U.S. PRESIDENTIAL CANDIDATE ANALYSIS

Narvaez Betzabeth

Department of Applied Information Engineering, Global Leaders College, Yonsei University, South Korea

ABSTRACT

Since the United States presidential election is coming up soon, citizens should be aware of and know more about the political climate and presidential candidates in the United States. By conducting an exploratory data analysis on Twitter tweets regarding the presidential candidates, Donald Trump and Joe Biden, we are able to see the prevalent topics the public is discussing regarding each presidential candidate. Furthermore, by conducting a sentimental analysis on the Twitter data, we are able to see the sentiment of the public towards each candidate, and thereby be able to make a prediction of the next president of the United States.

1. INTRODUCTION

As citizens of the United States of America, we all have a duty of voting for a presidential candidate in the upcoming presidential election to decide which candidate will become the next president of the United States of America. However, people can be currently uninformed about the political climate in the United States of America and know very little about the presidential candidates in the upcoming election which leads to those people being very hesitant to vote or deciding to not vote at all. [1]

Due to this, before casting our vote for a presidential candidate we know nothing about, it is imperative to conduct an analysis on data regarding the two presidential candidates, Donald Trump and Joe Biden, to learn more about the political topics and issues they are associated with. By doing this not only can we gain a deeper understanding about the two presidential candidates, but we can also be able to see the public sentiment of fellow U.S. citizens and get an idea of what they think about the candidates and therefore be able to predict which presidential candidate will most likely end up winning and becoming the next president of the United States of America.

2. METHODOLOGY

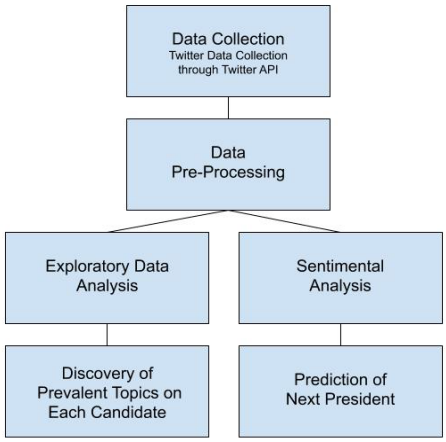

Figure 1. Diagram of Research Flow

This U.S. presidential candidate analysis project can be divided into two phases: the data collection and the data analysis. The data analysis part can be further divided into three parts: the data pre-processing part, to clean the data and get it ready to be analyzed, the exploratory data analysis part, where we will be able to see the prevalent political topics of each presidential candidate, and the sentimental analysis part, where we are able to use the average sentiment of the public on each candidate and determine which one will most likely win the presidential election.

2.1. Data Collection

The method of data collection for this project was conducted through web scrapping Twitter tweets by R. The collected data was saved in comma separated files, also known as csv files, to make the data pre-processing and data analysis more efficient.

2.1.1. Twitter

In order to collect the data, in this case the tweets regarding each presidential candidate, access to the Twitter API was required. After creating a Twitter account and converting it into a developer account, we were able to apply to be

granted permission to get access to and use the Twitter API. Our application was followed up by an email several hours later with directions on how to get started with the Twitter API.

For obtaining the Twitter tweets with the Twitter API, we used R and the R packages 'rtweet' and 'ROAuth' to contact the API and extract tweets from Twitter. To get tweets regarding the two presidential candidates, we used the two hashtags #Trump and #Biden. We decided that these two hashtags would be the least biased towards the presidential candidates since if we would have used hashtags like #Trump2020 or #CorruptBiden, the data collected would have been too biased and skewed in support or opposition of a candidate. By using the generic #Trump and #Biden, our data will include all sides of the political discourse, those in support or opposition of Donald Trump or Joe Biden.

For our data sets of tweets, we wrote in our R code to collect 10,000 tweets for each hashtag, however after looking at our saved csv files of the collected data, only about 7,900 tweets per hashtag were able to be scrapped with the API. Furthermore, in order to reduce replicas of the same tweets, we used the condition 'include_rts = FALSE', meaning that we did not want to include retweets when collected the data.

After running the script to collect the tweets, the results included many fields of data that we did not need, around 51 data fields. Since we only need the text of the tweets from the collected data, we saved only the text field, which is the text of the tweet, to a csv file.

2.2. Data Analysis

After obtaining all the necessary data for the project, the next step is the data analysis. All of this was conducted with R.

2.2.1. Data Pre-Processing

Before beginning to analyze and get meaning from the data, we prepared and cleaned the data. This process included converting the data into a corpus, converting all the text to lower case, removing unknown characters which resulted from emojis being used in the original tweet, removing links, removing English stopwords, removing extra white space, removing numbers, and stemming the data which is to convert multiple words of similar origin into one word. Then after doing all of that, we converted the cleaned data into a document term matrix to be used for the exploratory data analysis.

2.2.2. Exploratory Data Analysis

In order to discover the prevalent political issues of each presidential candidate, we conducted a frequency exploratory data analysis on our cleaned Twitter data. This was done through a word cloud and bar plot on the data regarding each candidate. The word cloud included the top 50 most frequent words for each hashtag and the bar plot included the top ten most frequent words in each hashtag.

2.2.3. Sentimental Analysis

To know the public sentiment of each presidential candidate and to later predict which candidate will most likely win the presidential election, we conducted a sentimental data analysis. We found two ways to conduct the sentimental analysis, one was by using the R packages 'tidyr' and 'tidytext', and the other by using the R package 'sentimentr'.

The first way, the harder of the two methods, was conducted with the R libraries 'tidyr' and 'tidytext' and the sentiments dataset called 'bing'. The 'bing' dataset classifies whether a word is positive or negative. This process involved first ungrouping the words in the text field, counting how many times a word was the same as a sentiment in the 'bing' dataset, and then doing an inner join with the 'bing' sentiments. This gave us a list of words and their sentiment and how many times it appeared in the text. Next, due to Donald Trump's last name having a positive meaning in the dataset, we had to remove the word 'trump', otherwise it would have skewed our data result. After that, to count the overall number of positive and negative words, we created a table from the sentiment values of the result from the inner join with 'bing'. Then we turned that table into a data frame and once we viewed the data frame, we got the number of positive and negatives for each presidential candidate.

The second method, the easier of the two, only required one R package, 'sentimentr'. After loading the library, we used the function 'sentiment_by' on our Twitter text data. The function ran for a long time, then after running, yielded the result which included four data fields: element_id, word_count, sd, and ave_sentiment. The one we focused on is the ave_sentiment field, which stands for average sentiment. We ran a summary function to get the mean of the ave_sentiment field for each data set of Donald Trump and Joe Biden. By doing so, we could see the median of the average sentiment for each presidential candidate.

3. RESULTS AND DISCUSSION

After conducting the data collection, and the data analysis, we were able to see the results for each of the analyses regarding each presidential candidate.

3.1. Exploratory Data Analysis Results

The results of the exploratory data analysis gave us insight on what type of topics Americans are discussing regarding each candidate.

3.1.1. Word Cloud

For #Trump, a word cloud was used to show prevalent topics. The results can be seen in figure 2 and 3.

Figure 2. #Trump Word Cloud with 'trump'

Figure 3. #Trump Word Cloud without 'trump'

The word cloud results for #Biden can be seen in figure 4 and 5.

Figure 4. #Biden Word Cloud with 'biden'

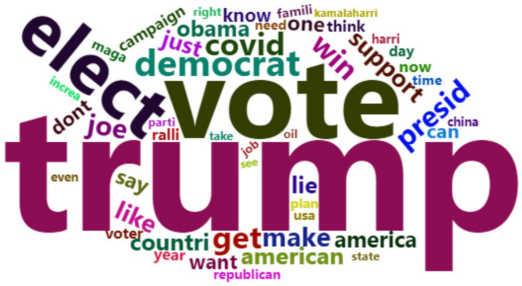

Figure 5. #Biden Word Cloud without 'biden'

I included two versions of the word clouds for each presidential candidate to show how skewed they can be if not preprocessed correctly. The first word cloud for each candidate includes the name of the candidate, which as shown is the most frequent word compared to others by a lot. In order to get rid of unnecessary information that my even cause confusion, a second word cloud is made for each candidate. The second word cloud of each candidate does not include the name of the respective presidential candidate, it does, however, include the name of the opposing candidate to show how relevant the opposing candidate is to the respective candidate.

A common theme found in each word cloud is that the name of the opposing candidate is mentioned a lot, followed by the words 'vote' and 'elect'. This demonstrates the prevalent notion of voting, that is being pushed now more than ever on to Americans. Furthermore, one clear difference that can be seen is that in the #Trump word cloud the word 'covid' is more prominent than that in the #Biden word cloud. This could be an indicator of the concern of Americans regarding how Donald Trump handled the COVID-19 pandemic in the United States.

3.1.2. Bar Plot

For the bar plots done for #Trump and #Biden, they can be seen in figures 6 and 7 respectively.

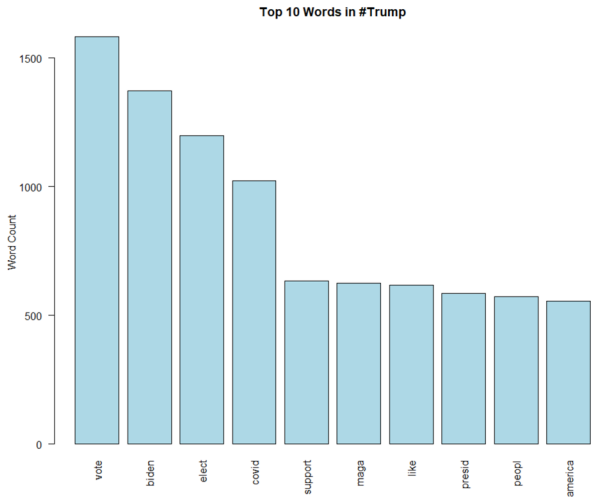

Figure 6. #Trump Bar Plot

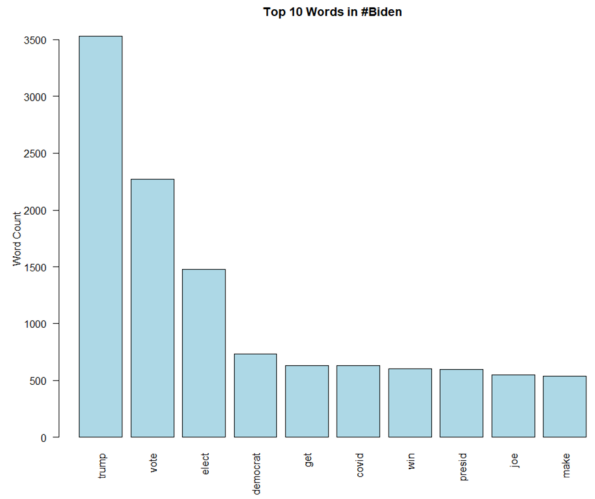

Figure 7. #Biden Bar Plot

As demonstrated by the word cloud, the first words include the opposing candidate's name, elect, and vote. Where the bar plot differs is that for #Trump the next top words are covid, support, and the abbreviation 'maga', meaning 'Make America Great Again', the slogan of Donald Trump. While for #Biden, the next top words are democrat and covid. This contrast can be indicative on how supporters of Donald Trump are only focused on him, while supporters of Joe Biden are not only focused on Joe Biden, but they are also supporting the democrat party in general.

3.2. Sentimental Analysis Result

The result of the sentimental analyses gave us an indication of the public sentiment on each presidential candidate. This public sentiment can be used to compare both presidential candidates and predict which one will win the presidential election.

3.2.1. Sentimental Analysis the 'Hard Way'

The 'hard' method of sentimental analysis yielded the result as follows in figure 8.

Overall Negatives and Positives for Biden	
Variable	Frequency
negative	1320
positive	580
Overall Negatives and Positives for Trump	
Variable	Frequency
negative	1417
positive	588

Figure 8. Overall Sentiment of Each Candidate

By adding the frequencies, the frequency for 'negative' would be a negative number and the frequency for 'positive' would be a positive number, we got that the overall sentiment score for Joe Biden is -740, and for Donald Trump it is -829. Comparing these two sentiment scores, we can see that Joe Biden has the more positive score, meaning he has the more positive public sentiment.

3.2.2. Sentimental Analysis the 'Easy Way'

The 'easy' method of sentimental analysis was conducted by using one R package called 'sentimentr' and by getting a summary of the average sentiment field. The result is as follows.

Summary of Biden Sentiment			
Min.	Median	Mean	Max
-0.91661	0.00000	-0.01580	0.83267
Summary of Trump Sentiment			
Min.	Median	Mean	Max
-1.61410	-0.03397	-0.04722	0.94589

Figure 9. Mean of Sentiment

By looking at the mean in each summary, we can tell the average sentiment for each presidential candidate. When comparing the two means, we can see that the mean for Joe Biden is less negative than that of Donald Trump, meaning that Joe Biden has a more positive sentiment from the public. Correspondingly, the more negative mean of Donald Trump means that he has a more negative sentiment from the public.

3.2.2. Prediction of Next President

The method of predicting the next president of the United States is to compare the results of the sentimental analyses and see which candidate has the more positive public sentiment. In both of our sentimental analyses conducted, Joe Biden had the more positive public sentiment, therefore we predict that the next president of the United States will be Joe Biden.

4. COMPARISON

Due to the nature of this analysis, once completed, we could then wait on the results of the actual presidential election and compare them to the results of this analysis. The United States presidential election was held on November 3rd, 2020. It took several days for all the votes to be counted in each state, which led to a delay in announcing a winner or even a projected winner. It was not until November 23rd, 20 days after the election, when it was officially announced Joe Biden as president-elect with Kamala Harris as the vice president-elect. This was a historically important election as it had the highest voter turnout since 1900, with Joe Biden receiving the most votes ever cast for a United States presidential election candidate at over 81 million votes.[5]

Comparing the result from this analysis and the presidential election, one can see that our prediction is correct in that Joe Biden will be the president-elect. It must be noted, though, that the votes of the Electoral College will not be casted until December 14th and officially counted on January 6th, 2021. [6] Due to this there is still a possibility, albeit extremely small, that Joe Biden might lose the presidential election if he does not win the Electoral College votes. If this were to happen, then the result of our analysis will be incorrect since we predicted that ultimately Joe Biden will become president of the United States. In order to certainly confirm whether our prediction is correct or incorrect, we will have to wait until January 6th of 2021 to confirm whether Joe Biden will become the next president of the United States of America.

5. CONCLUSION

Through this U.S. presidential analysis, we came to know some of the topics the public was discussing regarding the presidential candidates Donald Trump and Joe Biden. Furthermore, through sentimental analysis, we determined that Joe Biden had a more positive public sentiment as compared to Donald Trump. Due to this, we predicted that Joe Biden will become the next president of the United States. Additionally, we will not know if our presidential election prediction is correct or not until January 6th of 2021 as that is when the Electoral College will cast their votes and declare the next president of the United States.

11. REFERENCES

[1] Catherine Clifford, https://www.cnbc.com/2020/10/30/why-people-choose-not-to-vote.html

[2] David Locke, stackoverflow.com/questions/652136/how-can-i-remove-an-element-from-a-list.

[3] chemicalstatistician.wordpress.com/2015/02/03/how-to-get-the-frequency-table-of-a-categorical-variable-as-a-data-frame-in-r/..

[4] Sipra, Vajiha, towardsdatascience.com/twitter-sentiment-analysis-and-visualization-using-r-22e1f70f6967.

[5] Elizabeth Crisp, https://www.newsweek.com/joe-biden-crosses-80-million-votes-10-million-more-obama-got-1549612

[6] https://crsreports.congress.gov/product/pdf/IF/IF11641