer jahr sexuel	frauen gewalt opfer mä
26	csu cdu union seehof obergrenz	sude obergrenz seehof f
27	hamburg gipfel polizei demonstranten scholz	trump usa gipfel hamb
28	afd petri partei frauk fraktion	poggenburg sachsen an
29	verfassungsschutz deutschland bka maizièr sicherheitsbehörden	erdogan deutschland bu
30	daten informationen unternehmen mitarbeit bundesregierung	banken de ezb konzern
31	raf berlin schleyer terroristen entführung	verbrauch behörden da
32	flüchtling jahr zahl bamf deutschland	kommunen mietpreisbr
33	welt menschen politik deutschland land	menschen menschenrec
34	bundestag afd fraktion abgeordneten parlament	bundestag antrag deuts
35	muslim islam kirch feiertag muslimischen	kind herdt deutschland
36	prozess gericht bundesanwaltschaft nsu staatsanwaltschaft	türkei menschenrecht t
37	link linken partei wagenknecht spd	gabriel außenminist pol
38	zdf sendung talk bosbach politik	öffentlich rechtlichen ru
39	türkei erdogan deutschland türkischen türkisch	eu lambsdorff europäisc
40	höcke afd npd jahr neonazi	georg pazderski brandn
41	cdu spahn politik altmaier günther	tauber heiner politik ge
42	deutschland europa macron frankreich klimaschutz	deutschland klimaschut
43	eu deutschland europa menschen jahr	eu deutschland europa
44	familiennachzug flüchtling deutschland flüchtlingen union	deutschland familienna
45	cdu spd twesten niedersachsen abstimmung	de bundestag maizièr b
46	berlin jahr tag jahren ditib	opfer tag freiheit mens
47	august cdu spd prozent bundestagswahl	wahl hondura wähler m
48	ge ten be ver li	ge nen re ten ter
49	kinder deutschland studi prozent schulen	kinder bildung deutsch
50	terror köln ring rock festiv	endlich muslim terroris

Table 2: Most frequent words

A.4 Topic frequency

Topic	_overall	AfD	B90/GRÜNE	Bild.de	CDU	DIE LINKE	DIE WELT	FDP	FO
1	0.039	0.018	0.001	0.051	0.007	0.006	0.038	0.008	
2	0.023	0.015	0.006	0.038	0.001	0.004	0.021	0.009	
3	0.024	0.011	0.040	0.024	0.002	0.008	0.025	0.015	
4	0.046	0.008	0.006	0.064	0.007	0.018	0.050	0.030	
5	0.003	0.006	0.002	0.012	0.005	0.003	0.001	0.006	
6	0.017	0.021	0.058	0.015	0.050	0.039	0.007	0.050	
7	0.012	0.017	0.000	0.017	0.007	0.006	0.016	0.010	
8	0.018	0.008	0.002	0.026	0.000	0.014	0.018	0.002	
9	0.025	0.019	0.002	0.030	0.004	0.006	0.028	0.011	
10	0.017	0.015	0.021	0.019	0.042	0.031	0.014	0.014	
11	0.026	0.005	0.014	0.031	0.024	0.001	0.025	0.029	
12	0.015	0.015	0.023	0.006	0.006	0.028	0.018	0.001	
13	0.009	0.014	0.003	0.006	0.006	0.010	0.016	0.005	
14	0.039	0.019	0.007	0.043	0.015	0.013	0.055	0.012	
15	0.007	0.011	0.004	0.015	0.009	0.006	0.007	0.013	
16	0.010	0.006	0.004	0.010	0.012	0.008	0.010	0.003	
17	0.030	0.047	0.020	0.027	0.030	0.051	0.018	0.089	
18	0.010	0.001	0.004	0.009	0.012	0.002	0.013	0.002	
19	0.023	0.108	0.002	0.027	0.000	0.001	0.026	0.000	
20	0.035	0.046	0.019	0.012	0.011	0.034	0.051	0.030	
21	0.011	0.012	0.010	0.010	0.061	0.014	0.008	0.017	
22	0.015	0.007	0.014	0.013	0.038	0.010	0.010	0.019	
23	0.046	0.004	0.000	0.051	0.003	0.001	0.051	0.228	
24	0.024	0.045	0.030	0.016	0.033	0.097	0.015	0.019	
25	0.016	0.010	0.026	0.018	0.043	0.015	0.018	0.006	
26	0.024	0.011	0.002	0.026	0.001	0.004	0.027	0.010	
27	0.032	0.020	0.023	0.009	0.022	0.044	0.021	0.015	
28	0.021	0.047	0.005	0.024	0.000	0.001	0.019	0.001	
29	0.020	0.013	0.019	0.015	0.003	0.025	0.021	0.008	
30	0.013	0.012	0.020	0.014	0.003	0.041	0.007	0.015	
31	0.005	0.013	0.003	0.007	0.004	0.002	0.003	0.003	
32	0.028	0.014	0.011	0.022	0.030	0.023	0.040	0.012	
33	0.035	0.009	0.027	0.041	0.059	0.012	0.044	0.010	
34	0.025	0.033	0.023	0.025	0.013	0.013	0.024	0.005	
35	0.007	0.013	0.000	0.014	0.016	0.001	0.011	0.001	
36	0.030	0.011	0.027	0.010	0.041	0.029	0.026	0.017	
37	0.011	0.005	0.005	0.009	0.002	0.010	0.010	0.004	
38	0.008	0.012	0.002	0.033	0.002	0.001	0.006	0.006	
39	0.025	0.054	0.084	0.012	0.031	0.037	0.021	0.064	
40	0.013	0.057	0.006	0.008	0.002	0.003	0.015	0.001	
41	0.013	0.004	0.004	0.015	0.037	0.005	0.015	0.003	
42	0.014	0.018	0.060	0.012	0.014	0.020	0.008	0.021	
43	0.028	0.034	0.134	0.018	0.110	0.061	0.026	0.044	
44	0.019	0.065	0.001	0.014	0.019	0.007	0.026	0.016	
45	0.016	0.007	0.002	0.005	0.013	0.015	0.013	0.018	
46	0.019	0.021	0.032	0.028	0.030	0.007	0.022	0.003	
47	0.007	0.016	0.010	0.005	0.002	0.013	0.002	0.000	
48	0.003	0.002	0.001	0.013	0.000	0.000	0.002	0.002	
49	0.041	0.016	0.175	0.028	0.115	0.192	0.026	0.087	
50	0.004	0.005	0.006	0.002	0.001	0.006	0.003	0.002	

Table 3: Topic frequency

A.5 Topic correlation

AfD	B90/GRÜNE	Bild.de	CDU	DIE LINKE	DIE WELT	FDP	FOCUS Online
0.12							
-0.02	-0.02						
0.01	0.79****	0.00					
0.11	0.81****	-0.02	0.68****				
0.16	0.07	0.70****	0.11	0.08			
0.01	0.31*	0.36*	0.25	0.31*	0.37**		
0.07	-0.05	0.63****	-0.01	0.01	0.88****	0.20	
-0.03	0.86****	0.05	0.86****	0.81****	0.11	0.29*	0.01
0.09	-0.04	0.72****	-0.01	-0.01	0.73****	0.28*	0.78****
0.15	0.18	0.45***	0.10	0.29*	0.67****	0.38**	0.66****
0.11	0.38**	0.55****	0.37**	0.45**	0.72****	0.49***	0.72****
0.11	0.05	0.81****	0.04	0.02	0.75****	0.39**	0.67****