U.S. Sentiment Analysis and Random Forest Prediction of the 2020 Presidential Election

Kursten Reznik
Steven Bate
Bryce Robinette
Jacelyn Villalobos

11/01/2020

Contents

1 Introduction **3**

2 Data **3**

3 Methods **5**
3.1 Sentiment Analysis 5
3.2 Electoral Vote Data 6
3.2.1 Graphical Representation 6
3.3 Random Forest 7

4 Conclusion **7**
4.1 Conclusory Notes 7

5 Extension Note **8**

6 References and Sources **9**

7 Appendix **10**

1 Introduction

A major point of interest today is the outcome of the upcoming presidential election. Attempting to predict the presidential election is nothing new, however with the advent of social media, we have a new source of information from which to draw. In this report, we perform a *sentiment analysis* on data mined from twitter. Although we may not be able to extend our findings of the sentiment analysis to the broad population, it still yields insight into how the people might vote. Indeed, it is nevertheless, one more tool that can be used in our endeavors to not only predict the outcome of the 2020 presidential election, but to understand the feelings and emotional state of the citizens involved.
We then use our sentiment analysis in ensemble with a random forest algorithm in order to predict the election outcome.

Indeed, the outcome of a U.S. presidential election not only holds great ramifications for the U.S., but many other nations as well. As a heavy-hitter on the world stage, many governments, citizens, industries, economic experts, and policy makers are affected by the outcome of the U.S. presidential election. It is the intent of this report to contribute to the ongoing practice of statistics, and the methods used in predicting presidential elections.

2 Data

We obtained data from several sources for this report. We created the main dataset from data mining twitter in the form of word text. These tweets were scraped by using relevant hashtags, such as: #trump and #biden (all hashtags used can be found in the appendix). In doing so, we were able to form a collection of the most common words used, as well as the text that was to be analyzed for our sentiment classification.

We also obtained electoral college vote data for each state from 1976 up to the 2016 election. We curated the data we pulled in the endeavor to help us further understand how a state might vote in this election cycle. It also yields information that we may use in graphical representations of how the U.S. states have voted in the past. Furthermore, this dataset allows us to see if there may be trends on how a state has voted, i.e: if a previously red state started voting blue in more recent elections.

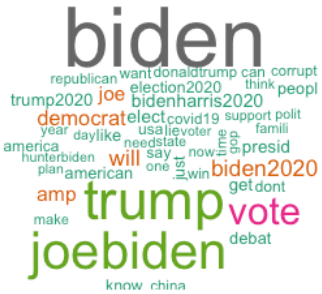


Figure 1: Word cloud: Biden

We then created databases in MySQL and performed our statistical analysis using the

programming language R. Furthermore, we imported our data into Tableau in order to create relevant visualizations as to the nature of some of our data.

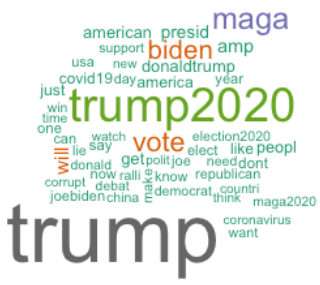


Figure 2: Word Cloud: Trump

Note:
The results of our sentiment analysis may not be adequate enough to extended to the entire U.S. population. Indeed, we mine one social media platform, Twitter, which is not used by every citizen. However, the demography of users may be sufficient enough to make conclusions about the larger population if the numbers fit the criterion of those who wish to use it. The age demography of the user data we retrieved from Twitter is as follows:
• 37% of 18–24 years of age. • 26% of 25–30 years of age.
• 22% of 30–49 years of age. • 15% of 50–64 years of age.

Note one more time:
We are retrieving data from a social media platform that not everyone uses. And even if an individual has Twitter, they may not be into politics and/or writing political hashtags, from which our data is derived. However, the percentages of the younger populations may indeed be relevant. The percentages of 37% and 26% may be considered ample sample sizes, but for the purposes of this report, we do not consider the data sample size's possibility for extended purview of the greater population.

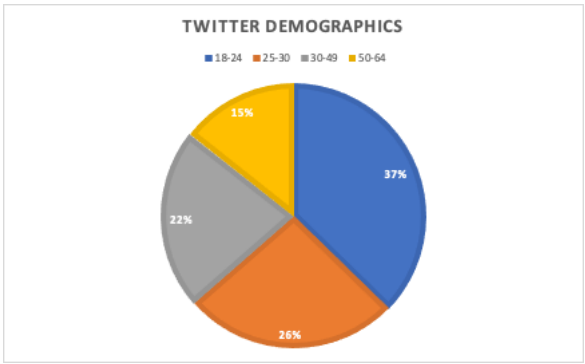


Figure 3

3 Methods

3.1 Sentiment Analysis

Sentiment analysis refers to using natural language processing, text analysis, and computational linguistics to assign values to the words that we use, so that we may ascertain whether the writing is of a particular nature of interest. That is to say, does the text have a positive, negative, or neutral connotation to it. This is referred to as *polarity* classification. In this report, we perform *beyond polarity* sentiment classification, which also takes into account emotional states. These include anger, surprise, joy, disgust, and so on.

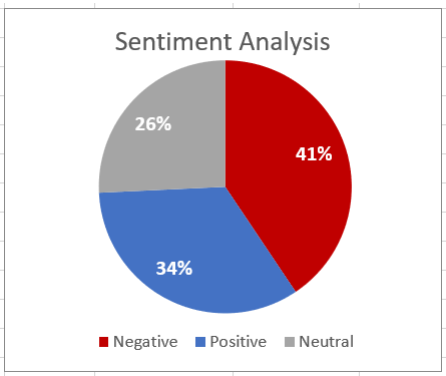


Figure 4

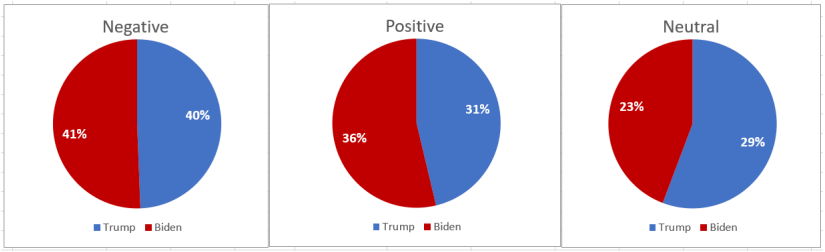
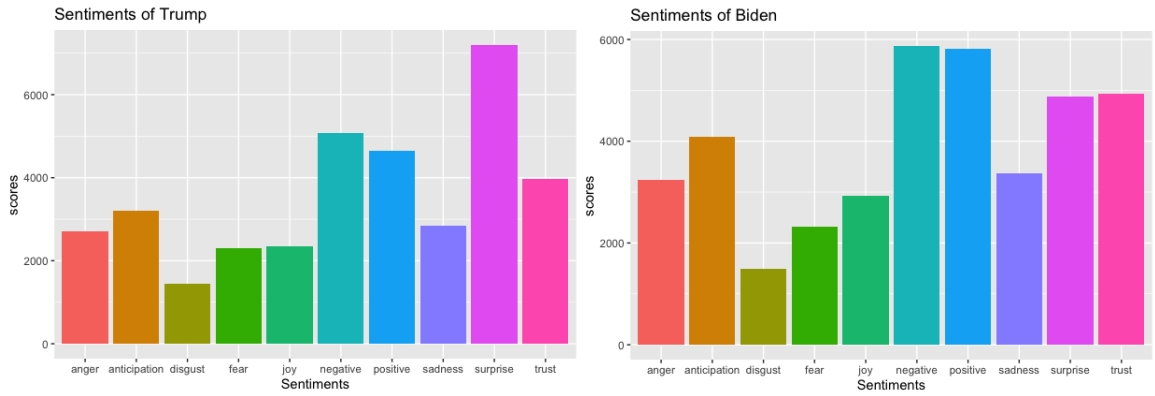


Figure 5: Age Demographic

After scrubbing 10,000 tweets per day, totaling at 75,368, and assigning value to the text, we end up with the resulting sentiment figures:



(a) Trump

(b) Biden

Figure 6: Presidential Sentiment

We can see that Biden has more positive sentiment than Trump. The sentiment analysis results in very similar outputs for both candidates with the demographic that we extracted our data. It should be noted that if we were to perform this analysis on subgroups by age, we could see different results.

The sentiment analysis does not necessarily predict who is going to become president, but is very important to gain insight of the emotional state of the voting public. The analysis performed shows the election is going to be a close race. The percentages of all three categories (negative, neutral, positive) show no significant difference between Trump and Biden. The word-cloud gives an interesting graphical representation of the words that appeared more frequently than others.

3.2 Electoral Vote Data

As was stated in earlier sections, we were able to obtain electoral college vote data from 1976 - 2016. It is the purpose of this data to give the reader insight into how states have voted in the past; and furthermore, gain insight into how the states may vote in the 2020 election cycle.

3.2.1 Graphical Representation

We were able to create an interactive map of the U.S. showing how the states voted from 1976 to the 2016 election. How the states voted is represented by the political party's color, red or blue. We have inserted a hyperlink that will allow the reader to interact with the map:

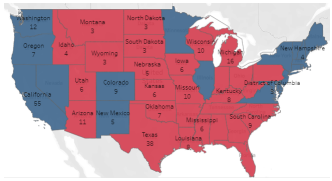


Figure 7

Click Me: U.S. Electoral College Vote Map.

3.3 Random Forest

A random forest is a statistical learning algorithm that utilizes a large number of decision trees to randomly select input features in order to predict an outcome. We utilized this algorithm in predicting the outcome of the 2020 presidential election. How we did it was we generated a model in order to predict whether or not a specific tweet belonged to either @RealDonaldTrump or @JoeBiden. We trained it on the candidates tweet information since the beginning of the year 2020. Once it was trained we ran tweets through the model, these tweets belong to other users accounts (Not the candidates). We then predicted these tweets to find who the tweets were most likely written by, either Joe Biden or Donald Trump. We worked under the assumption that if a tweet had a similar cadence, then that individual would be more likely to vote for that particular candidate.

Utilizing the random forest model on a dataset of over 20,000 tweets from all around the United States has yielded that Trump will win the 2020 presidential election.

4 Conclusion

From the various methods of analysis we were able to glean insight into the emotional state of many of the U.S. citizens concerning the 2020 election cycle. The sentiment analysis itself shows that there is no particular candidate that has a strong lead in the minds of the public. However, utilizing this analysis along side graphical representations on previous electoral vote data, as well as the random forest algorithm, has allowed us to formulate the statistical prediction that Trump will win the 2020 election.

4.1 Conclusory Notes

This election cycle, like so many of the past, has been particularly controversial. Therefore, true prediction is difficult to reign in. For example, in 1976, Jimmy Carter is elected president of the United States. In 1979, the Iran Hostage Crisis occurred, which ultimately lead to the downfall of Carter's administration. Incredibly difficult to predict events like these happen all over the globe, and they may have ramifications in our predictions. Another is Watergate, in the previous presidency to Carter's. A more modern example would state the implications of President Clinton's affair, and his subsequent impeachment. This is part of what makes statistical predictions of presidential elections almost nonviable. Yet, it makes it that much more of a goal for statisticians because of its insurmountable difficulty.

5 Extension Note

After all of our extensive research and scraping Twitter, we concluded that Trump would win the general election. However, we realized possible extensions and methods that may yield more accurate results. Instead of just scraping based off of hashtags, we could have scraped twitter using keywords. These keywords would include the hashtags as well. It is not accurate to conclude that all users will use hashtags. For example, someone could tweet "Trump is not fit for President," or "Biden is the best." Tweets such as these may not include a hashtag that our data mining program used to extract our data. However, they could be considered key components in future research. It could be difficult to ascertain results based only off of specific hashtags without using the keywords that run behind those hashtags.

6 References and Sources

Electoral College Vote Data was obtained from the *National Archives*:
https://www.archives.gov/electoral-college/results

7 Appendix

Hashtags used:

#trump
#republican
#donaldtrump
#maga
#teamtrump
#trump2020

#biden
#democrat
#joebiden
#teambiden
#biden2020