Political news coverage of online news

...

Franziska Löw*

November 2020

I Introduction

In democracies, the media fulfill fundamental functions: They should inform the people, contribute to the formation of opinion through criticism and discussion and thus enable participation. In recent decades, however, concern has grown about the role of the media in politics in general and in election campaigns in particular. They are criticized for influencing election results through their reporting and for helping populist parties in particular to flourish. After the 2017 federal elections in Germany, for example, the media were accused of contributing to the success of the right-wing populist AfD by increasingly including the party's content and using the same language in their articles as the AfD. Representatives of these media houses strongly opposed this accusation. The purpose of this study is to examine whether there is evidence that supports the accusation of biased media reporting, especially during election campaigns.

For advertising-financed media the battle for the attention of the recipients is at the center of economic decisions. Online media in particular, which offer their content to a large extent free of charge and generate their revenue through advertising space, compete for the scarce resource of attention. Consumers pay a non-monetary price providing their attention, which the media platform bundles and sells on to advertising customers. This business model corresponds to that of a platform market, in which media companies act as platforms that connect the market of advertising with the reader market to exploit the indirect network effects between them (Dewenter and Rösch 2014). A profit-maximizing publisher therefore directs its economic decisions according to what will attract the most attention.

This conclusion, derived from the economic theory of platform markets, corresponds to the notion of media logic, a central concept in the field of media and communication studies (Takens et al. 2013). The debate about media logic is embedded in the broader discussion about the interaction between the press, politics and the public. The underlying thesis is that the content of political news is the product of news values and narrative techniques that media use to attract audiences (Strömbäck 2008). According to Takens et al. (2013), three content attributes highly correspond with news values and influence how journalists interpret political events: 1) personalized content, i.e., the focus on individual politicians; 2) the framing of politics as a contest and 3) negative coverage. Similarly Blassnig et al. (2019) states that media primarily focus on news factors, i.e. the factors that turn an event into news worth reporting like conflict, drama, negativity, surprise or proximity. Likewise populist messages often co-occur with negative, emotionalized, or dramatized communication style, thus utilizing similar mechanisms as the media logic, respectively the attention economy. In fact, Blassnig et al. (2019) shows that populist key messages by political and media actors in news articles provoke more reader comments under these articles. Media competing for the attention of readers therefore have an incentive to pick up on the key messages of these parties. These, in turn, benefit from being able to place their agendas in the public debate (Druckman and Parkin (2005), Eberl (2018)). It is assumed, that smaller, non-established parties in particular benefit from placing their topics in the media in order to get them into the voters' heads. Here, the tendency of the reporting is irrelevant but rather the quantity is decisive.

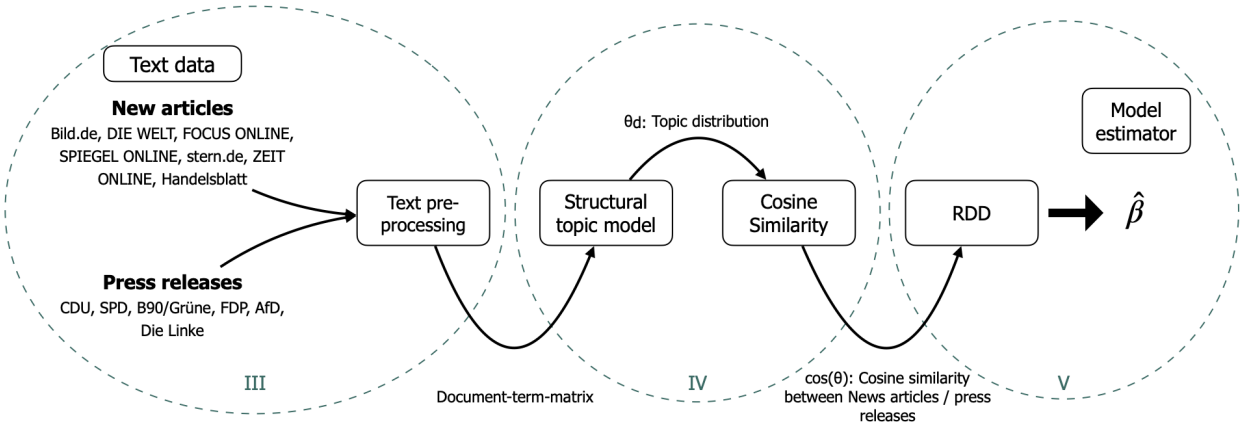
*Institut für Industrieökonomik, Helmut Schmidt Universität. Email: .

However, the causal relationship between reporting and voter preferences is not the subject of this study. Rather, it is intended to measure how online news coverage coincides with the messaging of different political parties. Especially during election campaigns political 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, Boomgaarden, and Wagner 2017). Parties instrumentalize their public relations 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). This paper therefore analyzes the period before and after the 2017 federal elections in Germany to answer the question whether political news report about similar topics and if this similarity changes after the election campaign is over.

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 - be it audio, video or text content. Studies that rely on quantifying media content for their analyses use, for example, visibility (how often political actors appear in the media) or tonality (how they are evaluated). 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 content similarity between online news and parties press releases.[1] To discover the latent topics in the corpus of text data, the structural topic model (STM) developed by M. E. Roberts, Stewart, and Airoldi (2016) is applied. This probabilistic text model results in a probability distribution for each document across all topics, which is then aggregated to calculate the degree of difference between the news articles of different media providers and the press releases of the parties.

The figure below gives a high level overview of the different steps that are taken to text data for the analysis. Section III describes the sources of data and how text data is processed in order to use it as input for the structural topic model. Section IV briefly explains how the STM uses this input in order to calculate topic probabilities for the different documents and how these probabilities are used to calculate the similarity between the documents. In the final section V, we use the similarity measures as our dependent variable in a regression discontinuity model. In order to set the context, the following section II provides an introduction to the political situation in Germany as well as the German news market during the time under investigation.

[1] For the sake of simplicity, both news articles and press releases will be referred to as documents in the following.



II Background information

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.[2] 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.

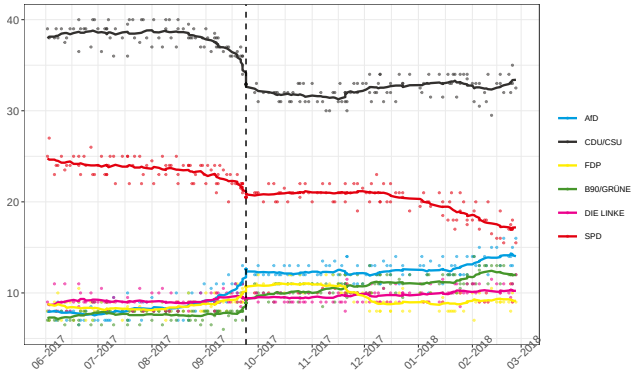


Figure 1: Election polls during the period under review

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, Handelsblatt, SPIEGEL ONLINE, stern.de, ZEIT ONLINE. As can be seen from Figure 2(a), expect for Handelsblatt (position 53), these media outlets are among the top 30 German online news providers in the period under review in terms of visits.[3]

The main source of income for these privately managed media houses 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 (N. 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[4] 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.[5] The results displayed in Figure 2(b) indicate that the media used for the analysis play a relevant role in their consumption.

III Text data

I conduct the estimation on a sample of 18,757 online news articles from the seven German news providers described in the previous section[6] about domestic politics and press releases of the seven parties that have been in the Bundestag since the 2017 federal elections[7]. Both news articles and press releases are dated from June 1, 2017 to March 1, 2018.

[2]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/.

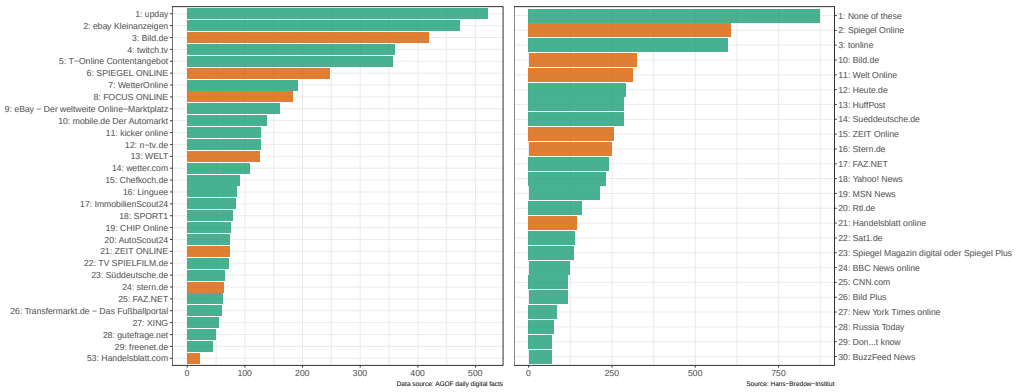
[3]The term visit is used to describe the call to a website by a visitor. The visit begins as soon as a user generates a page impression (PI) within an offer and each additional PI, which the user generates within the offer, belongs to this visit.

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

[5]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."

[6]Bild.de, DIE WELT, FOCUS ONLINE, SPIEGEL ONLINE, stern.de, ZEIT ONLINE, Handelsblatt

[7]CDU, SPD, B90/Grüne, FDP, AfD, Die Linke



(a) Total visits in million (Jan 2018) (b) Use of a brand to access news online

Figure 2: Selected german news brands

News articles scraped from the Webhose.io API.[8] In order to consider only news about national politics, the articles were filtered based on their URL. The press releases were scraped from the public websites of the political parties and parliamentary groups using an automated script.[9]

Figure 3 shows the distribution of the number of articles by date and media outlet. 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 lines).[10] Furthermore, Figure 3 shows that DIE WELT published the most articles on domestic policy, followed by stern.de, Handelsblatt and FOCUS ONLINE.

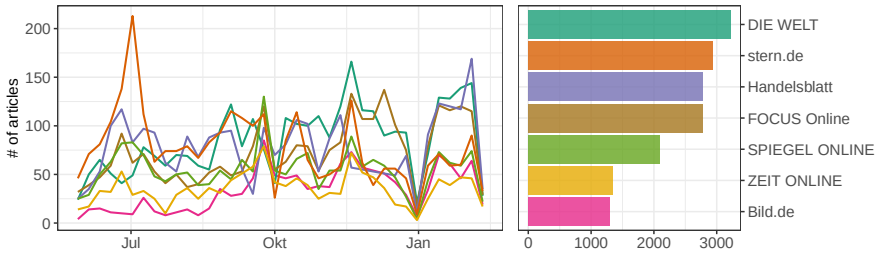


Figure 3: Distribution of news articles

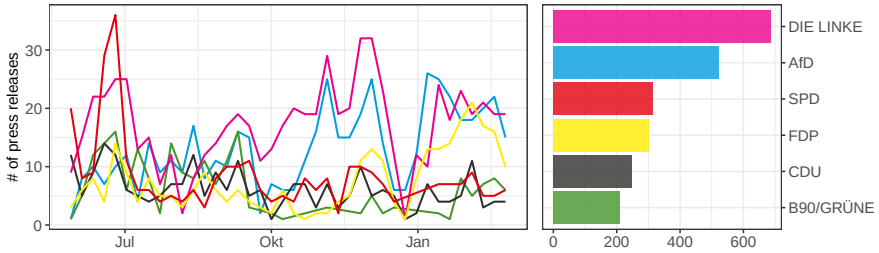


Figure 4: Distribution of press releases

[8] For more information see https://docs.webhose.io/reference#about-webhose.
[9] The scraping code was written in R and can be made available on request.
[10] The peak in July especially for stern.de is due to increased reporting about the G20 summit in Hamburg.

Table 1: Summary statistics of word counts

source	n	mean	sd	median	min	max
News articles						
Bild.de	1303	476.07	318.28	398.0	121	3710
DIE WELT	3222	509.57	612.06	380.0	121	14507
FOCUS Online	2780	393.89	317.05	297.5	121	5647
Handelsblatt	2785	589.51	495.82	488.0	121	6899
SPIEGEL ONLINE	2089	539.09	415.05	413.0	121	3466
stern.de	2943	514.66	616.55	373.0	121	9287
ZEIT ONLINE	1351	513.75	387.14	459.0	121	8015
Press releases						
AfD	523	212.83	72.16	196.0	103	553
B90/GRÜNE	211	229.32	63.37	219.0	104	399
CDU	248	274.54	106.08	256.0	100	1030
DIE LINKE	686	200.47	71.78	190.0	101	1048
FDP	301	161.90	83.78	144.0	100	999
SPD	315	213.41	56.16	208.0	103	429

Table 1 illustrates that on average, news articles have a higher word count than the parties' press releases.[11] While for news articles the average is between 394 (FOCUS Online) and 590 (Handelsblatt), with press releases the range is between 162 (FDP) and 275 (CDU). The article with the most words (14.507) was published by DIE WELT - the longest press release has 1.048 words and was published by DIE LINKE.

Text pre-processing

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 (M. Gentzkow, Kelly, and Taddy (2017), 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. 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 5, 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, and Taddy 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[12] 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.[13]

As an example, the following word clouds represent the most frequent words of the pre-processed articles for Bild.de (Figure 6(a)) and press releases of AfD (Figure 6(b)). It becomes evident that these are texts discussing domestic policy issues. The SPD in particular seems to be highly frequent for Bild.de.

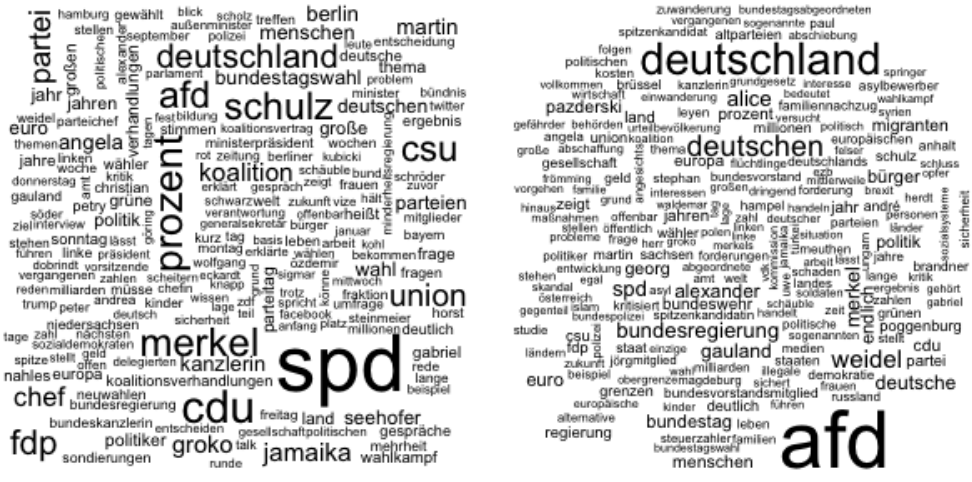
The next step is to divide the entire data set into individual documents and to represent these documents as a finite list of unique terms. In this setting, each news article and each press release represents a document d, whereby each of these documents can be assigned to a news website or a party. 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

[11] News articles with less than 120 words were filtered out in advance, as these were mostly reader comments. Similarly press releases with less than 100 words were filtered out.

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

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

Figure 5: Wordcloud before pre-processing



(a) Bild (b) AfD

Figure 6: Wordcloud after pre-processing

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. Table 2 provides a sample output of the document-term matrix used in this paper, where each document is represented by a unique id (the row name in the example below). This representation is often referred to as the bag of words model (M. Gentzkow, Kelly, and Taddy 2017), since the order in which words are used within a document is disregarded.

Table 2: Document-term matrix - sample values

	partnern	klassischen	patienten	oettinger	söder	linksunten	analysen
1401	0	0	0	0	0	0	0
15723	0	0	0	0	0	0	0
6187	0	0	0	0	0	0	0
11796	0	0	0	0	0	0	0
9645	0	0	0	0	0	0	0
6255	0	0	0	0	0	0	0
11633	0	0	0	0	0	0	0
6530	0	0	0	0	0	0	0
5133	0	0	0	0	0	0	0
2776	0	0	0	0	0	0	0

IV Estimate topic similarity of documents

A structural topic model to identify the latent topics

To discover the latent topics in the corpus of press releases and news articles, a structural topic modeling (STM) developed by (M. E. Roberts, Stewart, and Airoldi 2016) is applied. 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 is an extension of the standard topic modelling technique, labeled as latent Dirichlet allocation (LDA), which refers to the Bayesian model in Blei, Ng, and Jordan (2003) that treats each word in a topic and each topic in a document as generated from a Dirichlet-distributed prior.[14].

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 STM is an extension of the LDA process since it allows covariates of interest (such as the publication date of a document or it's author) to be included in the prior distributions for both topic proportions (θ) and topic-word distributions (ϕ). This way, STM offers a method of 'structuring' the prior distributions in the topic model, including additional information in the statistical inference procedure, while LDA assumes that θ Dirichlet(α) and ϕ Dirichlet(β), where α and β are fitted with the model.

In order to include the covariates in the statistical inference procedure, two design matrices of covariates (X and Z) are specified, where each row 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:

- 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;

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

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

One crucial assumption to be made for topic models like LDA or STM is the number of topics (K) that occur over the entire corpus. There is not a "right" answer to the number of topics that are appropriate for a given corpus (Grimmer and Stewart 2013). M. Roberts, Stewart, and Tingley (2016b) propose to measure topic quality through a combination of semantic coherence and exclusivity of words to topics. Semantic coherence is a criterion developed by Mimno et al. (2011) and is closely related to pointwise mutual information (D. Newman et al. 2010): it is maximized when the most probable words in a given topic frequently co-occur together.

Using the function *searchK* from the *stm* package several automated tests are performed to help choose the number of topics including the average exclusivity and semantic coherence as well as the held out likelihood (Wallach, Mimno, and McCallum 2009) and the residuals (Taddy 2012). This process revealed that a model with 40 topics best reflects the structure in the corpus. Furthermore, I use the author and bi-week dummies of a document as topical prevalence variable. In other words, I assume that the probability of a topic to be included in a news article or a press release depends on the author of that document and when it was published. I argue that these variables are best suited to capture temporal and publisher level variation in the documents.

```
## stm(documents = news_df_sparse, K = 40, prevalence = ~source +
##     s(year_biweek), data = covariates, init.type = "Spectral")
```

In general inference of mixed-membership models, such as the one applied in this paper, has been a thread of research in applied statistics 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 and the covariates influencing the prior distributions). 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 specifications (M. Roberts, 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, and Taddy 2017).

Results of the STM

As mentioned in the previous section, the generative process of the STM results in a topic distribution θ_d for each document d over all topics k and a word distribution ϕ_k for each topic over all terms in the vocabulary. The most probable words of each topic may help to understand what each topic is about.[15] However, since those most probable words are not necessarily the most exclusive words and they only represent a small fraction of the probability distribution, interpretation should be done very cautiously.

For our analysis, we use the topic distribution of each document to estimate the similarity of documents. Figure 7 illustrates such a topic distribution of two news paper articles. The red numbers display the topic probability (for probabilities >= 0.02). News article 1[16] shows a clear distribution towards topic 36, for which terms like Bundeswehr, Soldaten (soldiers), Nato, Verteidigungsministerin (defense minister) are among the most probable words. News article 2[17] does not show such a clear tendency towards a single topic. Topic 40, 18 and 5 are within a comparable range. However, for all three topics similar terms are among the top terms.

Similarly as for the news articles, Figure 8 illustrates the topic distribution for two press releases randomly chosen from the corpus. For press release 1[18] topic 24 is the most probable topic which contains terms about the G20 Summit, during which left-wing radicals caused considerable riots. Topic distribution of press article 2[19] shows peaks for topic 6 and 35. Top terms of topic 6 contain the words trump, us, usa, deutschland (Germany) and präsident (president). Similarly topic 35 seems to deal with German foreign policy, since top terms include words like eu, deutschland (Germany), europa and bundesregierung (Federal Government).

[15] Table 10 gives an overview of the most probable terms for each topic.
[16] Bundeswehr scandal: ex-commander attacks Von Der Leyen
[17] Bundestag elections: 42 parties want to be elected to parliament.
[18] Lars Herrmann: The danger for Germany and its Basic Law is also coming from the left
[19] Trump chooses the path to isolation

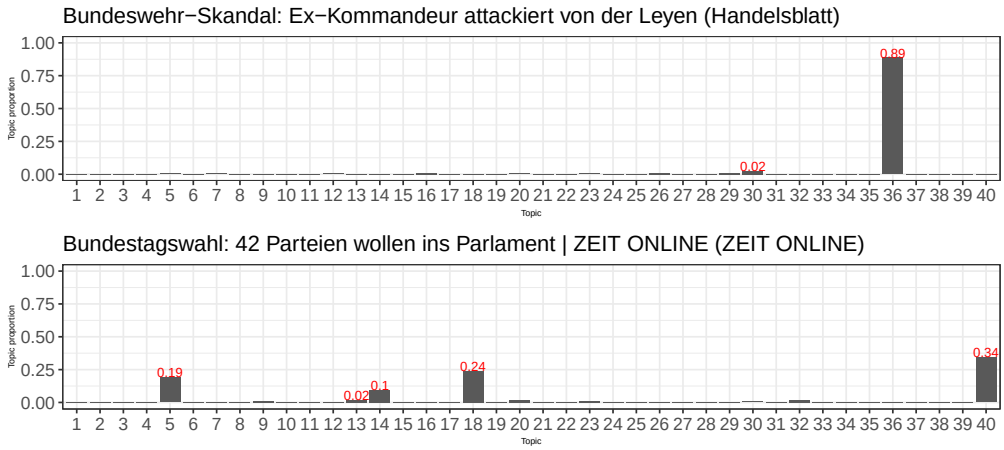


Figure 7: Topic probability of sample news articles

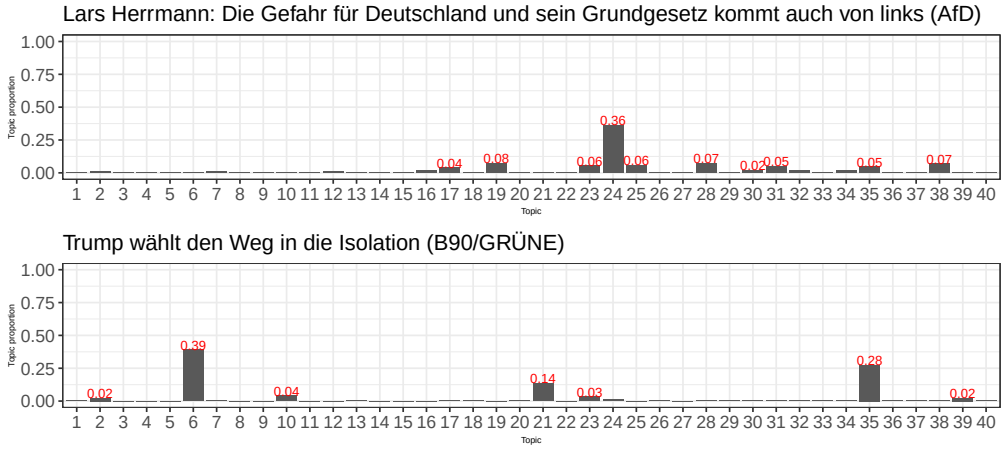


Figure 8: Topic probability of sample press releases

Since the source and publication date is known for each document, the probability of certain topics can be analysed, aggregated by this metadata. The left chart of Figure 9 is showing the 15 topics with the highest probability for press releases published by the AfD. The right side of the figure is aggregating the probability by source and time (in weeks) for two sample topics, displaying how they change over time in the AfD press releases compared to two sample news papers. It becomes clear, that topic 9[20] is systematically more likely in the AfD's press releases compared to the two newspapers. There is a noticeable increase in probability during the election campaign period and ends in a peak on election day itself. For Handelsblatt and Bild.de, too, a slight increase around election day is discernible and the probability of this topic shows some peaks for Bild.de both before and after the election. The top words of Theme 38[21] suggest that it addresses refugees - a topic the AfD has a very clear position on. The probability of this topic increases in the AfD's press releases until about a month before the election and then levels off somewhat. A similar trend can be seen for the news articles from Bild.de. The curve from Handelsblatt is rather flat and shows no apparent difference between before and after the election.

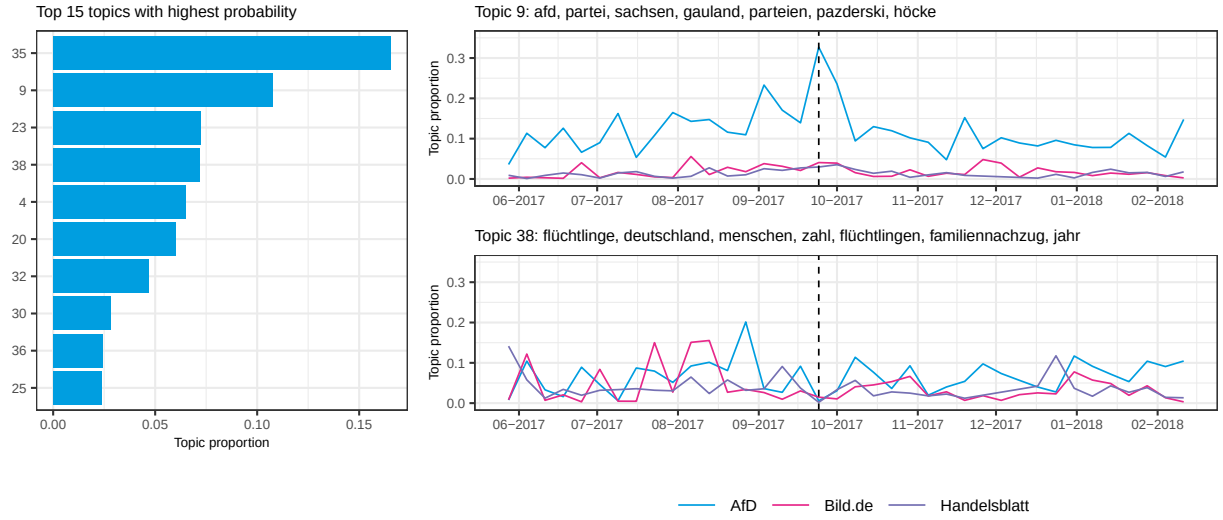


Figure 9: Comparison of topic probability - sample topics AfD

Figure 10 is doing a similar analysis for the aggregated topic distribution in press releases of the FDP. The chart on the left illustrates that, as in the case of the AfD, the topic 35 has the highest probability in the press releases of the FDP. Unlike in the case of the AfD, however, this is followed by topics that have clearer temporal peaks, as shown on the right using two example topics. Topic 10[22] has a clear peak for both the news papers and the FDP press releases around august 2017. At that time, there was a debate about whether and where driving bans for diesel cars would be introduced. After the states of Baden-Württemberg and North Rhine-Westphalia initially filed a lawsuit against this, the court proceedings that would decide whether driving bans are permissible began in mid-February 2018. The temporal curve of the FDP shows a further increase in topic probability at this time, which can also be detected at Handelsblatt. At Bild.de, however, the topic is apparently only taken up once briefly in August 2017, as only a very low topic probability can be seen thereafter. The peak of the probabilty of topic 39[23] in all three sources right after the election is reflecting the exploratory talks on the possible formation of a Jamaica coalition, which officially failed on November 19, 2017 after the FDP announced its withdrawal from the negotiations.

Cosine similarity

Next, the cosine similarity measure is used in order to compare the retrieved topic distribution of documents. Cosine similarity is a measure for the distance between two vectors and is defined between zero and one;

[20] translation: afd, party, saxony, gauland, parties, pazderski, höcke
[21] translation: refugees, germany, people, number, refugees, family reunion, year
[22] translation: diesel, enterprises, germany, cars, german, industry, driving bans
[23] translation: fdp, grünen, jamaika, cdu, union, grüne, cdu

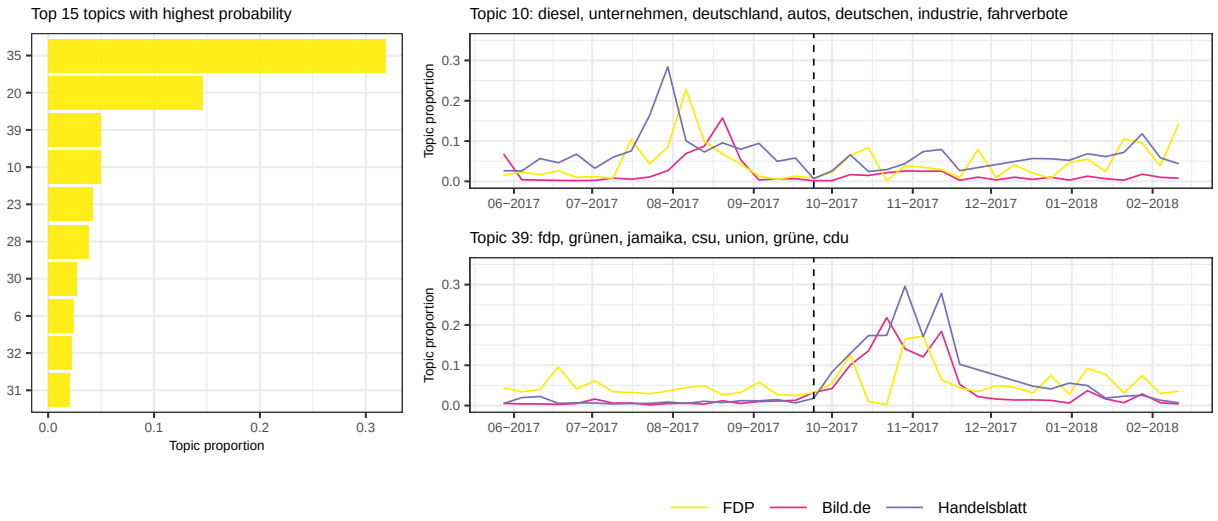


Figure 10: Comparison of topic probability - sample topics FDP

values towards 1 indicate similarity. As topic proportions per document are vectors of the same length, the cosine similarity allows a comparison of the topic distribution between two documents.[24]

$$\text{CS} = \cos(\theta) = \frac{a * b}{||a||||b||}$$

For example, Table 3[25] displays the most similar documents of the document with the title *Parteitag, Koalitions-Krimi, Ur-Wahl - Woran kann die GroKo jetzt noch scheitern? (Bild.de)*.[26]

Table 3: Most similar documents

title	source	cos_sim
Asylzahlen 2017 - schwarz-rot führt Obergrenze durch die Hintertür ein	DIE LINKE	0.542
Jörg Meuthen: Nicht neue GroKo, sondern LoKo – Loser Koalition	AfD	0.522
Rentenpläne von Union und SPD schlimmer als erwartet	FDP	0.478
Union und SPD stabilisieren die krasse soziale Ungerechtigkeit in diesem Land	DIE LINKE	0.401
Jürgen Pohl: Union und SPD verabreden Politik zu Lasten von Rentnern und Ostdeutschen	AfD	0.388

In the next step, for each news paper, the cosine similarity between all topic-document distribution pairs between the news papers articles and the press releases is calculated if that press releases was published within 7 days before the publication date of the news article. This means the topic distribution of news article 1 is compared to press release 1, 2, 3 and so on as long as the press release was published within 7 days before the news article. Table 4 illustrates a sample subset of the data for DIE WELT.

[24] For applications of cosine similarity to compare of topic model outcomes see e.g. Rehs (2020) and Ramage, Dumais, and Liebling (2010)

[25] Translations: 1) Asylum figures 2017 - black-red (synonym for GroKo) introduces upper limit through the back door (DIE LINKE) 2) Jörg Meuthen: Not a new GroKo, but LoKo - Loser Coalition (AfD) 3) Pension plans of Union and SPD worse than expected (FDP) 4) Union and SPD stabilize the extreme social injustice in this country (DIE LINKE) 5) Jürgen Pohl: Union and SPD agree on policy at the expense of pensioners and East Germans (AfD)

[26] Translation: Party conference, coalition thriller, primal election - what can fail the GroKo now? (Bild.de)

Table 4: Dataset structure step 1 - DIE WELT

title1	title2	cosine_sim	source1	source2	date1	date2
Angela Merkel: We...	Behördenplagiat: ...	0.01	DIE WELT	B90/GRÜNE	2017-09-17	2017-09-15
SPD-Kanzlerkandid...	Georg Pazderski: ...	0.20	DIE WELT	AfD	2017-08-05	2017-08-01
IS-Terror in Pari...	Deutschlands Enth...	0.01	DIE WELT	DIE LINKE	2017-11-12	2017-11-09
Knapp 400.000: Me...	Welt-AIDS-Tag: Pr...	0.49	DIE WELT	DIE LINKE	2017-12-03	2017-11-30
Haftanstalt Tegel...	Das Gesundheitssy...	0.05	DIE WELT	SPD	2018-02-10	2018-02-08

Next, the mean cosine similarity for each news article publication date (date1) and party (source2) is estimated to obtain the final data frame (see Table 5).

Table 5: Final dataset structure - DIE WELT

date1	source1	source2	cos_sim
2017-09-10	DIE WELT	AfD	0.15
2017-11-14	DIE WELT	DIE LINKE	0.17
2017-09-05	DIE WELT	AfD	0.21
2017-12-26	DIE WELT	DIE LINKE	0.13
2017-09-17	DIE WELT	FDP	0.08

V Regression model

In order to find out if the similarity between press releases and news articles is different for certain party-news publisher pairs and/or change over time, the cosine similarity measure from the previous section is used as the dependent variable in a regression model.

Three different model specifications are estimated for each news publisher: (1) First, a OLS regression is estimated including party dummies. (2) Second, a simple regression discontinuity specification is estimated (without dummies) followed by (2) a specification that includes dummy variables for parties to check for differences in news publisher/party pairs.

OLS dummy regression

To measure whether there is a significant difference in the topic similarity for each party for a news publisher, a simple OLS regression is estimated, where the similarity score ($\ln(\text{CS}_t)$) on day t between the news articles of that news publisher and the press releases of a political party k is the dependent variable and the political party dummies are the independent variables.

$$\ln(\text{CS}_t) = \beta_0 + \beta_n D_{t,k-1} + \epsilon_t,$$

with t = date[27], k = political party[28]

OLS dummy results

Figure 11 plots the mean cosine similarity between the news articles and each parties press releases over time for Handelsblatt and Bild.de (See Figure 15 for all news papers). Based on these figures, a few observations can be made. The similarity with the FDP, e.g., seems to increase over time for both Handelsblatt and Bild.de. In the case of Bild.de, this applies to all parties, except for the AfD, where a slight downward trend is discernible.

[27] date1 in Table 5
[28] source2 in Table 5

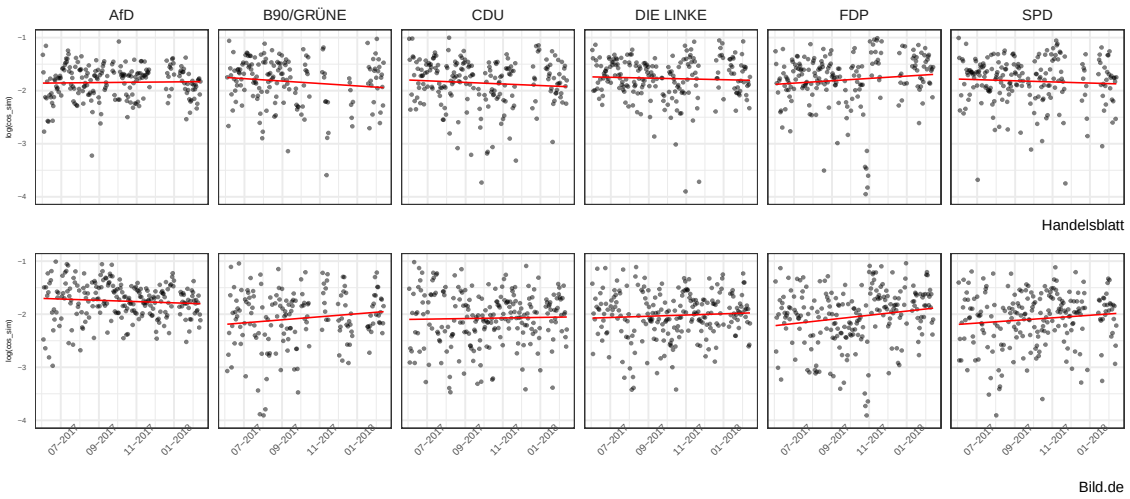


Figure 11: Log of daily mean cosine similarity between newspaper/press articles pairs

The columns in Table 6 report the results for each new publisher.[29] Since the dependent variable is log transformed, the the % impact of D on Y can be estimated as $exp(\hat{\beta}) - 1$. E.g. since the transformed coefficient for $D_{B90/GR\ddot{U}NE}$ in the first column - representing the model for DIE WELT - is -0.232, a switch from from 0 to 1 can be interpreted as a 20.6 decrease of topic similarity for B90/GRÜNE compared to AfD (the base dummy group), holding everything else equal. In other words: Compared to AfD, the similarity of topics of DIE WELT and B90/GRÜNE is 20.6 lower. Figure 12 plots the transformed coefficients for all news paper / party pairs (insignifcant coefficients are shown with a low opacity). In general, for all news papers - except for Handelsblatt - the topic similarity is significantly less between the news articles of that news paper and press releases of political parties when compared to press releases of the AfD. The biggest difference for all parties exist for Bild.de followed by FOCUS ONLINE and DIE WELT. In the case of Handelsblatt not significant difference can be found for CDU/CSU, B90/GRÜNE and SPD (compared to AfD). However, the positive coefficients of FDP and DIE LINKE indicate, that the topics discussed in press releases of these parties and the articles of Handelsblatt are significantly similar when compared to the press releases of the AfD.

Table 6: Results from the OLS model

	Dependent variable:						
	Cosine similarity of topic distribution						
	DIE WELT	stern.de	ZEIT ONLINE	Handelsblatt	FOCUS Online	Bild.de	SPIEGEL ONLINE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
B90/GRÜNE	−0.232***	−0.237***	−0.220***	0.033	−0.297***	−0.319***	−0.202***
	(0.036)	(0.045)	(0.057)	(0.050)	(0.038)	(0.055)	(0.045)
CDU	−0.274***	−0.207***	−0.280***	−0.003	−0.300***	−0.347***	−0.240***
	(0.033)	(0.042)	(0.053)	(0.047)	(0.036)	(0.051)	(0.043)
DIE LINKE	−0.268***	−0.140***	−0.197***	0.078*	−0.290***	−0.272***	−0.208***
	(0.033)	(0.042)	(0.053)	(0.047)	(0.036)	(0.051)	(0.042)
FDP	−0.247***	−0.139***	−0.201***	0.080*	−0.281***	−0.292***	−0.221***
	(0.033)	(0.042)	(0.053)	(0.047)	(0.036)	(0.051)	(0.043)
SPD	−0.263***	−0.180***	−0.228***	0.026	−0.301***	−0.337***	−0.267***
	(0.034)	(0.042)	(0.054)	(0.048)	(0.036)	(0.051)	(0.043)
Constant	−1.765***	−1.991***	−1.764***	−1.856***	−1.831***	−1.746***	−1.878***
	(0.024)	(0.030)	(0.038)	(0.033)	(0.025)	(0.036)	(0.030)
Observations	1,372	1,378	1,404	1,206	1,428	1,321	1,424
R^2	0.069	0.027	0.023	0.005	0.072	0.047	0.035
Adjusted R^2	0.065	0.023	0.019	0.001	0.069	0.044	0.031
Residual Std. Error	0.365 (df = 1366)	0.458 (df = 1372)	0.589 (df = 1398)	0.484 (df = 1200)	0.401 (df = 1422)	0.549 (df = 1315)	0.473 (df = 1418)
F Statistic	20.144*** (df = 5; 1366)	7.532*** (df = 5; 1372)	6.540*** (df = 5; 1398)	1.188 (df = 5; 1200)	22.193*** (df = 5; 1422)	13.048*** (df = 5; 1315)	10.136*** (df = 5; 1418)
Note:							*p<0.1; **p<0.05; ***p<0.01

[29] All regression output tables are created using Hlavac (2018)

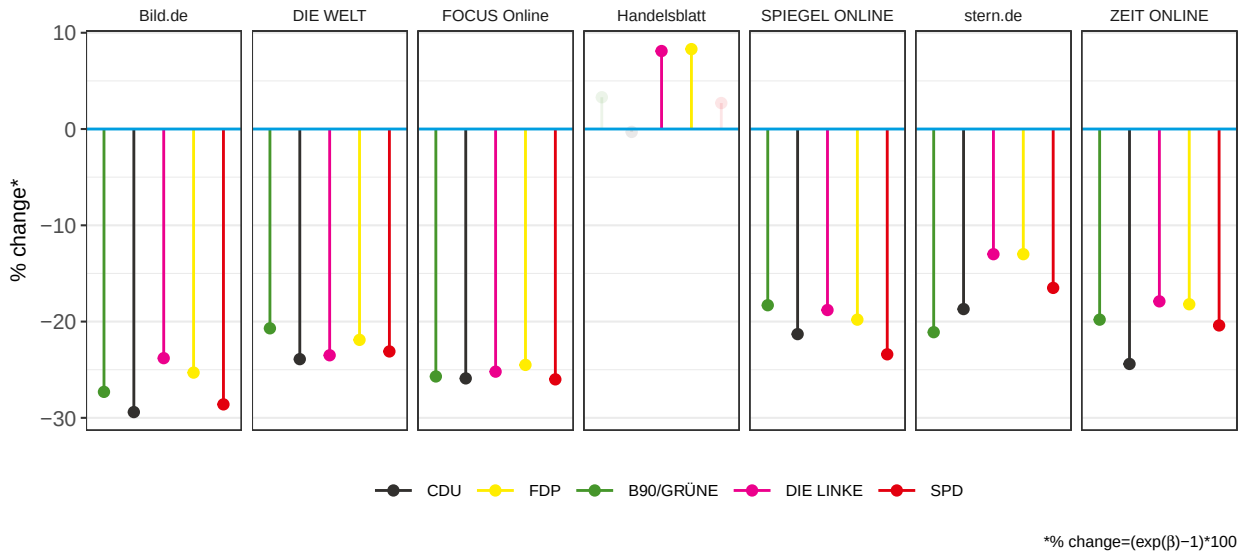


Figure 12: Coefficients of OLS dummy regression

Regression discontinuity model

We assume that news publisher report differently during election campaigns and that the election day introduces a change in the reporting. The underlying dynamic of this assumption coincides with the basic idea of regression discontinuity design (RDD). Therefore, a RDD is applied to identify the short-term effect of the election on the topic similarity between news paper articles and press releases. The RDD was designed by Thistlethwaite and Campbell (1960) and formalized by Hahn, Todd, and Van der Klaauw (2001) to measure the effect of a treatment in a nonexperimental setting, where the treatment defined as discontinuous function of a continuous, observed variable (the 'running' or 'forcing' variable). Like Thistlethwaite and Campbell (1960), who estimated the effect of receiving the National Merit Scholarship on future academic outcomes, early studies that rely on RD designs estimate the effects of certain thresholds of a running variable on educational outcomes (i.e. financial aid (Klaauw (2002)) or class size (Angrist and Lavy (1999))). Following these early studies in the area of education, the RDD has received attention in a wider range of the economic literature, including labor economics, political economy, health economics, and environmental economics. Compared to alternative quasi-experimental estimators like difference-in-difference and matching techniques, RDD is considered as the estimator with the greatest internal validity (Lee and Lemieux (2010)).

While RDD was originally applied in cross-sectional studies, an increasing number of studies, especially in the field of environmental and energy economics, has adapted the framework to time series applications.[30] In these studies, time is the running variable and treatment begins at a particular threshold in time. An important conceptual difference between regression discontinuity (RD) and regression discontinuity in time (RDiT) exists in the possible interpretation. The fact that in RDiT the running variable of time is not random, eliminates the interpretation of local randomization in which it is assumed that treatment status within a small neighborhood around the threshold can essentially be compared to a roll of the dice. As noted by Jacob et al. (2012), although some researchers have focused on this interpretation of local randomization (Lee and Lemieux (2010)), others have instead emphasized that RD is characterized by discontinuity at a threshold (Hahn, Todd, and Van der Klaauw (2001)). Thus, to the extent that the RD framework is simply another quasi-experimental framework (one that uses discontinuity), RDiT is conceptually similar to RD.

In this paper date is the running variable, the election day is the treatment and news publishers are the units that receive the treatment. Since the running variable (date) completely determines the treatment (election day), a sharp regression design is used, such that the probability that a news publisher receives a treatment jumps from 0 to 1 at the cutoff. Specifically, I estimate the equation

[30] See Hausman and Rapson (2018) for examples of this regression discontinuity in time (RDiT).

$$\ln(\text{CS}_t) = \beta_0 + \beta_1 T_t + f(W_t) + \epsilon_t$$

where

$$T_t = \begin{cases} 1, & \text{if date } \geq \text{election date} \\ 0, & \text{if date } < \text{election date} \end{cases}$$

and the running variable W_t is the time difference between date i and the election date (in days), such that β_1 is the average treatment effect for observations with $W_t =$ election date. In other words, β_1 gives the average change of the similarity between the content of news publisher and press releases after the election day. Identification in the RD model comes from assuming that the underlying, potentially endogenous relationship between ϵ_t and the date is eliminated by the flexible function $f(.)$. In particular, the relationship between ϵ_t and the date must not change discontinuously on or near the election date.

Following Imbens and Lemieux (2008) I estimate a local linear regression of the form:

$$\ln(\text{CS}_t) = \beta_0 + \beta_1 T_t + \beta_2 W_t + \beta_3 W_t * T_t + \epsilon_t$$

In this specification (results are shown in RDiT Results), the function $f(W_t)$ is specified as $\beta_2 W_t + \beta_3 W_t * T_t$, where by $W_t * T_t$ is assumed that in addition to the intercept (captured by the treatment effect T_t), the slope also changes after the election day. Both this interaction term and W_t should absorb any smooth relationship between the date and ϵ_t in the days surrounding the election day. If the RD assumption is valid (i.e., ϵ_t does not change discontinuously at the election day) the estimate of β_1, the coefficient of interest, will be unbiased even without further controls. However, in section RDiT dummy results I include dummy variables for each party k ($D_{t,k-1}$) assuming that the effect of the election day on topic similarity differs for different parties. Again, the interaction term $T_t * D_{t,k-1}$ is included to allow for a slope change depending on the party. Thus β_5 gives the average treatment effect for each newspaper / party pair.

$$\ln(\text{CS}_t) = \beta_0 + \beta_1 T_t + \beta_2 W_t + \beta_3 W_t * T_t + \beta_4 D_{t,k-1} + \beta_5 T_t * D_{t,k-1} + \epsilon_t$$

I specify a uniform kernel (Lee and Lemieux (2010)) and use a bandwidth of 115 days on each side of the election day threshold. The election took place on September 24, 2017, so the sample includes dates between June 1, 2017 and January 17, 2018. Since the identification strategy only attempts to estimate β at $W_t = 0$ (the election day) no additional dates beyond the 115 day bandwidth enter the sample. Alternative specifications with varying bandwidths led to similar results.

RDiT Results

Table 7 shows the results of the estimation.

The negative and significant values for β_1 indicate that - holding the time constant -, the election day is associated with a decrease in topic similarity for DIE WELT, stern.de, ZEIT ONLINE and Handelsblatt. Since we are interested in the treatment effect at the cutoff point (remember that $W_t = 0$ for the election day) and since

$$\frac{\Delta Y}{\Delta T} = \beta_1 + \beta_3 W,$$

we can interpret β_1 as the change in topic similarity with respect to the election day. Similarly to the interpretation in OLS dummy results, the the % impact of T on Y can be estimated as $exp(\hat{\beta}) - 1$. Figure 13 shows the transformed coefficients (of $beta_1$) for all newspapers: The negative effect of the election day on the topic similarity is strongest for Handelsblatt (~21.5% decrease), followed by DIE WELT (~15.6% decrease), ZEIT ONLINE (~13.5% decrease) and stern.de (~11.1% decrease).

Table 7: Results from the RDiT model

	Dependent variable:						
	Cosine similarity of topic distribution						
	DIE WELT	stern.de	ZEIT ONLINE	Handelsblatt	FOCUS Online	Bild.de	SPIEGEL ONLINE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
T	−0.169***	−0.118**	−0.145**	−0.242***	−0.058	−0.032	0.035
	(0.045)	(0.057)	(0.065)	(0.060)	(0.048)	(0.066)	(0.055)
W	−0.001*	−0.0005	−0.002***	−0.0003	−0.001	−0.001	−0.002***
	(0.0004)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.001)	(0.001)
TTRUE:W	0.003***	0.003***	0.002*	0.005***	0.002**	0.003***	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	−1.972***	−2.181***	−1.927***	−1.812***	−2.091***	−2.080***	−2.160***
	(0.029)	(0.036)	(0.045)	(0.043)	(0.033)	(0.046)	(0.038)
Observations	1,212	1,218	1,244	1,046	1,268	1,161	1,264
R^2	0.031	0.011	0.059	0.035	0.005	0.013	0.009
Adjusted R^2	0.028	0.009	0.056	0.032	0.003	0.011	0.007
Residual Std. Error	0.377 (df = 1208)	0.471 (df = 1214)	0.586 (df = 1240)	0.491 (df = 1042)	0.427 (df = 1264)	0.579 (df = 1157)	0.495 (df = 1260)
F Statistic	12.805*** (df = 3; 1208)	4.655*** (df = 3; 1214)	25.810*** (df = 3; 1240)	12.532*** (df = 3; 1042)	2.134* (df = 3; 1264)	5.154*** (df = 3; 1157)	3.820*** (df = 3; 1260)
Note:							*p<0.1; **p<0.05; ***p<0.01

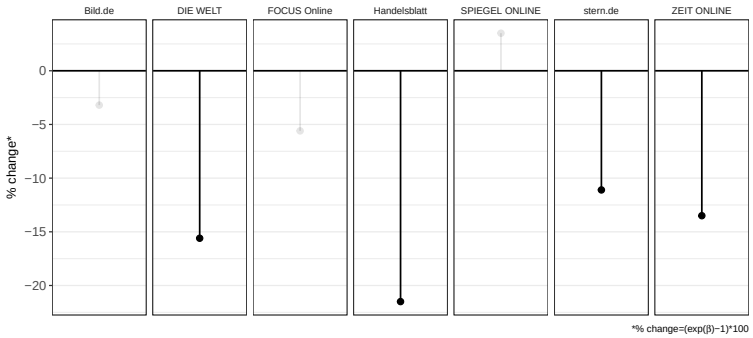


Figure 13: Treatment effect of RDiT regression

RDiT dummy results

The visual representation of the treatment effect for two sample news publisher Handelsblatt and Bild.de in Figure 14 reveals that the topic similarity for Handelsblatt decreases after the election for CDU, FDP and B90/GRÜNE. A less clear to no effect can be seen for SPD, DIE LINKE and AfD. In contrast, the treatment effect at Bild.de seems to exist only for the AfD (see Figure 16 for all news publisher).

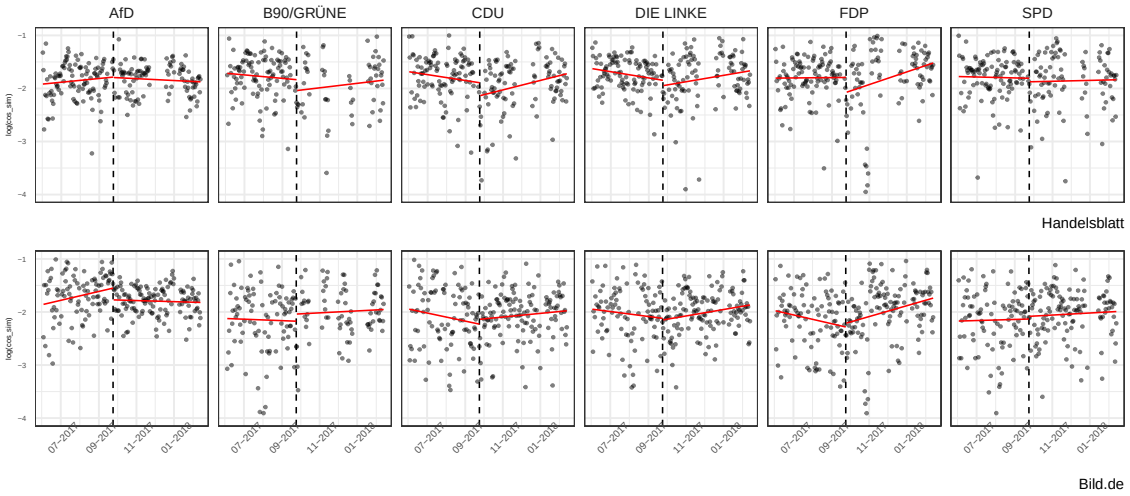


Figure 14: Log of mean cosine similarity between newspaper/press articles pairs - with cutoff value

Table 8: Results from the regression discontinuity model

	Dependent variable:						
	Cosine similarity of topic distribution						
	DIE WELT	stern.de	ZEIT ONLINE	Handelsblatt	FOCUS Online	Bild.de	SPIEGEL ONLINE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
T	−0.167***	−0.048	−0.038	−0.091	−0.072	−0.187*	0.076
	(0.063)	(0.081)	(0.096)	(0.090)	(0.068)	(0.097)	(0.081)
W	−0.001*	−0.0005	−0.002***	−0.0003	−0.001	−0.001	−0.002***
	(0.0004)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.001)	(0.001)
B90/GRÜNE	−0.267***	−0.215***	−0.180**	0.105	−0.337***	−0.445***	−0.186***
	(0.048)	(0.062)	(0.078)	(0.067)	(0.054)	(0.081)	(0.064)
CDU	−0.259***	−0.115*	−0.232***	0.098	−0.288***	−0.450***	−0.170***
	(0.048)	(0.061)	(0.078)	(0.067)	(0.054)	(0.081)	(0.064)
DIE LINKE	−0.251***	−0.109*	−0.160**	0.134**	−0.270***	−0.342***	−0.156**
	(0.048)	(0.061)	(0.078)	(0.067)	(0.054)	(0.081)	(0.064)
FDP	−0.291***	−0.195***	−0.201***	0.069	−0.355***	−0.441***	−0.292***
	(0.048)	(0.061)	(0.078)	(0.067)	(0.054)	(0.081)	(0.064)
SPD	−0.278***	−0.143**	−0.201***	0.098	−0.339***	−0.479***	−0.266***
	(0.048)	(0.061)	(0.078)	(0.067)	(0.054)	(0.081)	(0.064)
TTRUE:W	0.003***	0.003***	0.002*	0.004***	0.002**	0.003***	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
TTRUE:B90/GRÜNE	0.058	−0.137	−0.249**	−0.260**	0.055	0.223*	−0.032
	(0.079)	(0.101)	(0.124)	(0.120)	(0.087)	(0.124)	(0.103)
TTRUE:CDU	−0.080	−0.222**	−0.161	−0.272***	−0.073	0.149	−0.165*
	(0.070)	(0.090)	(0.111)	(0.104)	(0.078)	(0.112)	(0.092)
TTRUE:DIE LINKE	−0.062	−0.089	−0.122	−0.162	−0.074	0.092	−0.134
	(0.070)	(0.090)	(0.111)	(0.104)	(0.078)	(0.112)	(0.092)
TTRUE:FDP	0.066	0.078	−0.083	−0.061	0.104	0.219*	0.107
	(0.071)	(0.090)	(0.111)	(0.104)	(0.078)	(0.112)	(0.093)
TTRUE:SPD	0.020	−0.093	−0.087	−0.173	0.062	0.244**	−0.010
	(0.071)	(0.091)	(0.112)	(0.105)	(0.079)	(0.113)	(0.093)
Constant	−1.747***	−2.051***	−1.765***	−1.896***	−1.826***	−1.721***	−1.982***
	(0.042)	(0.053)	(0.067)	(0.061)	(0.047)	(0.069)	(0.056)
Observations	1,212	1,218	1,244	1,046	1,268	1,161	1,264
R^2	0.111	0.052	0.091	0.047	0.089	0.069	0.053
Adjusted R^2	0.101	0.042	0.082	0.035	0.080	0.059	0.043
Residual Std. Error	0.362 (df = 1198)	0.463 (df = 1204)	0.578 (df = 1230)	0.490 (df = 1032)	0.410 (df = 1254)	0.564 (df = 1147)	0.486 (df = 1250)
F Statistic	11.491*** (df = 13; 1198)	5.087*** (df = 13; 1204)	9.506*** (df = 13; 1230)	3.915*** (df = 13; 1032)	9.423*** (df = 13; 1254)	6.584*** (df = 13; 1147)	5.364*** (df = 13; 1250)
Note:							*p<0.1; **p<0.05; ***p<0.01

Example interpretation for *Bild.de* (see Table 8)

The coefficient β_4 for $D_{SPD} = 1$ shows the difference of the treatment effect for SPD. To illustrate this, we can compare the model equation for $D_{SPD} = 1$ and $D_{SPD} = 0$

$D_{SPD} = 1$:
$$\ln(\text{CS})(D_{SPD} = 1) = -1.647 + (-0.189)T + (0.001)W - 0.479 + (0.252)T$$
$$\ln(\text{CS})(D_{SPD} = 1) = -2,126 + (0.001)W + (0.252 - 0.189)T$$

$D_{SPD} = 0$:
$$\ln(\text{CS})(D_{SPD} = 0) = -1.647 + (-0.189)T + (0.001)W$$
$$\ln(\text{CS})(D_{SPD} = 0) = -1.647 + (0.001)W + (-0.189)T$$

- When D_{SPD} switches from 0 to 1, the treatment effect increases by 0.252.

To interpret this coefficient, we have to transform it (our dependent variable is log transformed).

For $D_{SPD} = 0$: $\exp(-0.189) - 1 = -0.172$

After the election day the topic similarity between *Bild.de* and AfD decreased by ~17.2%

$D_{SPD} = 1$: $\exp(-0.189 + 0.252) - 1 = 0.065$

After the election day the topic similarity between *Bild.de* and SPD increased by ~6.5% compared to AfD.

General interpretation:

- News articles of *DIE WELT* and *Bild.de* show a significant decrease of topic similarity with AfD after the election. No such effect can be found for the other news papers.
- News articles of *Handelsblatt* are significantly less similar with B90/G & CDU after the election (compared to AfD). The opposite is true for *Bild.de* where new articles are more similar to ALL party press releases (except for DIE LINKE) when compared to AfD.

The results show, that for some news papers (*DIE WELT* & *Bild.de*), the there is a significant change in the topic similarity between their news articles and the press releases of parties.

VI Discussion and conclusion

This paper investigates whether political reporting of news papers is similar for all political parties. Results show, that the news articles of all news papers (except for *Handelsblatt*) are significantly more similar to press releases of the AfD than any other party.

Furthermore, it was assumed, that this reporting differs between periods of election campaign. The results show a significant effect of the switch between "before" election (election campaign period) and "after" election for some news papers.

- For *DIE WELT* and *Bild.de* the election date has a significant effect on the similarity with the AfD: The similarity between news articles and press releases of AfD decreases

Annex

Table 9: Online sources for press releases

	Party	Parliamentary Group
CDU	cdu.de	presseportal.de
SPD	spd.de	spdfraktion.de
FDP	fdp.de	fdpbt.de
B90/Die Grünen	gruene.de	gruene-bundestag.de
DIE LINKE	die-linke.de	die-linke.de/start/presse/aus-dem-bundestag
AfD	afd.de	afdbundestag.de

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: So, Sep 19, 2021 - 16:18:28

Table 10: 7 most probable terms per topic

	Top Terms
1	a, the, s, of, u, brexit, großbritannien
2	merkel, angela, kanzlerin, bundeskanzlerin, cdu, merkels, deutschland
3	spd, union, cdu, csu, koalitionsvertrag, koalitionsverhandlungen, schulz
4	afd, weidel, gauland, alice, alexander, politiker, äußerungen
5	stimmen, wahlkreis, kandidaten, afd, wahl, gewählt, fdp
6	trump, us, usa, deutschland, präsident, donald, berlin
7	cdu, union, peter, politiker, spahn, altmaier, schäuble
8	spd, koalition, union, groko, große, koalitionsverhandlungen, parteitag
9	afd, partei, sachsen, gauland, parteien, pazderski, höcke
10	diesel, unternehmen, deutschland, autos, deutschen, industrie, fahrverbote
11	ge, ten, be, le, ver, lambsdorff, te
12	gericht, prozess, urteil, richter, staatsanwaltschaft, verfahren, jahre
13	berlin, deutschen, osten, o, tag, jahr, millionen
14	august, cdu, spd, prozent, bundestagswahl, wahl, parteien
15	kohl, helmut, kohls, einheit, kanzler, tod, deutschen
16	spd, nahles, andrea, partei, scholz, schulz, schwesig
17	csu, seehofer, horst, söder, obergrenze, bayern, chef
18	prozent, umfrage, spd, union, fdp, cdu, afd
19	polizei, stadt, menschen, polizisten, täter, verletzt, angaben
20	euro, milliarden, jahr, millionen, prozent, bund, geld
21	grünen, linke, linken, özdemir, partei, wagenknecht, göring
22	cdu, niedersachsen, spd, grünen, rot, fdp, landtag
23	welt, politik, menschen, jahren, lange, frage, fragen
24	g, hamburg, gipfel, polizei, hamburger, demonstranten, scholz
25	deutschland, is, verfassungsschutz, syrien, gefährder, islamisten, staat
26	steinmeier, schmidt, russland, frank, bundespräsident, glyphosat, walter
27	afd, petry, partei, fraktion, frauke, meuthen, gauland
28	berliner, berlin, amri, maizière, innenminister, behörden, daten
29	gabriel, sigmar, außenminister, spd, schröder, amt, gerhard
30	bundestag, spd, abgeordneten, abgeordnete, parlament, abstimmung, fraktion
31	türkei, erdogan, türkischen, deutschland, bundesregierung, türkische, deutsche
32	frauen, deutschland, kinder, studie, eltern, muslime, antisemitismus
33	fdp, jamaika, lindner, koalition, neuwahlen, spd, grünen
34	facebook, maas, twitter, gesetz, internet, netz, heiko
35	eu, deutschland, europa, bundesregierung, europäischen, deutschen, menschen
36	bundeswehr, soldaten, leyen, nato, ursula, einsatz, verteidigungsministerin
37	schulz, spd, martin, kanzlerkandidat, wahlkampf, bundestagswahl, partei
38	flüchtlinge, deutschland, menschen, zahl, flüchtlingen, familiennachzug, jahr
39	fdp, grünen, jamaika, csu, union, grüne, cdu
40	bundestagswahl, afd, wahl, prozent, partei, bundestag, parteien

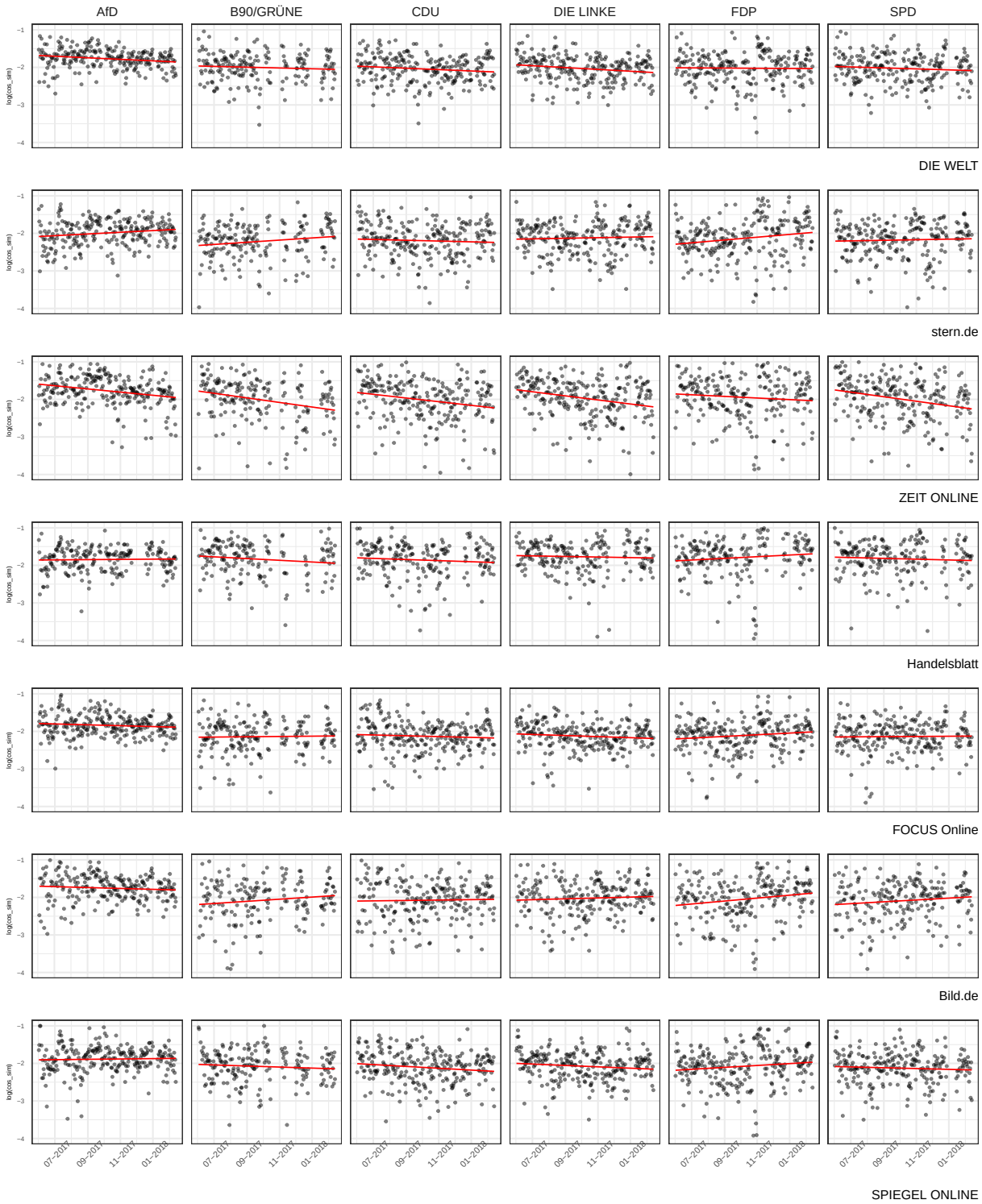


Figure 15: Log of daily mean cosine similarity between newspaper/press articles pairs

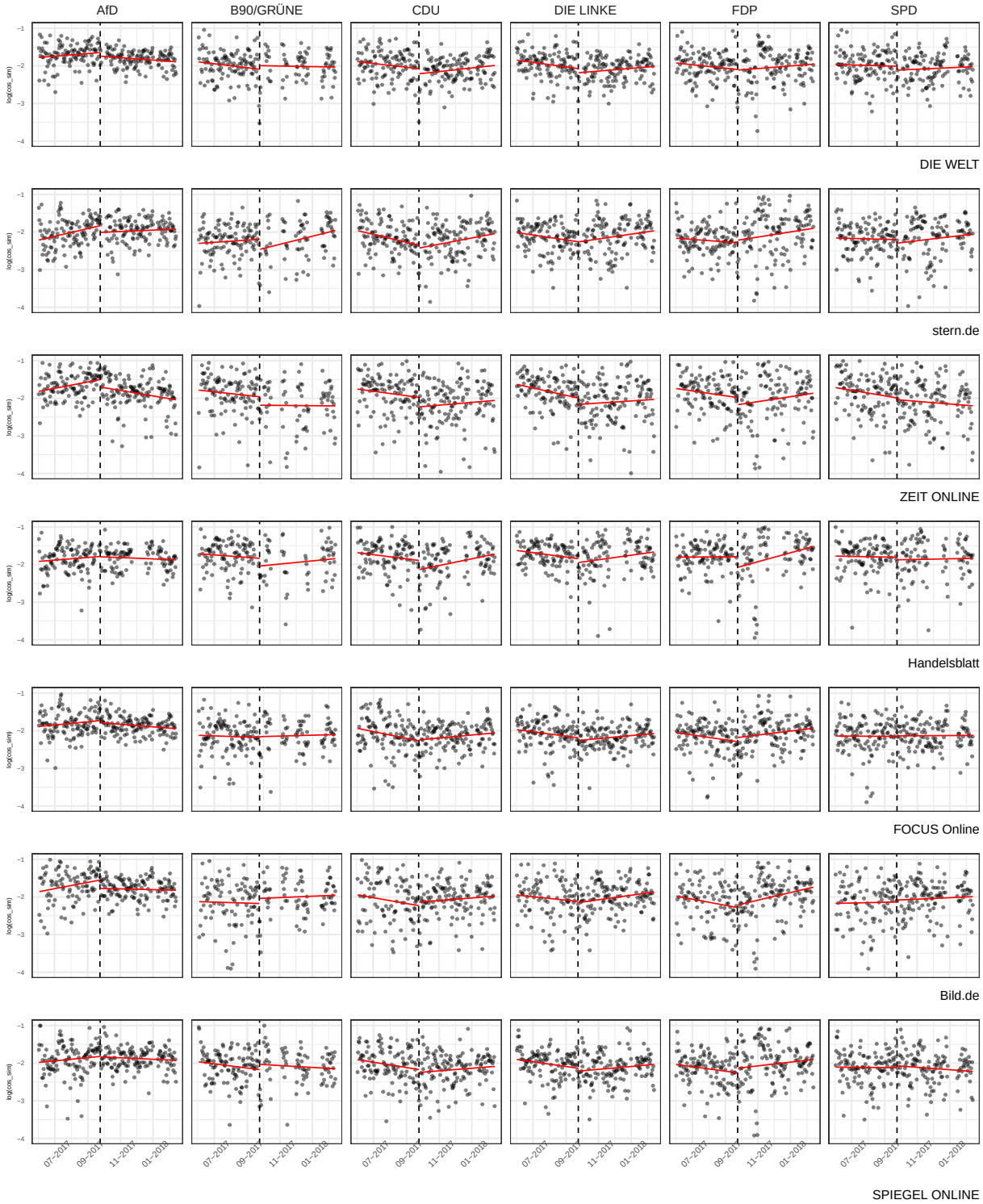


Figure 16: Log of daily mean cosine similarity between newspaper/press articles pairs - with cutoff value

References

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114 (2): 533–75. https://www.jstor.org/stable/2587016.

Bholat, David M., Stephen Hansen, Pedro M. Santos, and Cheryl Schonhardt-Bailey. 2015. "Text Mining for Central Banks." *SSRN Electronic Journal*, June. http://www.academia.edu/13430482/Text_mining_for_central_banks.

Blassnig, Sina, Sven Engesser, Nicole Ernst, and Frank Esser. 2019. "Hitting a Nerve: Populist News Articles Lead to More Frequent and More Populist Reader Comments." *Political Communication*, August, 1–23. https://doi.org/10.1080/10584609.2019.1637980.

Blei, David M., Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (January): 993–1022.

Braun, Michael, and Jon McAuliffe. 2010. "Variational Inference for Large-Scale Models of Discrete Choice." *Journal of the American Statistical Association* 105 (489): 324–35. https://doi.org/10.1198/jasa.2009.tm08030.

Dewenter, Ralf, and Jürgen Rösch. 2014. *Einführung in die neue Ökonomie der Medienmärkte: Eine wettbewerbsökonomische Betrachtung aus Sicht der Theorie der zweiseitigen Märkte*. Springer-Verlag.

Druckman, James N., and Michael Parkin. 2005. "The Impact of Media Bias: How Editorial Slant Affects Voters." *The Journal of Politics* 67 (4): 1030–49. https://doi.org/10.1111/j.1468-2508.2005.00349.x.

Eberl, Jakob-Moritz. 2018. "Lying Press: Three Levels of Perceived Media Bias and Their Relationship with Political Preferences." *Communications*, March. https://doi.org/10.1515/commun-2018-0002.

Eberl, Jakob-Moritz, Hajo G. Boomgaarden, and Markus Wagner. 2017. "One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences." *Communication Research* 44 (8): 1125–48. https://doi.org/10.1177/0093650215614364.

Erosheva, Elena, Stephen Fienberg, and John Lafferty. 2004. "Mixed-Membership Models of Scientific Publications." *Proceedings of the National Academy of Sciences* 101 (June): 5220–27. https://doi.org/10.1073/pnas.0307760101.

Gentzkow, Matthew A., and Jesse M. Shapiro. 2004. "Media, Education and Anti-Americanism in the Muslim World." *Journal of Economic Perspectives* 18 (3): 117–33. https://doi.org/10.1257/0895330042162313.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2017. "Text as Data." Working Paper 23276. National Bureau of Economic Research. https://doi.org/10.3386/w23276.

Griffiths, Thomas L., and Mark Steyvers. 2002. "A Probabilistic Approach to Semantic Representation." *Proceedings of the Annual Meeting of the Cognitive Science Society* 24 (24). https://escholarship.org/uc/item/44x9v7m7.

———. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (June): 5228–35. https://doi.org/10.1073/pnas.0307752101.

Grimmer, Justin, and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21: 267–97.

Groseclose, Tim, and Jeffrey Milyo. 2005. "A Measure of Media Bias." *The Quarterly Journal of Economics* 120 (4): 1191–1237. https://www.jstor.org/stable/25098770.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1): 201–9. https://www.jstor.org/stable/2692190.

Hausman, Catherine, and David S. Rapson. 2018. "Regression Discontinuity in Time: Considerations for Empirical Applications." *Annual Review of Resource Economics* 10 (1): 533–52. https://doi.org/10.1146/annurev-resource-121517-033306.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables. R Package Version 5.2.2.* https://CRAN.R-project.org/package=stargazer.

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

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, The regression discontinuity design: Theory and applications, 142 (2): 615–35. https://doi.org/10.1016/j.jeconom.2007.05.001.

Jacob, Robin, Pei Zhu, Marie-Andrée Somers, and Howard Bloom. 2012. *A Practical Guide to Regression Discontinuity.* MDRC. https://eric.ed.gov/?id=ED565862.

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*, edited by Juliana Raupp and Joachim Klewes, 113–24. Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-83381-5_9.

Klaauw, Wilbert van der. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43 (4): 1249–87. https://www.jstor.org/stable/826967.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355. https://doi.org/10.1257/jel.48.2.281.

Lott, John R., and Kevin A. Hassett. 2014. "Is Newspaper Coverage of Economic Events Politically Biased?" *Public Choice* 160 (1): 65–108. https://doi.org/10.1007/s11127-014-0171-5.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–72. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2145432.2145462.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. "Automatic Evaluation of Topic Coherence." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1857999.1858011.

Newman, Nic, Richard Fletcher, Antonis Kalogeropoulos, David Levy, and Rasmus Kleis Nielsen. 2018. "Reuters Institute Digital News Report 2018." Reuters Institute for the Study of Journalism. http://media.digitalnewsreport.org/wp-content/uploads/2018/06/digital-news-report-2018.pdf?x89475.

Ramage, Daniel, Susan Dumais, and Daniel Liebling. 2010. *Characterizing Microblogs with Topic Models.*

Rehs, Andreas. 2020. "A Structural Topic Model Approach to Scientific Reorientation of Economics and Chemistry After German Reunification." *Scientometrics* 125 (2): 1229–51. https://doi.org/10.1007/s11192-020-03640-0.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515): 988–1003. https://doi.org/10.1080/01621459.2016.1141684.

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.

———. 2016b. "Stm: R Package for Structural Topic Models." *Journal of Statistical Software* forthcoming (January).

Strömbäck, Jesper. 2008. "Four Phases of Mediatization: An Analysis of the Mediatization of Politics." *The International Journal of Press/Politics* 13 (3): 228–46. https://doi.org/10.1177/1940161208319097.

Taddy, Matt. 2012. "On Estimation and Selection for Topic Models." In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*.

Takens, Janet, Wouter Atteveldt, Anita van Hoof, and Jan Kleinnijenhuis. 2013. "Media Logic in Election Campaign Coverage." *European Journal of Communication* 28 (June): 277–93. https://doi.org/10.1177/0267323113478522.

Thistlethwaite, Donald L., and Donald T. Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51 (6): 309–17. https://doi.org/10.1037/h0044319.

Wallach, Hanna M., David M. Mimno, and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." In *Advances in Neural Information Processing Systems 22*, edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, 1973–81. Curran Associates, Inc. http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.pdf.