Sentiment Analysis of Trump's Tweets Before Election

Zhenmin Li

October 20, 2020

Abstract

2020 US presidential election is one of the impactful events in the world. And whether the current president, Donald Trump, will win the election and still serve as US president is hard to predict. In this case it is important to know his attitude toward the 2020 election. This project visualized the result of sentiment analysis of Donald Trump's Tweets before 2016 and 2020 election.

1 Description of Dataset

For this project, I will use Donald Trump's tweets in the year of 2020 and 2016 before election. I will only use two properties: content of the tweets for data mining and date for filter. I will probably use Twitter API to get more data if I need some other properties (e.g. who will be retweeted by Trump).

https://www.kaggle.com/austinreese/trump-tweets

This dataset has 20 versions till now. The latest version contains tweets from 7/18/2020 to 10/8/2020. I cannot download the previous version now so I cannot get the data prior to 7/18 in 2020. In the previous version it use a different format which contains all the trump's tweets from 2019. So I can get the data in 2016. It seems Kaggle doesn't provide it anymore (1 month before it can be downloaded) but you can get it in the "source_data" folder.

2 Data Exploration

2.1 Data Cleaning

I parsed the time in the data because we want to use Vcorpus, which has a strict requirement in data formatting. I only need tweets from 07.18 to 10.17 in 2016 data.

Then I removed the line end, amp, url and icon from the text of tweets. To Normalize the corpus I changed uppercase to lowercase, removed numbers, stop-words, punctuations, and strip the words.

For the corpus for sentiment analysis I also removed the name of candidates, otherwise these names will have the top frequency. The word ”sleepy” is also removed because Trump like to use the phrase ”sleepy joe”.

The Top 20 words for each corpus is in figure 1.

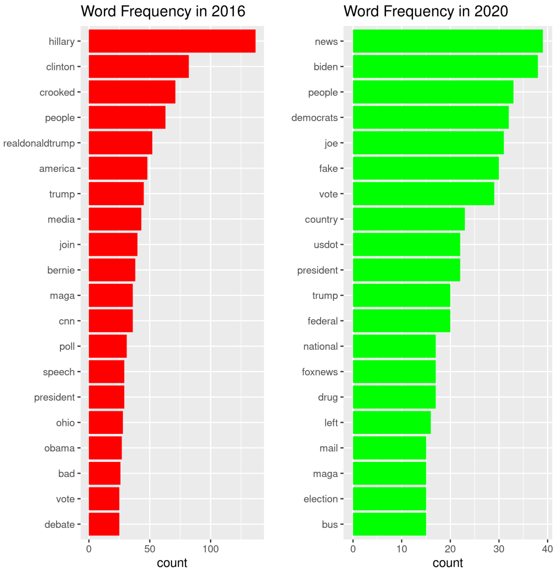

Figure 1: Top 20 words with or without names

And I also made some wordclouds for corpus in figure 2 and 3.

2.2 Sentiment Analysis

First I use bing. Bing is a corpus with words labeled positive and negative. The histogram of top positive and negative words is in figure 4.

Then I plot the sentiment of the twitter in time series (figure 5). You may notice some lines are not connected and there is a obvious long straight line in latter half of sentiment 2020, this is because lacking of data in that period.

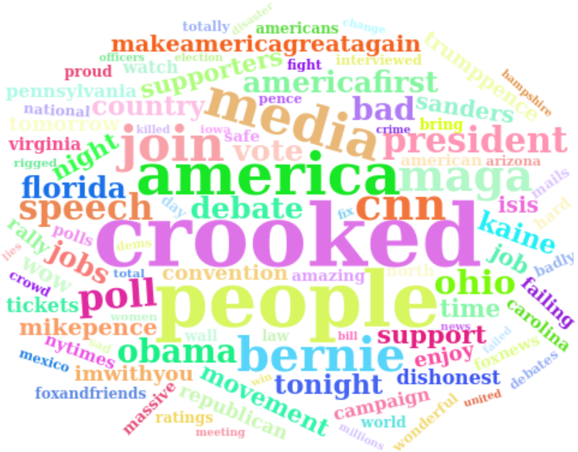

Figure 2: Wordcloud for 2016 tweets

3 Conclusion

In 2016 Hillary Clinton appears in Trump's tweets far more than other words. In 2020 it seems Biden didn't got enough attention as Hillary. His top focus now is fake news.

Trump's tweets most times stays neutral. In 2016 his attitude may be more positive than what it shows in the figures because of the word "crooked", which he use to describe Hillary, take accounts some of the negative sentiments.

It seems there's an important date in near August 12 which makes his attitude positive, but I cannot figure it out. I will inspect the text that day later.

Despite many hard times in 2020, it seems Trump's feeling is not greatly impacted.

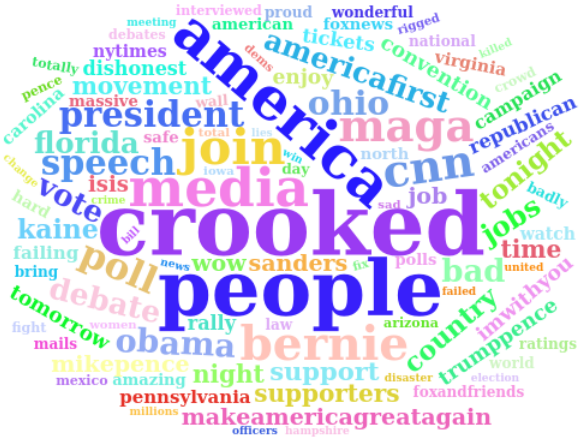

Figure 3: Wordcloud for 2020 tweets

4 Suggestions for Further Analysis

An obvious defect is that tweets of 3 month are not dense enough for time series analysis. I cannot get enough tweets because Twitter's API (I tried tweepy) can only provides limit amounts of tweets each month which is not enough for analysis. The dataset from Kaggle is better than API but the amount is still limited. I think expand the data scale to 1 year may help identify more findings.

I also want to know what really happened when dramatically changes occur in time series figure. And how long will the impact of these events to Trump's sentiment last.

Another potential improvement is to analyze the sentiment with the words that have high frequency (e.g. fake news) and check Trump's attitude.

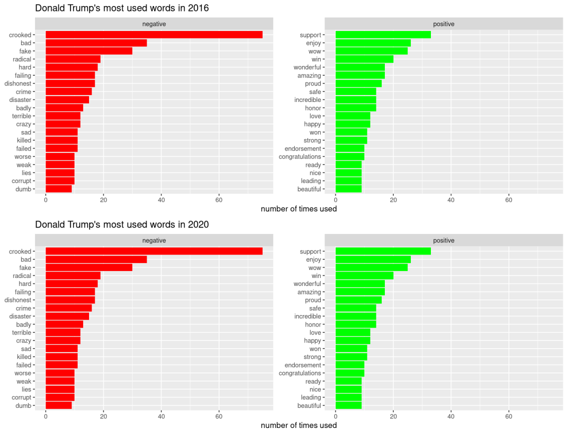

Figure 4: Most used positive and negative words for 2016 and 2020

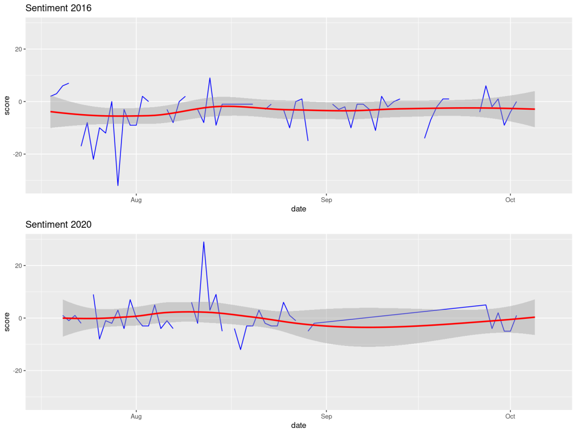

Figure 5: Sentiment analysis in time series using "bing"