3_data

Patrick Schulze, Simon Wiegrebe

June 2020

Contents

1 Data

1.1 Data Collection

The current political landscape of Germany consists of six parties: the Christian Democrats (*CDU/CSU*), the Social Democrats (*SPD*), the right-wing *AfD*, the Greens (*Bündnis 90/Die Grünen*), the Left Party (*Die Linke*), and the liberal *FDP*. These parties are represented in the German parliament (*Bundestag*) according to the votes obtained during the 2017 German federal election (*Bundestagswahl*), which took place on September 24, 2017. The legislative period amounts to 4 years, thus ending around September 2021. The parliament currently contains a total of 709 seats. The large majority of members of the German parliament (*(Bundestags-)Abgeordnete*) are assigned a single electoral district (*Wahlkreis*), the remaining ones do not have one.

In order to analyze German political entities based on text data, we constructed a broad database containing personal and Twitter data on an MP level as well as socioeconomic and election data on an electoral-district level, as detailed in the rest of this section. While parts of this database were used in the subsequent topic model analysis, it is also to be used in future text-based analyses regarding German politics.

As a first step in constructing the database, we gathered personal information on all German MPs. Using Python's *BeautifulSoup* web scraping tool as well as a selenium webdriver, we gathered data such as name, party, biographical information, electoral district, and social media accounts from the official parliament website for all of the 709 members of the German parliament during its 19th election period, elected on September 24, 2017.

(Footnote: As of March 30, 2020, the official parliament website contained information on 730 MPs. This is because MPs who resigned or passed away since the beginning of the election period are also listed on the website. These MPs were manually excluded from further analysis.)

An additional source of personal MP-level information would be the MPs' personal homepages. However, after inspecting some of these personal homepages at random, we found that there is no systematic way to scrape all of these websites. Furthermore, hardly any of these websites contain any informative text data comparable to tweets or Facebook posts. As a consequence, we decided against further pursuing this potential source of information. Due to difficulties and recent restrictions when scraping Facebook data, we also discarded Facebook as source of text data and focused solely on Twitter data.

Since information on social media profiles was scarce and incomplete on the official parliament website, we

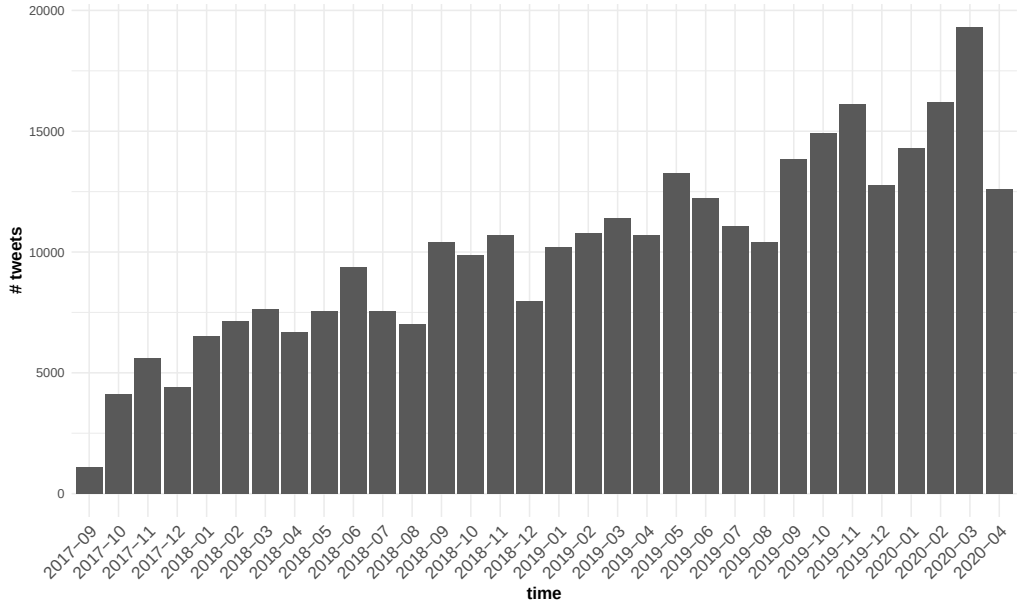
additionally scraped official party homepages of all of the six political parties represented in the current parliament. MPs who did not provide a Twitter account either on the official parliament website or on their party's official homepage were excluded. Using Python's *tweepy* library to access the official Twitter API, we scraped all tweets by German MPs from September 24, 2017 through April 24, 2020, i.e., during a total of 31 months. The *tweepy* library offers a variety of additional features to be extracted apart from the mere tweet texts, such as the number of followers of an account, retweets, or how many times a tweet was like or retweeted. We included the most relevant of these additional features in our database, for use in future analyses.

(Footnote: tweepy restricts the total number of tweets retrievable to 3,200. For those MPs who tweeted more than 3,200 tweets during our period of analysis, the most recent 3,200 tweets were taken into account. However, this only applied to two MPs.)

This initially yielded 342542 tweets from a total of 470 members of parliament.

To complement personal and Twitter data, we also gathered socioeconomic data such as GDP per capita and unemployment rate as well as 2017 election results on an electoral-district level for all of the 299 electoral districts from the official electoral website. After removing the only MP labeled as independent (*fraktionslos*) on the official electoral website as well as -3.23271×10^5 MPs without a specific electoral district assigned to them (for matchability with socioeconomic data), the final dataset counted 450 MPs. Overall, 63% of all 709 MPs were thus included in the analysis. The corresponding total number of tweets amounted to 323740. For those MPs without elecotoral district, electoral district-level socioeconomic variables could potentially be imputed by using state averages or values of nearby / similar districts. However, given that this only applies to -3.23271×10^5 out of the remaining 450 MPs and since imputing covariates would introduce further uncertainty, we decided to exclude those MPs.

The table below shows total monthly tweet frequencies for our period of analysis, September 24, 2017 through April 24, 2020. As can be seen, tweet frequencies - though fluctuating - show an increasing trend over time, peaking at almost 20,000 in March 2020.



Next, data were grouped and tweets were concatenated on a per-user level (thus aggregating tweets across the entire 31 months) as well as on a per-user per-month level, yielding a user-level and a monthly dataset. This means that a document represents the concatenation of *all* of a single MP's tweets for the user-level dataset, while it represents a single MP's *monthly* tweets for the monthly dataset. This also means that MP-level metadata such as personal information and socioeconomic data (through the electoral district matching) can be used as document-level covariates. For the monthly dataset, the temporal component (year and month)

constitutes an additional covariate. Since it is reasonable to assume that the importance of topics varies over time and due to resulting documents being shorter and more easily interpretable, we chose the monthly dataset for further analysis.

(Footnote: For instance, as stated in section 4.2, one topic is about COVID-19, which is clearly a relatively recent topic. The monthly dataset allows for tracing the development of this topic's relevance over time: a flat curve until January 2020 and a sharp increase subsequently. The user-level dataset, on the other hand, would simply assign a low overall proportion to this topic.)

At this point, the data preparation was completed, marking the starting point of the preprocessing required for topic analysis, which is identical for both the user-level and the monthly dataset.

1.2 Data Preprocessing

We used the *quanteda* package within the R programming language for preprocessing. As a first step, we built a quanteda corpus from all documents, already transcribing German umlauts *ä/Ä*, *ö/Ö*, *ü/Ü* as well as German ligature *ß* as *ae/Ae*, *oe/Oe*, *ue/Ue*, and *ss* and removed hyphens. Next, we transformed the text data into a quanteda document-feature matrix (DFM), which essentially tokenizes texts, thereby converting all characters to lowercase. From the DFM, we removed an extensive list of German stopwords, using the stopwords-iso GitHub repository, as well as English stopwords included in the *quanteda* package. Moreover, hashtags, usernames, quantities and units (e.g., *10kg* or *14.15uhr*), interjections (e.g., *aaahhh* or *ufff*), terms containing non-alphanumerical characters, meaningless word stumps (e.g., *innen* from the German female plural declension or *amp*, the remainder left after removing the ampersand sign, *&*) were removed. Terms with less than four characters and terms with a term frequency (overall number of occurrences) below five or with a document frequency (number of documents containing the word) below three were excluded. Finally, we manually removed overly frequent terms that would diminish the distinguishability of topics, such as *bundestag* or *polit* (see *semantic coherence* in section 4.1).

We also performed word-stemming, which means cutting off word endings to remove discrepancies arising purely from declensions or conjugations - of particular importance for the German language. Due to the nature of the German language, the additional gains of lemmatization (which aims at identifying the base form of each word) would only be small as compared to the large increase in complexity, which is why we decided to use stemming only. Another issue when dealing with German language documents are compound words, which are sometimes hyphenated, basically leading to a distinction where semantically there is none. We addressed this issue by removing hyphens in the very beginning of the preprocessing and converting all terms to lowercase, thus "gluing together" compound words; this way, terms like *Bundesregierung* and *Bundes-Regierung* are both transformed into *bundesregierung* (and, after stemming, into *bundesregier*). Finally, automatic segmentation techniques were not necesssary for the German language (Lucas et al. (2015)).

As the result of preprocessing, one empty MP-level document was dropped, so that a total of 10998 MP-level documents were eventually analyzed, each one associated with 90 covariates.

Bibliography

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–77.