AIT590 NLP Project Report

Team9 - section 002

2020 U.S. Presidential Election Sentimental Analysis: Trump Vs. Biden Tweets

Team Members: Elham Jafarghomi, Saravana Vallaban, Haniehsadat Taghavi

Instructor Dr. Liao

December 2020

GEORGE MASON UNIVERSITY

Table of Contents

Abstract

Twitter is a popular micro-blogging social media platform. For the 2020 US Presidential election, both candidates expressed their opinions and point of views on this popular social media platform. Given that sentiment analysis is a text processing technique used to measure views, opinions, and attitudes, we first used VADER, a simple rule-based model for general sentiment analysis and compared its effectiveness to machine learning algorithms, we built in this project. The models we built in this project include Logistic Regression and Naïve Bayes Classifiers. Logistic Regression model gave the highest accuracy among all other models with 73.43% on testset. Therefore, this model was used to predict Donald Trump and Joe Biden Tweets sentiments. The result of the analysis suggested that Joe Biden had more positive sentiments in his tweets than Donald Trump during the election cycle.

Introduction

The 2020 U.S. Presidential election was one of the most controversial in the nation's history. This Presidential Election was characterized by an intense competition of candidates stemming from a noticeable divide in the nation over social and economic issues. Furthermore, Donald Trump and Joe Biden were trying to contrast their views on twitter as an effective soundbite platform.

In this Project, we performed sentiment analysis by applying Natural Language Processing (NLP) techniques to analyze Donald Trump and Joe Biden Tweets during the 2020 election cycle. The motivation behind using NLP to do polarity analysis on Twitter data is to observe a single measure of sentiment over the high frequency short messages that each candidate sent over a wide range of issues within the relative short period of runup from May 1st until November 2nd.

Using NLP tools, we tried to capture insight into how their attitudes were evolving, getting closer to the election day. The way we approached this problem, was by training several machine learning models on a pre-labeled general tweets dataset and using the best performing model on candidates' tweets datasets for predicting their tweets sentiments.

Related Work

Social media has gained remarkable importance as a public engagement tool for different topics including political purposes. Rapid spread of information through social media platforms such as Twitter, provides politicians the ability to broadcast their messages to their audience directly.

Ussama Yaqub, Nitesh Sharma, Rachi Pabreja, Vijayalakshmi Atluri and Jaideep Vaidiya from Rutgers Business School collaborated with Soon Ae Chun from City University of New York and have performed sentiment analyses of Twitter location data. They used two case studies: US presidential elections of 2016 and UK general elections of 2017. For US elections, they plot state-wise user sentiment towards Hillary Clinton and Donald Trump. They discovered similar tendencies in Twitter sentiment towards political candidates and parties regardless of the methodology adopted for data collection. In our work, instead of state-wise user sentiment towards presidential candidates, our focus is going to be on candidates' tweets sentiments itself. [2]

Ramteke et al. tried to estimate the election result based on the number of tweets. They collected tweets for Donald Trump and Hillary Clinton, two candidates who participated in the 2016 USA Election, on March 16-17, 2016, containing the names of the parties and candidates. They used 60 thousand tweets they obtained for analysis. They analyzed the tweets containing the hashtags used by the supporters of a party/candidate with the assumption that it could be positively tagged for that candidate/party. With this assumption, hashtags used at high frequency in the Tweet dataset (minimum 20 in this study) were manually tagged as positive or negative. For example, the #MakeAmericaGreatAgain hash has been labeled as positive support for Trump. In their analysis with the Phyton software, they first labeled the emotional polarity of the sentences with the VADER algorithm. Then, by applying the Multinomial Naive Bayes and Support Vector machines machine learning models with the Python Scikit-learn package, they classified the tweets as positive and negative for the candidates. The estimation of the election result by dividing the total number of positive tweets sent about the candidate by the total number of tweets related to that candidate turned out to be incorrect. [3, 4]

Anuta et al. conducted a study comparing the election forecast model based on Twitter data with polls. They used previously collected tweet data about the 2016 US elections in the study. They filtered the tweets in the data set according to the location (USA) and the language of the Twitter user. They saved 3 million tweets of 750,000 users to the PostgreSQL database for analysis.

VADER application was used for sentiment analysis. According to the tweets of the people about the candidates, they decided whom they would vote for and estimated the votes by proportioning the number of people. As a result of the study, they stated that Twitter was more biased than surveys and the prediction was more erroneous with data based on tweets. However, this may be due to errors in their prediction methods based on Twitter data they use. As a matter of fact, if two candidates are mentioned in a tweet, one candidate may be praised and there may be a negative judgment against the other candidate. It is wrong to evaluate this tweet positively and analyze it as positive vote for both candidates. [4, 5]

Another example of related work is Brandon Joyce and Jing Deng analysis in which the main goal was to calculate the sentiment expressed by tweets from the 2016 US Presidential election, where many people expressed their likes or dislikes for a presidential candidate. They used a lexicon and Naive Bayes Machine Learning Algorithm to calculate the sentiment of political tweets collected one-hundred days before the election. they used manually labeled tweets as well as automatically labeled tweets based on hashtag content/topic. For our project, we hope to incorporate a similar approach using the 2020 election candidates tweets directly. [6]

Objectives

The primary objective of this project was to gain insight about the overall attitude of the 2020 U.S. presidential election candidates. The sentiment analysis on Donald Trump and Joe Biden tweets focused on polarity of their tweets they have been sending in the runup to the election day. Where polarity here can tell whether their tweets has a negative or positive tendency.

Datasets

Selection and Description:

For this project we first obtained a pre-labeled tweet dataset from Kaggle, which contains 99982 general random tweets with corresponding binary labels, 0 for Negative and 1 for Positive sentiments. We used this dataset as the training dataset for implementing both baseline solution using Vader-Sentiment and training our classifiers.

Afterwards datasets of each candidate's tweets were compiled by interacting with Twitter. For this, we applied for a developer account and had it approved by Twitter. We first tried to pull our data

directly from twitter accounts of Donald Trump (@realDonaldTrump) and Joe Biden (@JoeBiden) using API. Since the standard API only allowed to retrieve tweets up to past 7 days, and we needed tweets from beginning of May, we tried multiple other ways for obtaining the tweets. Finally, we used "snscrape.modules.twitter", a built-in python module, to obtain tweets from their official twitter accounts. We limited the time to six months' period form May 1st to November 2nd . The number of tweets we collected for this period, was 2570 and 578 tweets from Donald Trump and Joe Biden twitter accounts, respectively. We used these two datasets as our test datasets to predict the overall polarity of their tweets and sentiment changes over time using the model, we trained on the train dataset.

Datasets schema

The overview of the pre-labeled train dataset, along with Trump and Biden tweets datasets can be found in *Figure*s 1, 2 and, 3 respectively.

```
#Display train dataset summary
traindf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99982 entries, 0 to 99981
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   ItemID         99982 non-null  int64
 1   Sentiment      99982 non-null  int64
 2   SentimentText  99982 non-null  object
dtypes: int64(2), object(1)
memory usage: 2.3+ MB
```

```
#Display the first 6 row of the train dataset
traindf.head(6)
```

	ItemID	Sentiment	SentimentText
0	1	0	is so sad for my APL friend.....
1	2	0	I missed the New Moon trailer...
2	3	1	omg its already 7:30 :O
3	4	0	.. Omgaga. Im sooo im gunna CRy. I've been at this dentist since 11.. I was suposed 2 just get a crown put on (30mins)...
4	5	0	i think mi bf is cheating on me!!! T_T
5	6	0	or i just worry too much?

Figure1- Overview of the training dataset

	Date	Tweet
0	11/2/2020	Just landed in Traverse City, Michigan. Big crowd!
1	11/2/2020	As Christians throughout this great Country celebrate All Souls Day, let's remember those who went before us and built this great nation. May their legacy inspire us as we keep our nation what it has always been: blessed and great!
2	11/2/2020	Joe Biden is promising to delay the vaccine and turn America into a prison state-locking you in your home while letting far-left rioters roam free. The Biden Lockdown will mean no school, no graduations, no weddings, no Thanksgiving, no Christmas, no Fourth of July, and...
3	11/2/2020	Joe Biden is a globalist who spent 47 years outsourcing your jobs, opening your borders, and sacrificing American blood and treasure in endless foreign wars. He shuttered your steel mills, annihilated your coal jobs, and supported every disastrous trade deal for half a century...
4	11/2/2020	I gave Maine everything that Obama/Biden took away from it. 5000 square miles, Lobster, Fishing, ended tariffs from China and E.U. and much more. Vote Trump Maine!
5	11/2/2020	Landing in Scranton, Pennsylvania!

Figure2- Overview of Donald Trump tweets dataset

	Date	Tweet
0	11/2/2020	Donald Trump is the most corrupt president in modern history. Donald Trump is the most racist president in modern history. Donald Trump is the worst jobs president in modern history. Why would we give him another four years?
1	11/2/2020	Here's the truth: Donald Trump inherited a growing economy from President Obama and me. And just like everything else he's inherited in life, he squandered it.
2	11/2/2020	When America votes, America will be heard. And when America is heard, I believe the message is going to be loud and clear: It's time for Donald Trump to leave the White House.
3	11/2/2020	A Biden-Harris administration will: - Implement nationwide mask mandates - Ensure access to regular, reliable, and free testing - Accelerate the development and distribution of safe and effective treatments and vaccines We won't waste any time getting this virus under control.
4	11/2/2020	Together, we're going to rebuild our economy. And when we do, we'll not only build it back - we'll build it back better.
5	11/2/2020	In public, President Trump compared COVID-19 to the flu and suggested people inject bleach to treat it. In private, he told Bob Woodward it was deadlier than the flu and that he wanted to downplay it. It's unthinkable.

Figure3- Overview of Joe Biden tweets dataset

Pre-processing

Tweets like any other texts data are unstructured and contain punctuations, stopwords, numbers, mentions, URLs and etc, which may not add any value for the determination of sentiments. These noises need to be eliminated from text data before any analysis.

For the Pre-processing, we defined separate functions for tokenization, lemmatization, punctuation and stopwords removal. For punctuation removal, in addition to using the pre-defined punctuation list from string library, we converted the tweets to lower case, included steps for removing all references and @mentions, digits and numeric characters, re-tweets (RT), links and URLs. For stopwords removal we chose to use Spacy due to its completeness in stopwords dictionary. However, we customized it by eliminating some words from the dictionary and adding some other words to it. Finally, we removed all words with less than or equal to two characters that remained in our dataset after all text cleaning steps, since by a more detailed exploration, we found that these words have no contribution to the sentences sentiments.

The system

System Architecture

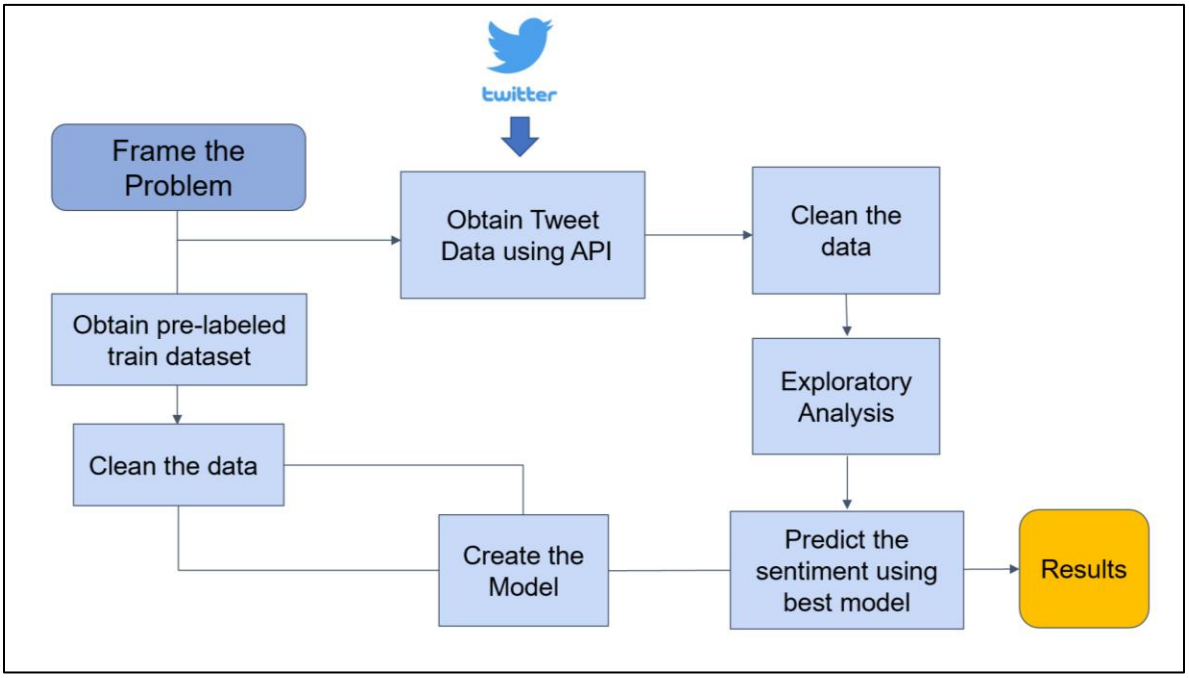

Figure4. System Architecture and Framework

Figure4 is displaying the system architecture and our project framework. In this section, we will briefly review each part as following.

- ***Frame the Problem**:* As it was described in the introduction, for the 2020 U.S. Presidential election, Donald Trump and Joe Biden had an intense competition and they were trying to contrast their views on social media from months before the election. In this project we tried to use NLP tools to understand the sentiment of candidate tweets and get insight into how their attitudes change during the six months period before the election day. To approach this problem, we chose to first use a baseline sentiment analysis technique and then several machine learning techniques to obtain a better result.

- ***Obtain the train and test data**:* As the project directed towards conducting a sentiment analysis on presidential candidates' tweets, we obtained our test data from their official twitter accounts using API along with python twitter scrapping module. Along with data acquisition for test data, we obtained a pre-labeled data from Kaggle as our train dataset to train our models on that by using supervised machine learning techniques.

- ***Clean the data:*** All cleaning steps described in pre-processing section were applied to both train data and candidates' datasets.

- ***Create models:*** After cleaning the data and implementing a baseline solution, using VaderSentiment, we trained several models on pre-labeled train dataset and compared the accuracies resulted from evaluating each model on the holdout set. The model with the higher accuracy was selected.

- ***Exploratory data analysis:*** We did exploratory data analysis on both train and test dataset including designing wordcloud of most frequent positive and negative words on train data, and wordcloud for Trump and Biden frequently used words.

- ***Predict the Sentiments:*** Using the best performing model created during model training, we predicted the presidential candidates' tweets overall sentiments.

- ***Results:*** The results are presented through visualizations and analysis.

NLP and Data Analytics Approaches

After initial inspection of data, we first tried implementing sentiment analysis using existing NLP algorithms. Among existing resource-efficient rule-based algorithms, we selected VADER as our baseline solution since it does not suffer severely from a speed-performance tradeoff. Vader captured the percentage of positive, negative, and neutral sentiment and calculated a normalized score called compound score for each tweet. We used these scores which are ranging from -1 to +1 for creating a new sentiment column for positive and negative tweets. In this process, scores greater than 0 were labeled as positive and less than 0, labeled as negative. We dropped the 0 scores tweets as they had neutral sentiments and they were not the purpose of this project. After that we checked the accuracy of VaderSentiment prediction. The accuracy obtained from the baseline solution was 69.7 %.

Next, we tried using machine learning algorithms to obtain higher accuracy. For this, we vectorized the dataset using TF-IDF vectorization method before training our machine learning classifiers. Before model building, the train-test split procedure was performed via train_test_split function with allocating 80% to trainset for fitting the model and 20% for evaluating the performance of the algorithm.

First, we fitted a Linear Logistic Regression model from sklearn library, and we obtained 73.39% accuracy on the testset. We then trained a Multinomial Naïve Bayes model. This model applies Bayes theorem with a Naïve assumption of no relationship between features. The accuracy resulted from Multinomial NB was 72.89%. Finally, we trained a Bernoulli Naïve Bayes model that with accuracy of 72.70% gave a close result to Multinomial NB.

Although, we could not train machine learning models with more promising accuracy, but we could achieve a higher accuracy by 4% increase compared to baseline solution. We chose our best performing model, Linear Logistic Regression with 73.39% accuracy, for using in sentiment prediction on Trump and Biden tweets datasets.

Hardware and Software Development Platforms

Hardware used for this project is as follow:

- Windows operating system with at least 16GB of RAM with a dual core i7 processor

Softwares used for conducting this analysis are as follow:

- Twitter API
- Python
- Jupyter NoteBook from Anaconda Navigator
- Microsoft Office 365 PowerPoint, Word, and Excel

Experimental Results and Analysis

With the help of NLP techniques for text processing and WordCloud visualization , we could visualize the most frequently used words by presidential candidates. Although, word clouds do not offer the most sophisticated analysis, but they can provide a quick snapshot from thousands of tweets. As we can see from the wordclouds presented, each candidate had been using a handful of focus-grouped words and phrases in their tweets. For instance, Donald Trump repeatedly used word “fake”, “sleepy Joe”, “great”, “people”, “vote”, “job” and, etc. Joe Biden used more optimistic words such as “believe”, “nation”, “together”, “care”, “live”, ”president” and, etc. However, there are some words that are used by both candidates such as “people”, “America”, and “vote”.

Donoald Trump Most repeated words *Joe Biden Most repeated words*

Using Logistic regression model that we trained earlier in this project, we performed sentiment prediction on Trump tweets dataset, and we found out that the overall polarity of trump tweets is

not significantly skewed towards positive or negative sentiments. The result of sentiment prediction showed that Trump overall polarity has 53.37% positive sentiment. Figure5, shows that his tweets negative and positive sentiments ratio does not change significantly during the six months period before election. We only can observe a significant difference in his attitude in July. This may be due to temporary decrease on number of COVID-19 cases in United States, or due to the Independence Day (4th of July), which made the sentiment of his tweets more positive.

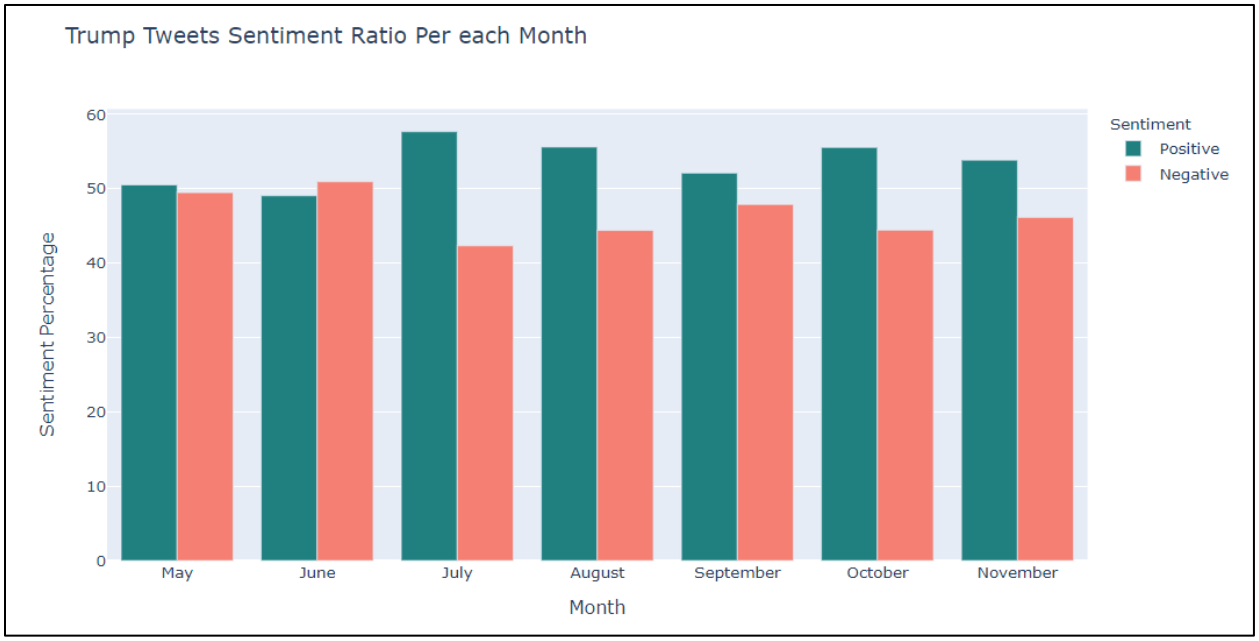

Figure5- Donald Trump Tweets Sentiment Ratio per each month

We also used the Logistic Regression model to perform sentiment prediction on Biden tweets dataset, and we observed that the overall polarity of Biden tweets more skewed towards positive sentiment. The result of sentiment prediction showed that Biden overall polarity has 63.15% positive sentiment, suggesting approximately 10% higher than Donald Trump. As shown in Figure6, Joe Biden tweets negative and positive sentiments ratio does change during the six months period before election. In some months we can observe a lot more positive sentiment in his tweets that suggests a more positive attitude. We can see this fact significantly in Month of June.

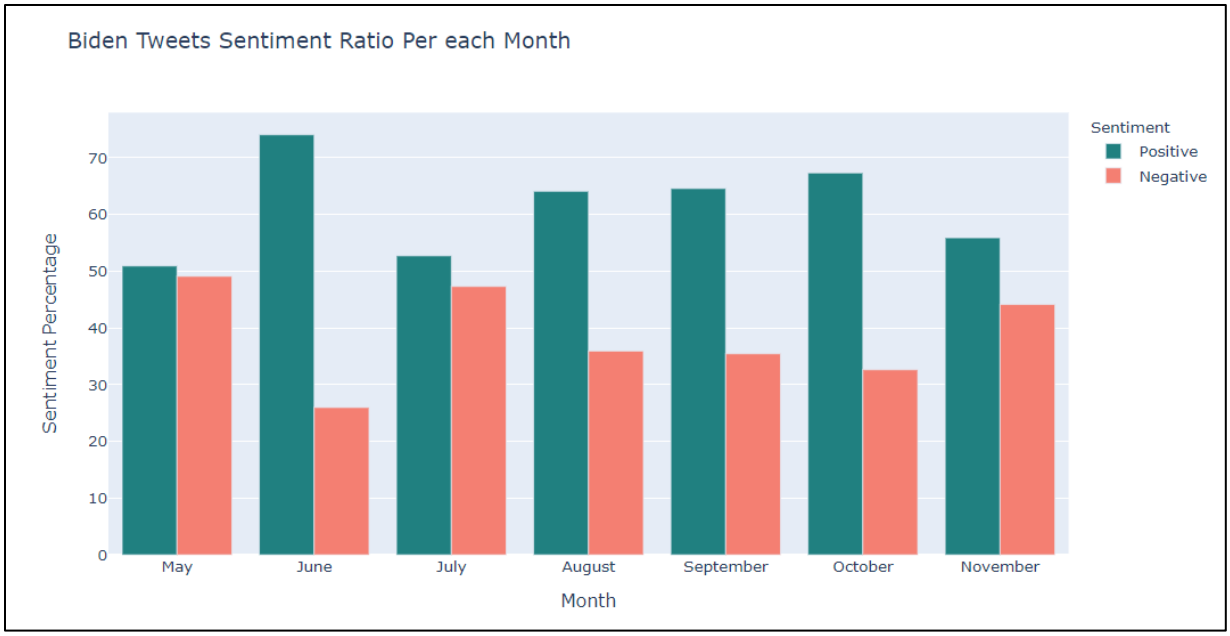

Figure6-Joe Biden Tweets Sentiment Ratio per each month

As the model results for sentiment prediction on each candidate tweets dataset suggested in our a nalysis, the polarity of Joe Biden is 10% more positive than Donald Trump overall polarity senti ment with 63.15% and 53.37% for Biden and Trump, respectively. The sentiment analysis result can be seen in Figure7.

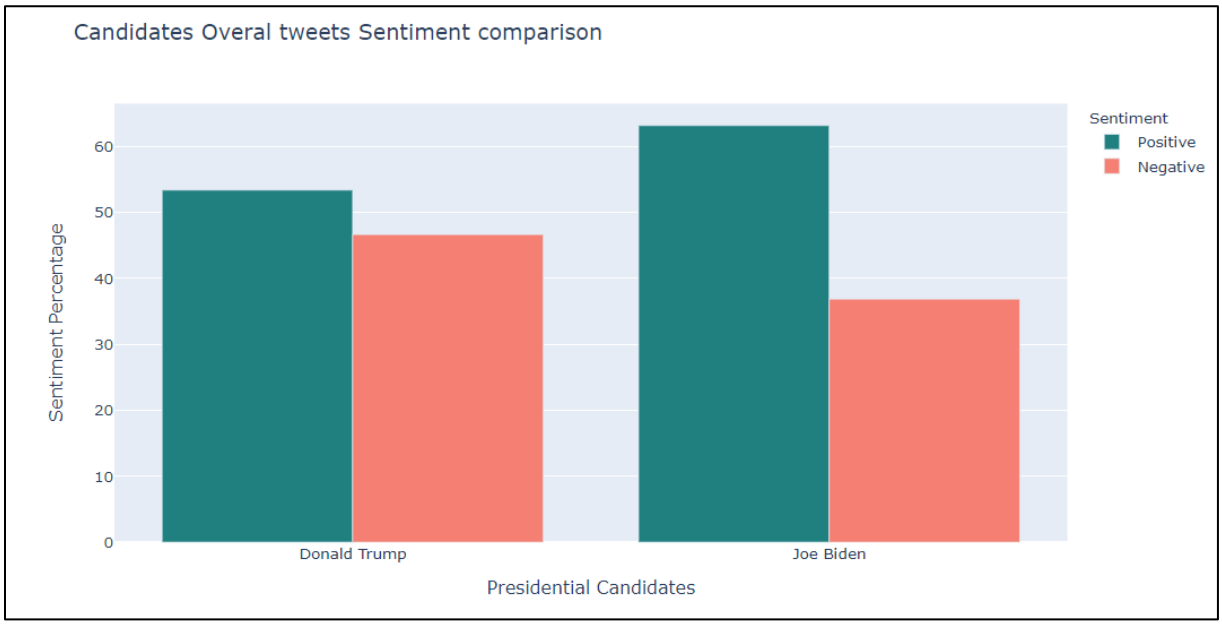

Figure7-Candidates overall tweets sentiment comparison

Finally, for summarizing our sentiment analysis in this project and comparing both candidate overall attitude changes during the 2020 election cycle, we created the line plot shown in Figure8, which indicates that Joe Biden tweets always had more positive sentiment from beginning of May to election Day, except for the month of July that Donald Trump positive sentiments exceeds as it was explained above.

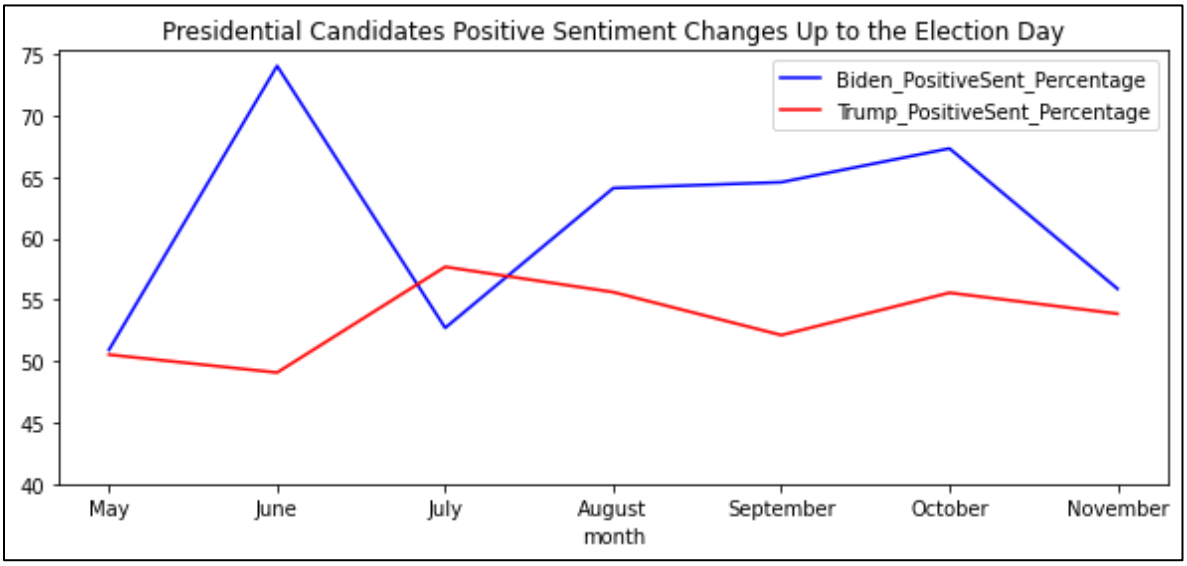

Figure8- 2020 Presidential candidates attitude change and sentiment comparison over time

Conclusion

In this project, we performed sentiment analysis on U.S. 2020 presidential candidates tweets during the election cycle to find out their polarity and changes in sentiment over 6 months period. We observed that Joe Biden has more positive sentiments in his tweets than Donald Trump. However, despite the assumption made by most people regarding the use of more negative words than positive by Donald Trump, his polarity is not significantly skewed towards positive or negative and it is almost consistent over time.

In the future, we would like to expand our study by performing more detailed sentiment analysis by including topic-based analysis to find out their attitudes difference toward specific subject.

References:

1. Erik Bruin, (2018). *Text mining the Clinton and Trump election Tweets*, Kaggle. Retrieved from: https://www.kaggle.com/erikbruin/text-mining-the-clinton-and-trump-election-tweets
2. Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. (2020). *Location-based Sentiment Analyses and Visualization of Twitter Election Data*. *Digit. Gov.: Res. Pract.* 1, 2, Article 14 ,19 pages, Retrieved from: https://doi.org/10.1145/3339909.
3. Ramteke, J. et al. (2016). *Election result prediction using Twitter sentiment analysis.* International Conference on Inventive Computation Technologies (ICICT) 1 (2016): 1-5. Retrieved from: https://www.semanticscholar.org/paper/Election-result-prediction-using-Twitter-sentiment-Ramteke-Shah/55487833519a73bd3322dc971dbbd019ebe6a713
4. C.J. Hutto Eric Gilbert. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Georgia Institute of Technology, Atlanta, GA 30032. Retrieved from: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf
5. David Anuta, Josh Churchin, Jiebo Luo. (2017). *Election Bias: Comparing Polls and Twitter in the 2016 U.S. Election*. Cornell University. Retrieved from: https://arxiv.org/abs/1701.06232
6. Brandon Joyce, Jing Deng. (2017). *Sentiment analysis of tweets for the 2016 US presidential election*. IEEE. Retrieved from: https://ieeexplore.ieee.org/document/8284176
7. Ronen Feldman, James Sanger. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.* Retrieved from: http://dl.icdst.org/pdfs/files/25a6d982ee80e1db7a4ebf7eeca4e0ec.pdf
8. Yohanssen Pratama, Anthon Roberto Tampubolon, Liana Diantri Sianturi, Rifka Diana Manalu and David Frietz Pangaribuan. (2019). *Implementation of Sentiment Analysis on Twitter Using Naïve Bayes Algorithm to Know the People Responses to Debate of DKI Jakarta Governor Election*. 1st International Conference on Advance and Scientific Innovation (ICASI). Retrieved from: https://iopscience.iop.org/article/10.1088/1742-6596/1175/1/012102/pdf
9. Ussama Yaqub, Soon Ae Chun, Vijayalakshmi Atluri, Jaideep Vaidya. (2017). *Sentiment based Analysis of Tweets during the US Presidential Elections*. the 18th Annual International Conference. Retrieved From: https://www.researchgate.net/publication/317272393_Sentiment_based_Analysis_of_Tweets_during_the_US_Presidential_Elections
10. Ankur Agrawal, Tim Hamling. (2017). *Sentiment Analysis of Tweets to Gain Insights into the 2016 US Election.* Columbia Science Journal. Retrieved from: https://journals.library.columbia.edu/index.php/cusj/article/view/6359
11. Ravi Parikh, Matin Movassate. (2008). *Sentiment Analysis of User-Generated Twitter Updates using Various Classication Techniques.* Researchgate. Retrieved from: https://www.researchgate.net/publication/242660794_Sentiment_Analysis_of_User-Generated_Twitter_Updates_using_Various_Classication_Techniques

12. Sri Hari Deep Kolagani, Arash Negahban, Christine Witt. (2017). *Identifying Trending Sentiments in The 2016 U.S. Presidential Election: A Case Study Of Twitter Analytics*. Issues in Information Systems, Volume 18, Issue 2, pp. 80-86. Retrieved from: https://iacis.org/iis/2017/2_iis_2017_80-86.pdf
13. Caetano, J., Lima, H., Santos, M. *et al.* (2018). *Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election.* Journal of Internet Services and Applications 9, 18. Retrieved from: https://jisajournal.springeropen.com/articles/10.1186/s13174-018-0089-0
14. Pre-Labeled Training Data: https://www.kaggle.com/imrandude/twitter-sentiment-analysis