Topic and tone of political news articles in German online media.

Franziska Löw *

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

November 19, 2018

Abstract

However, the concept of media bias actually encompasses different subtypes: Visibility bias is the salience of political actors, tonality bias the evaluation of these actors, and agenda bias the extent to which parties address preferred issues in media coverage. Most literature on media bias focuses on only one type of bias, mainly disregarding agenda bias, as the operationalization is somewhat challenging. Additionally, studies that analyse the effect of media content primarily employ manual content analysis and are therefore much more time-consuming and susceptible to errors. Using automated data mining tools this paper provides an approach that allows an efficient and objective analysis to measure the different types of media bias. In order to investigate whether the media landscape transmits a biased reality of the political landscape, online news articles are analysed together with party press releases.

*Electronic address: loewf@hsu-hh.de

Contents

1 Introduction

The importance of the internet as a source of information for political topics has grown strongly in recent years. 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). The expansion of the internet as a new method of communication provides a potential challenge to the primacy of the traditional media and political parties as formers of public opinion (Savigny, 2002).

However, the discussion about the influence of media on the political opinion-forming process has not only been studied in the literature since the growing importance of the internet. The general hypothesis is that a biased media reporting in political news may have a profound influence on voter opinions and preferences (Eberl et al., 2017; Ferree et al., 2002; McCombs, 2005). It can therefore be argued, that one central responsibility of the media is to supply voters with balanced and objective information on relevant political issues and actors (Eberl et al., 2017; Strömbäck, 2008). The concept of media bias actually encompasses different subtypes (D'Alessio and Allen, 2000; Eberl et al., 2017; Junqué de Fortuny et al., 2012): (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). Most of the literature on media bias focuses on one type of bias and 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 the media, one would have to know the universe of news stories at a given point in time (D'Alessio and Allen, 2000). I adopt the approach used in Eberl et al. (2017), in using parties' campaign communication as an approximation of the potential universe of news stories. To identify the underlying topics in the text corpus a structural topic model is applied. To measure coverage and tonality bias computational text-mining approaches are applied.

The data analyzed in this study 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).

While the coverage and tonality bias can be calculated using simple counting methods, more sophisticated approaches are required to operationalize the agenda bias, since the topics covered have to be identified. To discover the latent topics in the corpus of text data, the structural topic modeling (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?). In the next step the topics addressed in campaign communication (i.e., the party agenda) are compared with the topics the parties address in media coverage (i.e., the mediated party agenda).

Building on the insight that media bias can have three different forms, this study

makes two key contributions. (1) I use a combination of different text-mining techniques to analyze a large text corpus 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. (2) I present a new approach to measure agenda bias by combining approaches from political and communication sciences with current text analysis techniques.

The remaining course of the paper is as follows: The following section provides an overview of the related literature. Section 3 gives a introduction to the political trends in the past six month (June 2017 to March 2018). The data used to conduct the model is described in section 4. 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

The general hypothesis of the literature on the influence of media bias on on the political opinion-forming process is that a biased media reporting in political news may have a profound influence on voter opinions and preferences (Eberl et al., 2017; Ferree et al., 2002; McCombs, 2005). In the literature on the influence of media bias on the political opinion-forming process usually encompasses different subtypes of media bias (D'Alessio and Allen, 2000; Eberl et al., 2017): (1) Visibility bias, (2) tonality bias and (3) agenda bias. These three concepts measure how often political actors appear in the media (visibility 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). However, these benchmarks do not allow for a comparison between media outlets. 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. Other empirical studies that measure the influence of media visibility on the electoral behaviour of voters are usually based on regional differences in the reach of certain media (DellaVigna and Kaplan, 2006; Enikolopov et al., 2011; Snyder and Strömberg, 2010). 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. However, in Eberl et al. (2017) the coverage bias effect can not be confirmed, as the effect of visibility bias is positive but not statistically significant.

In addition to visibility, the tonality of the reporting is also important, as it provides consumers with a certain understanding of politics. For example, 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, the concept of valence framing suggests that positive or negative aspects of an object are highlighted in the media, which affects the attitudes towards a certain subject match the context value of the information received the validity of these aspects in the consciousness of the public (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.[1] 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 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.

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

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

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.

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 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 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

[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 Gentzkow et al. (2017) give an overview of text mining applications in economics.

and simultaneously allow the analysis to be reproduced.

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.

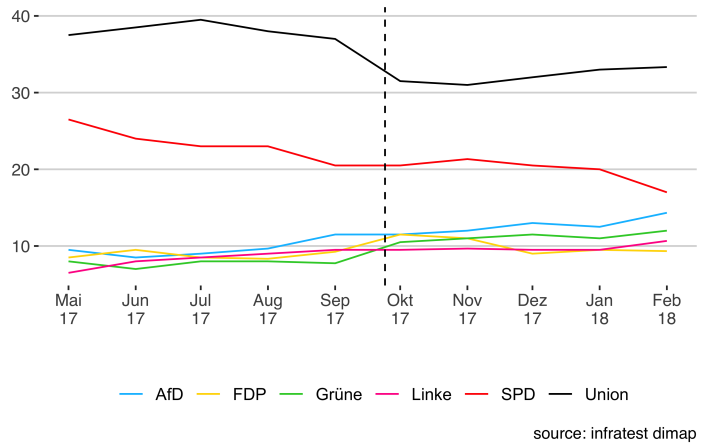
On the right side of the political spectrum, AfD (alternative for Germany) managed to be elected to the German Bundestag for the first time in 2017. The political debate about the high refugee numbers of the past years brought a political upswing to the AfD, which used the dissatisfaction of parts of 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. 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.

Figure 1: Election Polls



4 Data

medium	total_articles	share	visits
Bild.de	3828	0.28	156867922
DIE WELT	18663	0.14	52962285
FOCUS Online	9235	0.22	68730873
SPIEGEL ONLINE	5590	0.31	96139323
stern.de	15804	0.15	18192672
tagesschau.de	2261	0.41	75400000
ZEIT ONLINE	5467	0.23	30600773

Table 1: News sources used for the analysis

I conduct the estimation on a sample of 11,880 online news articles from seven German news providers about domestic politics[4]. The articles are dated from June 1, 2017 to March 1, 2018. I first extract all online articles using the Webhose.io API.[5] Overall, the selected news providers are among the top ten German online news providers - in terms of site visits[6] - in the period under review, with only Tagesschau.de belonging to the public media. The reason for this is that the content structure of Tagesschau.de is most similar to that of the private providers. Other public media offers provide their content in video (ZDF.de) or audio (Deutschlandfunk (DLF))) format, which make them difficult to compare. In order to limit the analysis to articles on domestic politics, all articles which mention at least one of the major parties[7] have been filtered out.

Figure 2a shows the distribution of the number of articles by date. There is a high peak around the federal elections on September, 24th and another one shortly after the failure of the Jamaica coalition talks on November, 19th (indicated by the red dotted

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

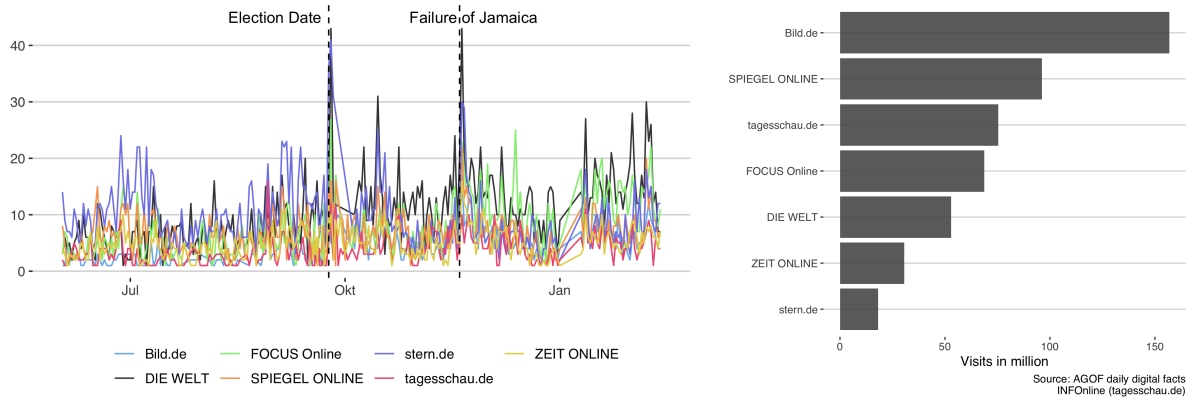
[5]For more information see https://docs.webhose.io/v1.0/docs/getting-started. The scraping code was written in Python and can be made available on request.

[6]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.

[7]The exact search terms were "CDU, CSU, Union", "SPD", "FDP", "Grüne", "Linke", "AfD".

lines). Figure 2b shows that DIE WELT published the most articles on domestic policy, followed by stern.de and FOCUS ONLINE.

Figure 2: Article distribution...



(a) ...by date

(b) ... by news provider

5 Estimate bias

1. Estimate different bias types 2. Estimate a fixed effect regression, to analyze the impact of the different biases that vary over time.

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. In this section the measurement methodology for each of the three types of bias defined in the literature (visibility bias, tonality bias and agenda bias Eberl et al. (2017) and Junqué de Fortuny et al. (2012)) is described. The strategy for calculating the bias follows the same pattern in all three cases: First, the value for which the bias will be determined is calculated for each party (i.e. visibility, tonality and agenda setting). This is done at a monthly level of the different news providers. The average value of all other parties in the month and the medium serves as the reference value to calculate the bias. To ensure comparability between the different bias metrics, they are standardized to range from 1 to 1, where a party would have a bias of 0 (neutral), when its visibility, tonality or agenda is equal to the mean visibility, tonality or agenda across all parties (in that media outlet) (Eberl et al., 2017).

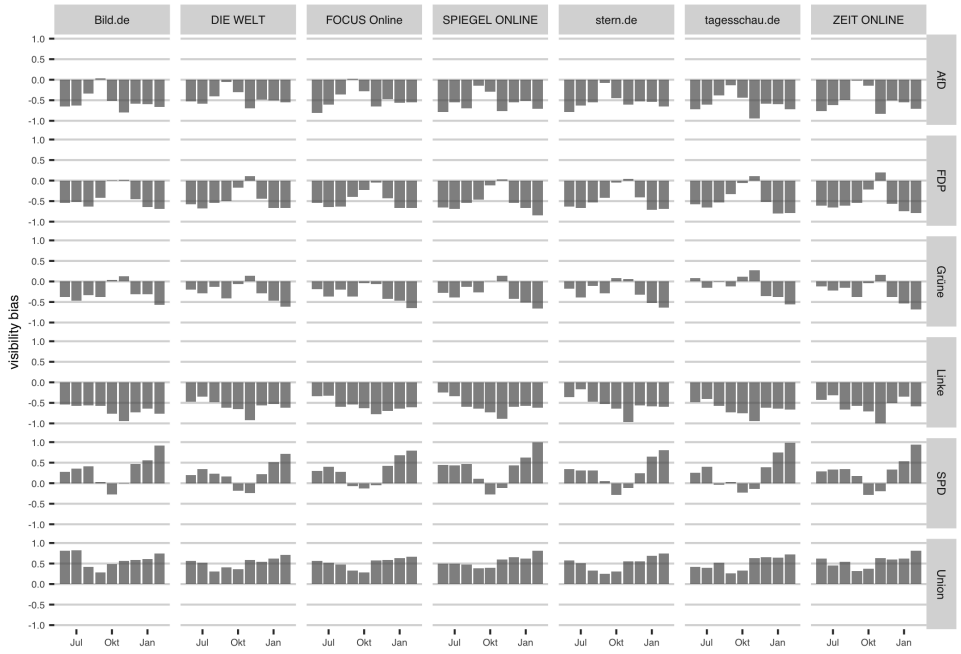
5.1 Visibility Bias

The overall visibility for each party in a news provider is defined as the number of news articles in which a party is named at least once, normalised on the total amount of articles by that news provider in the corpus. Next, to distinguish between balanced and biased reporting the average visibility of all other parties in a news outlet is used as a key benchmark.[8] The results displayed in Figure 3 do not show clear evidence for a difference between the different news provider: Overall, they are positively biased towards "Union"

[8]In another setting Junqué de Fortuny et al. (2012) uses the popularity in terms of votes as the a priori fair distribution. The visibility bias of a media source is the difference between the real distribution and the fair distribution.

and "SPD" and more negatively biased towards the other parties, with "DIE LINKE" having the most negative bias overall. The SPD has the greatest differences over time: The party has a negative bias during the Jamaica negotiations (Sep-Nov), turning positive following the failure of the Jamaica negotiations and the start of the coalition talks on the "GroKo". The opposite is true for "FDP" and "Grüne": Both show a rather positive or neutral bias during these month and a negative bias for the rest of the rime. Visibility of "AfD" is biased negatively during the time of analysis, showing less negativity during the election month (Sep).

Figure 3: Visibility Bias



5.2 Sentiment Analysis

To measure the tone (or sentiment) of an article a dictionary-based method is applied. To conduct such an analysis, a list of words (dictionary) associated with a given emotion, such as negativity is pre-defined. The document is then deconstructed into individual words and each word is assigned a sentiment value according to the dictionary, where the sum of all values results in the emotional score for the given document. 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 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[9], 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. Table 2 shows ten examples entries of the dictionary. To obtain a more reliable correlation between the "target" (a political party) and the word's polarity score, sentiment words are counted in a window of two sentences before and after the mention of a political party (Junqué de Fortuny et al., 2012). The tonality score of an article is then calculated from the sum of the these words divided by the total number of words in that article. Again, the tonality bias for is then computed as the deviation of each party's specific tonality from the average tonality of all other parties in that outlet and standardized to range from 1 to 1.

	word	value
1	brisant	-0.00
2	makel	-0.18
3	vergöttern	0.00
4	gediegen	0.08
5	mühe	-0.00
6	untreue	-0.33
7	unhöflich	-0.00
8	erschweren	-0.49
9	addieren	0.00
10	fehlleistung	-0.00

Table 2: Sentiment Dictionary (Sample)

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

Figure 4: Tonality Bias

5.3 Agenda Bias

To allow for an operationalization of agenda bias, parties' press releases are used as an approximation of the potential universe of news stories (Eberl et al., 2017). The press releases were scraped from the website of each party[10] to compare the topics addressed with the policy issues in the news articles.

5.3.1 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). Intuitively the term frequency (tf) of a word is a measure of how important that word may be for the understanding of the text. As can be seen in Figure **??**, 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 (Gentzkow et al., 2017), are important to the grammatical structure of a

[10]https://www.afd.de/presse, https://www.spdfraktion.de, https://www.die-linke.de/start/presse/aus-dem-bundestag, https://www.fdp.de, https://www.gruene-bundestag.de/, https://www.presseportal.de/nr/7846

text, but typically don't add any additional meaning and can therefore be neglected. A common strategy to reduce the number of language elements is to pre-process the text by imposing some preliminary restrictions (e.g. stop-word removal and stemming) based on the nature of the data (e.g. twitter text, newspaper articles, speeches, etc.) (Gentzkow et al., 2017).

(a) Wordcloud (before pre-processing)

To remove distorting words, the pre-defined stop word list from the Snowball project[11] is used together with a customized 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.[12] After completing these steps 68,576 unique terms were left in the vocabulary. The word clouds in Figure **??** represent the most frequent words of the pre-processed articles for Bild.de and Tagesschau.de.[13] It becomes evident that these are texts discussing domestic policy issues. The SPD in particular seems to be highly frequent. However, at first glance, there are no obvious differences between the news providers.

check, if this is correct.

[11] http://snowball.tartarus.org/algorithms/german/stop.txt
[12] https://tartarus.org/martin/PorterStemmer/
[13] The wordclouds of the other parties can be found in the appendix A.2

Figure 6: Wordclouds

(a) Bild.de (after pre-processing) (b) Tagesschau.de (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 (Gentzkow et al., 2017), since the order in which words are used within a document is disregarded.

5.3.2 Structural topic model

To find out the latent structure 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. One crucial assumption to be made for such models is the number of topics (K) that occur over the entire corpus.

Each individual topic 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 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:[14]

1. for each document i, draw its distribution of topics θ_d depending on the metadata included in the model;

2. for each topic k, draw its distribution of words ϕ_k depending on the metadata included in the model;

3. for each word n, draw its topic z_n based on θ_i;

4. 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 in the past few years (Blei et al. (2003) Erosheva et al. (2004) Braun and McAuliffe (2010)). 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

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

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 (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 55 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, the corresponding news website or party of an article or press release influences the probability distribution of topics for that document. Additionally I assume that the topical content differs between news articles and press releases.

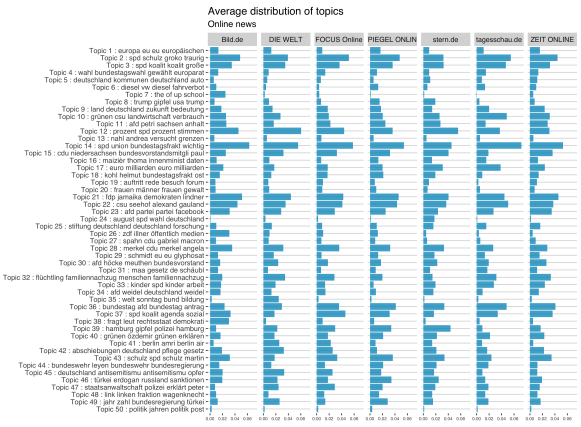
describe model specifications

5.3.3 Results

In order to get an initial overview of the results, Figure ?? displays the topics ordered by their expected frequency across the corpus. To assign a label to each topic, I looked at the most frequent words in that topic and the most representative articles (M. E. Roberts, B. M. Stewart, and E. M. Airoldi, 2016).

It becomes apparent that topic 4 about the coalition talks between CDU/CSU and SPD - the "Grand coalition" or "GroKo" - is the topic with the highest expected frequency in the whole corpus, followed by the topic about the so-called Jamaica parties (CDU/CSU, FDP and B90/Die Grünen), which was the first alternative to be negotiated directly after the elections.

Figure 7: Expected frequency



(a) Online news

../figs/wordclouds/topic_proportion_press.png

(b) Press releases

6 Fixed effects regression

This section seeks to examine the association between sentiment reflected in online news content and phone poll results in Germany. Specifically, it aims to find the extent to

which online sentiment and phone survey results correlate given a number of lags. I use the data from the "Sonntagsumfrage" (Sunday survey) from infratest dimap.[15] The institution regularly asks at least 1000 German citizens the question: "Which party would you choose if federal elections were to take place next Sunday?" The survey thus measures the current election tendencies and therefore reflects an intermediate state in the opinion-forming process of the electoral population.

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.

In the present paper, the relationship between monthly average of both the sentiment value of individual topics (x_t) and the survey value of the parties (y_t) is estimated using the cross correlation function (CCF). Thus, the CCF between x_{t+h} and y_t for $h \pm 1$,$h \pm 2$,$h \pm 3$ is computed. A negative value for h is a correlation between the topic sentiment value at a time before t and the survey value at time t. The correlation value for $h = 0$ indicates the contemporary correlation between the two time series. Based on the coefficients of the cross correlation estimation shown in Figure 8, the significant correlations between topic sentiment and survey value are evaluated for each party.[16] It is important to note that no causal relationships are described below, but only the correlation between the two

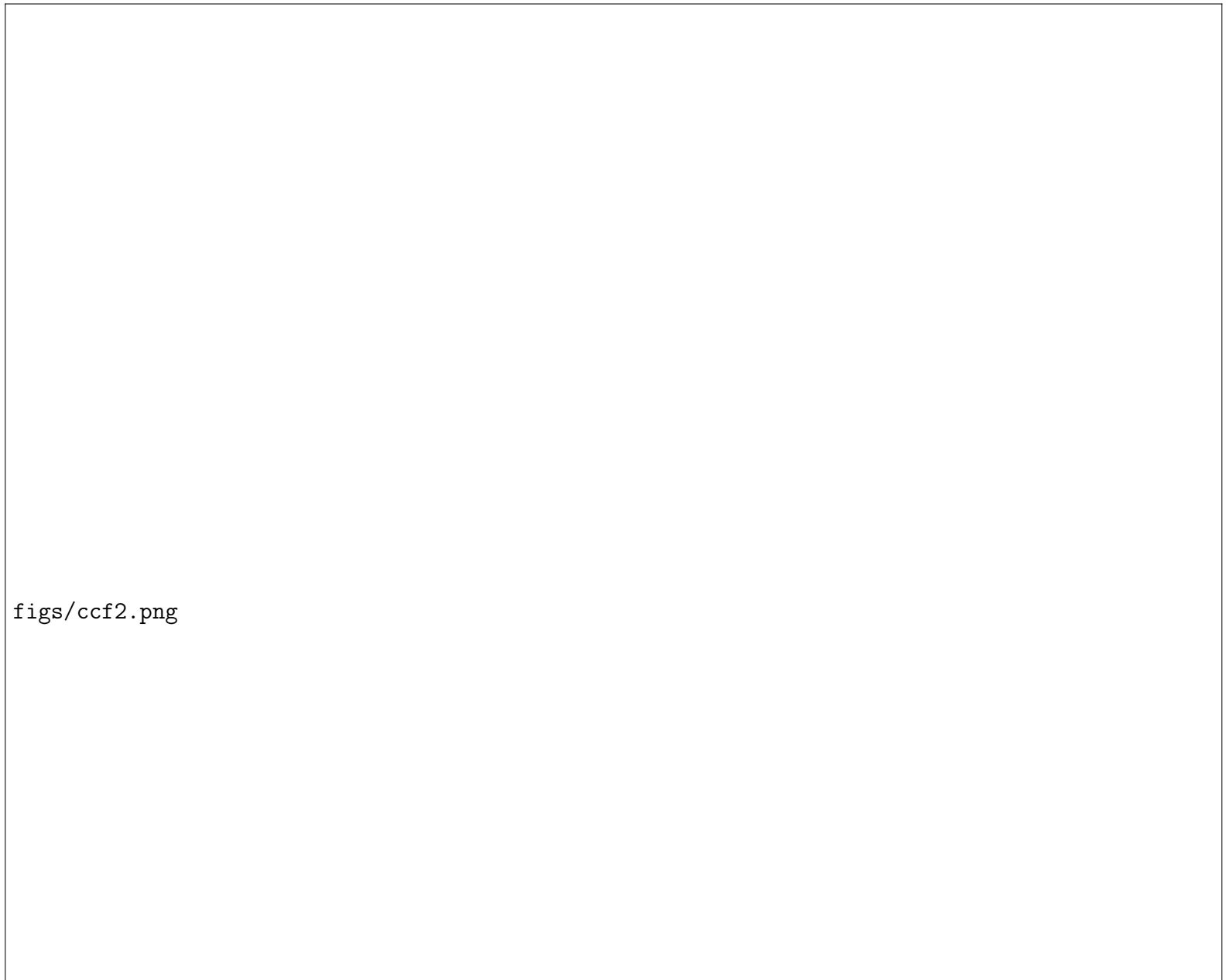
[15] https://www.infratest-dimap.de/umfragen-analysen/bundesweit/sonntagsfrage/

[16] The value of the cross correlation coefficients for lag 0 are listed in the appendix A.7

time series.

The survey results of the AfD correlate negatively with topics relating to the SPD (17, 22) at lag 0. Thus, if the SPD was more negatively reported, the poll value of the AfD increased in the same month (and vice versa). Another significant negative correlation exists between the reporting on the GroKo (4) and the survey value of AfD at lag -1 (x_{t-1}). So if the GroKo was more negatively reported in one month, the survey value of the AfD increased in the following month (and vice versa). For the FDP, too, only negative correlation coefficients can be detected, with the strongest negative correlation existing for the topic relating to the CSU (23). If the CSU got off worse in the online news, the poll value of the FDP went up. Another interesting observation is that the FDP's poll results correlate negatively with issues relating to the Jamaica coalition at lag 1 (x_{t+1}). So if the poll results for the FDP rose in one month, the following month the FDP was reported more negatively. The Green Party survey results show no negative correlation with any of the topics, except topic 30 at lag 1. It is striking that there seems to be a strong negative correlation between the SPD topics (1, 17, 22) and the poll results of the left party (DIE LINKE). This means that the poll value of the left party has climbed if the topics related to the SPD were discussed more negatively. Same applies to the reporting on the GroKo (30) for lag -1. Conversely, the SPD's survey results correlate strongly positively with these topics, and also with topic 30 with a delay of one month. For the CDU/CSU, too, only significant negative correlations are discernible: the survey results correlate negatively with the topic of the Schulz v Merkel debate (10) and negatively with topic 30 with a delay of one month (x_{t+1}).

Figure 8: Cross-Correlation Coefficients



After the figures above have been analyzed, the following points can be summarized:

1. Only the survey results of the SPD correlate positively with the emotional value of the topics. There seems to be a strong correlation between the way topics concerning the SPD are discussed in the online news and the poll results.

2. The poll results of the Left Party, on the other hand, seem to correlate negatively with the reporting on the SPD.

3. Similar tendencies can also be seen with regard to the AfD, since here too the survey results correlate significantly negatively with the topics about the SPD and the grand coalition.

Summarizing the analyses from this and the previous section, it can be observed that the positive correlation between the emotional value of the reporting and the survey value of a party is particularly large if the reporting is conspicuously negative.

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 (1) whether the political reporting of different media differs in terms of topic frequency and topic tonality and (2) whether the emotional value of the reporting correlates with poll results.

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 in order to answer (1). After assigning a topic to each news article, the sentiment value of articles about contemporary political events is calculated using a dictionary-based method. In order to tackle (2), the results from the sentiment analysis are then compared to poll results.

Regarding (1), the analysis revealed that there are differences between the media considered, both in terms of topic prevalence and the way in which these topics are discussed. Although the topics are discussed negatively on average, differences can still be observed, especially regarding topics that deal with the coalition negotiations. The smallest differences were observed for topics concerning the AfD. However, no evidence has been found that the media systematically report more negatively on the AfD than on other parties. With regard to (2), the analysis has shown that the tonality of topics discussed by the SPD shows a strong positive correlation to current survey results. Overall, there seems to be a link between reporting on political issues and electoral preferences. The results of this study show evidence that the content of media could have an influence on the opinion-forming process of the voters and therefore underline the responsibility of media in the political context.

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).

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).

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya (2011). "Media and Political Persuasion: Evidence from Russia". In: *The American Economic Review* 101.7, pp. 3253–3285. URL: https://www.jstor.org/stable/41408737 (visited on 08/22/2018).

Erosheva, Elena, Stephen Fienberg, and John Lafferty (June 4, 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 (May 1, 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).

Ferree, Myra Marx et al. (2002). "Four Models of the Public Sphere in Modern Democracies". In: *Theory and Society* 31.3, pp. 289–324. URL: https://www.jstor.org/stable/658129 (visited on 10/20/2018).

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, 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).

— (June 4, 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.

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).

McCombs, Maxwell (Nov. 1, 2005). "A Look at Agenda-setting: past, present and future". In: *Journalism Studies* 6.4, pp. 543–557. URL: https://doi.org/10.1080/14616700500250438 (visited on 10/20/2018).

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. (Jan. 5, 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.

— (Jan. 12, 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.*

Savigny, Heather (Feb. 1, 2002). "Public Opinion, Political Communication and the Internet". In: *Politics* 22.1, pp. 1–8. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9256.00152 (visited on 08/14/2018).

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).

Strömbäck, Jesper (July 1, 2008). "Four Phases of Mediatization: An Analysis of the Mediatization of Politics". In: *The International Journal of Press/Politics* 13.3, pp. 228–246. URL: https://doi.org/10.1177/1940161208319097 (visited on 10/20/2018).

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

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.

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

A.2 Wordclouds

(a) DIE WELT

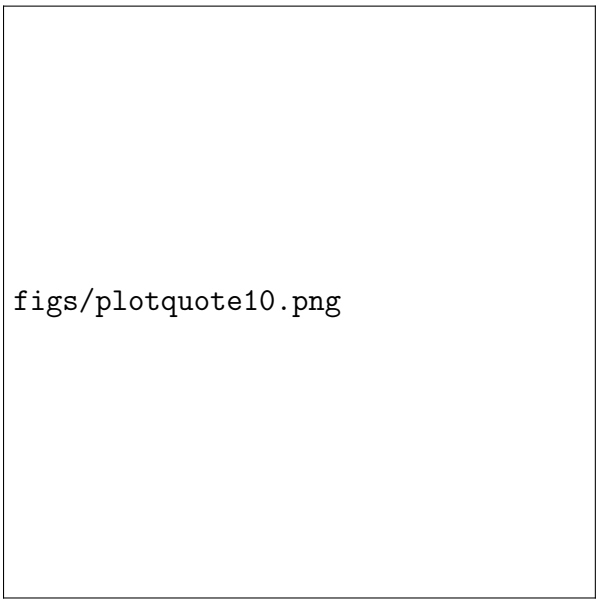
(b) FOCUS ONLINE

(c) SPIEGEL ONLINE

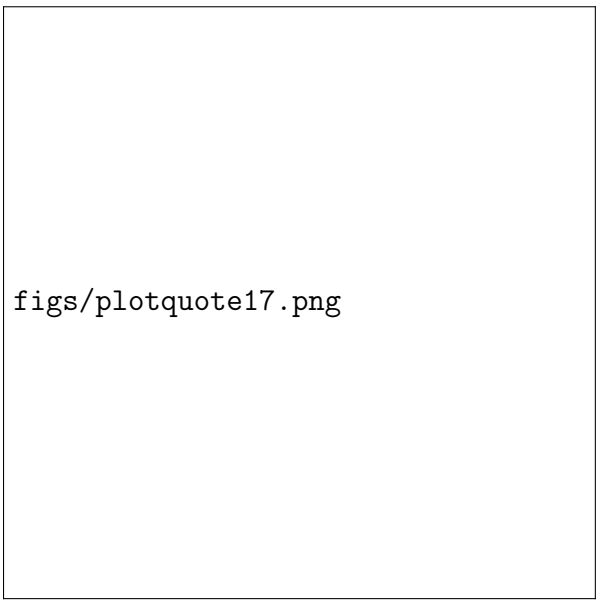
(d) Stern.de

(e) ZEIT ONLINE

A.3 Most frequent words




figs/plotquote13.png


figs/plotquote17.png


figs/plotquote20.png


figs/plotquote23.png


figs/plotquote26.png


figs/plotquote27.png


figs/plotquote30.png


figs/plotquote32.png

A.4 Regression Results

topic_name	parameter	Estimate	Std. Error	t value	p
1: SPD, M.Schulz	(Intercept)	0.036	0.003	11.460	0.000
1: SPD, M.Schulz	FOCUS ONLINE	-0.007	0.004	-1.777	0.076
1: SPD, M.Schulz	SPIEGEL ONLINE	-0.008	0.004	-1.965	0.049
1: SPD, M.Schulz	stern.de	-0.009	0.004	-2.340	0.019
1: SPD, M.Schulz	Tagesschau.de	-0.015	0.005	-3.311	0.001
1: SPD, M.Schulz	DIE WELT	-0.015	0.004	-3.834	0.000
1: SPD, M.Schulz	ZEIT ONLINE	-0.012	0.004	-2.686	0.007
2: B90/ Die Grünen	(Intercept)	0.018	0.003	7.057	0.000
2: B90/ Die Grünen	FOCUS ONLINE	-0.005	0.003	-1.515	0.130
2: B90/ Die Grünen	SPIEGEL ONLINE	0.001	0.003	0.260	0.795
2: B90/ Die Grünen	stern.de	0.005	0.003	1.689	0.091
2: B90/ Die Grünen	Tagesschau.de	-0.003	0.004	-0.805	0.421
2: B90/ Die Grünen	DIE WELT	-0.004	0.003	-1.160	0.246
2: B90/ Die Grünen	ZEIT ONLINE	0.004	0.004	1.035	0.301
4: Great Coalition debates	(Intercept)	0.064	0.005	13.210	0.000
4: Great Coalition debates	FOCUS ONLINE	-0.010	0.006	-1.698	0.089
4: Great Coalition debates	SPIEGEL ONLINE	0.003	0.006	0.487	0.626
4: Great Coalition debates	stern.de	-0.032	0.006	-5.472	0.000
4: Great Coalition debates	Tagesschau.de	-0.005	0.006	-0.841	0.401
4: Great Coalition debates	DIE WELT	-0.012	0.006	-2.183	0.029
4: Great Coalition debates	ZEIT ONLINE	0.002	0.006	0.400	0.689
10: M.Schulz, vs. A.Merkel	(Intercept)	0.009	0.002	4.131	0.000
10: M.Schulz, vs. A.Merkel	FOCUS ONLINE	0.001	0.003	0.192	0.848
10: M.Schulz, vs. A.Merkel	SPIEGEL ONLINE	0.002	0.003	0.565	0.572
10: M.Schulz, vs. A.Merkel	stern.de	0.007	0.003	2.860	0.004
10: M.Schulz, vs. A.Merkel	Tagesschau.de	0.003	0.003	0.991	0.322
10: M.Schulz, vs. A.Merkel	DIE WELT	-0.003	0.003	-1.117	0.264
10: M.Schulz, vs. A.Merkel	ZEIT ONLINE	0.012	0.003	3.798	0.000
13: A.Merkel, election campaign	(Intercept)	0.024	0.003	9.704	0.000
13: A.Merkel, election campaign	FOCUS ONLINE	0.002	0.003	0.550	0.582
13: A.Merkel, election campaign	SPIEGEL ONLINE	0.003	0.003	0.992	0.321
13: A.Merkel, election campaign	stern.de	0.009	0.003	2.709	0.007
13: A.Merkel, election campaign	Tagesschau.de	-0.006	0.003	-1.824	0.068
13: A.Merkel, election campaign	DIE WELT	0.001	0.003	0.297	0.767
13: A.Merkel, election campaign	ZEIT ONLINE	0.004	0.004	1.137	0.256
17: Debates within SPD	(Intercept)	0.035	0.003	10.632	0.000
17: Debates within SPD	FOCUS ONLINE	-0.005	0.004	-1.315	0.189
17: Debates within SPD	SPIEGEL ONLINE	-0.009	0.004	-2.186	0.029
17: Debates within SPD	stern.de	-0.001	0.004	-0.361	0.718
17: Debates within SPD	Tagesschau.de	-0.013	0.005	-2.809	0.005
17: Debates within SPD	DIE WELT	-0.010	0.004	-2.612	0.009
17: Debates within SPD	ZEIT ONLINE	-0.007	0.005	-1.579	0.114
20: AfD, right-wing radicalism	(Intercept)	0.017	0.003	5.704	0.000
20: AfD, right-wing radicalism	FOCUS ONLINE	-0.002	0.004	-0.607	0.544
20: AfD, right-wing radicalism	SPIEGEL ONLINE	-0.003	0.004	-0.724	0.469
20: AfD, right-wing radicalism	stern.de	-0.003	0.004	-0.784	0.433
20: AfD, right-wing radicalism	Tagesschau.de	-0.000	0.004	-0.033	0.973
20: AfD, right-wing radicalism	DIE WELT	0.002	0.004	0.512	0.609
20: AfD, right-wing radicalism	ZEIT ONLINE	-0.002	0.004	-0.538	0.590
22: SPD, stuffing debates	(Intercept)	0.019	0.003	7.408	0.000
22: SPD, stuffing debates	FOCUS ONLINE	-0.004	0.003	-1.170	0.242
22: SPD, stuffing debates	SPIEGEL ONLINE	0.010	0.003	2.960	0.003
22: SPD, stuffing debates	stern.de	-0.006	0.003	-1.943	0.052
22: SPD, stuffing debates	Tagesschau.de	-0.003	0.003	-0.903	0.367
22: SPD, stuffing debates	DIE WELT	-0.004	0.003	-1.308	0.191
22: SPD, stuffing debates	ZEIT ONLINE	0.005	0.003	1.449	0.147
23: CSU, Söder & Seehover	(Intercept)	0.035	0.004	9.191	0.000
23: CSU, Söder & Seehover	FOCUS ONLINE	-0.004	0.005	-0.887	0.375
23: CSU, Söder & Seehover	SPIEGEL ONLINE	0.006	0.005	1.094	0.274
23: CSU, Söder & Seehover	stern.de	-0.002	0.005	-0.365	0.715
23: CSU, Söder & Seehover	Tagesschau.de	-0.006	0.005	-1.214	0.225
23: CSU, Söder & Seehover	DIE WELT	-0.006	0.004	-1.238	0.216
23: CSU, Söder & Seehover	ZEIT ONLINE	0.002	0.005	0.357	0.721
26: Jamaica fail, Reelections or GroKo?	(Intercept)	0.035	0.003	11.775	0.000
26: Jamaica fail, Reelections or GroKo?	FOCUS ONLINE	-0.004	0.004	-1.137	0.256
26: Jamaica fail, Reelections or GroKo?	SPIEGEL ONLINE	-0.004	0.004	-1.076	0.282
26: Jamaica fail, Reelections or GroKo?	stern.de	-0.012	0.004	-3.463	0.001
26: Jamaica fail, Reelections or GroKo?	Tagesschau.de	-0.010	0.004	-2.584	0.010
26: Jamaica fail, Reelections or GroKo?	DIE WELT	-0.012	0.004	-3.333	0.001
26: Jamaica fail, Reelections or GroKo?	ZEIT ONLINE	-0.005	0.004	-1.356	0.175
27: Jamaica Coalition debates	(Intercept)	0.066	0.005	13.865	0.000
27: Jamaica Coalition debates	FOCUS ONLINE	-0.029	0.006	-5.152	0.000
27: Jamaica Coalition debates	SPIEGEL ONLINE	-0.009	0.006	-1.438	0.150
27: Jamaica Coalition debates	stern.de	-0.028	0.006	-5.088	0.000
27: Jamaica Coalition debates	Tagesschau.de	-0.018	0.006	-2.939	0.003
27: Jamaica Coalition debates	DIE WELT	-0.008	0.005	-1.416	0.157
27: Jamaica Coalition debates	ZEIT ONLINE	-0.003	0.007	-0.460	0.645
30: AfD, F.Petry & Meuthen	(Intercept)	0.026	0.003	7.579	0.000
30: AfD, F.Petry & Meuthen	FOCUS ONLINE	-0.005	0.004	-1.168	0.243
30: AfD, F.Petry & Meuthen	SPIEGEL ONLINE	0.002	0.004	0.430	0.667
30: AfD, F.Petry & Meuthen	stern.de	-0.004	0.004	-0.896	0.370
30: AfD, F.Petry & Meuthen	Tagesschau.de	-0.014	0.004	-3.035	0.002
30: AfD, F.Petry & Meuthen	DIE WELT	-0.011	0.004	-2.812	0.005
30: AfD, F.Petry & Meuthen	ZEIT ONLINE	0.004	0.005	0.914	0.361
32: AfD, Gauland & Weidel	(Intercept)	0.023	0.004	6.308	0.000
32: AfD, Gauland & Weidel	FOCUS ONLINE	0.001	0.004	0.191	0.848
32: AfD, Gauland & Weidel	SPIEGEL ONLINE	-0.001	0.005	-0.123	0.902
32: AfD, Gauland & Weidel	stern.de	0.007	0.004	1.593	0.111
32: AfD, Gauland & Weidel	Tagesschau.de	-0.006	0.005	-1.221	0.222
32: AfD, Gauland & Weidel	DIE WELT	0.006	0.004	1.378	0.168
32: AfD, Gauland & Weidel	ZEIT ONLINE	0.006	0.005	1.198	0.231
37: AfD & DIE LINKE in parliament	(Intercept)	0.024	0.004	6.652	0.000
37: AfD & DIE LINKE in parliament	FOCUS ONLINE	0.005	0.004	1.221	0.222
37: AfD & DIE LINKE in parliament	SPIEGEL ONLINE	0.013	0.005	2.984	0.003
37: AfD & DIE LINKE in parliament	stern.de	0.001	0.004	0.252	0.801
37: AfD & DIE LINKE in parliament	Tagesschau.de	0.005	0.005	1.130	0.258
37: AfD & DIE LINKE in parliament	DIE WELT	0.002	0.004	0.483	0.629
37: AfD & DIE LINKE in parliament	ZEIT ONLINE	0.011	0.005	2.143	0.032
46: CDU	(Intercept)	0.018	0.002	7.336	0.000
46: CDU	FOCUS ONLINE	-0.001	0.003	-0.245	0.806
46: CDU	SPIEGEL ONLINE	0.002	0.003	0.753	0.451
46: CDU	stern.de	-0.006	0.003	-1.942	0.052
46: CDU	Tagesschau.de	-0.011	0.003	-3.364	0.001
46: CDU	DIE WELT	-0.001	0.003	-0.403	0.687
46: CDU	ZEIT ONLINE	-0.000	0.003	-0.023	0.982

A.5 Sentiment Values (monthly aggregated)

A.6 Sentiment Values (aggregated by news website)

A.7 Cross Correlation Coefficient (lag 0)

	Var1	AfD	FDP	Grüne	Linke	SPD	Union
1	1: SPD, M.Schulz	-0.636	0.072	-0.438	-0.693	0.752	0.199
2	10: M.Schulz, vs. A.Merkel	0.438	0.566	0.505	0.341	-0.229	-0.758
3	13: A.Merkel, election campaign	0.098	0.350	0.440	-0.060	0.076	-0.507
4	17: Debates within SPD	-0.770	0.075	-0.676	-0.754	0.841	0.382
5	2: B90/ Die Grünen	0.253	-0.092	0.185	0.480	-0.329	-0.048
6	20: AfD, right-wing radicalism	0.225	0.369	0.224	0.037	-0.157	-0.319
7	22: SPD, stuffing debates	-0.710	-0.112	-0.465	-0.797	0.877	0.302
8	23: CSU, Söder & Seehover	-0.169	-0.897	-0.249	-0.283	0.145	0.565
9	26: Jamaica fail, Reelections or GroKo?	-0.251	-0.501	-0.388	-0.258	0.047	0.537
10	27: Jamaica Coalition debates	-0.197	-0.439	-0.266	-0.222	0.000	0.436
11	30: AfD, F.Petry & Meuthen	0.410	-0.032	0.351	0.296	-0.348	-0.224
12	32: AfD, Gauland & Weidel	-0.220	0.430	0.221	-0.315	0.336	-0.296
13	37: AfD & DIE LINKE in parliament	-0.217	-0.521	-0.418	-0.003	0.238	0.552
14	4: Great Coalition debates	-0.165	0.336	-0.361	0.292	-0.126	0.228
15	46: CDU	-0.083	-0.066	-0.108	-0.286	0.285	-0.007

Table 3: Cross Correlation at lag 0