**Sentiment and Topic Analysis with Textual Tweets Data of Donald Trump**

Zhewen Song (zs1356) | Chenjie Su (cs5998)

**Word count:2514**

**Introduction**

Donald John Trump is the 45th and current president of the United States. Before entering politics, he was a businessman and television personality. We may find some intriguing facts when we analyze Mr.President's thoughts and behaviours. In many cases, the president ’s remarks are very important. Those may affect people ’s views on social events, their attitudes toward other countries, and some major international affairs. President Donald Trump’s tweets are a great way to know what the president is thinking, doing or thinking of doing. “Tweets” are the messages that users post and interact with others on an American microblogging and social networking service. Donald Trump tweets have become a source of information. For some, Trump’s tweets are crossing a line since he is using the platform to manifest official actions and decisions. For others, he keeps using the tweets as he has done always: as he pleases no matter what. This research, which is based on Trump’s tweets data, completed basic sentiment analysis and applied topic model for a deeper understanding of Mr. President’s thoughts as well as how to utilize tweets data to solve some practical text as data problems.

**Literature**

There are many researches based on Trump’s words and speeches. Recently, three journalists for The New York Times reviewed more than 260,000 words spoken by President Trump during the recent pandemic(Jeremy, Elaina, and Maggie, 2020). They found out the self-regard, the credit-taking, the audacious rewriting of recent history to cast Trump as the hero of the pandemic rather than the president who was slow to respond after analyzing White House briefings and other presidential remarks. There is also an article published by New York Times investigating how Trump reshaped the presidency in over 11,000 tweets (Michael, et al., 2019) and comparing his real policies versus tweets in different time periods.

We were inspired by the previous research and decided to study on tweets data before the announcement of the presidential election, after the announcement, after becoming the president and during the epidemic to achieve our research purposes.

**Theory and Hypotheses**

For sentiment analysis, we used sentiment scores to evaluate the attitudes of all the tweets. We assumed that the overall sentiment scores might be slightly different when we analyze tweets in different periods of time. But we think the sentiment scores we get should basically not have particularly huge fluctuations.

In the second part, we implemented topic modelling for the contents of Trump’s tweets in the 4 periods of time that we had chosen and as a whole using the LDA (Latent Dirichlet Allocation) model. We expected to see the rank of top topics that were most likely to be the tweets’ topic changing over time along with the party’s political strategy changed at different stages. For example, we assumed Trump’s tweets would focus on different topics during the election period when he needed to build the most positive self image, and after he won the election and needed to deal with those practical issues. And we also expected to see new topics rising during specific time intervals to detect those hot political events related to Trump, such as building walls on the Mexican border, Russian secret agent, coronavirus, and so on.

**Data and Methods**

We found the original dataset from Kaggle (Kaggle, 2020). The raw data contains all tweets from twitter account @realdonaldtrump in the time period from May 4th, 2009 to April 15th 2020. The dataset includes 42296 records and 8 features which are id, link, content, date, retweets, favorites, mentions, and hashtags. We preprocessed the raw dataset by the following steps. Firstly ,we extracted the tweet records by date from June 16th, 2014 to April 15th, 2020 and splitted it by 3 special dates - June 16th, 2015 when Trump announced to participate into the president election, November 9th, 2016 when he won the election, January 22nd, 2020 when he published the first tweet about coronavirus, so that we can analyze the tweet information in some specific political stages in deep and track the trend, as well as for a particular social event like the spread of the coronavirus. Secondly, to deal with those irrelevant information on tweets data such as hashtags, mentions, website address, and retweets, we applied a few cleaning functions to remove all strings that followed ‘@’, ‘#’, ‘http’ and records started with ‘RT’. Thirdly, when converting text corpus into dfm, we removed numbers, punctuations, english stopwords, and transformed all words into lower cases.

First of all, we feature engineered day of week and time of day to get an insight into Mr Trump’s tweeting habits and computed his daily and weekly tweet frequency from June 16th, 2014 to April 15th, 2020. We added hour and week day columns. For calculating daily frequency and weekly frequency, we used two loop functions to generate the number of total tweets of each day and each week. We can see he tweets on average 12 times a day and 87 times a week. He tweets a lot especially during the time before he won the election. The trend slows down during the second half of his campaign. By analyzing the tweet frequency by hour, we found out he really does like post messages in the morning especially on Monday morning.

Then we tried to figure out what words Mr.Trump tweet most frequently. Word clouds in different time periods can help us to know about this more directly. Before Trump announced his participation in the presidential election, the most words he used were trump, president, great, will, won, and donald. All of these shows he was pretty confident and he will participate in the election. During his campaign, the most popular words were trump, great, will, america, people, hillary and makeamericagreatagain. Now we all know that Hillary was his strongest competitor and “make america great again” is his slogan. After he won the election, there were some new words that appeared, like fake, news, democrats. From he tweeted the first tweet about coronavirus to April 15th, we noticed among the most frequent words, there are words like coronavirus, impeachment, hoax, and some officers’ names like mike and bernie. Overall, he talked a lot about himself and sometimes attacked someone or something.

In sentiment analysis, we implemented positive and negative words dictionaries that were discussed in the Hu and Liu’s (2014) customer review research. We generated a sentiment score for each review based on the number of positive words minus the number of negative words. When the sentiment score is bigger than zero, we classify this tweet record as a positive record. And others are negative tweets. We attempted to visualize the daily sentiment score and sentiment count, but we couldn’t attain clear patterns whatever on the whole dataset or subsets. So we aggregated average sentiment scores and sentiment count by hour of the day.

For the topic analysis section, we applied the LDA (Latent Dirichlet Allocation), one of the most common algorithms for topic modeling, which was introduced by Blei in 2003 (Blei, Ng, and Jordan, 2003). The statistical model treats each document as a mixture of topics, and each topic as a mixture of words, which can help us find the underlying key points for a collection of tweet texts.

We separately fed the model with the tweet text data for the 4 time intervals that were mentioned above - one year before election, election period, after election, and coronavirus period, with 12 topics, 3000 iterations and method of “Gibbs”. And we output the top 10 terms for each topic and the rank of top topics based on the number of tweets for which they are the primary topics. Here we also attempted to use mean probabilities of topics over the entire collection of tweets to rank the topic but found this ranking algorithm favors more to those with general semantic coherence rather than specific contents. So we chose the first ranking algorithm for better event detection. Figure x shows the top 10 terms for the generated 12 topics for the 4 time periods. Table x shows the top 5 topics for each time period with topic number, top words, and labels that we gave based on the top words.

In addition to that, we trained the model with tweets data from all 4 time periods to find out the overall pattern. The top 12 topics and top terms for each topic were shown in figure x. For each topic, we also counted the number of tweets for which they are the primary topics in each year and plotted the counts throughout the 7 years to see how the rank of topics changed in different years and political stages, as shown in figure x.

|  |  |  |  |
| --- | --- | --- | --- |
| Topic Number | Tweet Counts | Words | Labels |
| 2 | 1194 | president run country need please vote needs mr running save | Vote |
| 1 | 1032 | new golf course hotel beautiful amazing today chicago national | Golf |
| 5 | 787 | apprenticenbc show apprentice im celebapprentice tonight celebrityapprentice looking watch season | Reality show |
| 3 | 727 | dont know never entrepreneurs work go keep give way success | entrepreneurs |
| 7 | 603 | obama us now ebola stop isis must obamacare fix mexico | Obama |

Table y. Top 5 topics and labels in period 1 (before election/2014.6-2015.6)

|  |  |  |  |
| --- | --- | --- | --- |
| Topic Number | Tweet Counts | Words | Labels |
| 1 | 872 | get vote time can win go cant going republican keep | Vote |
| 5 | 856 | thank makeamericagreatagain iowa support big crowd see speech carolina amazing | election |
| 2 | 825 | president one like love need mr best truth man next | president |
| 3 | 820 | cnn foxnews debate last much night nice watch thanks really | Public debate |
| 4 | 680 | us country jobs must illegal right immigration take border strong | immigration |

Table y. Top 5 topics and labels in period 2 (during election/2015.6-2016.11)

|  |  |  |  |
| --- | --- | --- | --- |
| Topic Number | Tweet Counts | Words | Labels |
| 2 | 1310 | thank today american day honor first law love americans incredible | American |
| 1 | 1089 | great big vote job military state win strong total crime | military |
| 3 | 1033 | news fake media new bad story corrupt even fact york | Media |
| 11 | 946 | witch hunt russia collusion fbi obama hillary campaign mueller election | Hillary |
| 4 | 817 | us many back china trade deal will usa dollars coming | China |

Table y. Top 5 topics and labels in period 3 (after election/2016.11-2020.1)

|  |  |  |  |
| --- | --- | --- | --- |
| Topic Number | Tweet Counts | Words | Labels |
| 1 | 152 | bernie mini mike democrat even can hard iowa now | Bloomberg |
| 2 | 148 | democrats nothing hoax impeachment win never left dems senate radical | Impeachment |
| 4 | 145 | thank house today white back pm work conference americans time | White house |
| 3 | 110 | just working many good president job china country one coronavirus | virus |
| 9 | 89 | news fake media said much must cnn msdnc night story | media |

Table y. Top 5 topics and labels in period 4 (COVID-19/2020.1-2020.4)

**Results and Discussion**

We could notice that all the hourly average sentiment scores are between 0.2 and 1.0 in the whole period of time that we focus on, which means the overall sentiment of tweets is positive to some degree. Especially during period 1(before election/2014.6-2015.6) and period 2 (during election/2015.6-2016.11), where the average scores are higher than other periods. However, during periods 3 and 4 (after election), there are some negative average sentiment scores in the day. The results allowed us to verify that even though there is some content to blame someone or something, in general, positive tweets are slightly more than negative tweets during the day. Citizens prefer to hear positive and inspiring remarks from the president. Positive leader knows that positive energy unlocks human potential and therefore inspires people to get involved in things happening in this country. If the public can obtain more positive information from the president, it means that the president believes he has achieved good accomplishments during his term of office, which will also increase the popular support rate to a certain extent. Perhaps too much attack and accusation of others is not a smart thing.

As for the topic modelling results, by looking at the top topic labels in the prior-election period, which included ‘vote’, ‘Obama’, ‘reality show’, and ‘entrepreneurs’, we can have a basic image of Trump’s identity as a successful entrepreneur and the cast of a reality show at that time and his early intention to participate in the presidential election. The top topics for the election period are ‘vote’, ‘election’, ‘president’, ‘public debate’ and ‘immigration’, which shows the importance of his twitter as a ‘smokeless battlefield’ of the election and his focus on the immigration issue. The top topics after he won the election are ‘American’, ‘military’, ‘media’,

‘Hillary’ and ‘China’, where we can detect those political events that Trump discussed most frequently on his twitter including the constant investigation of Hillary Clinton’s former emails which is criticized as a “witch hunt”, America’s trade war with China, and Trump’s arguments to the media’s counterview. And in the period of 2020, the top topics are ‘Bloomberg’, ‘impeachment’, ‘white house’, ‘virus’, and ‘media’. We can see Trump’s preparation for the new round of election and his attack on Michael Bloomberg as ‘mini mike’. Looking deeper to the top terms related to coronavirus - ‘just’, ‘working’, ‘good’, ‘China’, we can see Trump’s optimistic attitudes towards the virus in the early stage and his blame to China. But for this issue, checking the relevant original tweets will give us a more accurate view.

From the resulting primary topic counts through years, we can see that the rank of topic 11 and topic 7, which are labeled as immigration/border and Hillary/Russian, rises significantly in 2017 and 2018, corresponding to the shifts of Trump’s speech focus from these 2 years. The rank of topics about ‘vote’ and ‘election’ explicitly drops after 2016. And topic 4 which is labeled as ‘China’ constantly ranks at 5th position from 2016 to 2020.

From above, the LDA topic modelling performed well on detecting those political events that Trump talked about the most on twitter and tracking his shifts of public speech focus over time. Perhaps these issues are too obvious for us because we have very well known the president from the daily news and conversations. But we can infer that the method is a good way to quickly capture the focus of attention of someone who we don’t know well from his/her tweets.

**Future Work and Summary**

We have noticed that twitter data is the most comprehensive source of live, public conversation worldwide. There are a lot of functions for users, like retweet, comment and like. We included Trump’s retweets as the whole tweets data in our research project. If we can analyze Trump’s retweet separately, we can get what content Trump agrees or disagrees with. Also, in the future, we can focus on some specific tweets which get the most public attention and analyze some comments data to attain how people think about the specific events. Besides, due to the characteristics of tweets, which are relatively shorter and more colloquial compared with real articles, quite a few topics generated from topic modelling contain more general terms rather than indicating specific events or attitudes, which we could find ways to mitigate that impact in the future work.

**Appendix**

*Figures*

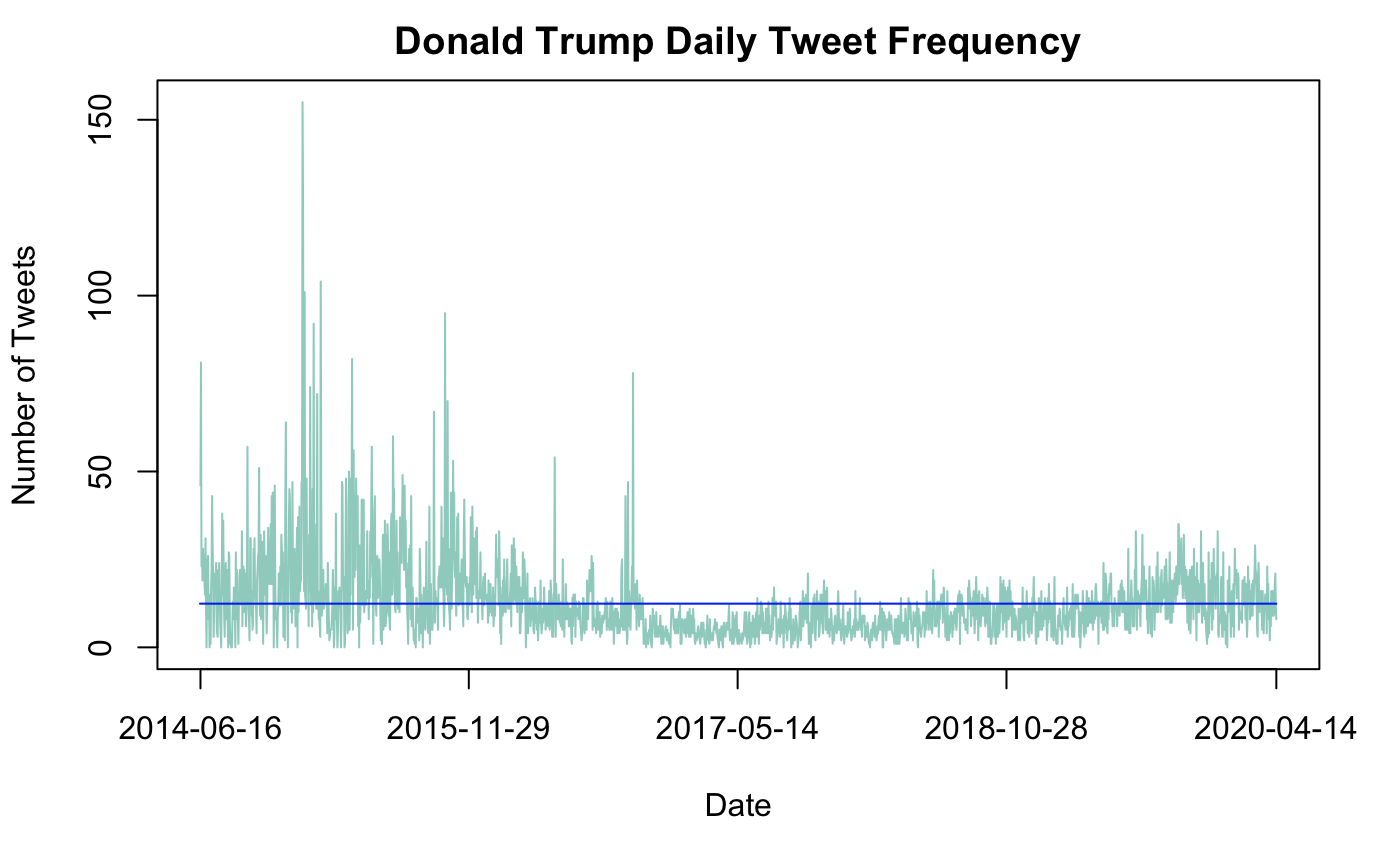


Figure 1. Donald Trump Daily Tweet Frequency

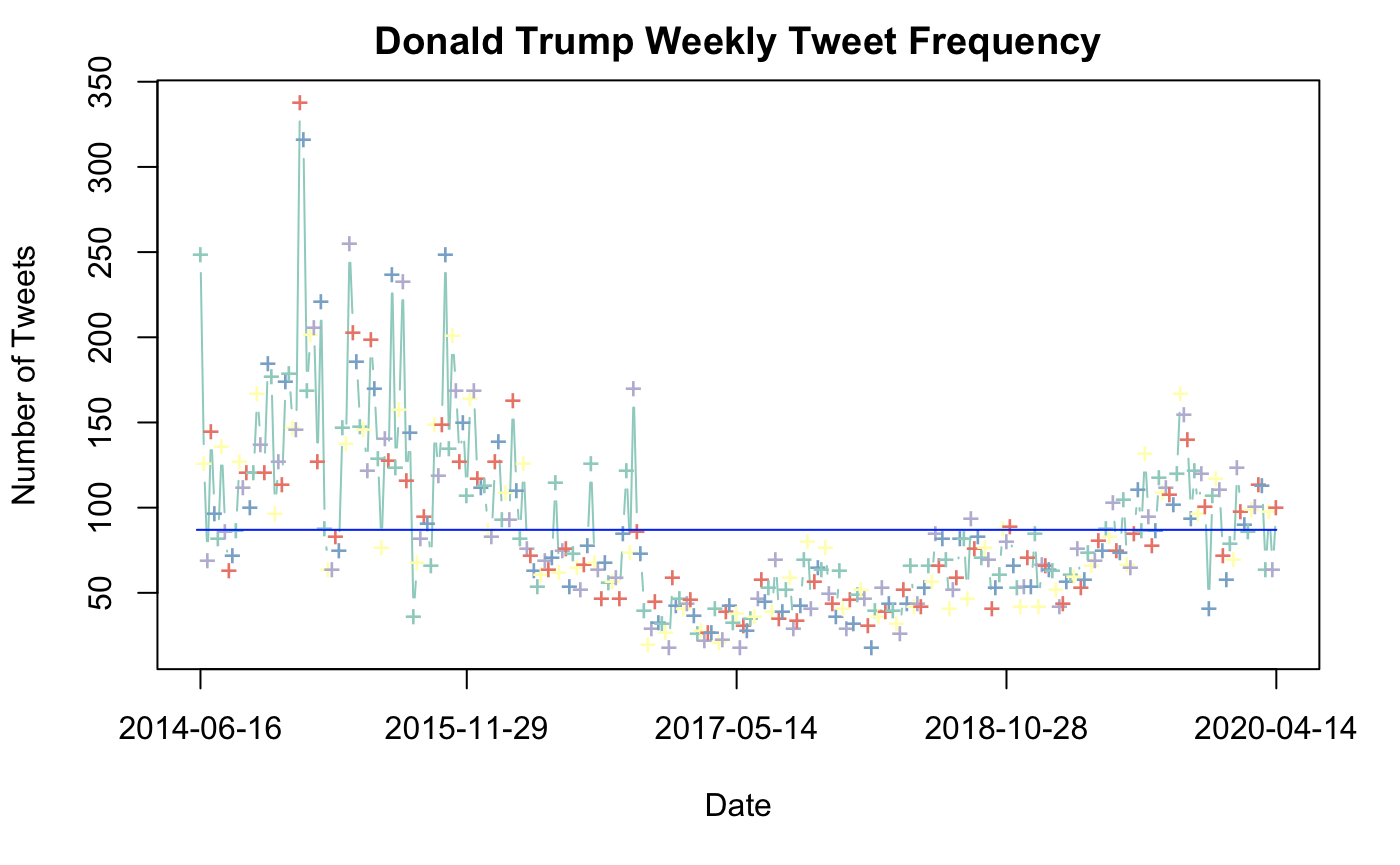


Figure 2. Donald Trump Weekly Tweet Frequency

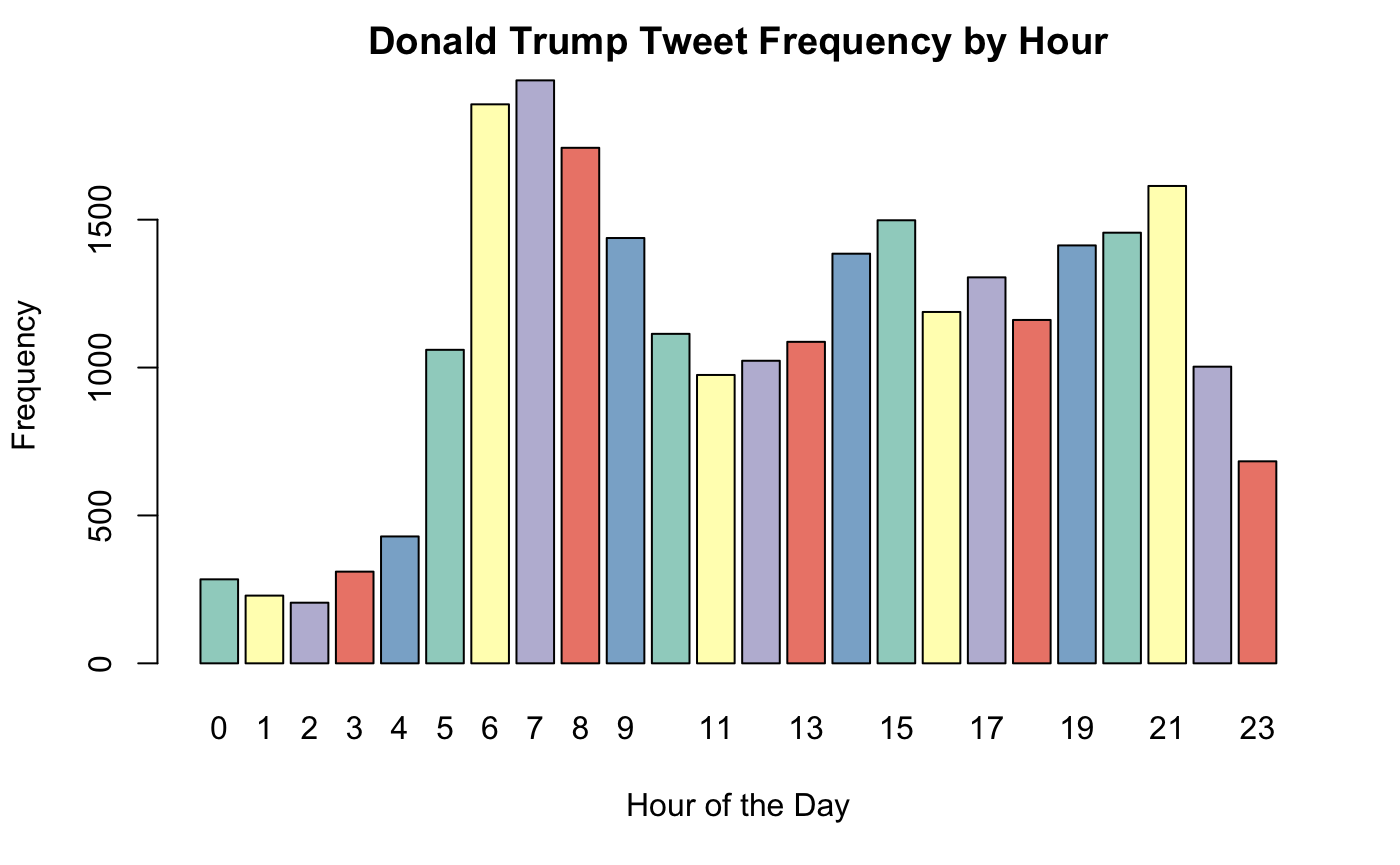


Figure 3. Donald Trump Weekly Tweet Frequency

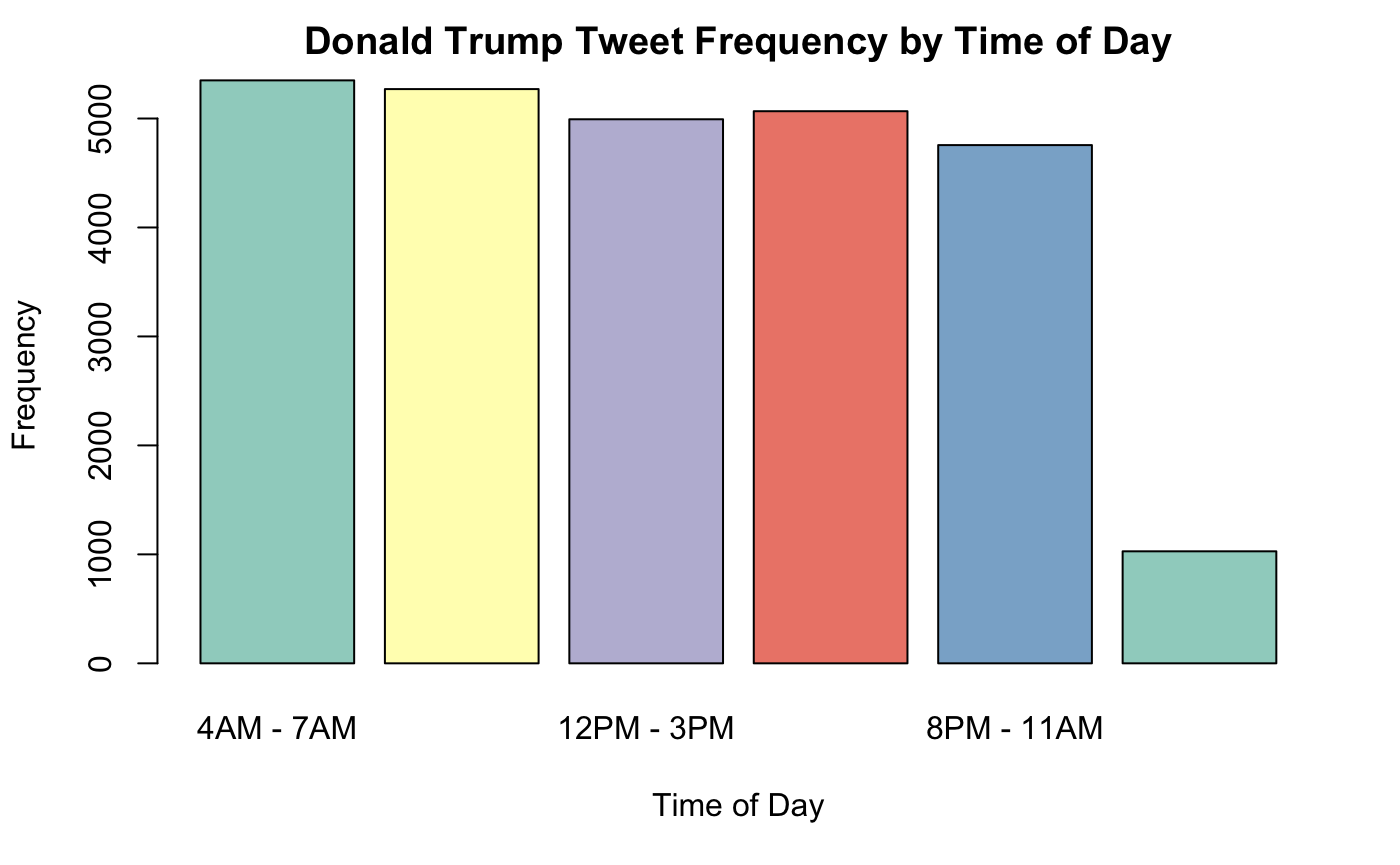


Figure 4. Donald Trump Tweet Frequency by Time of Day

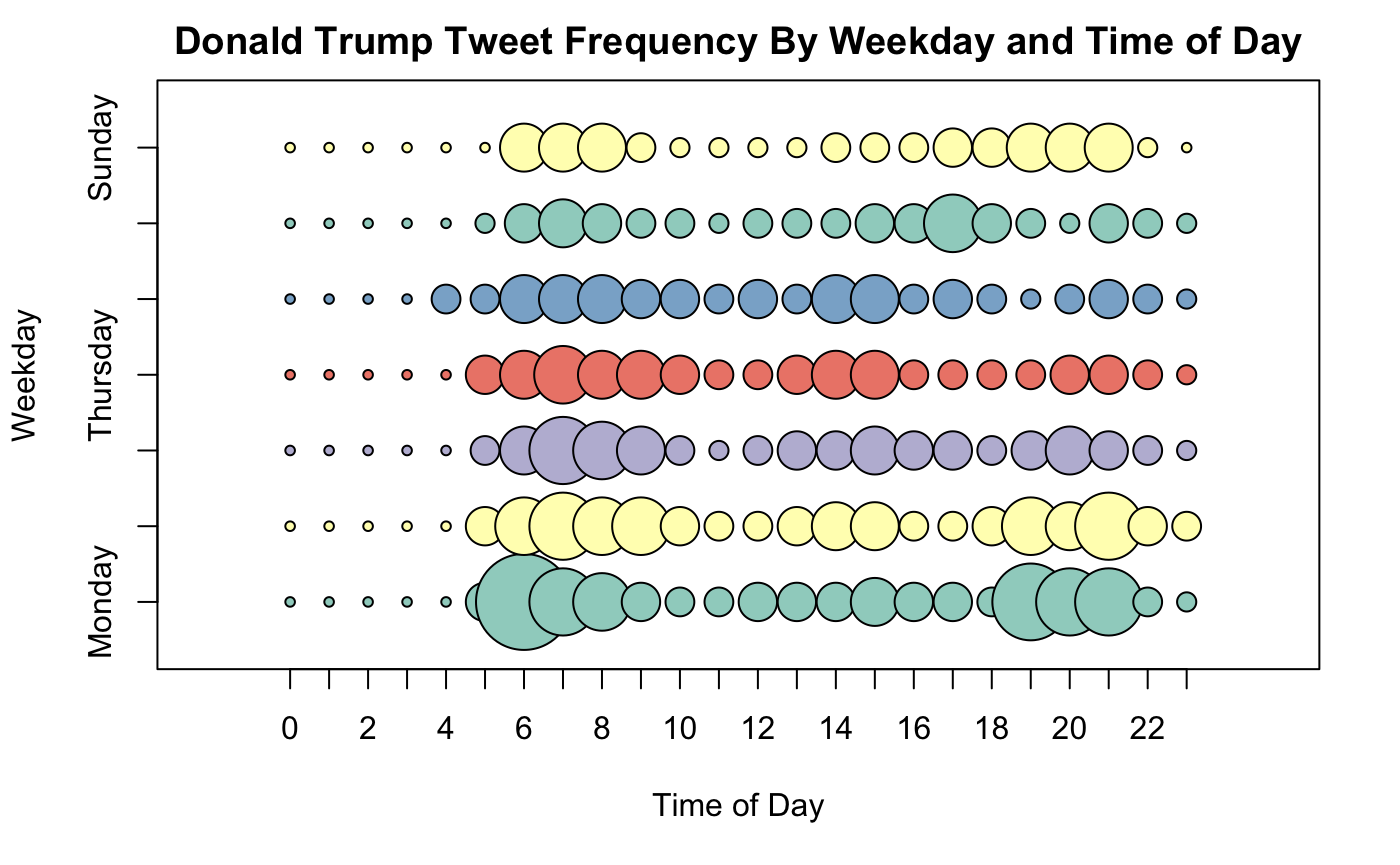


Figure 5. Donald Trump Tweet Frequency By Weekday and Time of Day

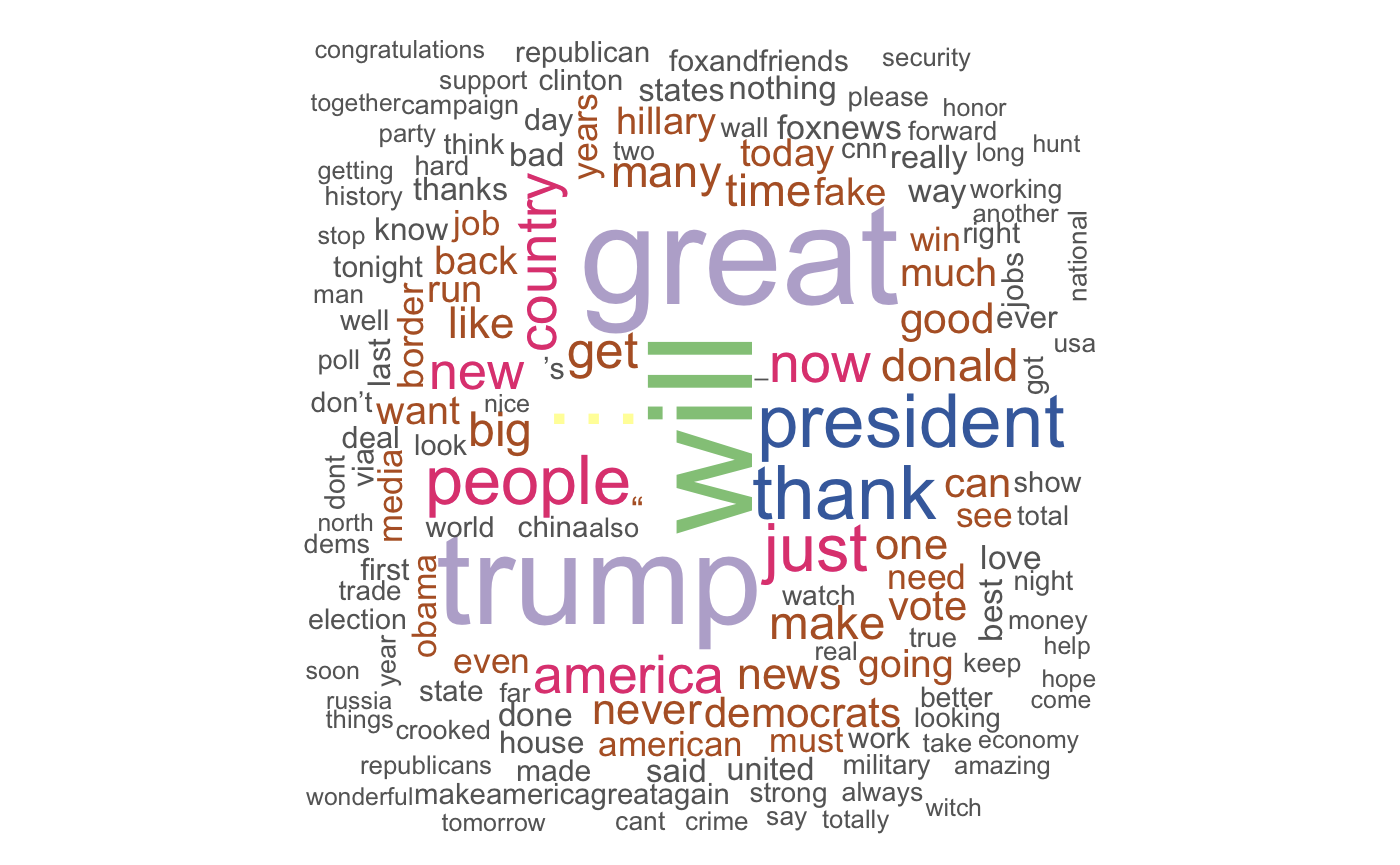


Figure 6. Word Cloud for whole periods of time

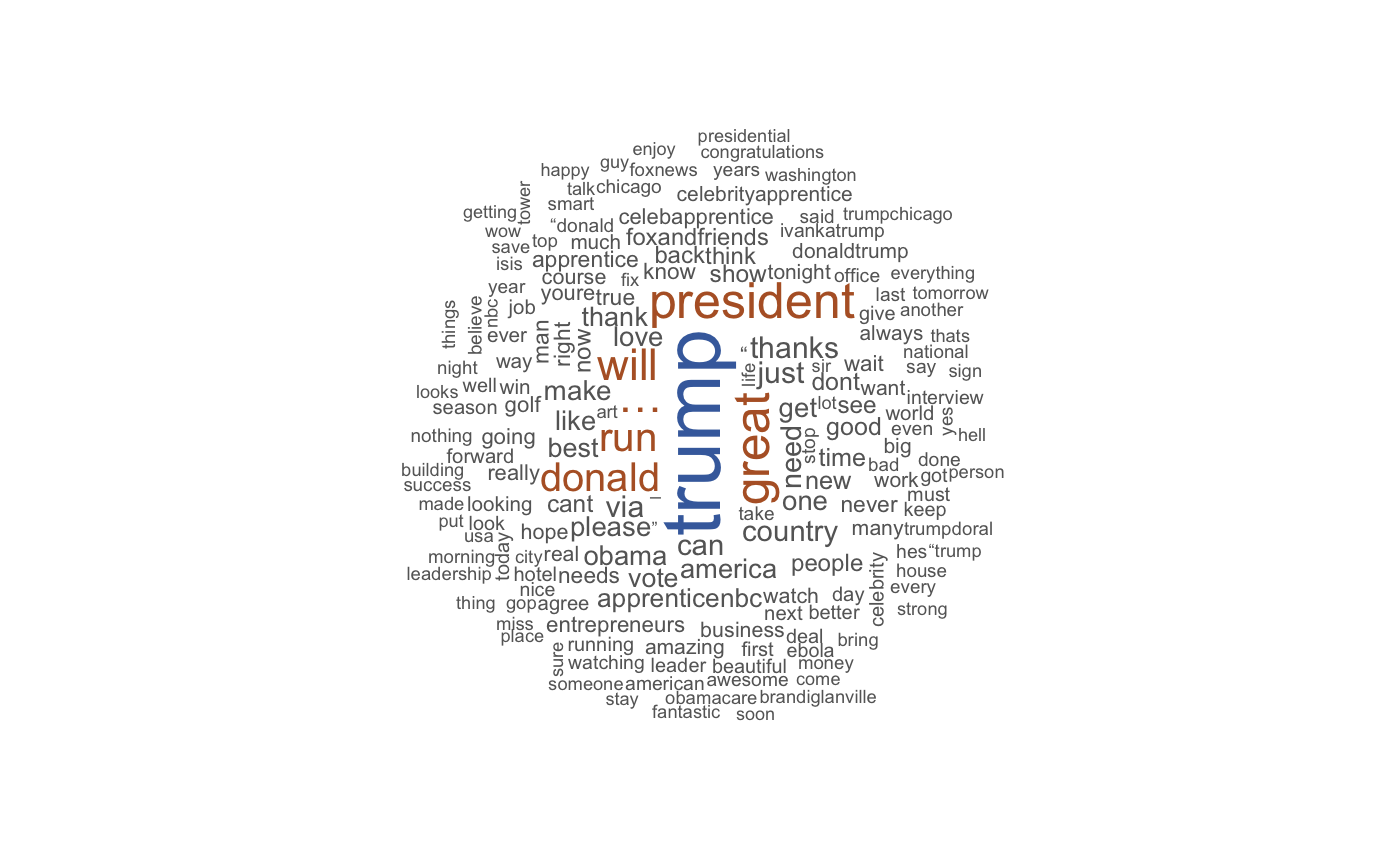


Figure 7. Word Cloud for period 1 (before election/2014.6-2015.6)

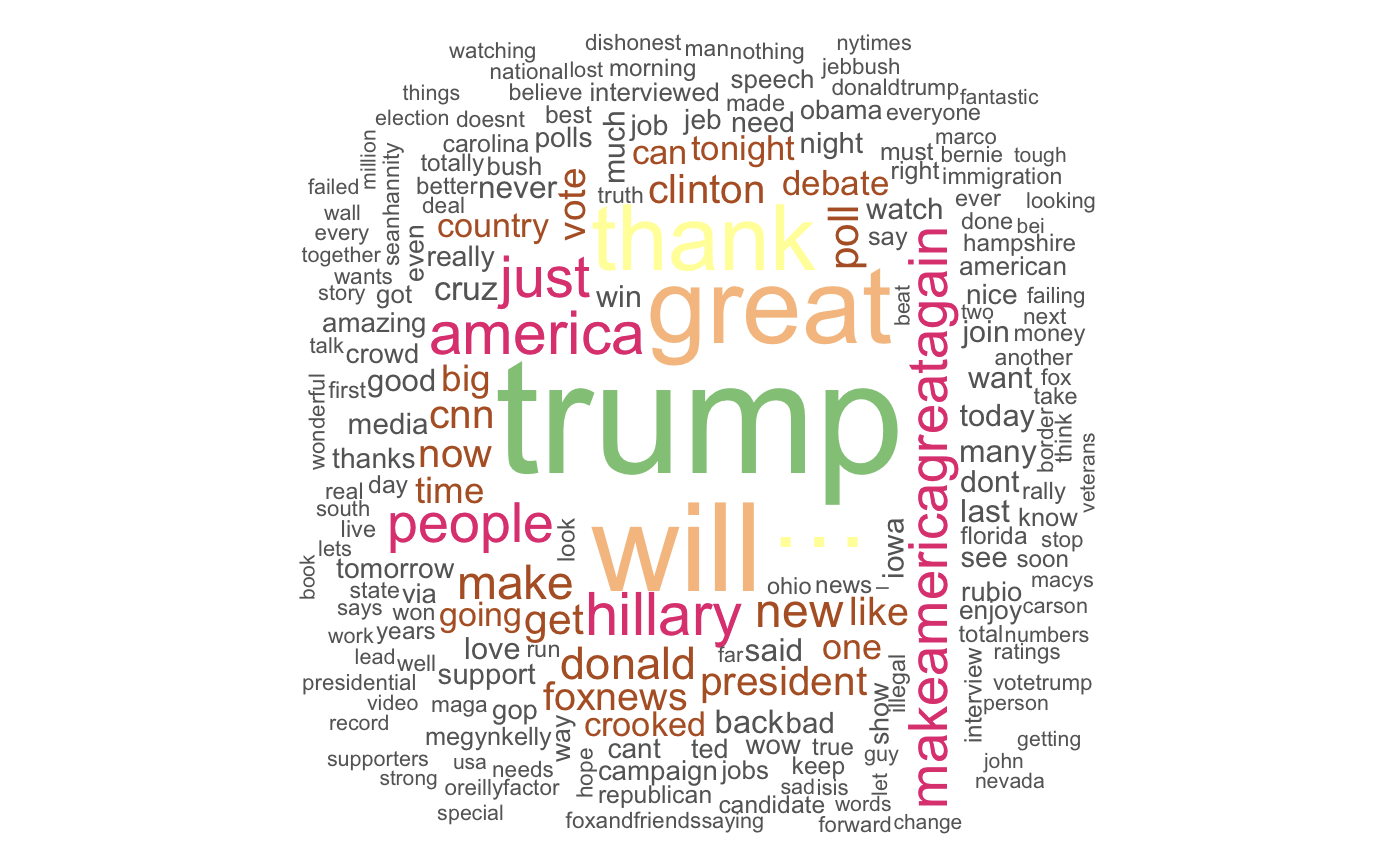


Figure 8. Word Cloud for period 2 (during election/2015.6-2016.11)

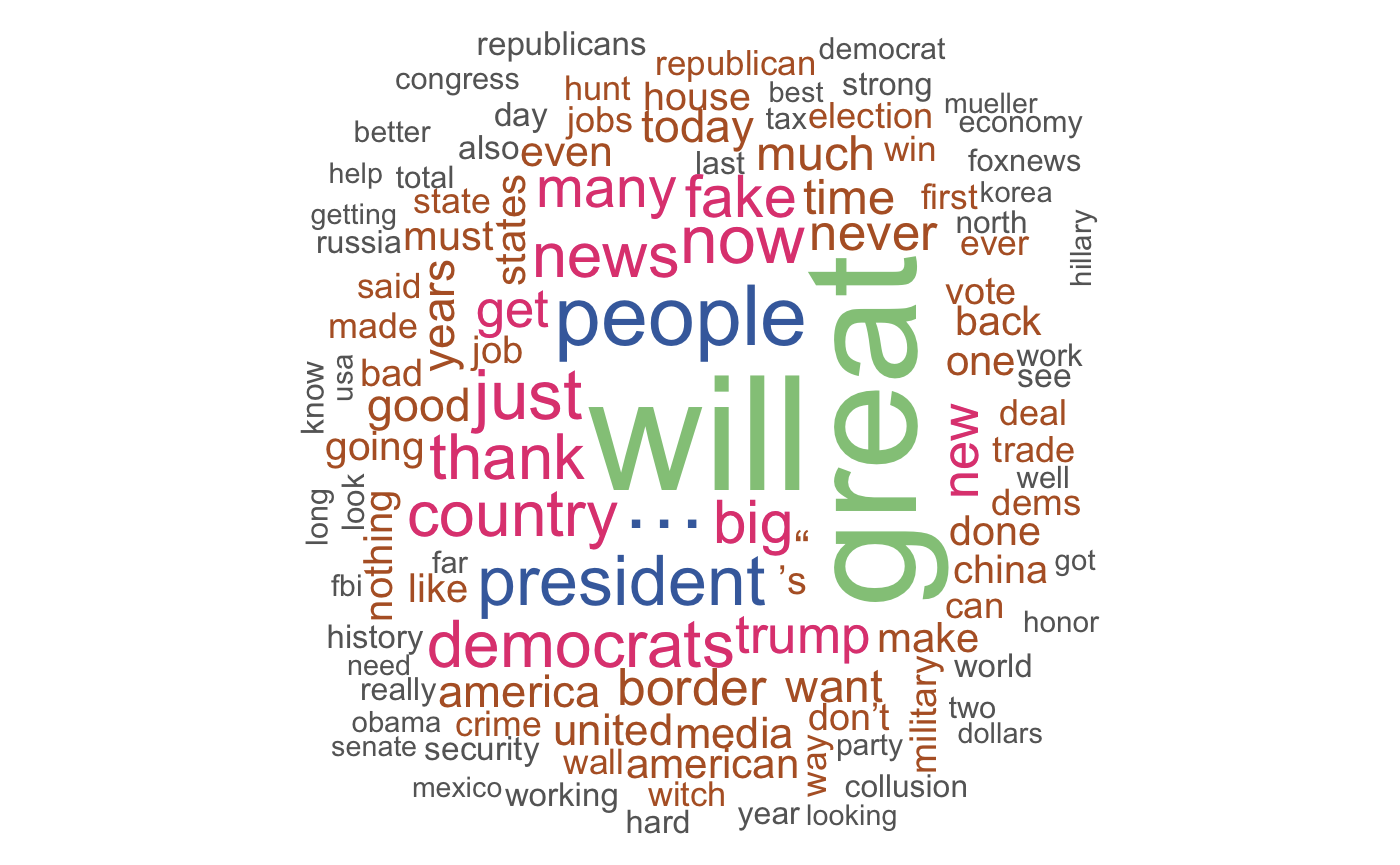


Figure 9. Word Cloud for period 3 (after election/2016.11-2020.1)

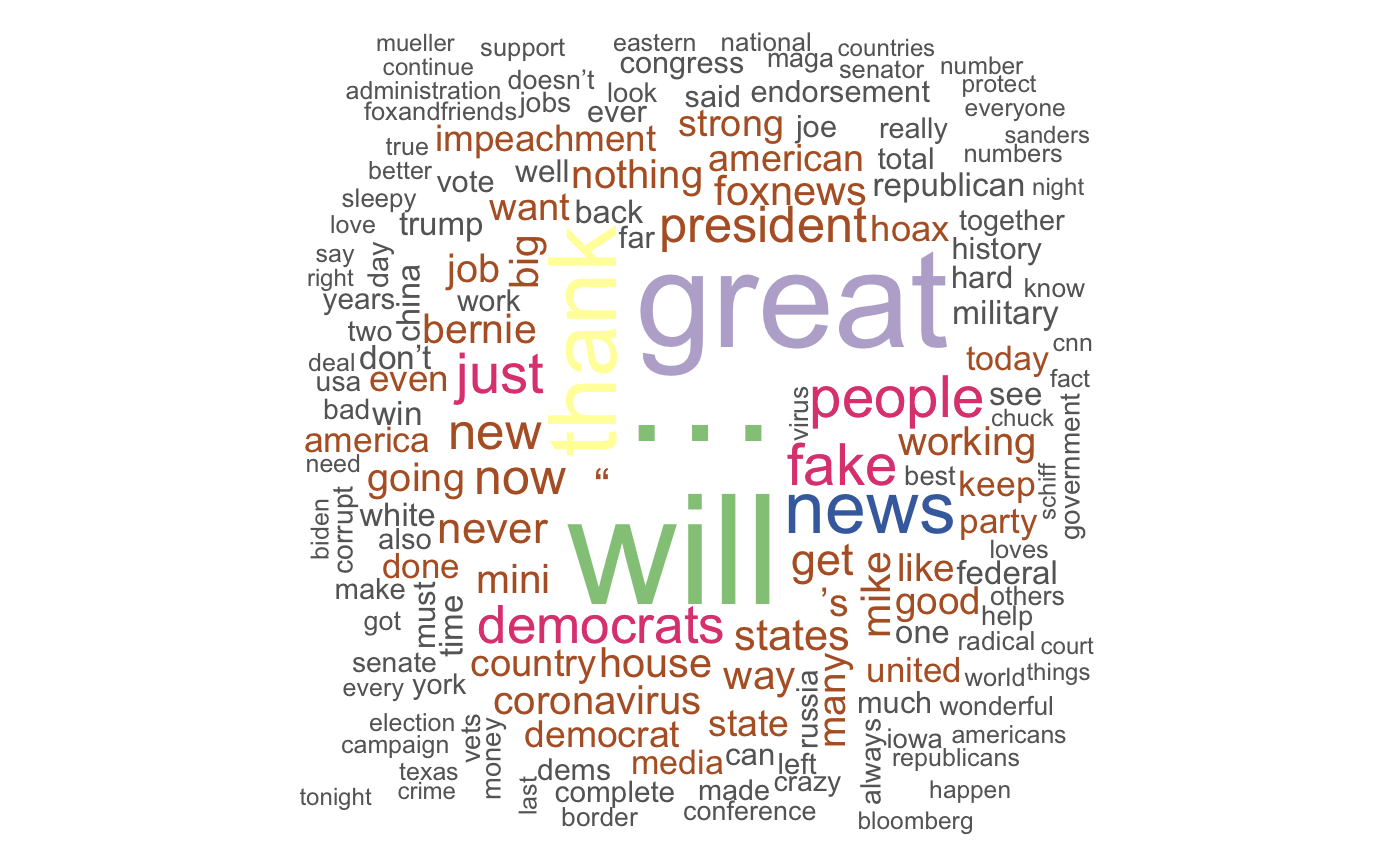


Figure 10. Word Cloud for period 4 (COVID-19 period /2020.1-2020.4)

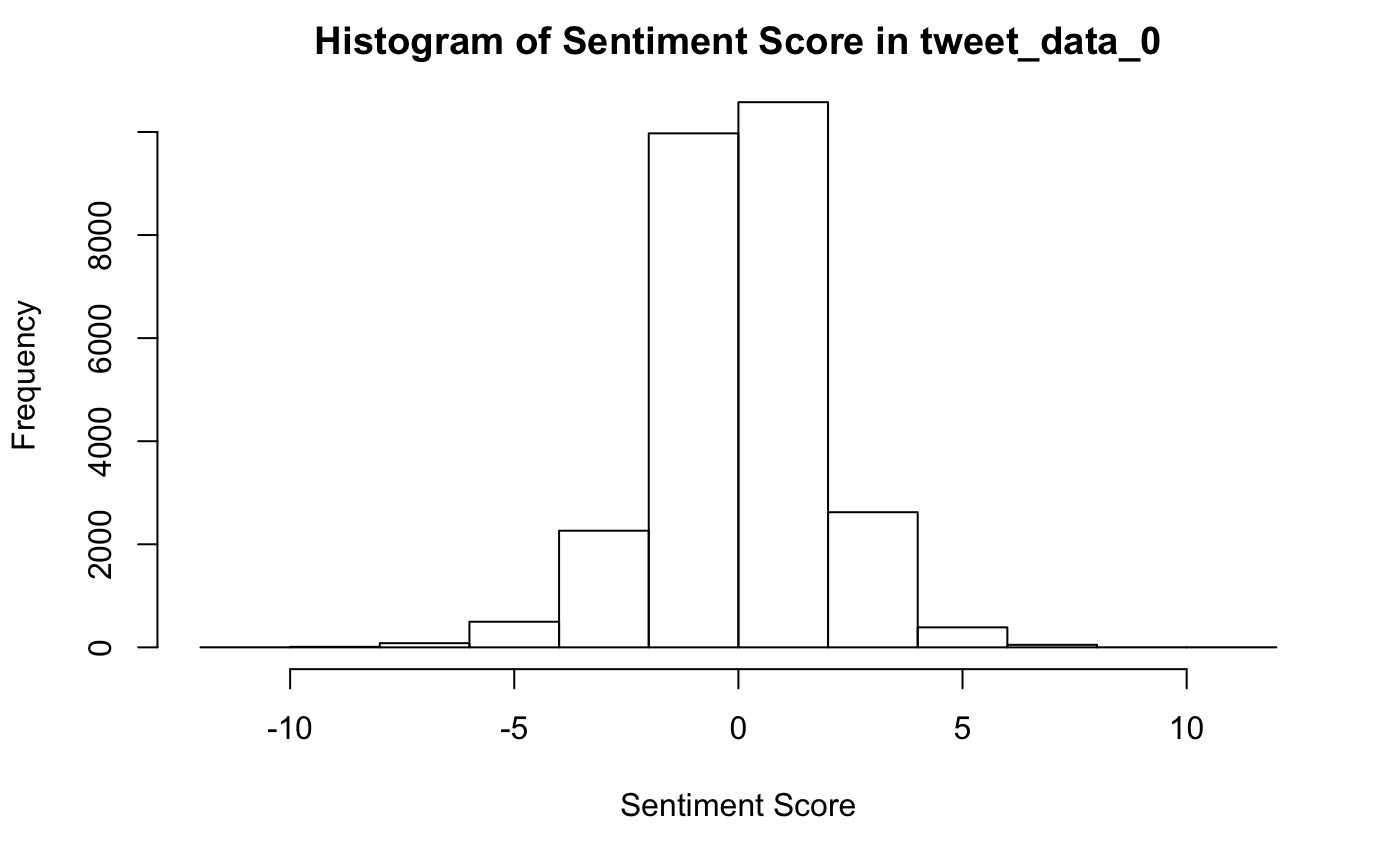


Figure 11. Histogram of sentiment score for all tweets records we used

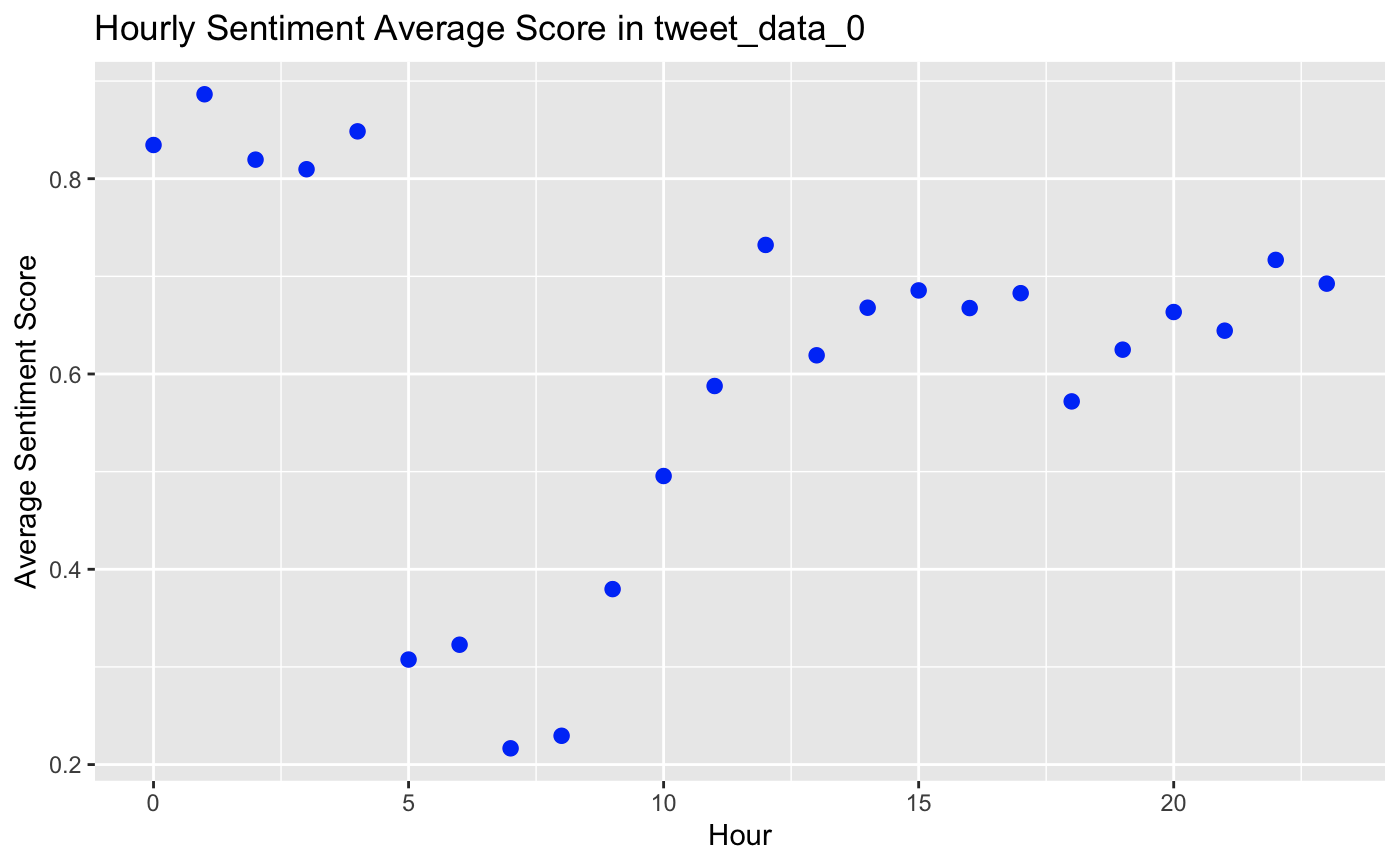


Figure 12. Hourly Average Sentiment Score for whole

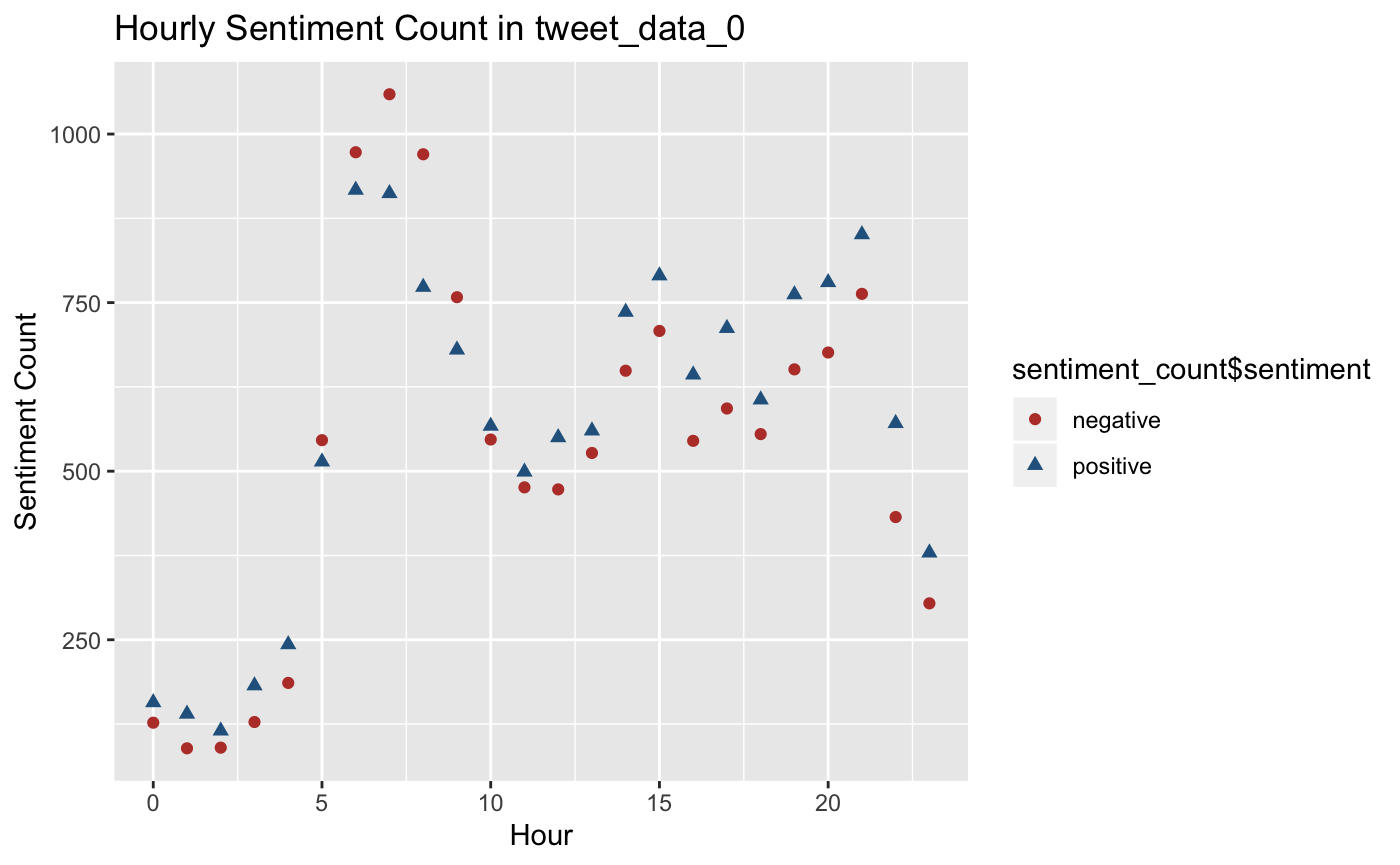


Figure 13. Hourly Sentiment Count for whole

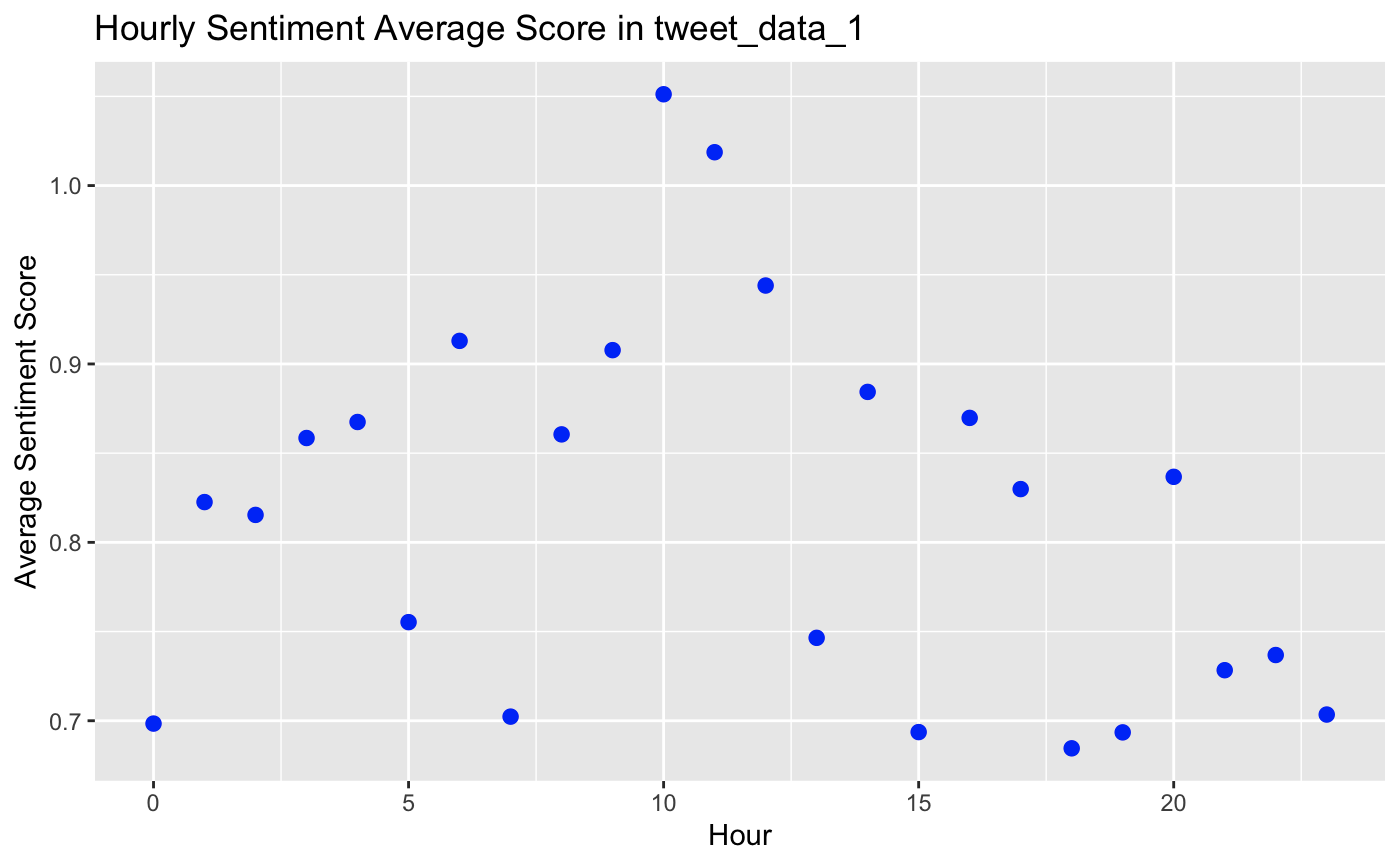


Figure 14. Hourly Average Sentiment Score for period 1 (before election/2014.6-2015.6)

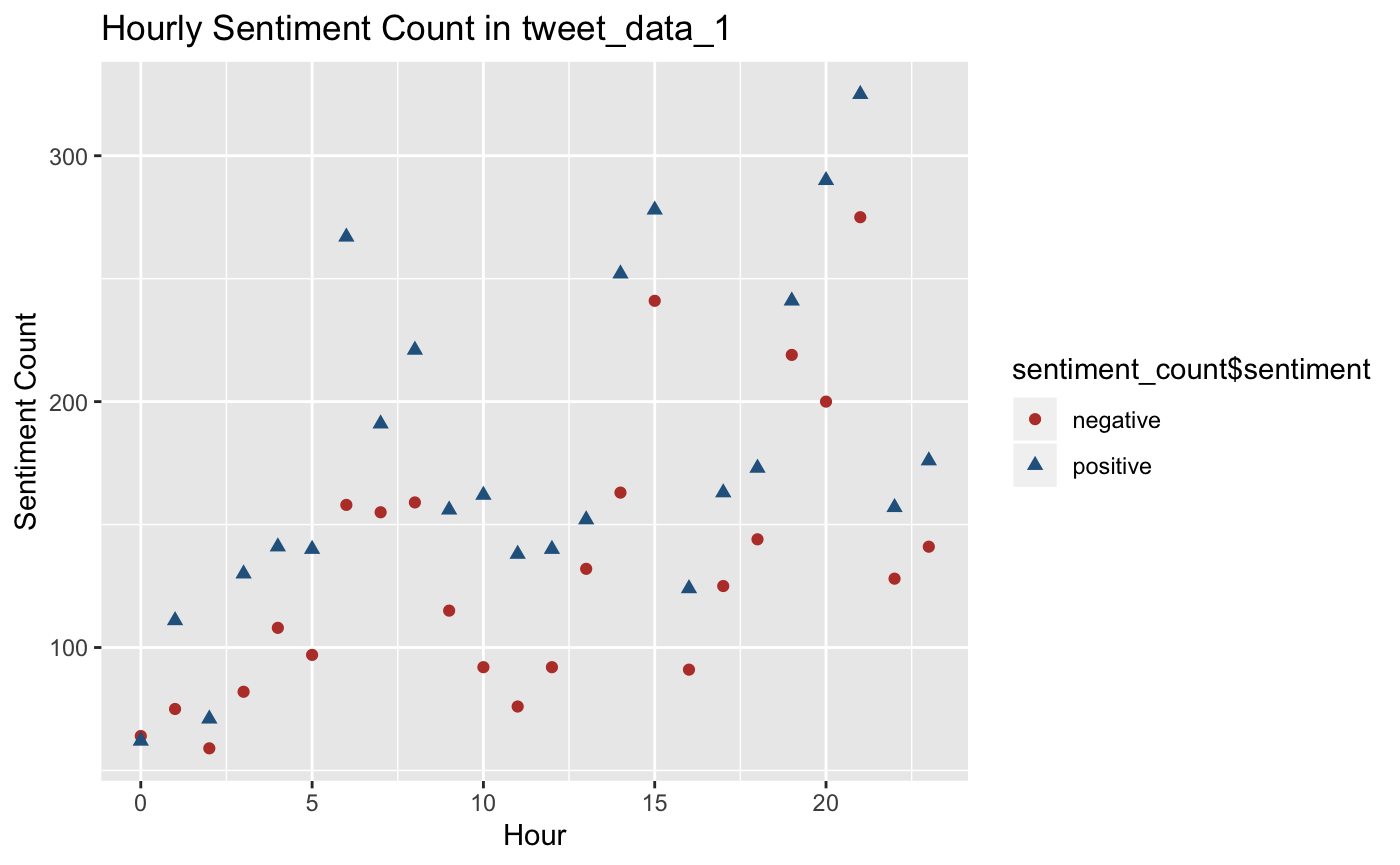


Figure 15. Hourly Sentiment Count for period 1 (before election/2014.6-2015.6)

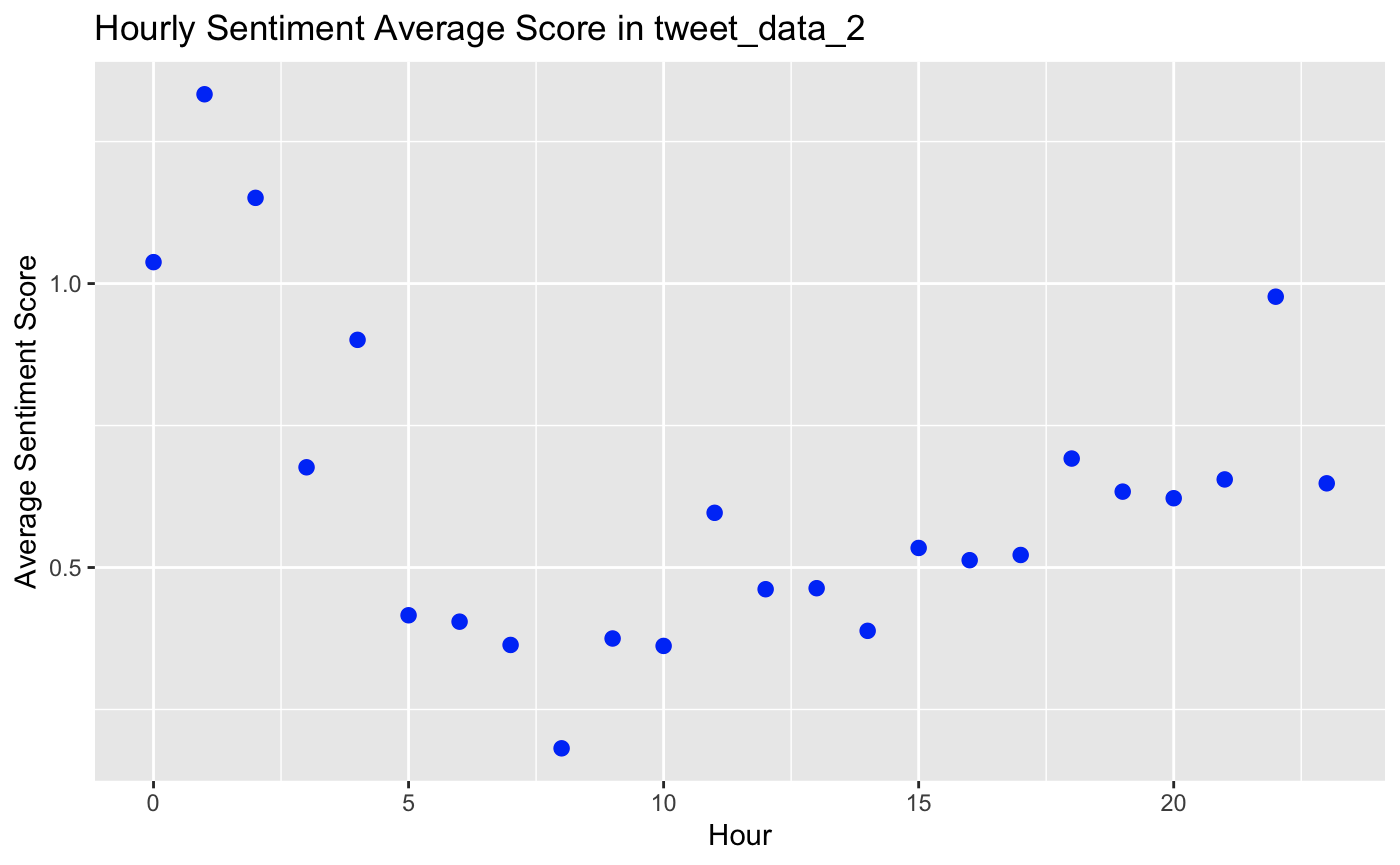


Figure 16. Hourly Average Sentiment Score for period 2 (during election/2015.6-2016.11)

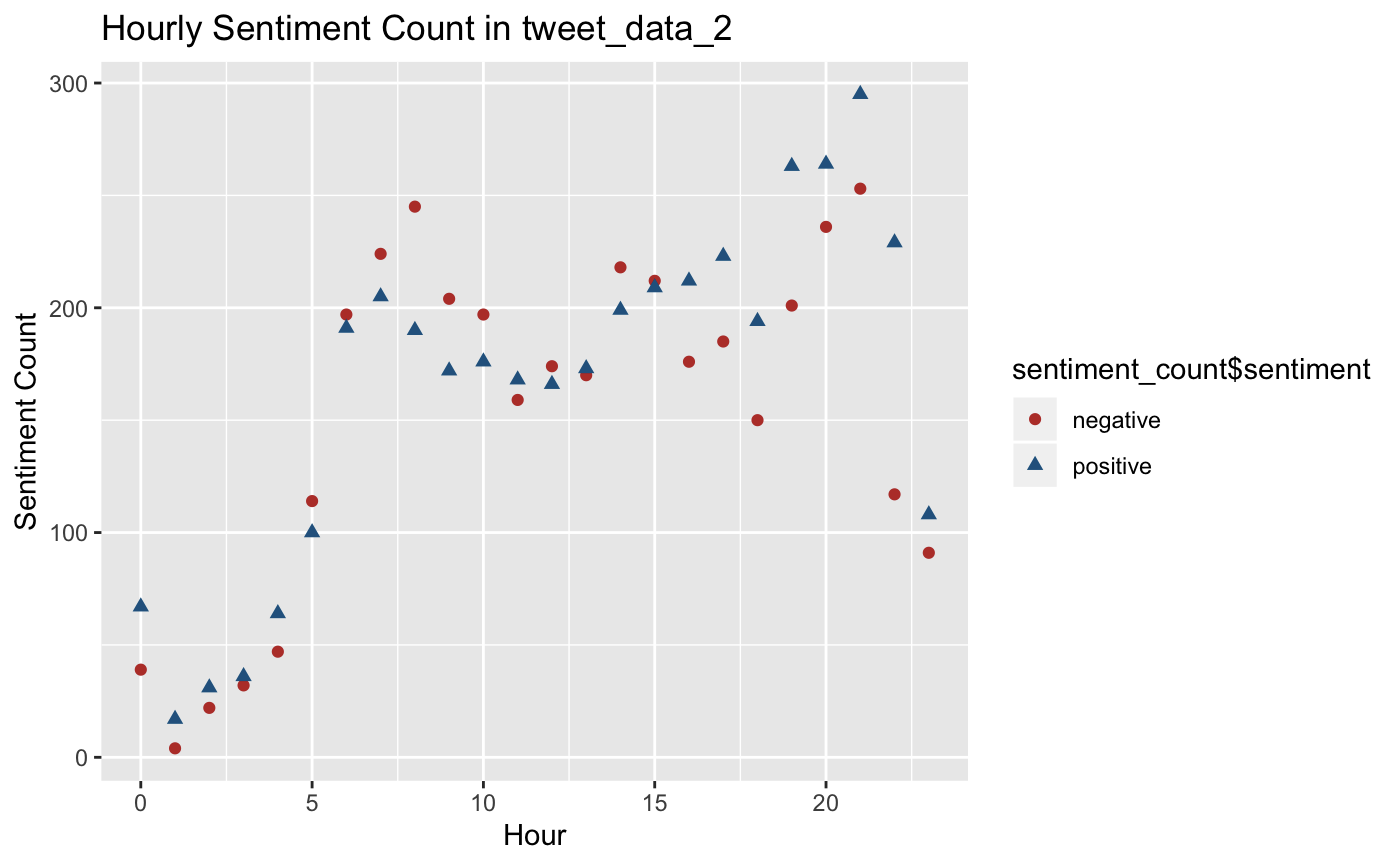


Figure 17. Hourly Average Sentiment Score for period 2 (during election/2015.6-2016.11)

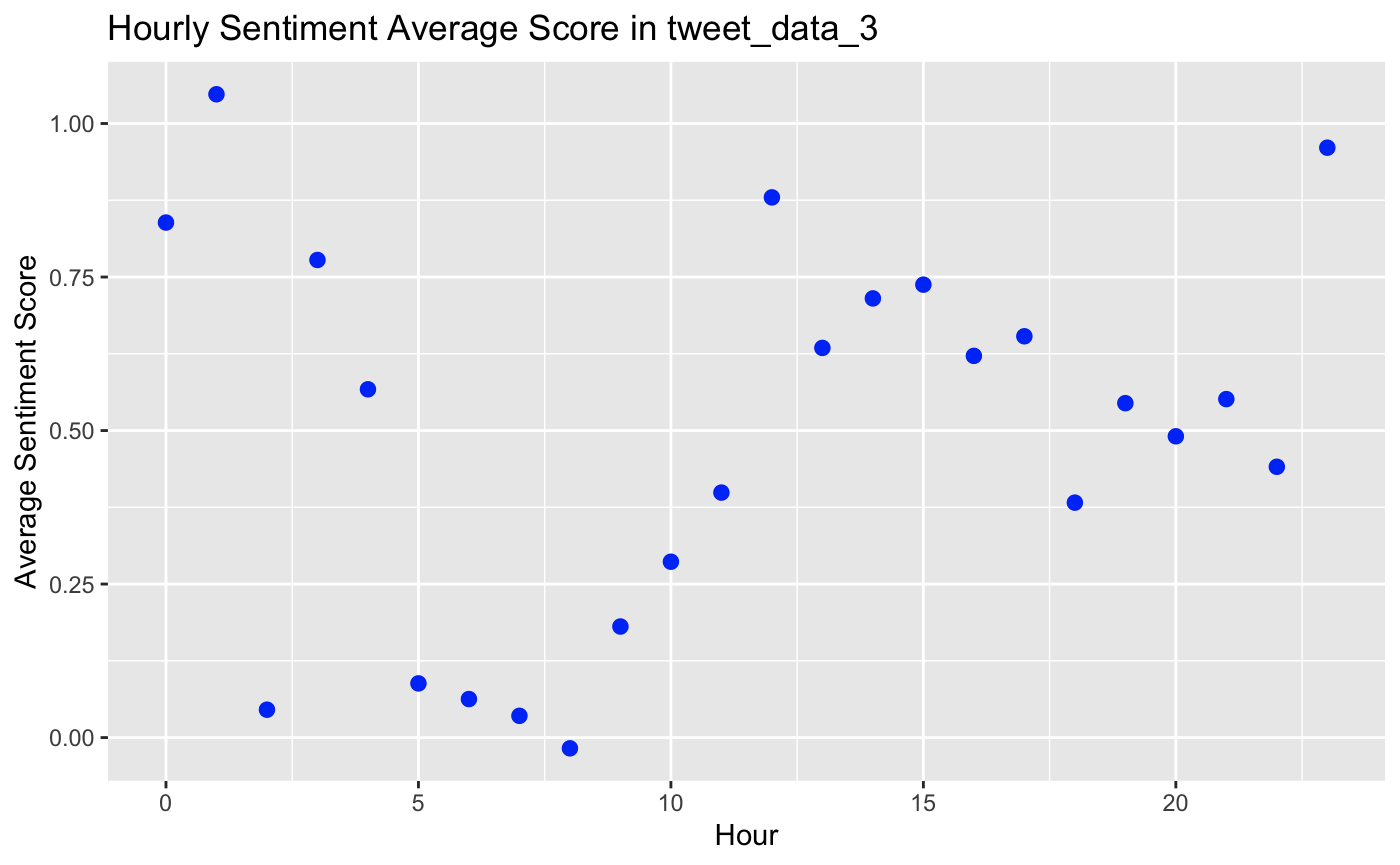


Figure 18. Hourly Average Sentiment Score for period 3 (after election/2016.11-2020.1)

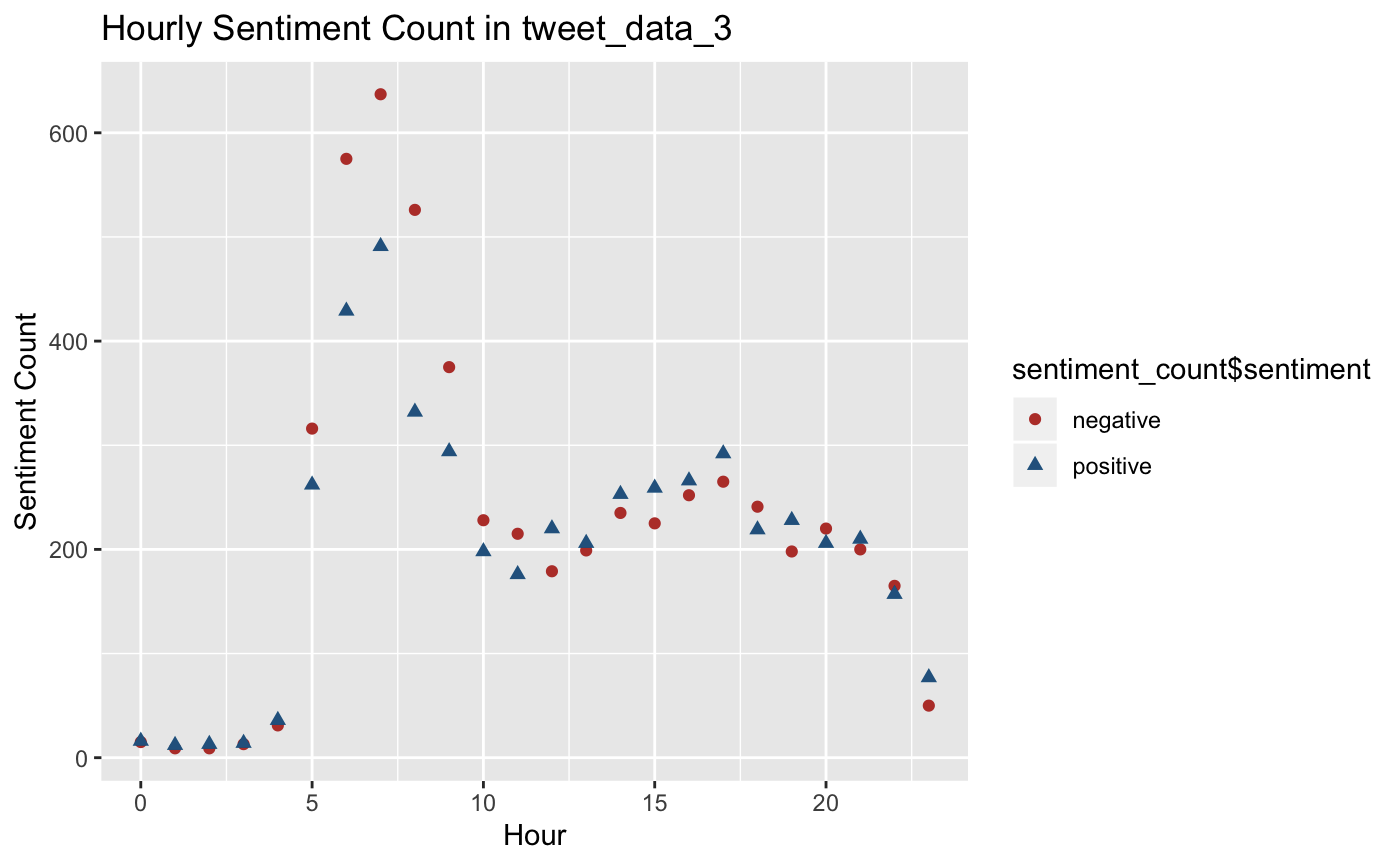


Figure 19. Hourly Sentiment Count for period 3 (after election/2016.11-2020.1)

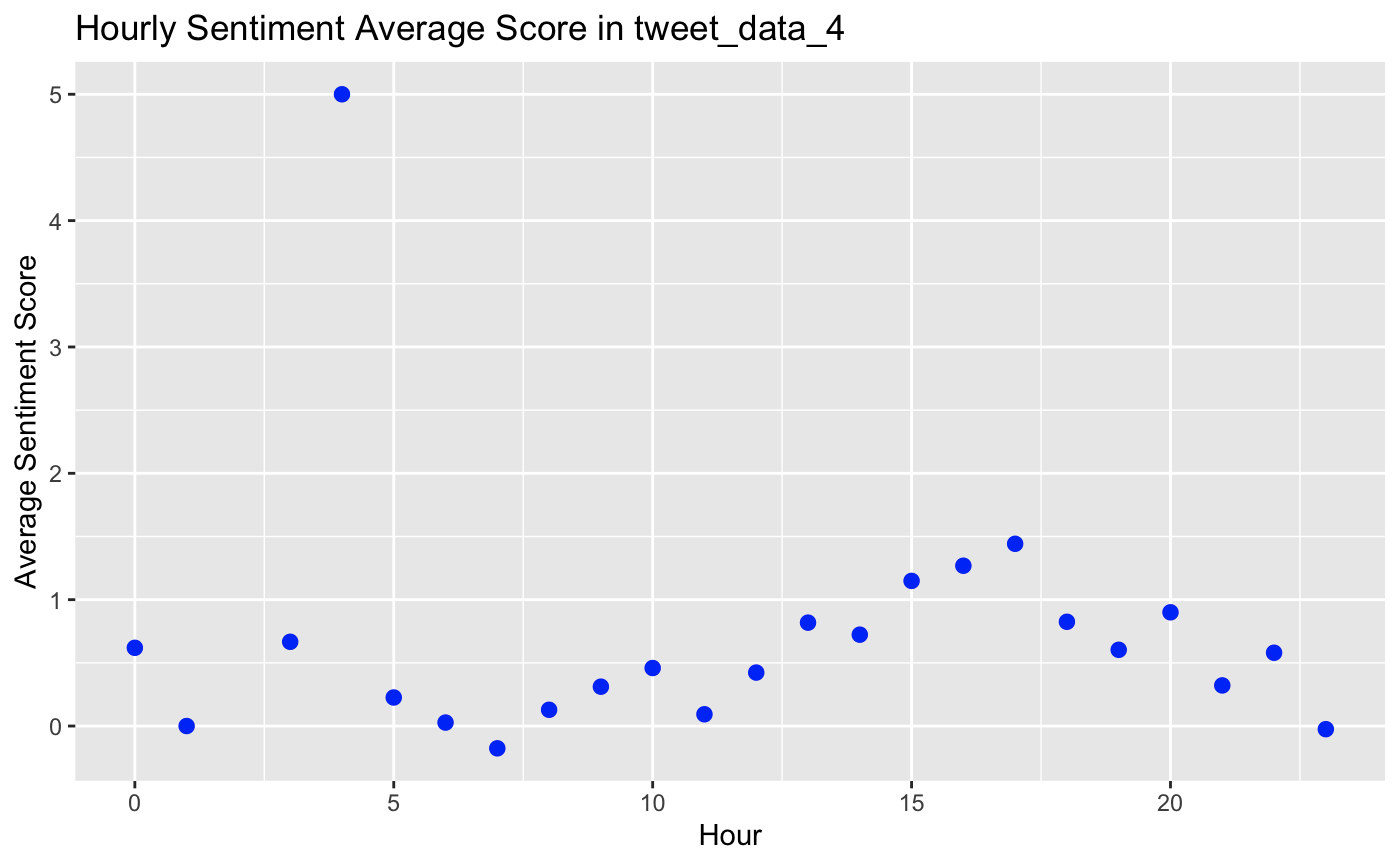


Figure 20. Hourly Average Sentiment Score for period 4 (COVID-19 period /2020.1-2020.4)

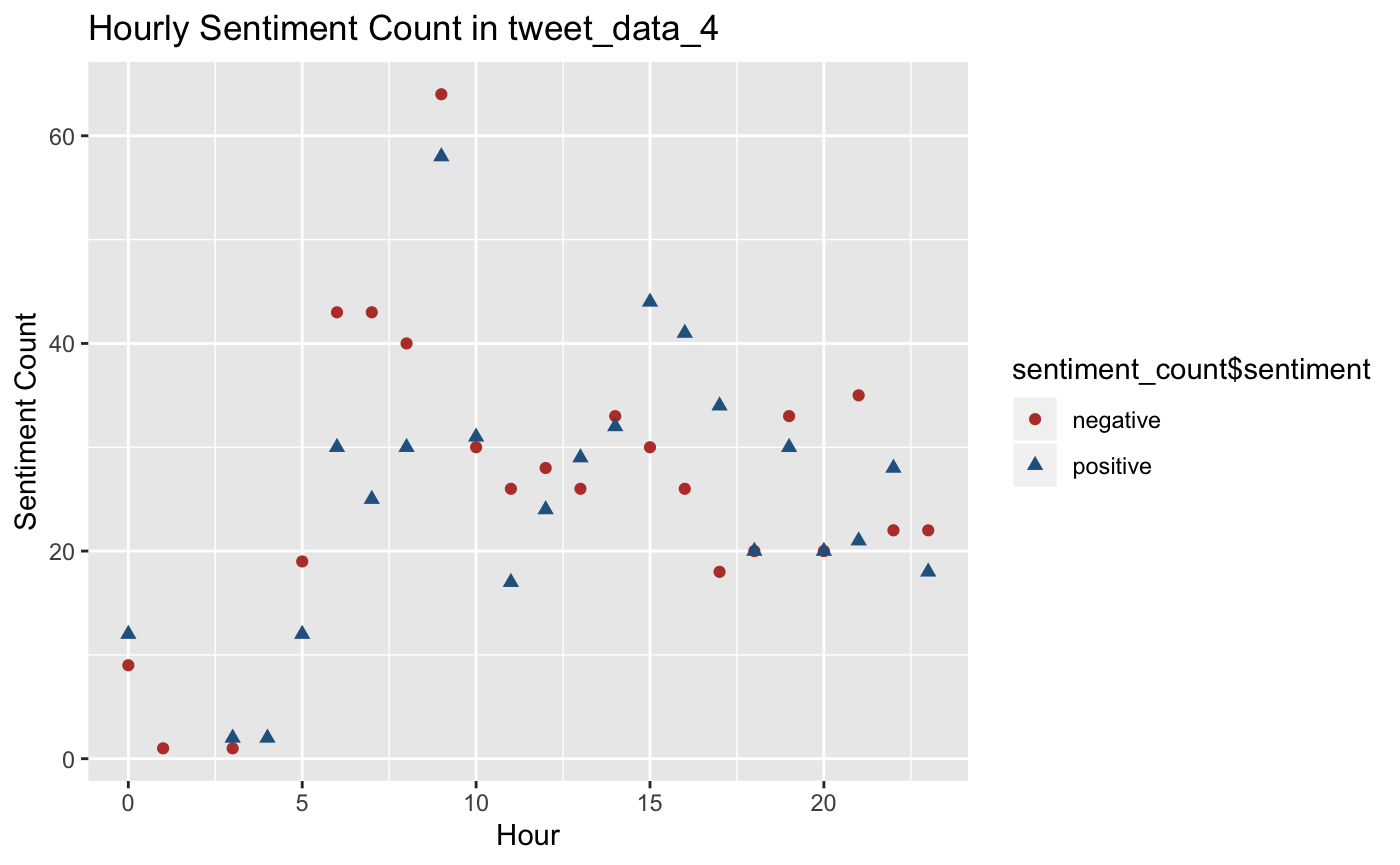


Figure 21. Hourly Sentiment Count for period 4 (COVID-19 period /2020.1-2020.4)

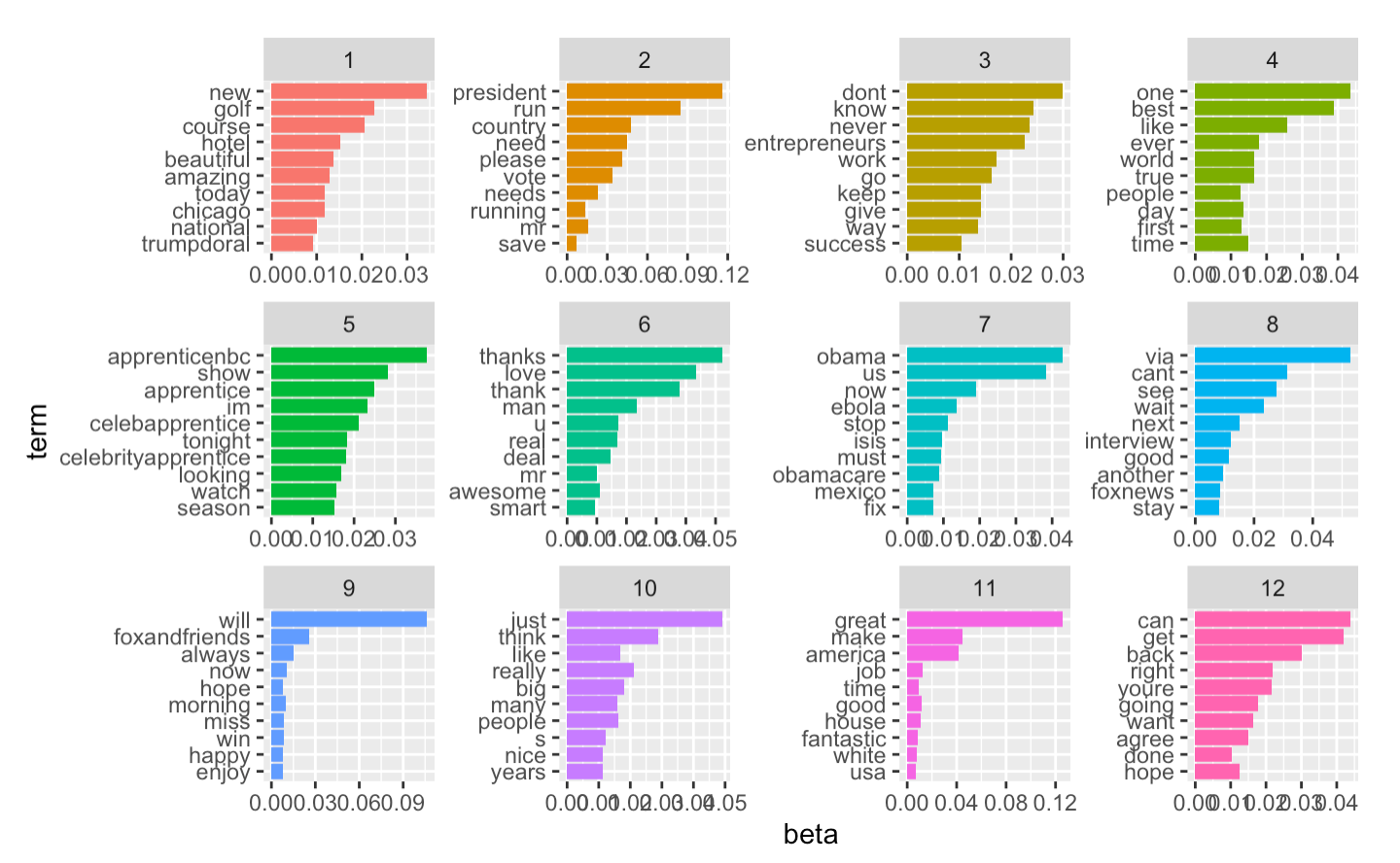


Figure 21. Top 12 topics and terms for period 1 (before election/2014.6-2015.6)

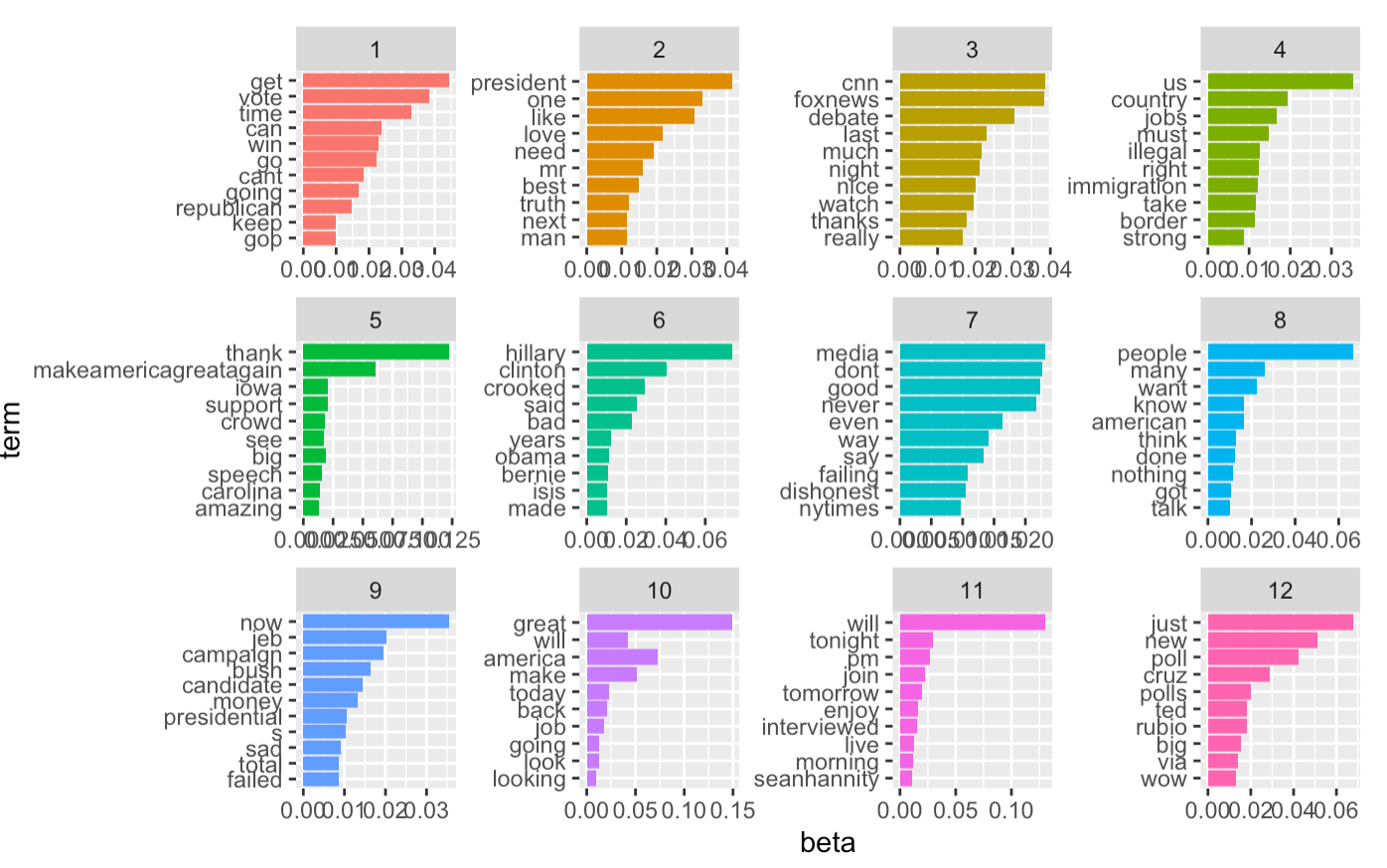


Figure 22. Top 12 topics and terms for period 2 (during election/2015.6-2016.11)

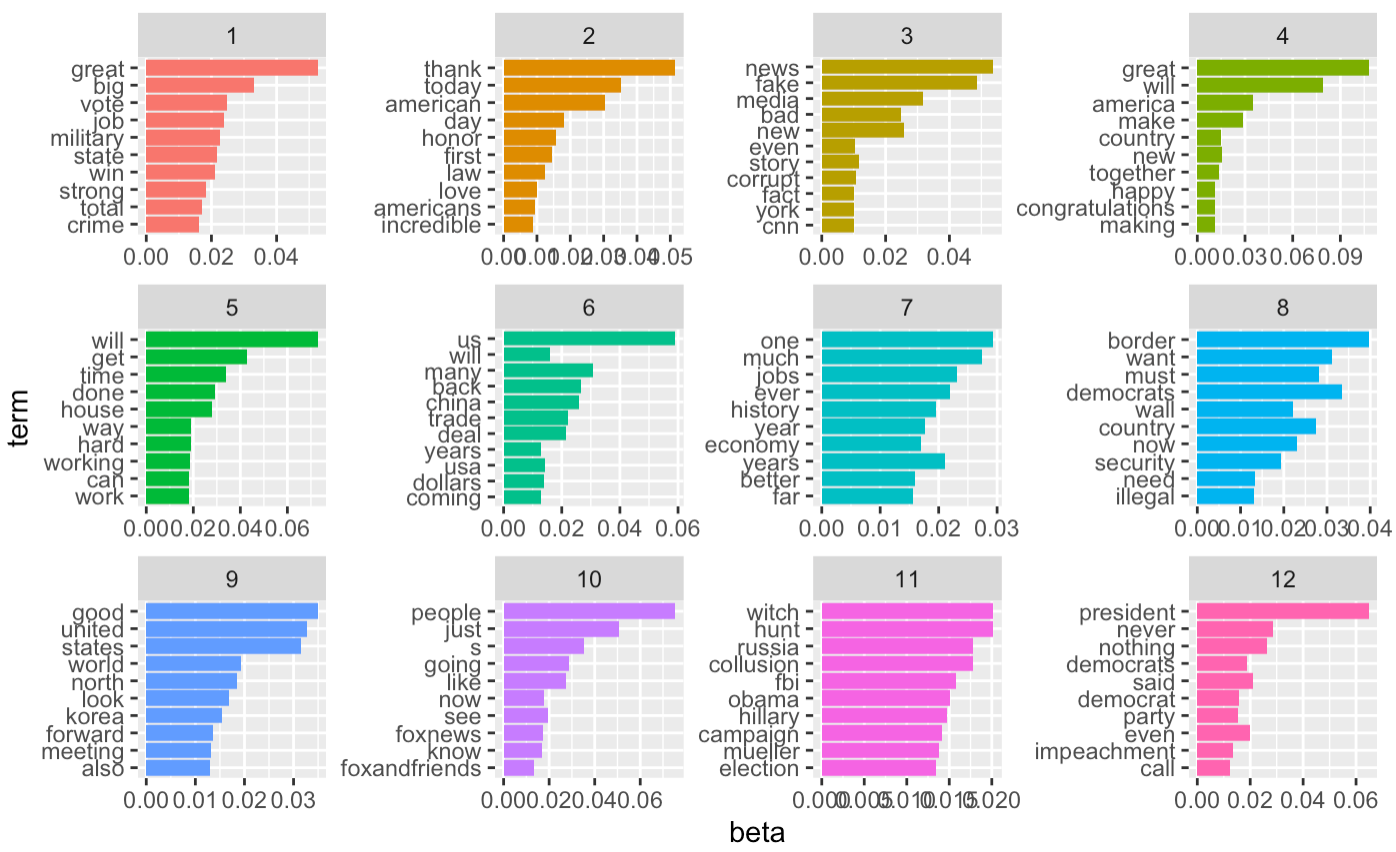


Figure 23. Top 12 topics and terms for period 3 (after election/2016.11-2020.1)

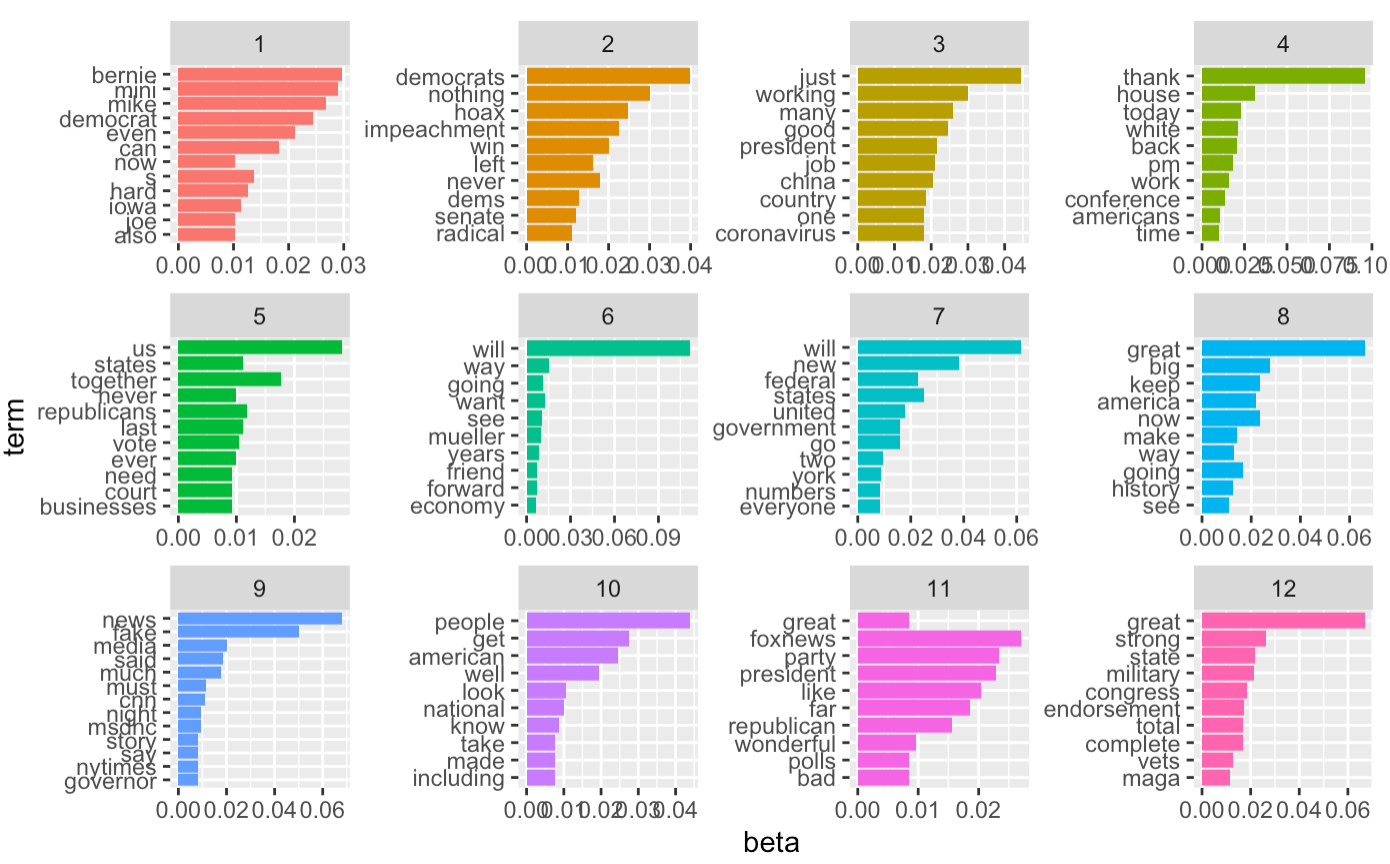


Figure 24. Top 12 topics and terms for period 4 (COVID-19 period /2020.1-2020.4)

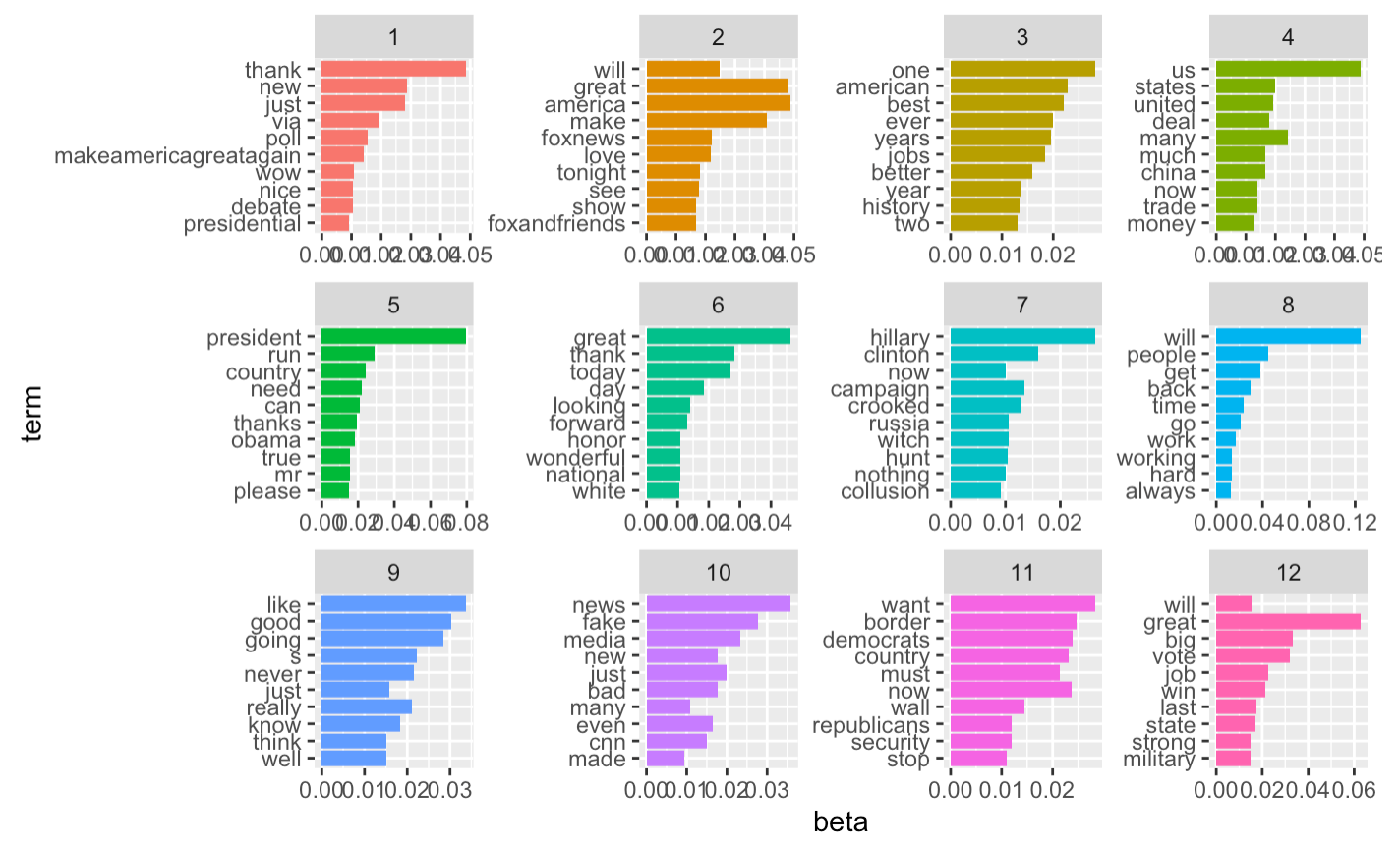


Figure 25. Top 12 topics and terms for all period (2014.6-2020.4)

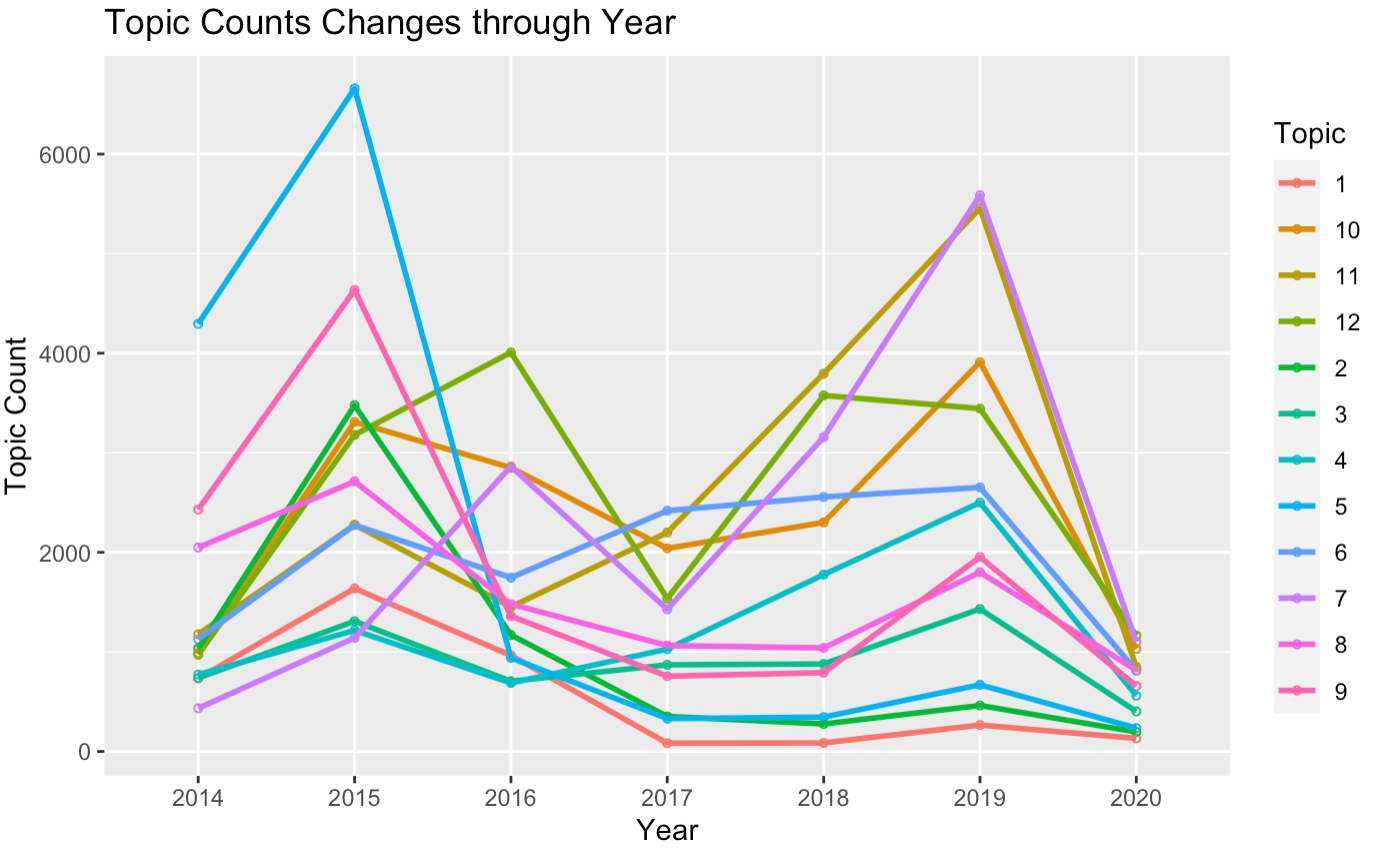


Figure 26. Primary topic counts through years (2014.6-2020.4)

References:  
1. 260,000 Words, Full of Self-Praise, From Trump on the Virus, Retrieved April 26, 2020, from

<https://www.nytimes.com/interactive/2020/04/26/us/politics/trump-coronavirus-briefings-analyzed.html>

2. Blei D.M., Ng A. Y., and Jordan M. (2003, March). I. Latent Dirichlet Allocation. Journal of Machine Learning Reseach[J]. 993 - 1022

3. How Trump Reshaped the Presidency in Over 11,000 Tweets, Retrieved November 2, 2019 from <https://www.nytimes.com/interactive/2019/11/02/us/politics/trump-twitter-presidency.html>

4. Kaggle, (2020), Trump Tweets, Retrieved from <https://www.kaggle.com/austinreese/trump-tweets>

5. Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREc (Vol. 10, pp. 1320-1326).