Facebook Data Scandal: English and German Text Analysis

Tiffany Cheng

May 7, 2018

Introduction

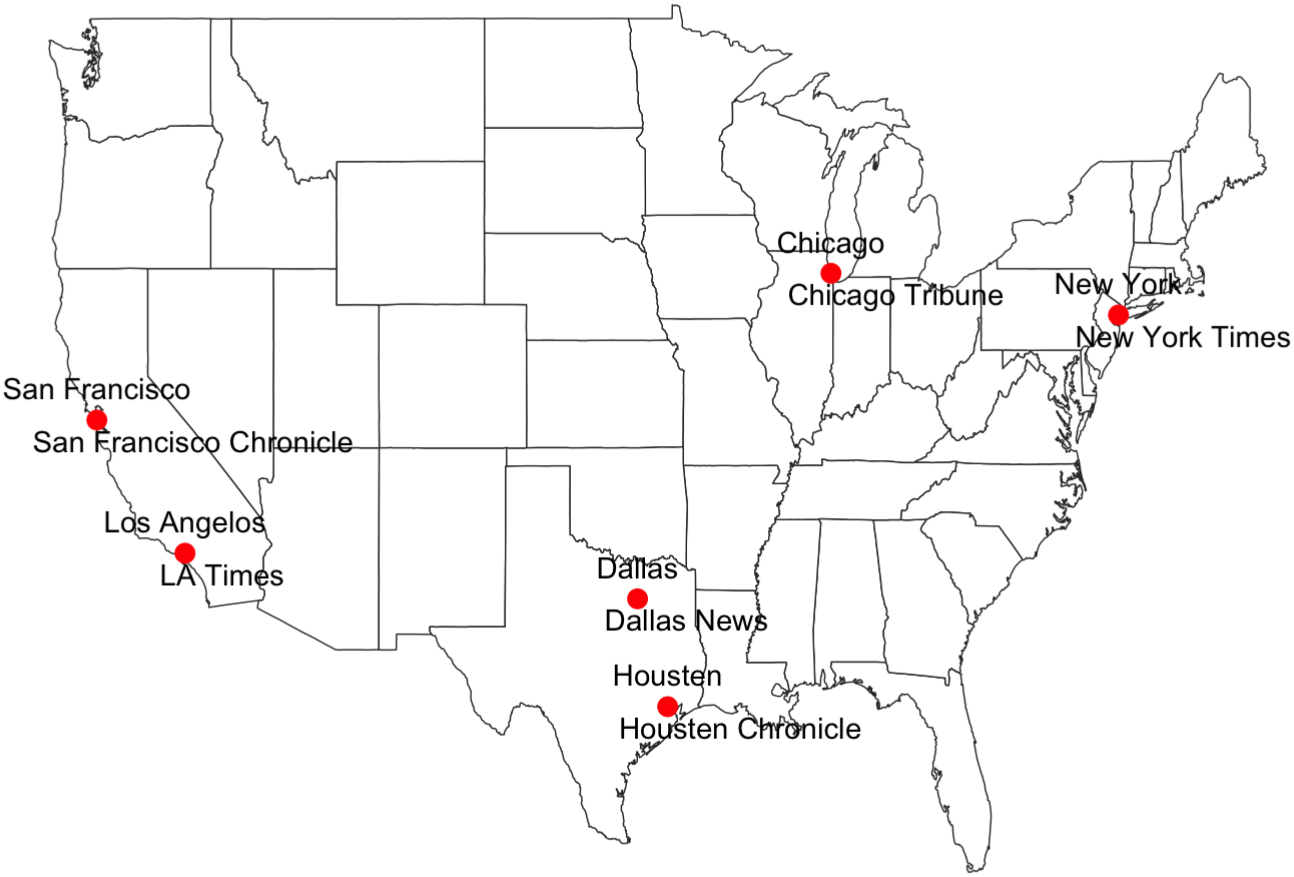
Inspired by the two times I've studied abroad in Germany, I decided that I wanted to apply and extend my English text analysis skills to the German language. From my experience, I found that Germans are very conscious of their personal data and take great precautions to keep it as secure as possible. When news of the Facebook and Cambridge Analytica data scandal of late March to early April 2018 broke out, I thought that this would be the perfect opportunity to investigate and compare reactions between English and German texts.

Data Collection Methods

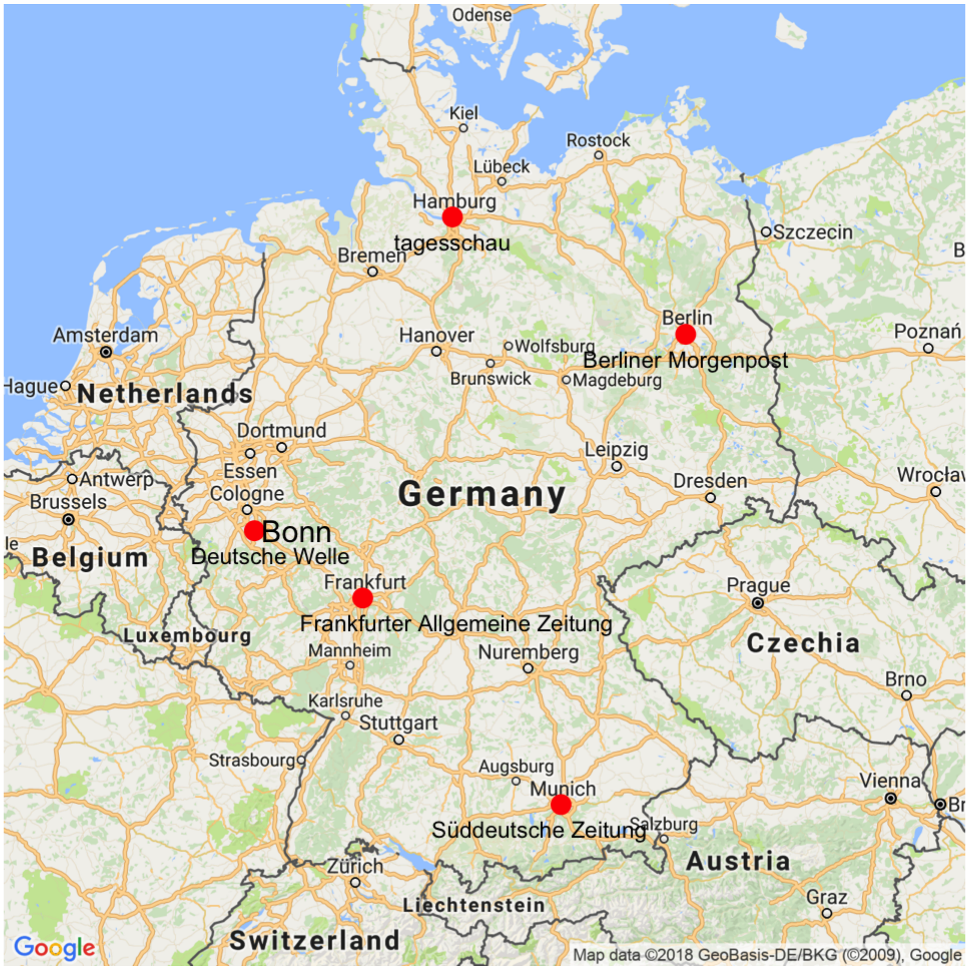
This study includes the use of Twitter tweets as well as text scraped from news articles found online. The Twitter API was used to search for tweets both in English and in German with the following hashtags: `#MarkZuckerberg`, `#Zuckerberg`, `#FacebookDataBreach`, and `#Facebook` together with the word `Zuckerberg` (adding Zuckerberg filtered out any unrelated tweets). There was an additional hashtag used for German, which was `#Datenskandal` (meaning data scandal) due to the small sample of German tweets. Retweets were filtered out during the search process. These tweets were used as real-time data and were collected in the time frame from April 10, 2018 to April 21, 2018. For news articles, a selection of five news companies were chosen throughout the United States and Germany to get a range of different journalistic styles and opinions. In total, five articles were chosen from each company. There was a sixth news company chosen in the United States because one company (The Houston Chronicle) only published three articles. Therefore, the Dallas News was chosen to provide the remaining two articles. The news articles serve as historical data and they were published between the dates of March 23, 2018 and April 11, 2018.

Location of News Companies

The names of the six US news companies are as follows: New York Times, Chicago Tribune, Dallas News, Houston Chronicle, LA Times, and San Francisco Chronicle. Their locations on the map are seen below.



The five German news companies chosen are as follows: tagesschau, Berliner Morgenpost, Deutsche Welle, Frankfurter Allgemeine Zeitung, and Süddeutsche Zeitung. tagesschau is the most popular source of news in Germany and Deutsche Welle is a public broadcaster and features channels in different languages. The Berliner Morgenpost, Frankfurter Allgemeine Zeitung, and Süddeutsche Zeitung are popular regional newspapers, although they are read widely in Germany. The locations of the companies are located on the map below.

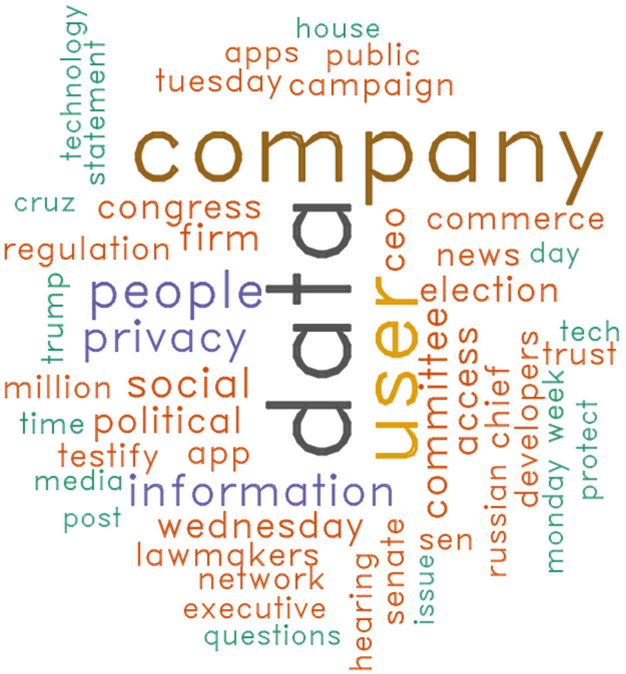


Word Frequency Analysis for News Articles

To get acquainted with the text, the most frequently used words in news articles were counted and can be seen in the word clouds below. As a note, words of the same color appear roughly the same amount of times and bigger words appear more often than smaller words do in the articles.

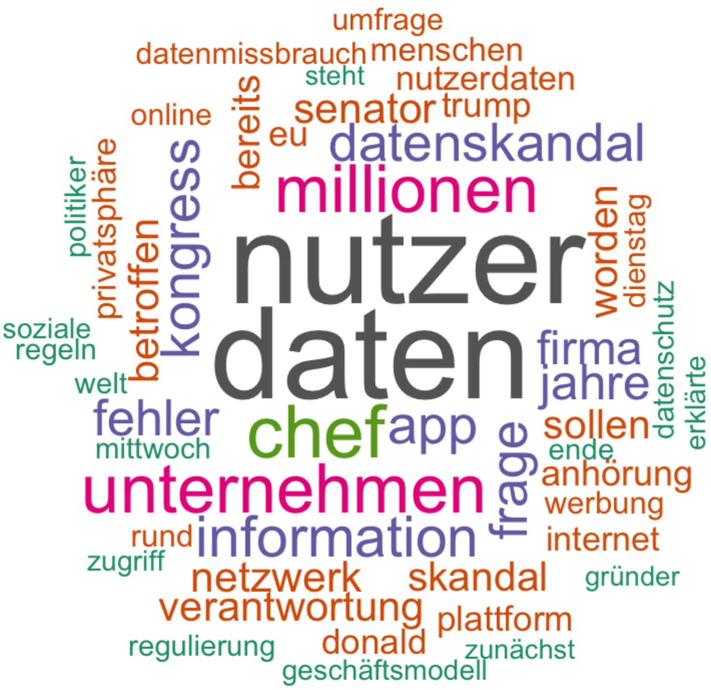
English

From this word cloud, it is evident that the words “data”, “company”, and “user” appear the most. This is followed by words related to the upcoming testimony in front of Congress (“tuesday”, “wednesday”, “senate”, “testify”, “house”) as well as the broader implication of the data breach (“million”, “protect”, “information”, “personal”, “russian”).



German

The most frequently used words in German news were counted and reproduced below. There is a decent among of overlap when compared to the English news in the words used and their frequency. Words like “nutzer”, “daten”, and “unternehmen” (“user”, “data”, “company”) appear the most, which matches with the English news. Interesting to note is that words such as “privatsphäre”, “datenmissbrauch”, “skandal”, “fehler”, and “regeln” (“privacy”, “data misuse”, “scandal”, “mistake”, “rules”) are used to discuss the implication of this event, which are stronger and more negative than the words used in English news articles.

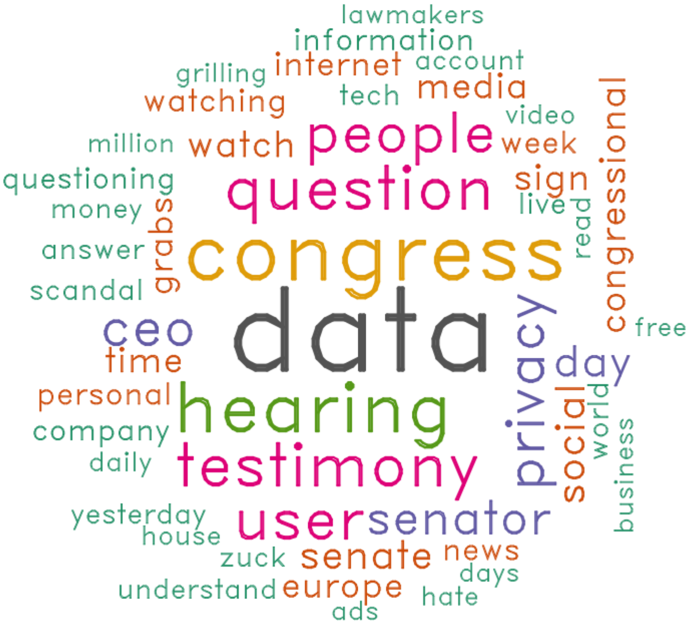


Overall, there is not a significant difference in word choice, as many of the words overlap between the two languages. However, the German words are more descriptive in regards to the consequences of this news.

Word Frequency Analysis for Twitter

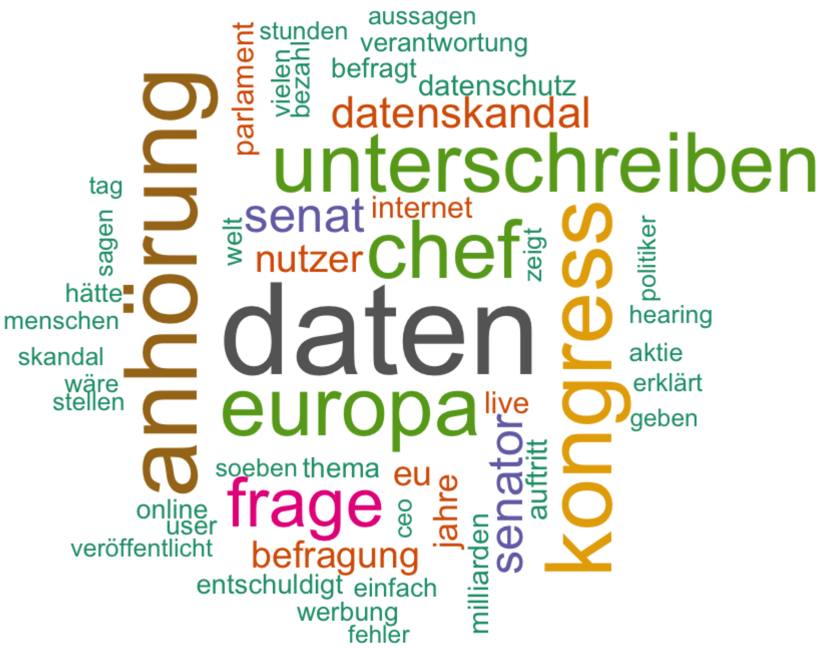
English

Since Twitter tweets were used as real-time data, most of the words pertain to the testimony, such as “hearing”, “watch”, and “question.” This differs from the words used in the news articles because words like “regulation”, “campaign”, and “russian” have disappeared. Interesting to note is that “zuck” has appeared in the tweets due to the informal manner in which tweets are written. Some other words including “hate” and “grilling” have appeared, which are also more informal words.



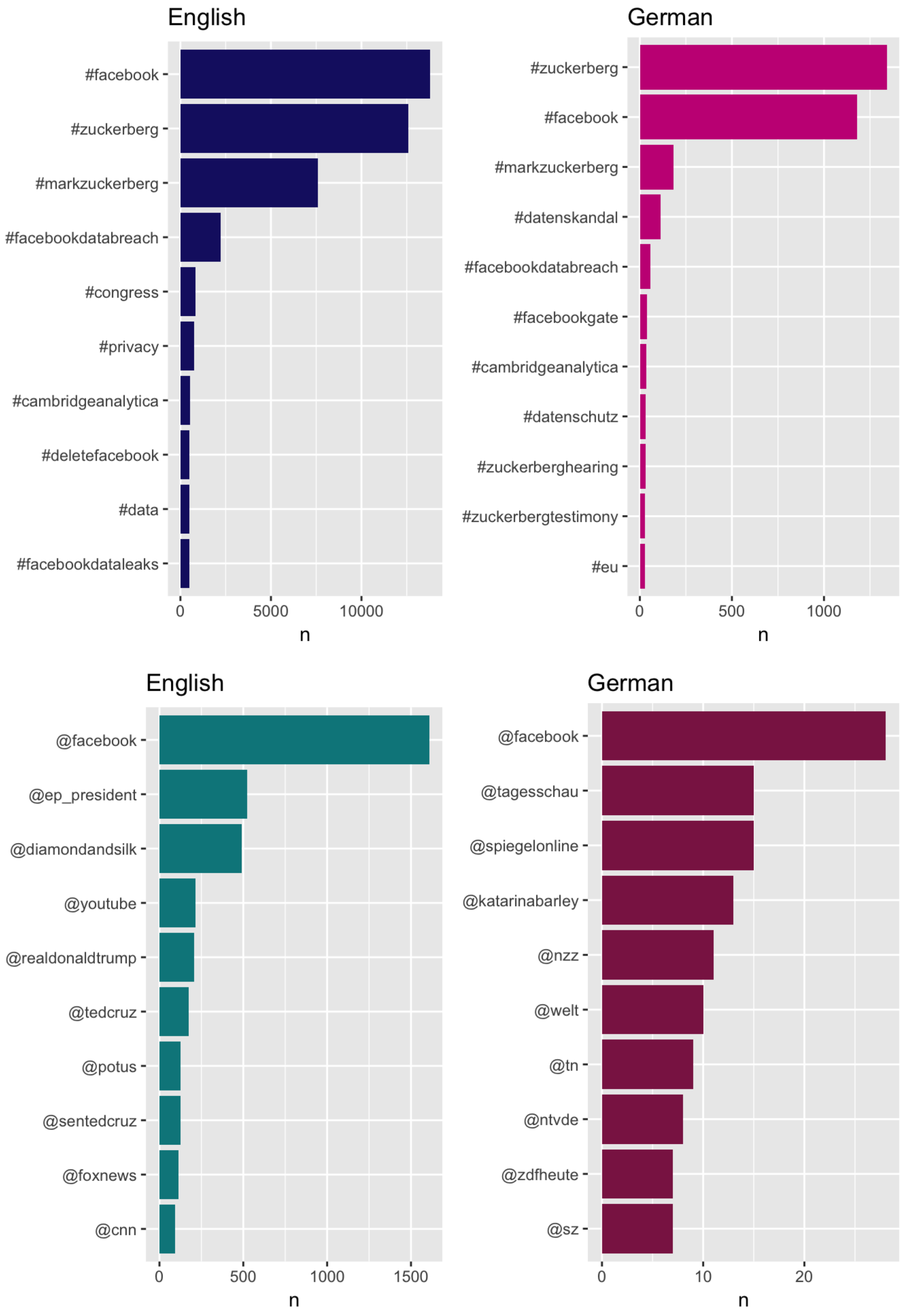
German

The German tweets include verbs in the subjunctive tense, such as “hätte” and “wäre”, meaning “would have” and “would be.” This might indicate that the tweets are mocking Zuckerberg by saying he could have done something or something would have been better. In addition, “fehler”, “entschuldigt”, and “datenschutz” come up, meaning “mistake”, “apologizes”, and “data privacy.” These words did not appear in the English tweets. It is also interesting to note that “europa” (“europe”) was quite frequent. This could be because the European Parliament also wanted Zuckerberg to testify there.



Overall, the word choice remained much the same between news articles and tweets, with the exception of some informal words.

The two following graphs were made to compare hashtags and mentions used between English and German tweets. There is not a significant difference in hashtags used, except some German tweets included the hashtag `#eu`, possibly referring to the European Parliament also wanting Mark Zuckerberg to testify in Europe. However, there is a noticeable difference in mentions. German tweets overwhelmingly mention Twitter pages of news companies such as `@tagesschau`, `@spiegelonline`, and `@welt`. English tweets, on the other hand, mention the Twitter pages of politicians such as the President (`@potus` and `@realdonaldtrump`) and Senator Ted Cruz (`@tedcruz` and `@sentedcruz`).


English
#facebook
#zuckerberg
#markzuckerberg
#facebookdatabreach
#congress
#privacy
#cambridgeanalytica
#deletefacebook
#data
#facebookdataleaks
0
5000
10000
n
German
#zuckerberg
#facebook
#markzuckerberg
#datenskandal
#facebookdatabreach
#facebookgate
#cambridgeanalytica
#datenschutz
#zuckerberghearing
#zuckerbergtestimony
#eu
0
500
1000
n
English
@facebook
@ep_president
@diamondandsilk
@youtube
@realdonaldtrump
@tedcruz
@potus
@sentedcruz
@foxnews
@cnn
0
500
1000
1500
n
German
@facebook
@tagesschau
@spiegelonline
@katarinabarley
@nzz
@welt
@tn
@ntvde
@zdfheute
@sz
0
10
20
n

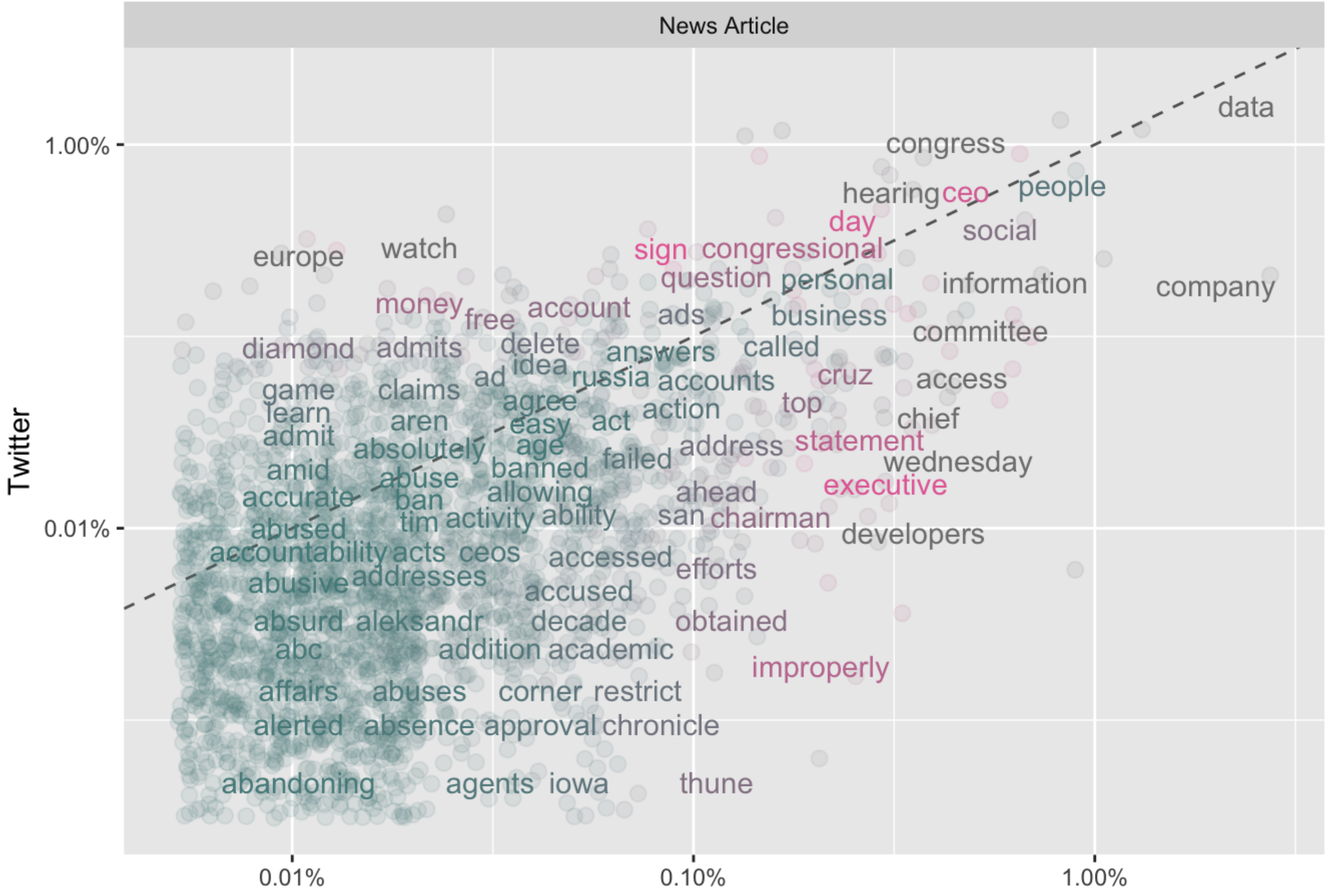
Comparisons Among the Same Language

The relative frequency of words used in tweets and news articles are plotted together to compare usage. Words close to the dashed line indicate that similar proportions of the word was used, while those farther away from the line and that have a pinkish color represent words that differ more in usage.

English

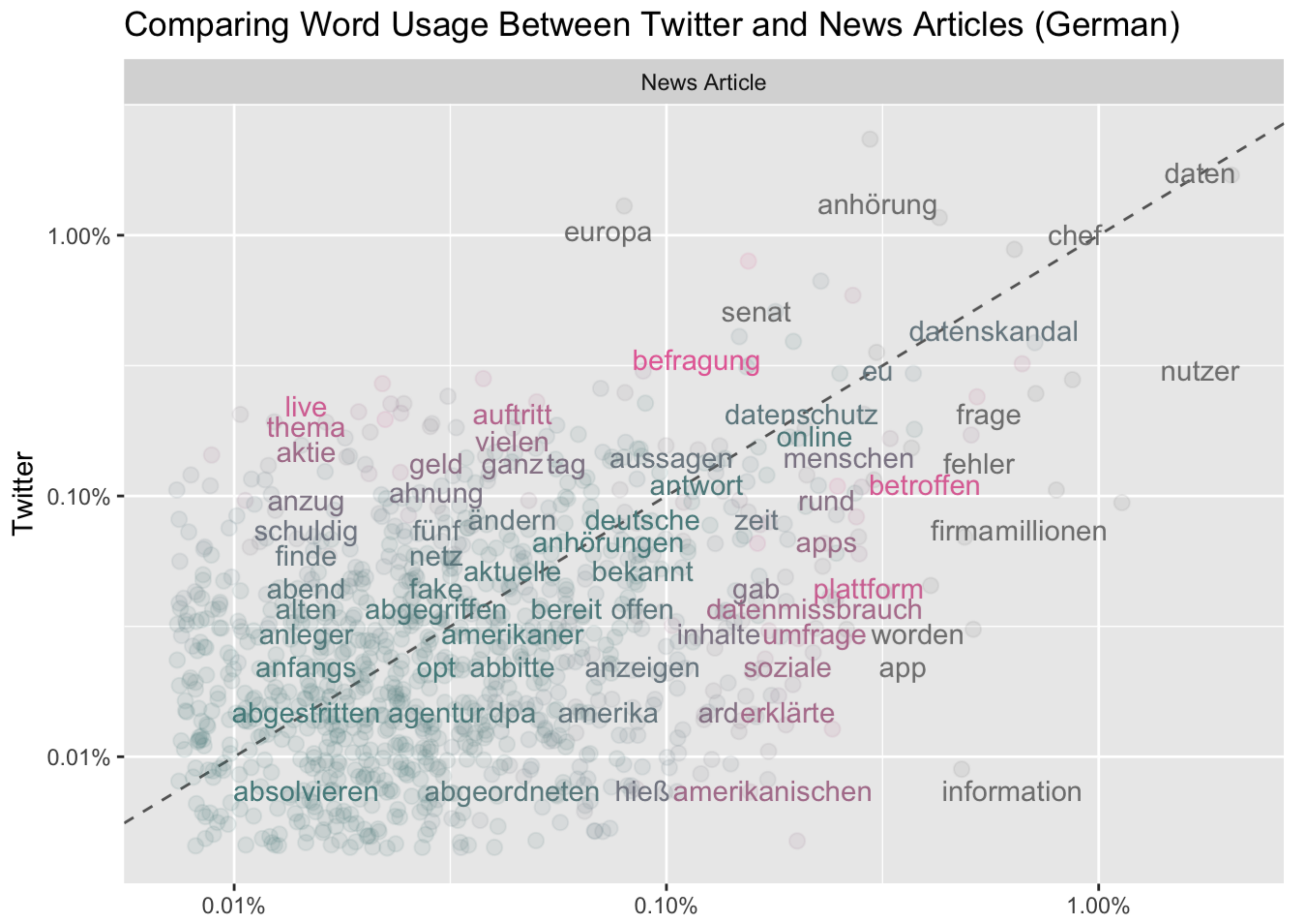
Words like “improperly”, “presidential”, and “executive” appear more in news articles than in tweets. On Twitter, words such as “europe”, “watch”, and “admits” are used more frequently. To assess correlation between the two types of text, a correlation t-test was carried out. The resulting correlation is 0.6538692 with a p-value of 1.585974710^{-305}. This leads to the conclusion that there is a statistically significant correlation between the words used. This supports the analysis above with the word counts because the words used remained roughly the same between news articles and tweets.

Comparing Word Usage Between Twitter and News Articles (English)



German

Words like “amerikanischen”, “datenmissbrauch”, and “information” (“american”, “data misuse”, “information”) appear more in news articles than in tweets. On Twitter, words such as “live”, “aktie”, and “befragung” (“live”, “stock”, “questioning”) appear more. This makes sense because people tweeting at the same time they are watching the testimony and sharing their thoughts. Once again, a correlation t-test was carried out and the resulting correlation is 0.594824 with a p-value of 1.585974710^{-305}. The conclusion is that there is a statistically significant correlation between the words used in news articles and in tweets, which confirms the analysis presented in the section above.



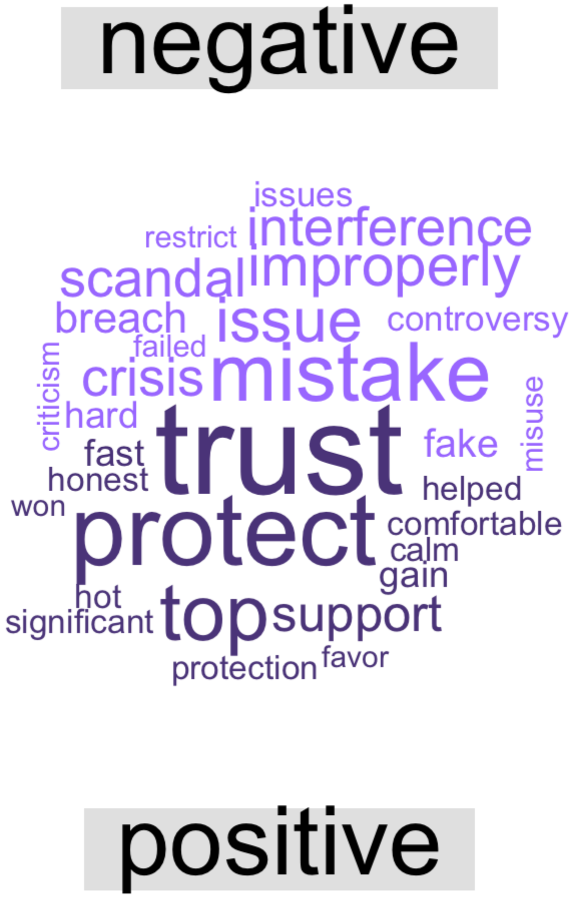
Sentiment Analysis for News Articles

English

The `bing` sentiments lexicon was used as well as the `afinn` sentiments lexicon. The `bing` lexicon groups words into binary categories of positive and negative and the `afinn` sentiments lexicon assigns a score to each word between -5 to 5 with negative scores representing negative sentiments and positive scores representing positive sentiments.

Using the `bing` sentiment lexicon results in the following table and the next figure representing the top 30 most commonly used negative and positive words. The table indicates that the overall sentiment is negative for English news articles, meaning that more negatively associated words were used than positively associated words. It is important to note that the `bing` sentiment lexicon categorizes “trump” as a positive sentiment, however, that word takes on a different meaning in this context, so it was anti-joined out of this particular data set. The `afinn` sentiment lexicon results in a sentiment score of -262.

negative	positive	sentiment
606	396	-210

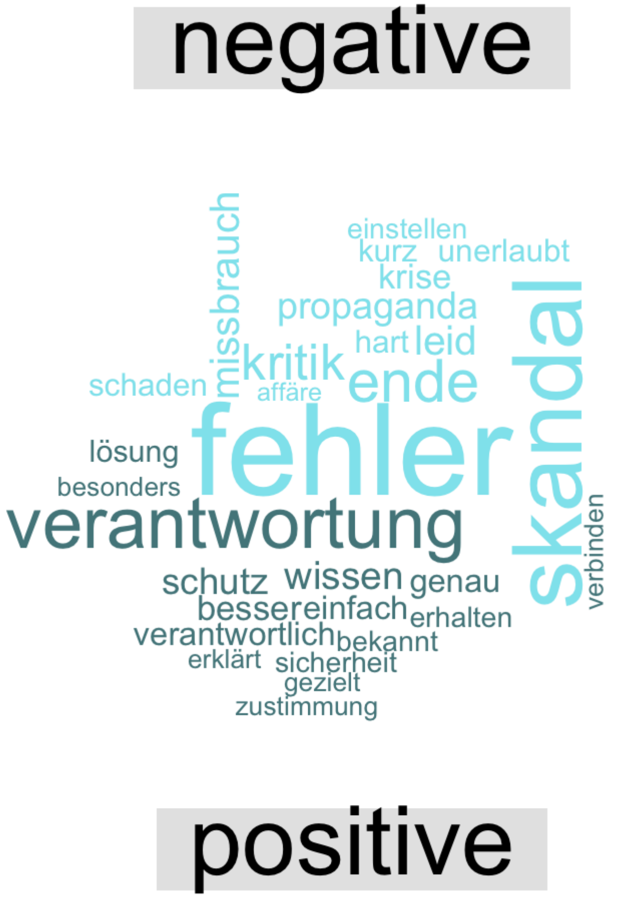


German

Neither the `bing` sentiment lexicon nor the `afinn` sentiment lexicon are available. Therefore, the SentimentWortschatz (or SentiWS) from the University of Leipzig was used. Words in this sentiment collection are scored from -1 to 1 with negative scores representing negative sentiment and positive scores representing positive sentiment. What is particularly helpful about this collection is that it includes inflected words next to their original stem words and scores them as well, which is incredibly useful when doing text analysis on a heavily inflected language like German. Lastly, the original collection did not sort the words into binary groups of positive and negative. That part was coded in when the sentiment text files were read into R.

The following table summarizes the count for negative and positive words found and the overall sentiment is positive, unlike the English news, which was negative. The graphic that follows plots the top 30 positive and negative sentiments. Using sentiment assigned to each word results in a sentiment score of -56.43515. This seems to contradict the results from the binary categories, but SentiWS could assign a score that is larger in magnitude to the negative words than the positive words, which could result in in this negative score. This could possible be an indication that the words used in the German news articles are stronger, thereby warranting a "larger" negative score.

negative	positive	sentiment
247	368	121



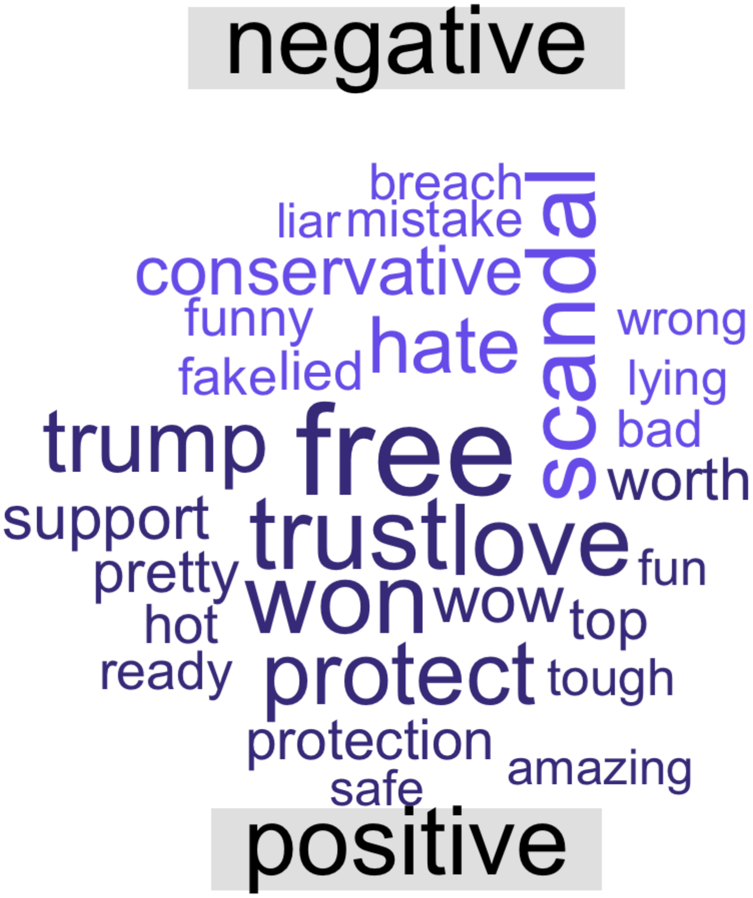
In comparing the two graphs for positive and negative sentiments, the languages have some overlap in words used. For negative words, “mistake”/“fehler”, “misuse”/“missbrauch”, “scandal”/“skandal”, “criticism”/“kritik”, and “crisis”/“krise” are common between the languages. For positive words, only “protect”/“schutz” are the same and German news write more about “responsibility”, “solution”, and “safety” (“verantwortung”, “lösung”, “sicherheit”).

Sentiment Analysis for Twitter Tweets

English

The `bing` sentiments lexicon was used as well as the `afinn` sentiments lexicon once again for the following tweet analysis. The `bing` sentiment lexicon results in the following table and following graph of top 30 words for tweets. Overall, the sentiment for the tweets is very negative. Sometimes people post on social media when they want to complain about something and express their thoughts in uncensored ways and this may just be reflective of those actions. Using the `afinn` sentiment lexicon results in a sentiment score of -8176, which is also very negative and reflects the results of the `bing` lexicon.

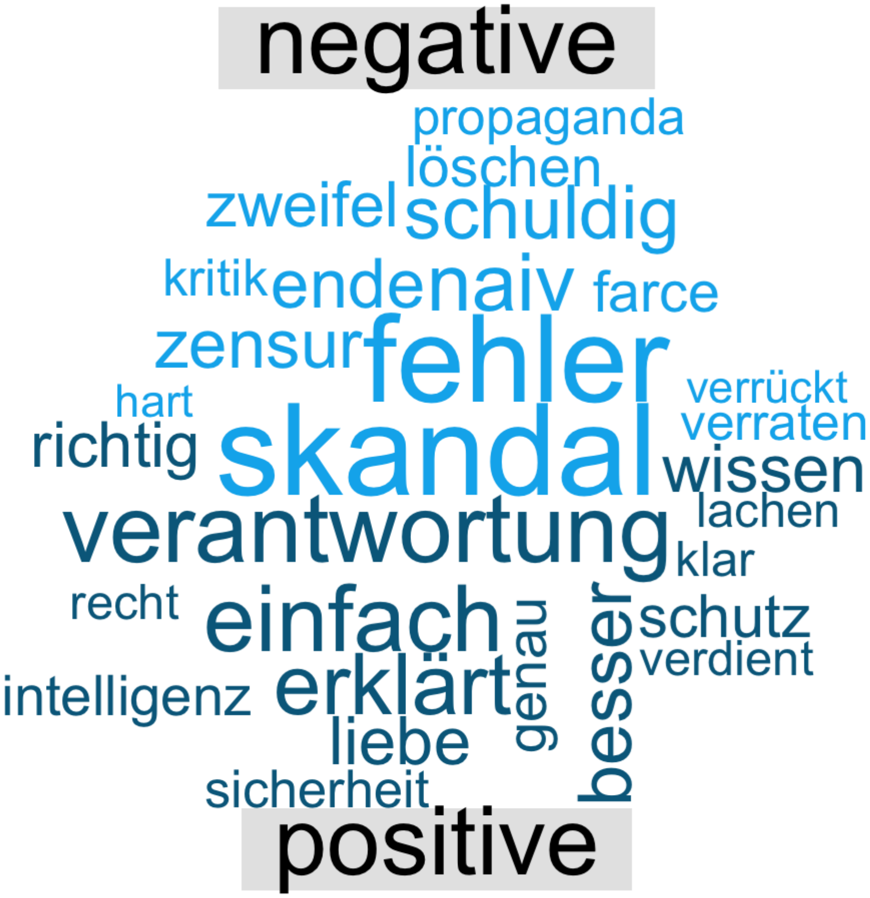
negative	positive	sentiment
13622	7264	-6358



German

The SentiWS was used once again. German tweets were a little harder to gather than English tweets. However, the sentiment is still generally positive. Lastly, using the sentiment score assigned to each word results in a score of -53.16455.

negative	positive	sentiment
247	368	121



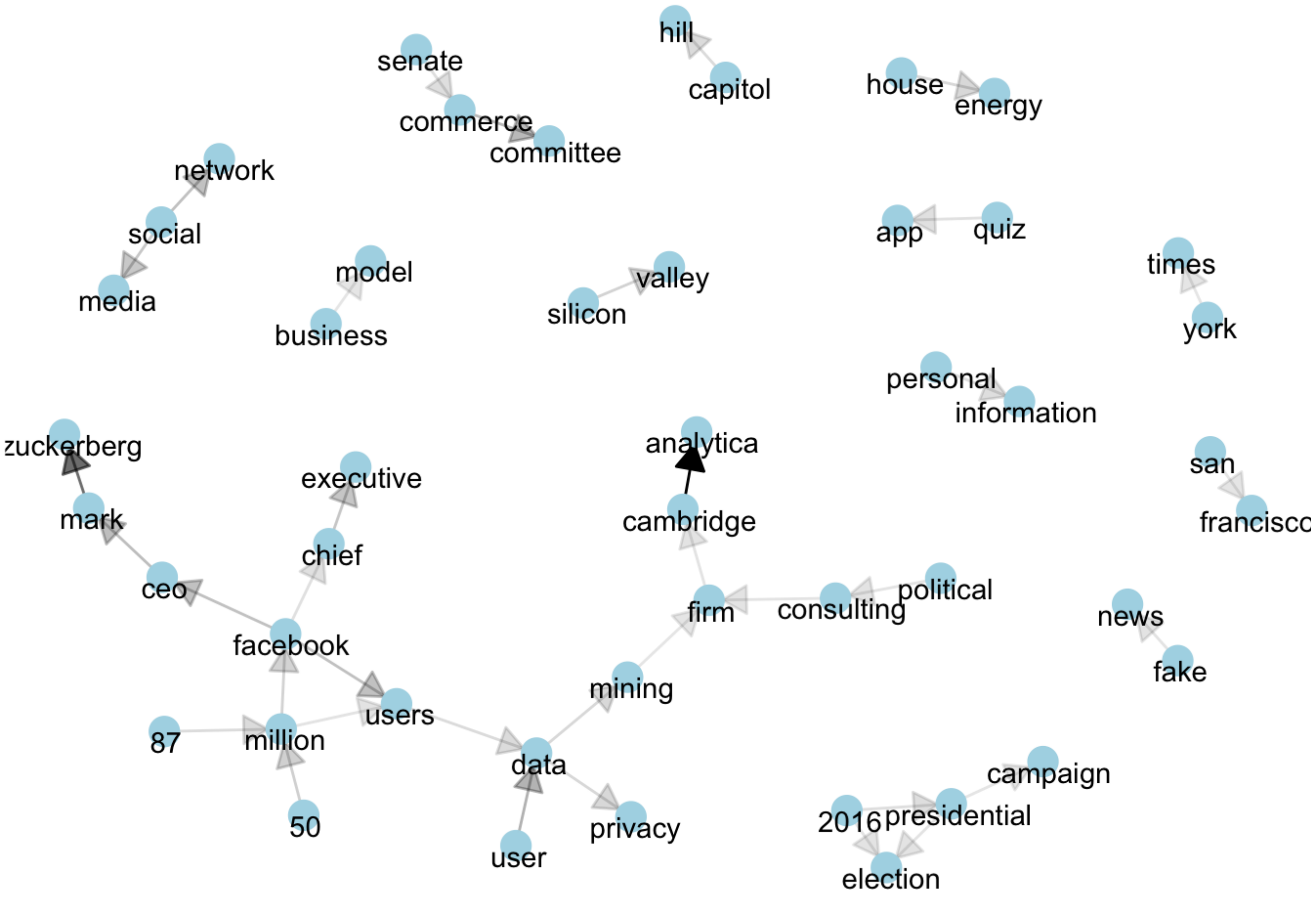
There is not as much overlap of words used among the tweets as used in the news articles. The overlapping negative words are, “mistake”/“fehler”, “scandal”/“skandal”, and “bad”/“schlecht”, and the overlapping positive words are, “protect”/“schutz” and “love”/“liebe.” Some other German negative words include “zweifel”, “verraten”, and “at fault” (“doubt”, “betray”, “schuldig”) and some German positive words include “verantwortung”, “einfach”, and “erklärt” (“responsibility”, “easy”, “explains”).

n-grams

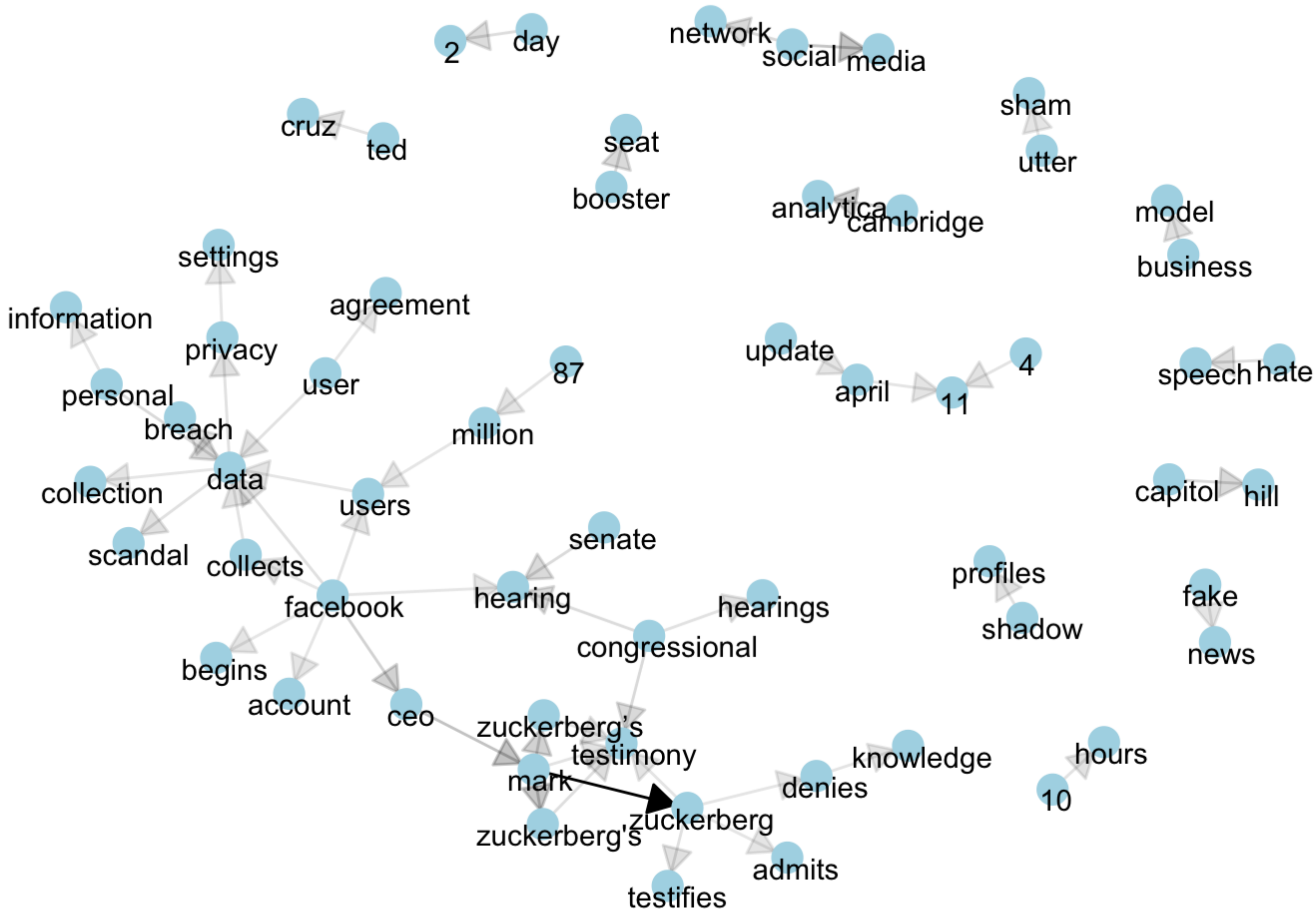
Bigrams were plotted for news articles and tweets to see the relationships between common pairs of words. The shade of the arrows symbolize the frequency of that pair of words with darker arrows meaning more frequent.

English

The most popular pairs for news articles seem to be “cambridge analytica” and “mark zuckerberg.” This is understandably so since both of then are the key players in this data scandal. From this bigram plot, it is also evident that Cambridge Analytica was involved in political consulting due to the connections between the words “political consulting firm”, “data mining firm”, and “2016 presidential campaign.”

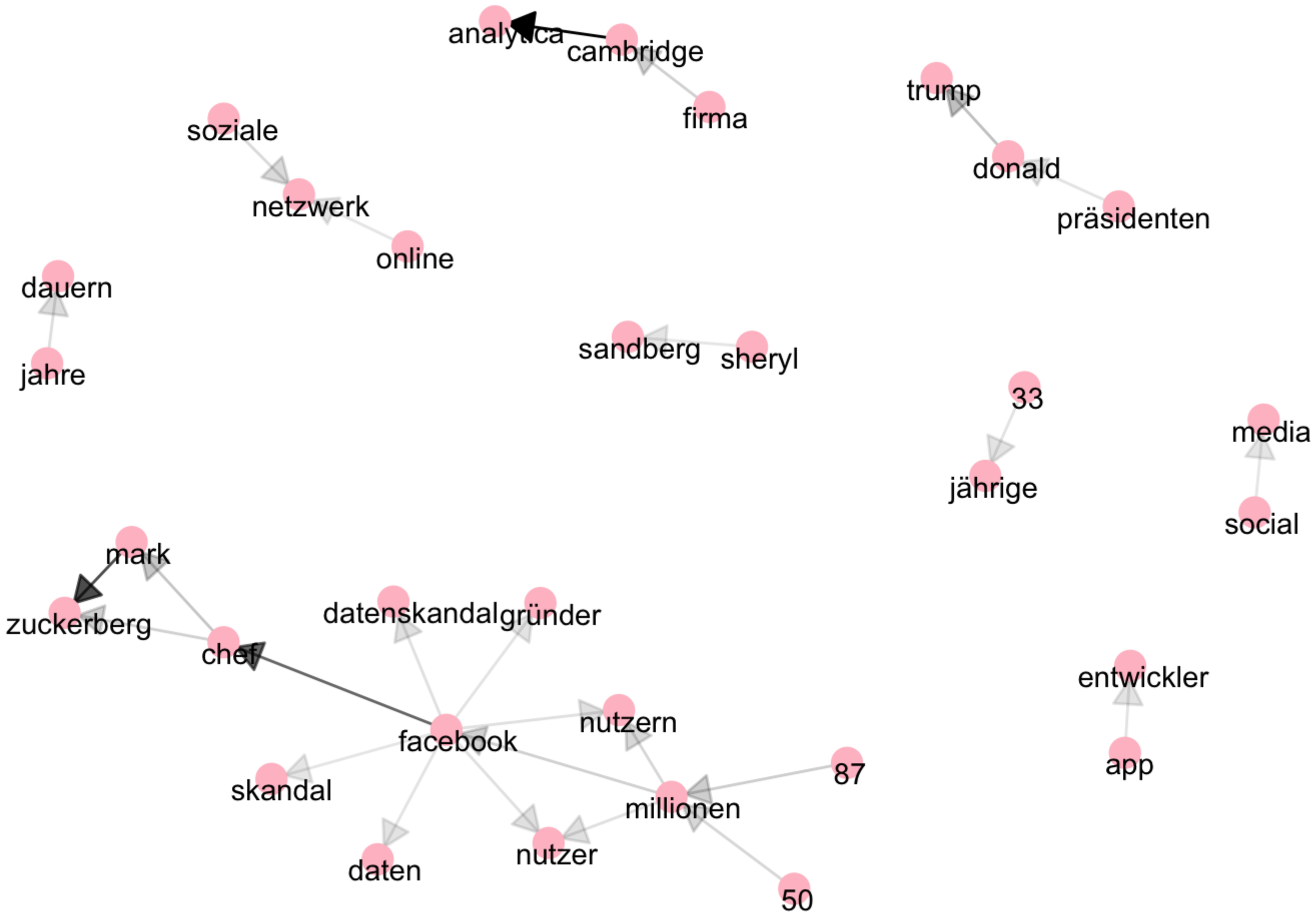


The most frequent pairs of words for tweets refer much more to Mark Zuckerberg's testimony in front of Congress. There are also a significant amount of numbers mentioned, from the "2 day" testimony that lasted "10 hours" and ended on "april 11" to the "87 million" users who were affected. The language used also changed. Words like "utter sham" and "booster seat" appear for the first time. Lastly, there is stronger clustering than in the news articles, as seen in the lower left-hand side.

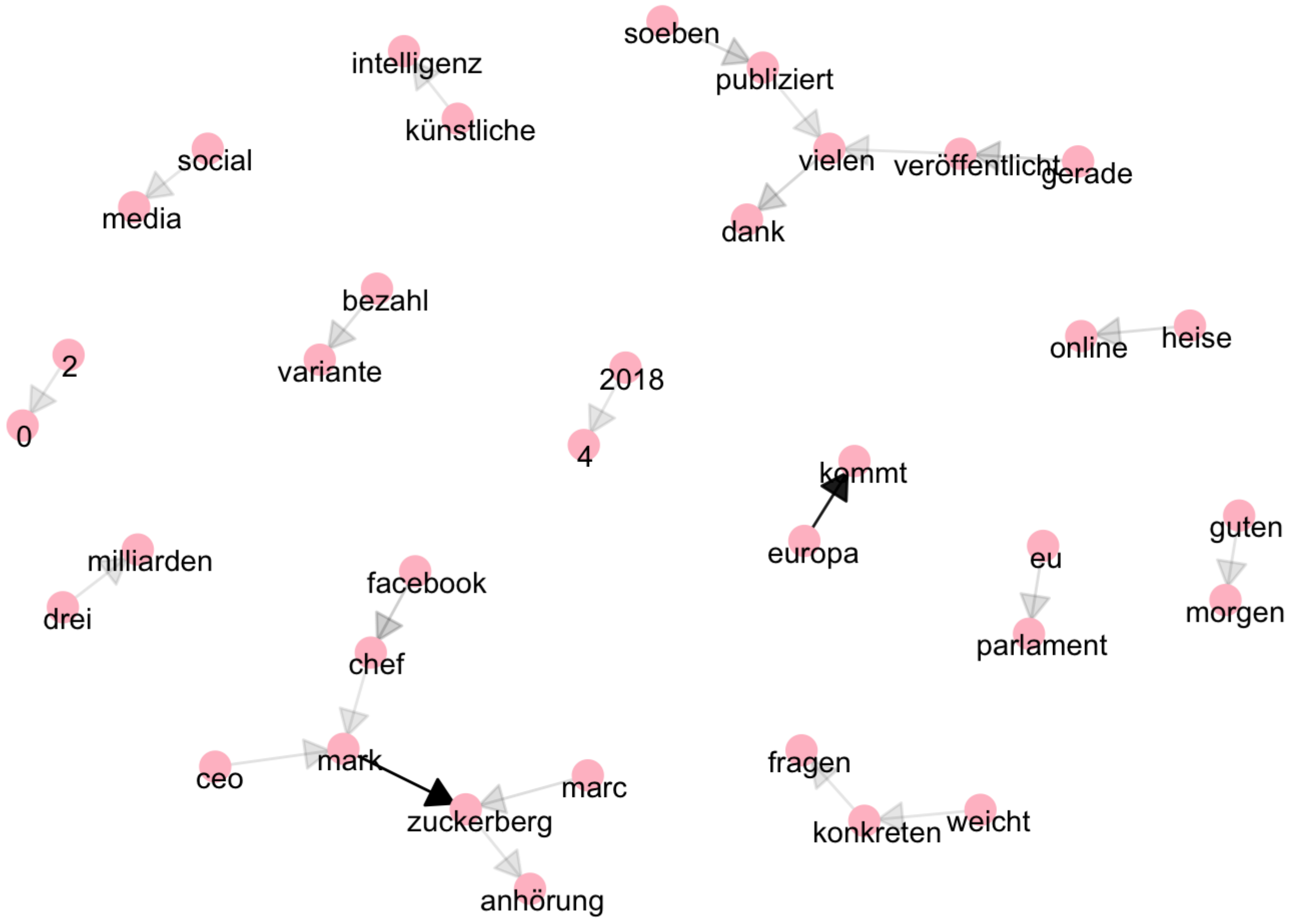


German

Much like the English news, “mark zuckerberg” and “cambridge analytica” appear the most often, followed by some background information about Zuckerberg, for instance, “facebook chef”, “facebook gründer”, and “33-jährige” (“facebook boss”, “facebook founder”, “33-year-old”).



The German tweets also reference more numbers and most of the words refer to what is happening at the testimony, such as “konkreten fragen”, “soeben publiziert”, and “gerade veröffentlicht” (“specific questions”, “just published”, “just released”).



An application of n-grams are Markov Chains, which rely on the previous event in determining the next event. Since bigrams group words into groups of two, the general trend for which words are most likely to be grouped together can be calculated. Therefore, an entire sentence or paragraph can be composed based on a sample of text. An example of this can be found in my Shiny application (please see "Further Resources below").

Conclusion

In conclusion, the English texts are more negative than the German texts and there is not a significant difference in word usage between tweets and news articles. I, honestly, thought the German texts would be more negative since I had my hypothesis that Germans would have stronger opinions about their data being spread around without their consent. However, the `bing` sentiment library for English is much larger (6788 words vs. 3468 words) than the German sentiment library, which probably left a decent amount of German words out of the analysis. I am happy though that people care so much about the data scandal that occurred last month and are actively voicing their opinions about it.

Further Resources

To explore the data further, I recommend my Shiny App (https://tcv14.shinyapps.io/final-project/). There, you will have the opportunity to explore a larger list of frequently used words and compare the positive and negative words side by side. There is also an application of n-grams in my app where you will be able to generate your own text using Markov Chains. I hope you enjoy it!