draft data section

Simon

May 2020

Introduction

- TBD
- define / translate / introduce German political concepts (Abgeordnete(r), Bundestag, Legislaturperiode, Wahlkreis, Ausschüsse, Parteienlandschaft)

Theoretical Framework

- Introduce topic analysis vocabulary (documents, tokens, . . .)

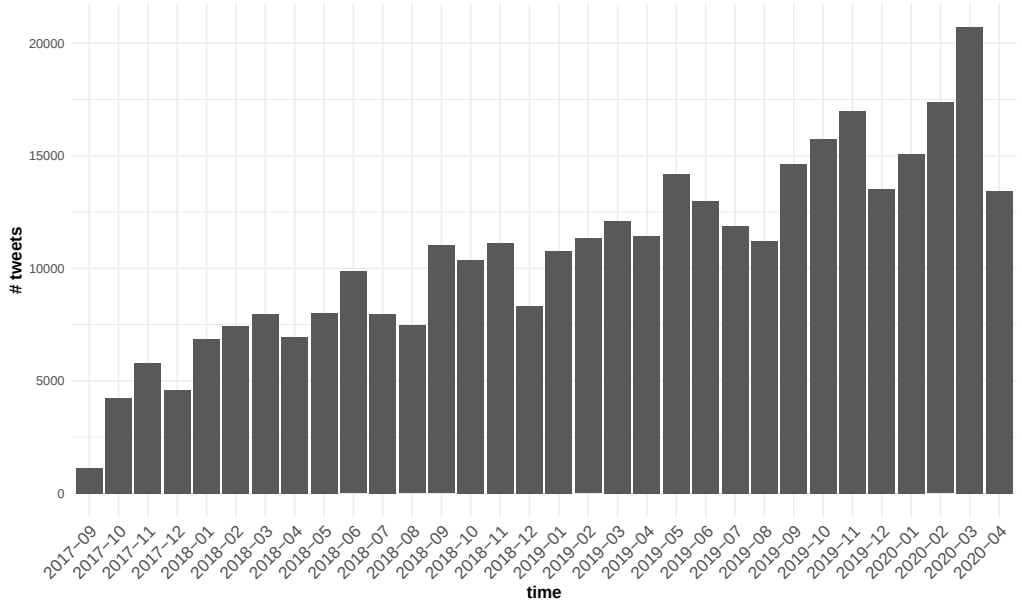
Data

Data (2-3 Seiten): * sentiment analysis: training corpus

As a first step towards applying the STM to German political entities, we constructed a database with personal information about all German MPs. Using Python's BeautifulSoup web scraping tool as well as a selenium webdriver, we gathered data such as name, party, and electoral district 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: MPs who resigned or passed away since this date were also listed on the website and thus included initially; they were manually excluded from further analysis.)

Since information on social media profiles was scarce and incomplete on the official parliament website, we scraped official party homepages for each 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. (Footnote: tweepy restricts the total number of retrievable tweets to 3,200. For those MPs with a larger number of tweets, the most recent 3,200 tweets are taken into account. However, this only affects very few MPs.) This initially yielded 342542 tweets from a total of 470 members of parliament.

To complement personal 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 independent MPs as well as MPs without a specific electoral district assigned to them (for matchability with socioeconomic data), the final dataset counted 450 MPs. The corresponding total number of tweets amounted to 323740. 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 - increase over time, peaking at almost 25,000 in March 2020.



Next, data were grouped and tweets 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 (user-level) monthly dataset. This means that a document represents the concatenation of *all* of a single MP's tweets for the user-level dataset and 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. At this point, the data preparation was completed, thus marking the starting point of the preprocessing required for topic analysis, which is identical for both datasets.

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 convering 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 are excluded. Finally, we manually removed over-frequent terms that would diminish the distinguishability of topics, such as *bundestag* or *polit*.

We also performed German word-stemming, which means cutting off the endings of words to remove discrepancies arising purely due to declensions or conjugations, which is particularly important for the German language. Due to the nature of the German language, the 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 address this issue by removing hyphens in the very beginning 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 are not necesssary for the German language (Lucas et al. (2015)).

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.