Analysis on Donald Trump's Tweets

Chen Yuxuan

July 18, 2018

1 Introduction

Our research is to do sentiment analysis on Tweets of Donald Trump, the 45th and current President of the United States, whose Twitter account is @realDonaldTrump, which means we try to determine the attitude and emotion of Donald Trump's Tweets. Our method includes extracting twitter data of Trump (All 34,335 Tweets until Monday Jul 16 08:56:58) using Perl with module Net::Twitter, doing basic statistics and sentiment analysis on the corpus, and finally visualizing our results using multiple data visualization tools.

In this report, we will include different kinds of sentiment analysis: including positive-negative word frequency analysis, racism word frequency analysis, swear word frequency analysis, markov chain generation. And we also use Python to do part of the visualization.

And finally, the reason why we are choosing this research question is that there are many criticisms that Trump's Tweets contain much negative, anti-social, racism contents, and our research will analyze Trump's Tweets to check whether his tweets contain many negative emotions, anti-social expressions or racism contents. We will try our best to analyze the data in a neutral way.

We also published the dataset, codes and this report on GitHub and Kaggle so that everyone can get access to it. [1][2]

2 Corpus

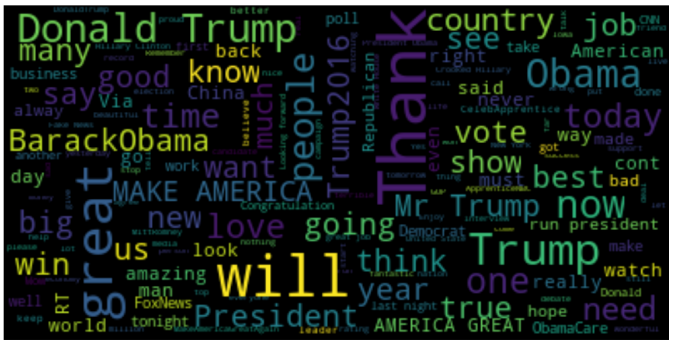


Figure 1: Word Cloud of Trump's Tweets

The Corpus we crawled from Twitter using module Net::Twitter[3] (crawl.pl), and we crawled 34,335 Tweets (TrumpTweets.xml) until Mon Jul 16 08:56:58, the raw data contains 'source', 'data', 'text', 'id' and different notations with the size 9,991,442 Bytes.

Using Regular Expressions (text.pl), we extracted the Tweets from the raw data and the size is 3,730,150 Bytes (puretweets.txt).

From the Word Cloud Algorithm[4], we used Python (wordcloud.py) to build a Word Cloud as Figure.1. We can see that Trump mentioned a lot about himself, "Obama", "great", "America", "people", "true", "FoxNews", "vote", "amazing".

3 Methods and tools

3.1 Positive-negative word frequency analysis

We have used two dictionaries [5] of Positive Words (positive.txt) and Negative Words (negative.txt) to calculate the frequency of the words in the corpus, and we drew the Word Cloud of the results.

3.2 Racism word frequency analysis

We crawled the data of Racial Slur Database [6] (racism.txt) to calculate the frequency of the words in the corpus, and we drew the Word Cloud of the results.

3.3 Swear word frequency analysis

We wrote many swear words (swearWords.txt) to calculate the frequency of the words in the corpus.

3.4 Word Cloud Analysis

We used the Word Cloud Analysis (wordcloud.py) to draw Word Cloud Graph to get a better understanding on the dataset.

3.5 Sentiment Analysis

We used Python to do Sentiment Analysis on the dataset, and we have drawn diagrams to have a better understanding on the results.

3.6 Tweets by Time

We will redo the above analysis to check the differences of the results by Time (montext.pl, tuetext.pl, wedtext.pl and more). The time period we choose contains a week, a month or a day.

3.7 Markov Chain to generate Trump-like-Tweets

This part, we used Markov Chain (markovchain.pl) to generate Trump-like-Tweets. Our method use 5-gram with 1000000 times of iterations for generation.

4 Findings

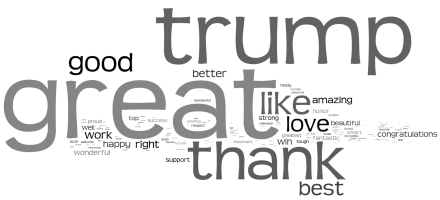
Using the method of 3.1, we find out that Trump Tweets have 13,060 negative words and 29,068 positive words. We can see from Figure.2(a) that Trump mentioned a lot of "Fake", "Bad", "Illegal" which are used to criticize the Liberals and the liberal media.

And we can learn from Figure.2(b) that the positive he mentioned are "great", "Trump", "work", "love", "thank", which are often used in Trump's policies, and traditional American Value.

Using the method of 3.2, 3.3, we can get there are 114 swear words and 0 racist words. As shown in Figure.2(c), "Ass" is the word Trump uses the most often, and he sometimes uses other swear words to express his emotions.

And Trump did not use racist words in his Tweets, this might be his PR team has drawn a clear line about what can be sent and what cannot.

(a) Negative Word



(b) Positive Word

(c) Swear Word

Figure 2: Word Cloud

And in 3.4, we did sentiment analysis on his tweets, we found out that most of the Tweets' sentiment are neutral.

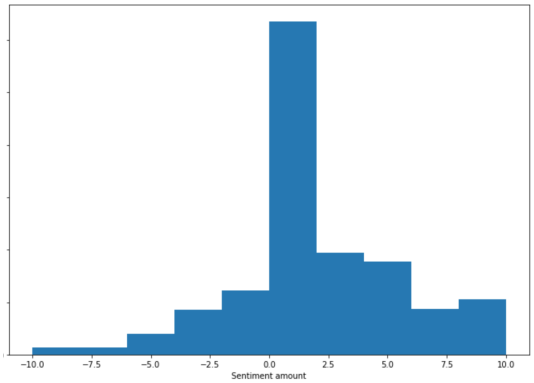


Figure 3: Sentiment Analysis

And using the method in 3.5, we extracted the data of weekdays, different time of a day, and do the

Sentiment Analysis on the dataset again by time. We can learn from Figure.4 that Trump is most negative on Thursday and Sunday, and he is most negative between 10:00 to 13:00.

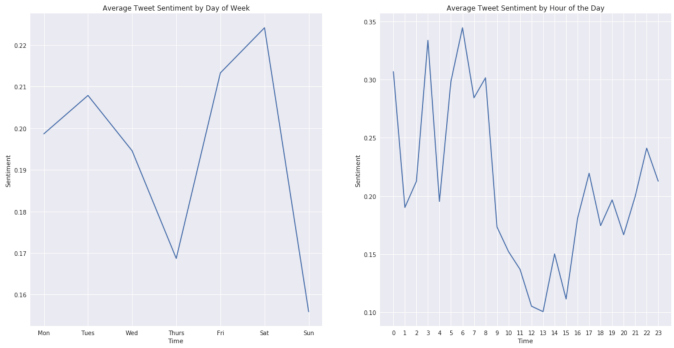


Figure 4: Sentiment Analysis by Time

Here are some Trump-Like-Tweets examples using method of 3.7:
"I can't want to beat Hillary!"
"I will go nowhere. He should change."
"Wow, the Fake News Media, and I feel is going to blow them to vote for Trump!"
"I am very proud!"
"Now arriving jobs back to Washington!"
"I can't understand what is going to the Law Enforcement!"
We can see that the results have many exclamation marks, and the results are quite good and similar t Trump's Tweets.

5 Limits

5.1 Limit On The Data

The dataset we crawled have only short messages according to the limits of the length of Twitter. If we can get access to all the speeches that Trump have given, we can draw more conclusions. And the Tweets might be edited by the PR teams, and some sensitive contents might be deleted by himself, his PR team or the Twitter Official Administrators.

5.2 Limit On The Methods

We have tried to use Perl as much as we could, including the crawler, lexer, parser and Markov-Chain generator. However, Perl is not as useable as it is in analyzing texts in visualization. So we turned to Python to do all the data visualization part. And this dataset is only around 10MB, it is allowed to do all the analysis in a PC and not using Data parallelism. If we are trying to analyze huge amount of data, we can use MapReduce or other Distributed System Algorithms to proceeded with our results.

Our method can be extended with other dataset as well. However, for Japanese, Chinese contents that the sentences haven't been divided into words, Perl is hard to do NLP operations on the datasets.

And the generations of Trump-Like-Tweets using Markov-Chain are limited, and we could use one2sequence Algorithm or other Machine Learning Algorithms to gain better results.

5.3 Limit On The Findings

Our findings might be biased and may not necessary show what is the true attitudes of Trump. The findings can only be a reflection of Trump that he wanted show to the people.

5.4 Future Study

We are considering tracking the comments of the Tweets of Trump, and keeping records of Trump's Tweets and finding out what kind of Tweets might be deleted. We can find out the patterns when Trump is using his iPad, his iPhone or other devices, and what contents are different from each other in different devices.

And we might switch to Python instead of Perl because Python has more modules that can be easily used, and we will keep records of everything we do on GitHub as well.

6 Conclusions

In conclusion, Trump's Tweets are quite Neutral. And he uses more Positive words than Negative words. He sometimes uses swear words and he uses "Ass" most often, and he does not use racism words or sexism words in his Tweets. The media he mentioned most is "FoxNews". His opinions might be biased, but according to the sentences and words he uses, his words are neutral-positive.

However, he uses words like "Fake", "Bad", "Illegal" to criticize the Liberals, and its supporters, related media. And he is most negative on Thursday, Sunday, and from 11:00 to 13:00 everyday.

Our future research will focus on the attitudes of the comments of Trump's Tweets. We will switch our method and use more universal tools, and we will keep open-sourcing our codes, results in GitHub[1].

References

[1] Sandy Chen (Chen Yuxuan). (GitHub). https://github.com/sandy2008/Sentiment-Analysis-on-Trump-s-Tweets-Using-Perl

[2] Sandy Chen (Chen Yuxuan). (2018, July 16). Tweets of Trump | Kaggle. Retrieved from https://www.kaggle.com/sandy2008/tweets-of-trump

[3] Mims, M. (2018, January 17). Net::Twitter. Retrieved from https://metacpan.org/pod/Net::Twitter

[4] Cui, Weiwei, et al. "Context preserving dynamic word cloud visualization." Visualization Symposium (PacificVis), 2010 IEEE Pacific. IEEE, 2010.

[5] Song, Jong-Seok, and Soo-Won Lee. "Automatic Construction of Positive/Negative Feature-Predicate Dictionary for Polarity Classification of Product Reviews." Journal of KIISE: Software and Applications 38.3 (2011): 157-168.

[6] Races. (n.d.). Retrieved from http://www.rsdb.org/