Twitter Sentiment Analysis for 2020 US Presidential Candidates

Akshay Punwatkar (ap509)

12/9/2019

Summary

Sentiment analysis of tweets about the 2020 US Presidential Election candidates was performed. The tweets were streamed from Twitter using Spark streaming with Tweepy, and the streamed tweets were filtered based on the keywords and hashtags mentioned in the tweets corresponding to a candidate. The sentiment behind the tweets were computed using sentiment analysis tool VADER and Bag of words model. Subsequently, data were analyzed and visualized using R for further analysis.

Introduction

In the age of social media, the fight for the contest, such as US general elections, starts long before the actual event. Even before the major debates and prime-time interviews, the campaign has already begun on Twitter, which is one of the biggest social media platforms of the current age. Furthermore, with the last general elections surrounded by controversies of social media influence on the election results, the effect of such platforms cannot be ignored. **Considering these effects of social media platforms on elections, this project is aimed towards a similar analysis**. Each of the major candidates being analyzed has a plethora of followers on Twitter, and Twitter acts as a platform for these candidates to showcase their views, plans, and promises. The most followed of them all are US Senators Bernie Sanders and Elizabeth Warren, with nearly 10 million followers each. Moreover, Twitter also serves as a platform for those millions of users who follow these leaders and reacts to their plans and views via tweets. The primary objective of this project is to analyze the sentiment behind those tweets made by the general audience, and by doing so, develop a generalized view of each candidate among the general public. Sentiment analysis is a sub-field of Natural Language Processing (NLP), which attempts to recognize and extract opinions within a given text. The objective of sentiment analysis is to measure the emotions and sentiments of the writer based on the subjectivity of the text.

Primarily, this research aims towards answering the following questions:

- **What is the sentiment of the public towards the presidential candidates ?**
- **Is there any topic which contributes to a certain type of sentiments ?**
- **Are certain sentiment biased based on a certain device type (apple/android) ?**
- **How the sentiment varies across states ?**

The scope of this sentiment analysis is limited to 6 candidates mentioned below:

1. **Bernie Sanders**
2. **Elizabeth Warren**
3. **Joe Biden**
4. **Andrew Yang**
5. **Pete Buttigieg**
6. **Cory Booker**

Data

Data Streaming

The tweets were streamed in real-time from Twitter over a week using Spark streaming and Tweepy Application Programming Interface (API). Similarly, tweets made by each of the candidates over the past year was streamed using another twitter's API in Python. In total, **65,000** tweets were streamed over a week. Another **26,000** made by the **6** candidates over the past year was also streamed. Data cleaning and processing were performed in Python, followed by an analysis of sentiment using the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool provided by the NLTK (Natural Language Toolkit) package in Python and Bag of words model using different schemes. Subsequent analysis and visualization were performed in R.

Spark is an open-source cluster computing framework licensed under the Apache software foundation. Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams. Similarly, Tweepy is an open-source Python API used to communicate with Twitter using user authentication. Spark Streaming and Tweepy were combinedly used for streaming tweets, which were filtered simultaneously using specific keywords relevant to each of the six candidates. The table below shows the list of keywords used for filtering tweets. Thereafter, the tweets were processed and cleaned for sentiment analysis.

Candidate	Filters
Bernie Sanders	berniesanders, BernieSanders2020, sanders ,berni2020
Pete Buttigieg	petebuttigiegforpresident, peteforamerica, petebuttigieg, buttigieg, mayorpete
Cory Booker	CoryBooker, corybooker, SenBooker
Andrew Yang	yang2020, AndrewYang, andrewyang, yanggang2020
Joe Biden	joebiden, biden2020, biden
Elizabeth Warren	ElizabethWarren, elizabethwarren, senwarren, SenWarren, ewarren

The streamed tweet contained several pieces of information. However, only a few of those variables were relevant for this analysis. Following variables were used:

1. **Tweet**
2. **Location Information of the user.**
3. **Device Information from which the tweet is made.**

Sample observation:

"@BernieSanders We love you, Bernie, but when asked about why you went around sucking HRC's cock in '16, you literal Twitter Web App Indiana Bernie_Sanders.

For the Bag of words model, **sentiment140** data (provided by Stanford NLP) was used for training the model. The data consists of 1.6 million tweets labeled positive or negative based on happy or sad emojis in the tweet.

Data Processing

Locations from the tweets were processed to extract State codes. However, only 50% of tweets ~30,000 appeared to have relevant location information available.

In order to create a Bag of words model, the tweets were tokenized and cleaned of any **mention**, **hashtags**, **urls**, **xml/html text** and **byte code**. Subsequently, **Stemming** was performed on the tokens to only keep the root words. Finally, cleaned tokens were joined again to create a tweet.

Model

For this analysis, sentiment analysis was performed using two models, Bag of words with different schemes (CounterVector, Hashing, and Tfidf) and VADER.

BAG OF WORDS

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval. In this model, a text (such as a sentence or a document) is represented as the Bag (multiset) of its words. Sentiment analysis can be performed over the data using a classification model (Navie Bayes Classifier in this case) using the Bag of words. The process of converting the text to a bag of words can be performed using three schemes (Countervecotrizer, hashing vector, and Tfidf).

1. CounterVector - Tokenization of word to builds a vocabulary of the words generating a sparse matrix.

2. Hasing vector - Hashing of words to integers to create a sparse matrix of hashed words.

3. TFIDF - *Term frequency - inverse term frequency* method uses Td-idf value. Tdidf is the product of the *how often the word appears in a document* and *how many documents have that word*. Using Tdidf, a sparse matrix of words could be created, which can then used for classification.

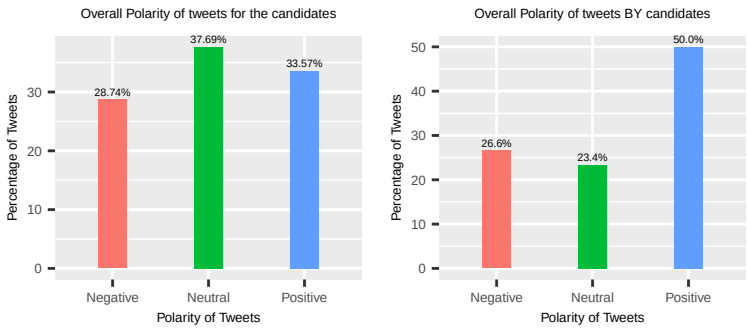
Due to the unavailability of the resources, the model could not be trained with the entire dataset (1.6 million tweets) using either of the schemes. Subsequently, using a fraction of the data to create a smaller sparse matrix resulted in poor model performance, with an accuracy of ~ 60%. Hence, this model was not used for the final analysis.

VADER

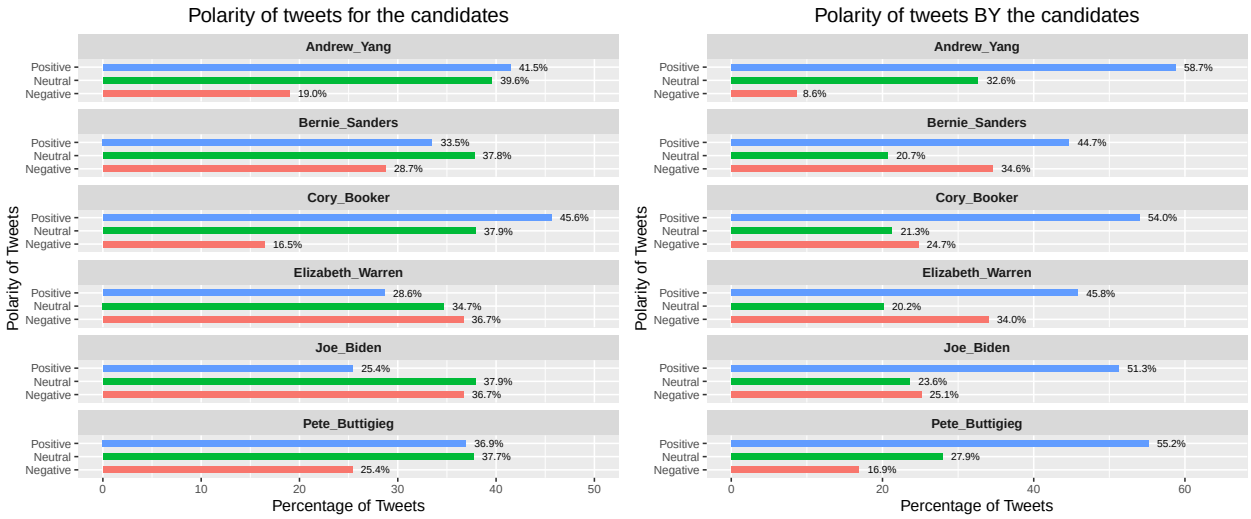
VADER is a lexicon and rule-based sentiment analysis tool that is specially developed to analyze opinions expressed in social media under the NLTK package in Python. It uses a composition of the lexicon (which is a list of lexicons, e.g., words, which are labeled according to their semantic orientation, i.e., Positive, Negative, or Neutral). For a given text, VADER provides a 4 set of values, the Compound Rating (CR), Positive Score (PoS), Negative Score (NeS), and Neutral Score (NuS). CR is the combined value of all the lexicon ratings in the normalized form, i.e., between -1 to 1, and PoS, NeS, and NuS is the measure of the proportion (probability) of a text belonging to each category. For this analysis, one of polarity (Negative, Neutral, and Positive) was assigned to the tweets using the CR values. The table below illustrates the range of CR and corresponding assigned polarity.

Results

The overall sentiment distribution highlighted that a major share (~37%) of tweets made by the user for the candidates was neutral, followed by 33% positive tweets and 28% remaining tweets. Overall the distribution of sentiments seemed balanced. However, the sentiment distribution for the candidates seems largely towards the positive sentiment. On average, 50% of the tweets made by the candidates had positive sentiment associated with them.



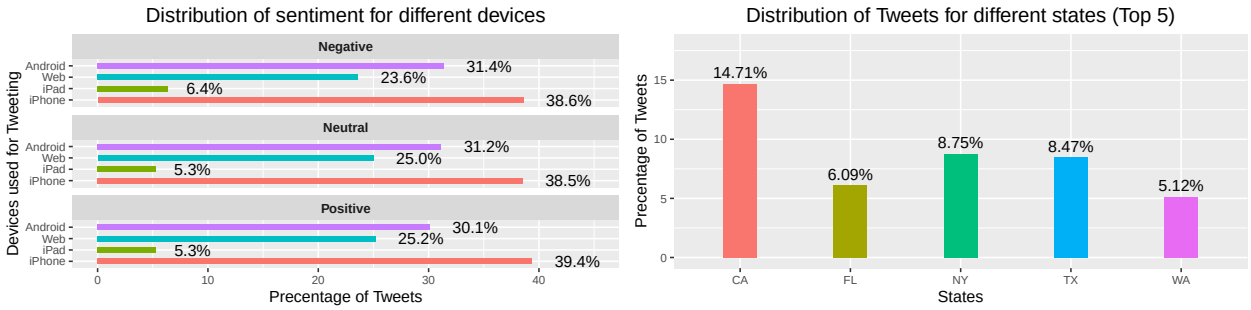
Analyzing the tweets tweeted about the candidates individually highlighted **Cory booker** had the highest share of **positive tweets** of all, with about ~**45%** positive tweets, followed by **Andrew yang** at around ~**41%**. On the contrary, **Elizabeth Warren** and **Joe Biden** appeared to have the highest share of **negative tweets** at about **35%**. Furthermore, analysis of the tweets **made by the candidates** revealed a significant fraction (45-50%) of tweets were associated with positive sentiments. However, 1/3 of the tweets made by Elizabeth Warren and Bernie Sanders appeared to have negative sentiment associated with them.



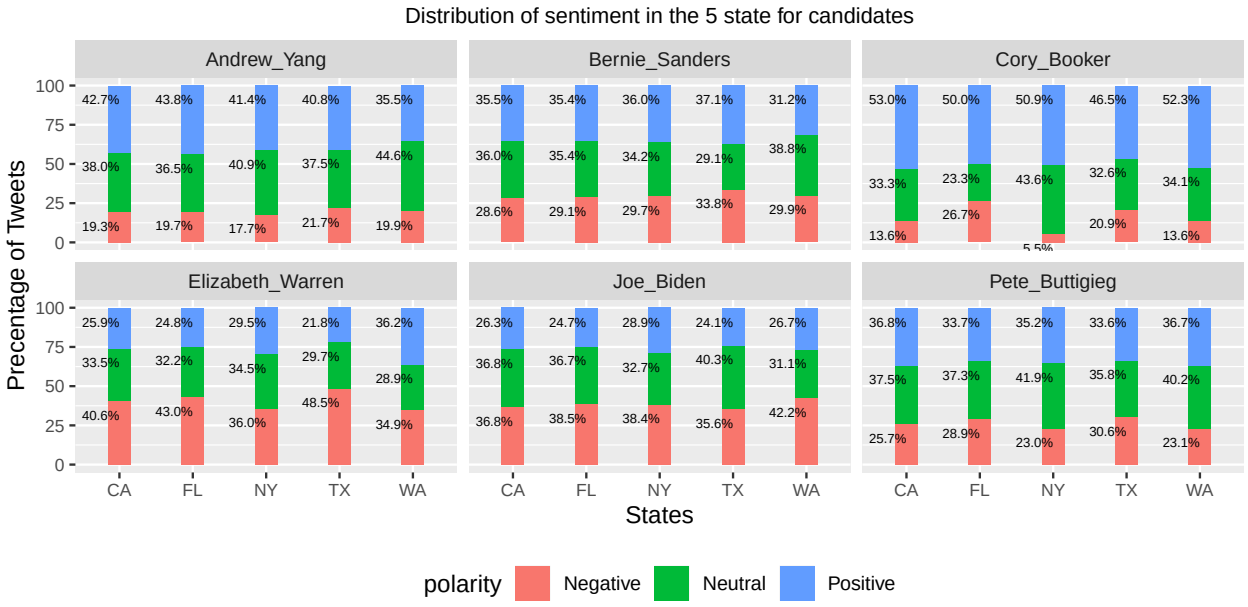
Analysis of the **WordCloud** (*Appendix*) of the most frequent words in the tweets made by the candidates provided a few interesting insights. Tweets mentioning **people**, **America**, **thank**, all of which were highly used in the tweets made by **Cory Booker**, can be associated with positive sentiment. It can be inferred that candidate talking about the American people and showing gratitude have more positive sentiment associated with them. On the contrary, candidates mentioning **President**, **Trump** such as **Bernie Sanders**, and **Joe Biden** have higher negative sentiment associated with them. Overall, it can be inferred that in the tweets made by the general public, certain topics generate a positive sentiment, while specific issues tend to change the sentiment towards negative. This might also be because of the current impeachment inquiry for the President, and the tweet might not be negative towards the actual candidate.
Few topics such as **Health care** and **gun violence** also appeared promptly in the tweets made by the candidates. This could be indicative of the popular agendas for the 2020 presidential campaigns

On checking the distribution of devices used for tweeting, it appeared Apple was the most used device followed by Android. However,**distribution of sentiment did not appear to be swayed towards a certain device.**

Analysis of the state-wise distribution of tweets, California appeared to have the highest share of tweets (~15%) followed by New York, Texas, Florida, and Washington. Texas and Florida appeared to be foremost for negative tweets. While Washington and New York were the front runners in positive tweets



Finally, the analysis of the state-wise sentiment distribution of the candidate provided a few interesting information. **Cory Booker** appeared to have the most significant fraction (~50%) of **positive tweets** from the state of New York. **Andrew Yang** also seemed to have considerable fraction of positive tweets from the state of California. Conversely, Elizabeth Warren seemed to have a significant fraction of tweets from the state of Texas towards negative sentiment. Similarly, Joe Biden also appeared to have 1/3 of the tweets associated to negative sentiment in all five states.



Conclusions

This analysis brings to light several trends among the sentiments towards different candidates competing in the 2020 US General Elections. Senator Elizabeth Warren, who has over 10 million followers on Twitter, did not seem to be much discussed in tweets, which was surprising. The most popular candidates on the platform were Cory Booker and Andrew Yang. Interestingly, Cory Booker, with just 5 million followers (comapred to other candidates), appeared to have the highest fraction of positive tweets. Also, the most popular device among users appeared to be Apple. California seemed to have the highest fraction of tweets.

The analysis provided some useful insights regarding the current popularity of the candidates. However, the study has several limitations. The trend in sentiments and the number of tweets varies day to day such that tweets collected on specific days might not reflect the actual sentimental overview of the general public. Additionally, due to the limitation of the number of tweets that can be streamed in a day, the dataset used for this analysis was relatively small. For better analysis, more data is required to be streamed over a sufficient period, especially during events such as debates and talks, followed by sentiment analysis. Also, the proposed Bag of words model could not perform well due to computing limitations, which could have provided better insights about the sentiments. Moreover, since both the models, VADER and Bag of words, translate the sentiment based on the lexicon value and does not consider the order of words in a sentence, the sentiment might not be accurately translated.

Analysis using a recursive deep learning model could be done in the future to analyze the sentiment more accurately. Moreover, the tweets could be classified using topic segmentation, which might be able to provide better insights into the topic that carries a positive or negative sentiment.

Git repository - https://github.com/akshaypunwatkar/IDS702

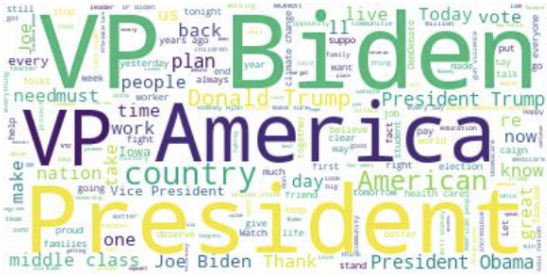


Figure 1: Word Cloud for Joe Biden

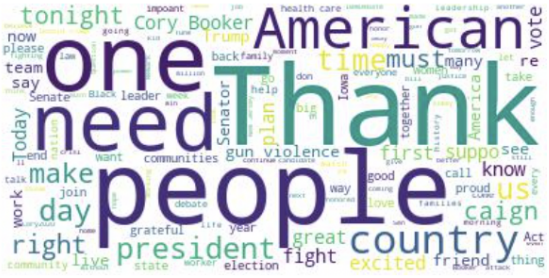


Figure 2: Word Cloud for Cory Booker

Figure 3: Word Cloud for Pete Buttigieg

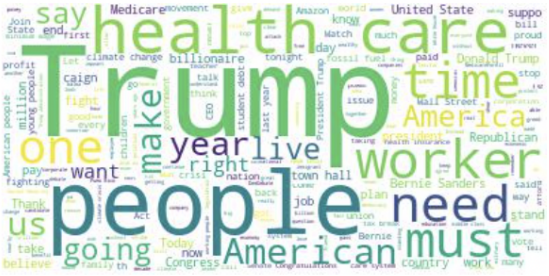


Figure 4: Word Cloud for Bernie Sanders

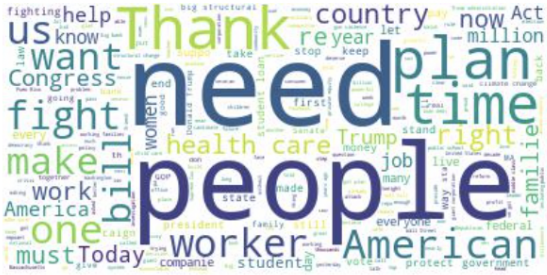


Figure 5: Word Cloud for Elizabeth Warren

Figure 6: Word Cloud for Andrew Yang

Figure 1: "Word Cloud for the Tweets made by the candidates"

Appendix