

# METHODOLOGY

## Twitter

### *Collected Data*

We collect tweets with the help of the public Twitter Streaming API. It allows us to retrieve data by providing key parameters. In our case a list of hashtags and users. For this project, we carefully selected 239 relevant hashtags and 13,633 users. The user list includes accounts from political parties, politicians, media portals, journalists, bloggers, editors, correspondents, consultants and other important political actors. We do not only collect the tweets from these accounts, but also the mentions and retweets these users get.

For the hashtag list, we selected four types of hashtags:

- German political parties, e.g. #cdu, #fdp, #grünen
- German politics e.g. #finanzpolitik, #groko, #bundestag
- Media sites from all the political orientations e.g. #spiegel, #welt, #bild
- German politicians, e.g. #merkel, #gauland, #steinmeier

As researchers, **we are objective about the results we are presenting**. We do not bias our data to a specific political ideology. We tried to have a balanced list of hashtags and users that represented all the German political parties. We do not include the complete list of hashtags and users in this documentation. For further details, please send us an enquiry.

Note: Twitter only provides a sample of the complete tweets and this biases the data. Nevertheless, we collect a significant number of tweets and hope to have a representative sample that can give plausible insights.

**Important: For all the plots, we only take into consideration tweets written in the German language.**

### *Plots Creation*

**Counter:** The counter on the dashboard shows the number of tweets we have in our database since we started the collection process in March 2017. **The results presented in the Dashboard are live.**

**URLs:** We preprocess all the URLs in the tweets by extracting the endpoint. For the dashboard, we do not consider URLs from other tweets, Facebook posts, YouTube videos, Instagram photos and Google/Bing searches. It is possible that sometimes URLs not related to media appear, this could be caused by spambots.

**Supporters:** Since March 2017, we have collected retweets from the main accounts of the German political parties. With this data, we can observe which users retweet the political accounts. From these users, we define a *supporter* as a user that has retweeted at least five times a political party. One user can be a supporter of more than one party. We do not explicitly collect all the tweets from these users. We obtain data only if they tweeted something that is politically relevant. For each party, we extract the hashtags used by its supporters in the last 24 hours.

*Spider plot:* To construct this plot, we use the information from the top ten used hashtags and the hashtags from the party supporters. We count the number of times each hashtag has been used by the supporters of each party in the last 24 hours. We normalize by the number of tweets collected for each party. We are interested in showing **percentages** and not absolute numbers. This allows us to compare the importance of a topic for the supporters of the parties. Each color-shape on the plot corresponds to a German party and it intersects the lines corresponding to the hashtags. For one hashtag, the dots closer to the outer circle tell us that the supporters of these parties are proportionally sharing the hashtag more often than the supporters of the other parties. If all the dots for one hashtag are close to the center, the supporters of all parties are discussing this hashtag in similar quantities.

### *Bavarian Dashboard*

For this dashboard, we additionally collect tweets with a different list of hashtags and users. In this case, the list is smaller than the general dashboard and we therefore include it here:

#### Users

@AfD\_Bayern, @Markus\_Soeder, @Gruene\_Bayern, @KathaSchulze, @LudwigHartmann, @LINKE\_Bayern, @AtesGuerpinar, @fdpbay, @realMartinHagen, @BayernSPD, @NataschaKohnen, @FREIEWAEHLER\_BV, @HubertAiwanger

#### Hashtags

Aiwanger, ausspekuliert, Bayern, BayernimHerzen, BesserWohnenmitKohnen, DasBestefuerBayern, fragkatharina, freie, Freiewaehler, freieWähler, frischesBayern, gruenebayern, Gürpinar, Hagen, jazuBayern, kathaunterwegs, Kohnen, KohnenPLUS, Itw18, Itw2018, LtwBayern, Itwby, LTWBY18, MehrfürdieMehrheit, MitunsdieZukunft, münchenistbunt, nopag, pag, PolitikmitCharakter, schulze, Seehofer, söder, soedermachts, Spurwechsel, ZukunftimKopf

Note: For the URLs in Bavaria, we explicitly removed URLs from the FC Bayern page. As we follow #Bayern, we get many tweets related to the football club.

## Media

### *Collected Data*

We collected online news media articles related to German politics with the help of the RSS feeds of the news media websites and the Python package BeautifulSoup. We retrieve data on articles in real time, including their title and full text. We filtered news that did not relate to German politics. For this project, we carefully selected the following 40 German online media sites:

- Ach Gut
- Allgemeine Zeitung
- Bild
- Bayerischer Rundfunk
- Cicero Online
- Compact-Magazin
- Der Spiegel
- Deutschlandfunk
- Die Presse
- Deutsche Welle
- Epochtimes
- Frankfurter Allgemeine Zeitung
- Focus
- Frankfurter Rundschau
- Freie Presse
- der Freitag
- Handelsblatt
- Huffington Post
- Journalistenwatch
- Junge Freiheit
- Jungle World
- Manager Magazin
- Mitteldeutscher Rundfunk
- N-tv
- Neues Deutschland
- Rheinische Post
- RT deutsch
- Stern
- Die Süddeutsche Zeitung
- Die Tagespost
- Der Tagesspiegel
- Tagesschau
- Die Tageszeitung
- Tichys Einblick
- Die Welt
- Wirtschaftswoche
- Die Zeit

Our selection procedure was based on three guidelines:

1. We wanted to cover media from all the different political orientations
2. We looked at the media included in the top sites of online traffic in Germany (Alexa)
3. We only considered the news media that an RSS feed. There are relatively few big media sources that do not have this service, for example, Reuters Germany.

**Important Note 1:** We tried to stay unbiased and selected news media from different political orientations. If you, the reader, think that some media sources should be added please send us an email to [contact@political-dashboard.com](mailto:contact@political-dashboard.com), and we will consider it.

**Important Note 2:** The full text from each article is under strict data protection, and should not be reproduced without the consent of the news media. The results we present are only aggregates and analysis on the data. We are not recreating the content in any way.

### Plots Creation

**Counter:** The counter on the dashboard shows the number of articles we have in our database since we started the collection process in January 2019. **The results presented in the Dashboard are live.**

**News Topics:** To automatically understand the relevant topics in the news we use topic modeling with LDA (Latent Dirichlet Allocation). It is a machine learning method to find topics in documents. Moreover, we use optimization algorithms to find the optimal number of topics. We used the `tmtoolkit` Python package for this.

The plot considers only the news articles in the last 24 hours. The articles have a given probability of belonging to each one of the topics. By summing the probabilities of all articles per topic, we calculate the importance of the topic. The percentage then shows the percentage of news articles that were reporting on the given topic.

For each topic, we present the top 8 words. The relevance of each word is calculated according to this formula (with  $\lambda$  0.3):

$$r(w, k | \lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right)$$

which is described on this paper:

<https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf>

**Spider plot:** To construct this plot, we use the information from the top seven topics from the topics plot. First, we categorize news sources in the database according to their political orientation: left, left-center, center, center-right and right. We then count the number of articles that belong to each group of political orientation. Finally, we calculate the relevance of each topic by summing the probabilities of each subset of articles. The topic proportions are not normalized as we report only seven topics and the algorithm calculates the optimal number of topics, which is normally larger than seven.

On the plot, we only present the top word for each topic. Each color-shape corresponds to media with a certain political orientation and it intersects the lines corresponding to the topics. For one topic, the dots closer to the outer circle tell us that the news media belonging to a political orientation are reporting more on this topic as other media with other orientations. If all the dots for one topic are close to the center, all the news media are discussing this topic in similar quantities.

**Party Attention:** We count the number of times the political parties' names were mentioned in the articles. For this, we analyze the text of the news articles.

**Facebook Shares:** To obtain the number of Facebook shares for an article, we use the service Shared Count (<https://www.sharedcount.com/>). The service has a limit per day for 10,000 links. For this reason, we are only updating the database every two hours. Moreover, we limit ourselves to obtaining the shares only for articles that were published in the last 5 days.

# Facebook

## *Collected Data*

We show the information of 102 Facebook political pages. These include the main German parties' pages and their regional pages (For every one of the 16 German states). We do not collect data on the city level political pages.

We obtain the post data from the **Crowdtangle** service (<https://www.crowdtangle.com/>), which was obtained thanks to the Social Science 1 project between the Political Data Science professorship and Facebook. The data does not include any personal data, such as users and their comments. We therefore only inform on the number of interactions and the contents of the created posts.

## Political Ads

We collect the German political ads on Facebook using its ad archive API. The analyses are done only on the active ads at the moment. For more information on the public API here: <https://newsroom.fb.com/news/2018/08/introducing-the-ad-archive-api/>

The political ads dashboard was developed for the research project "*Political advertisement and microtargeting - The importance of social media during the European elections in Germany*" in cooperation with the media authorities of North Rhine-Westphalia, Bavaria, Berlin-Brandenburg and Rheinland-Pfalz. We are thankful for this cooperation with the [Landesanstalt für Medien NRW](#), [Bayerischen Landeszentrale für neue Medien \(BLM\)](#), der [Medienanstalt Berlin-Brandenburg \(mabb\)](#) and the [medienanstalt rlp](#)

## *Plots Creation*

**Facebook Interaction Counter:** The counter takes into consideration all the posts from all the considered pages in the last seven days and quantifies the number of interactions. Interactions include likes, shares, comments, and the other five Facebook smileys.

**Mood Plot:** We run sentiment analysis on the texts originated in the posts. For this, we use the Spacy library and the SentiWS dictionary. We assign each word to a sentiment score, and then the scores are averaged to give the final result per party.

**Reactions Spider Plot:** The same idea applies as for the previous two spider plots. The number of reactions is first normalized by the total number of reactions for each political page to obtain percentages per party. Afterward, the count is normalized to the total number of reactions from all the parties. In this way, we account for the fact that some pages are more active than others.

**Wordcloud:** For this plot, we obtain the text in the posts from the pages we follow and extract the nouns. We then select the 20 most common nouns for each party. The nouns may be repeated and, therefore, less than 140 words appear on the word cloud. The color of the word corresponds to the party which used the word the most in their posts, and the size corresponds as a sum of the contribution of each party to the given words. The weight of a word per party is normalized by the number of nouns used in

each group of pages belonging to a political party. We take with this into consideration to tackle the problem of having pages posting more frequently than others. (See Post plot)

**Ads Counter:** The counter shows the number of active ads on Facebook. **The results presented in the Dashboard are live.**

**Targeting Map:** Every ad has a region distribution where the advertiser can decide which federal state should be targeted. For the main Facebook page from each political party and their regional pages, we collect the ads and average the percentages of the regional targeting. Political ads from regional political pages are not included. Therefore, this only represents federal targeting strategies. The intensity of the color on the map represents the intensity for each federal state.

## Licences

For the creation of the tables used in the different pages, we used the DataTables plug-in under the MIT license (<https://datatables.net/license/mit>)