Sentiment Analysis

and some stuff or whatever

Kursten Reznik
Steven Bate
Bryce Robinette
Jacelyn Villalobos

11/01/2020

Contents

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 finding 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.

It will be written in a fancy and understandable way the totality of what we do once we have come closer to the end. As we get closer and know what will hold for the sentiment, votes, map, and random forest, I'll fill in the introduction.

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.

More Yaba Yaba, if desired.

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. 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.

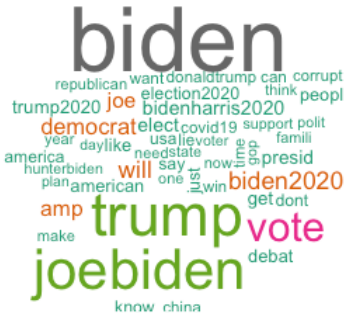


Figure 1: Word cloud: Biden

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.

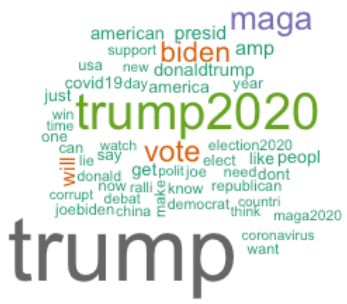


Figure 2: Word Cloud: Trump

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 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:
• 44% of 18–24 years of age. • 31% of 25–30 years of age.
• 26% of 30–49 years of age. • 17% of 50–64 years of age.

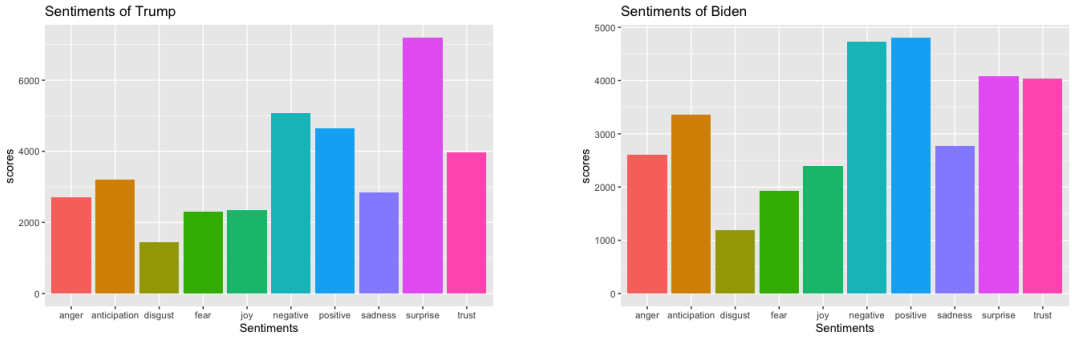
Note one more time:
We are retrieving data from a social media platform that not everyone uses. And even if an individual has twitter, they may not be into politics and/or writing political hashtags, from which our data is derived. However, the percentages of the younger populations may indeed be relevant. The percentages of 44% and 31% 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.

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.

After scrubbing 10,000 tweets (per frickin day?) and assigning value to the text, we end up with the resulting sentiment figures:



(a) Trump (b) Biden

Figure 3: Some Mother-Fuggin Figures

From this we see some stuff and say a lot more stuff about how close or not close the race may be. We could say some stuff about the relativity of the sentiments. Like, Biden has more negative sentiment than Trump, but also has more positive sentiment than Trump. Although upon further inspection, the "y" axis is not scaled so we may be accused of misrepresentation in initial viewing. Should something be said about the disproportionality of our data? I really would rather not.... That sounds exhausting, and I just want a good grade with a good report...

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 take you to the map:

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

3.2.2 LDA

Note sure this applies. We could use a weight analysis. As in, I have the percentgaes of each state since 1976 on how they voted. We could do a super simple thing where we take into account how the states voted in the past, and use that to predict how they will vote in 2020. Basic as hell, but it could be of import. From that we choose how the electoral college will conclude and compare it to our sentiment??? God its late...

3.3 Random Forest

A random forest is a machine learning method for classification regression. It constructs a number of decision trees in order to determine which classification is most likely to occur. From the different analyses that we performed, we implemented a decision tree that drew from these methods in order to for us to determine which party is most likely to win the election.

4 Conclusion

Damn dude... we did some stuff. Not all the stuff worked out the way we maybe had hoped, but its stuff nonetheless. From the results we obtained in the aforementioned sections, it is this reports predictions that.... somebody wins? is the most likely candidate to win this election cycle.

4.1 Conclusory Notes

Obviously this election cycle is f*cked... so 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 of statisticians because of its insurmountable difficulty.

5 References and Sources

6 Appendix

Hashtags used:

- #trump
- #republican
- #donaldtrump
- #maga
- #teamtrump
- #trump2020
- #biden
- #democrat
- #joebiden
- #teambiden
- #biden2020