**NLP Sentiment and Social Network Analysis for 2020 US Election Tweets Final Proposal**

By Enid Roman

For my final project I will discuss the prediction of the 2020 US election results based on Sentiment Analysis using tweets. Sentiment Analysis is a powerful technique with significant business implications such as predicting election results. We can forecast the outcome of the U.S. election by examining the sentiment behind the significant number of tweets by the general public toward participating candidates in the 2020 U.S. election.

While searching and researching for an interesting Natural Language Processing Project for my final project, I came across the 2020 Election Tweets in Kaggle’s website. In reading the 2 documents that were attached, https://arxiv.org/abs/1706.02271 and https://dl.acm.org/doi/fullHtml/10.1145/3339909, I got more interested using the datasets for my final project and seeing the results. Since we are again in 2024 approaching Presidential Election, I did try looking for a 2024 US Election Tweets dataset but was not successful in finding one.

I obtained the two datasets, hashtag\_joebiden.csv and hashtag\_donaldtrump.csv, from https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets/data. The data used is tweets that were collected using the Twitter API statuses lookup and sns scrape with #DonaldTrump and #JoeBiden used as keywords. The tweets were collected from 10/15/2020 to 08/11/2020. I will explore information such as the distribution of tweets by state and source; the sentiment of tweets for each candidate (negative, neutral, or positive); and the overall sentiment of tweets.

Questions I would like to answer are the following:

1. How do subjectivity scores for each candidate vary across states and who was mentioned in more subjective tweets?

2. How does sentiment of Twitter messages base on state location correlate with the real-world sentiment of the public towards the two candidates?

3. Which sentiment holds the most value and how it will affect the election results?

4. Which community leaders from the social network are the most important nodes in this network?

The following are the dataset features:

created\_at: Date and time of tweet posted

tweet\_id: Tweet’s unique ID

tweet: Full tweet text

likes: Number of likes

retweet\_count: Number of retweets

source: Utility used to post the tweet

user\_id: User ID of tweet creator

user\_name: Username of tweet creator

user\_screen\_name: Screen name of tweet creator

user\_description: Self-description by tweet creator

user\_join\_date: Join date of tweet creator

user\_followers\_count: Followers count on tweet creator

user\_location: Address was given on tweeter’s profile

lat: Latitude parsed from user\_location

long: Longitude parsed from user\_location

city: City parsed from user\_location

country: Country parsed from user\_location

state: State parsed from user\_location

state\_code: State code parsed from user\_location

collected\_at: Date and time tweet data was mined from Twitter

The libraries I plan to use are the following: pandas, numpy, matplotlib.pyplot, plotly.express. Libraries for Sentiment Analysis are: re, nltk,from nltk.corpus stopwords, and wordnet, from nltk.stem WordNetLemmatizer, TextBlob and WordCloud. TextBlob is used for tokenization as well as for finding the polarity and subjectivity of a text. Regular expression (re): removes unnecessary symbols that do not provide any meaning such as @, comma, hashtags, etc. Natural Language Toolkit (nltk) removes stop words such as and, is, the, in, or, etc.

Steps I will take are the following:

Data Assessment: For both datasets I will study the shape of data and what it tells, checking variables and their data types.

Data Preprocessing: This refers to cleaning (dropping missing values), transforming, and integrating data in order to make it ready for analysis. It is the most important step in an analysis to achieve accurate results with increased quality and reliability as the raw data can be inconsistent or incomplete in formatting. This process involves tasks such as performing data integration and transformation: After data cleaning it is preferable to convert it into formats that are easy to read and understand by software applications and algorithms. I can integrate and transform the data by condensing datasets to reduce the total amount of data by concatenating both datasets for easier analysis using the pandas’ concat function.

Exploratory Data Analysis (EDA): is a method used to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. This will include the following, Tweets for Candidates, Comparison of Likes, Top10 Country-wise tweets Counts, Tweet Counts for Each Candidate in the Top 10 Countries.

S**entiment Analysis for prediction of election results:** I shall only consider US citizens’ opinions here, as they are a crucial deciding factor in who becomes the 2020 US president. I will get polarity, subjectivity, and Analysis for both Biden and Trump to fetch sentiments from the data. First, I will find out if the sentence is subjective or objective. Subjective means sentence refers to a personal opinion and not facts whereas objective means facts. Polarity analysis is used to determine the emotional attitude of a text writer with respect to the topic under discussion. **Polarity** is represented by a float value that lies between -1 and 1, where 1 = positive sentiment, -1 = negative sentiment. **Sentiments** classify the sentiments as positive, negative or neutral*.*

I will then analyze the sentiments to predict the results for Donald Trump and Joe Biden.

I will attempt to create social network analysis of the 2020 U.S. presidential elections with users as nodes and retweets as the edges between them. the network analysis will represent the interaction between users tweeting about the 2020 U.S. presidential elections. Network Analysis will identify the community leaders from the social network which are the most important nodes in this network since they appear to be the center of the network and link most of the retweets. My main concern is that it will be too overcrowded and will have to limit to top 10 or 20.

I will use centrality analysis to measure the central node in the network, which is used to measure the importance of the network and the node’s reputation in the network. To look at the network structures and metrics in this study I will use the four measurements of centrality which includes, degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. The degree centrality will be utilized to explore the nodes who reach the highest degree and lowest degree within the network. Degree centrality can identify people who can directly reach many other people and show the number of edges and the total connections. The betweenness centrality will be used to measure how important a node is in the flow of information from one part to another network. The closeness centrality will be utilized to measure how quickly information can spread from a given node to all other nodes in the network. The high closeness centrality are often seen as influential or important in terms of communication or information flow. Eigenvector centrality will be utilized to measure the quality of connections rather than the quantity. The measurement identifies nodes that are influential because they are connected to other influential nodes.

In conclusion, this proposal outlines a comprehensive approach to exploring the predictive power of sentiment analysis using Twitter data for the 2020 U.S. Presidential election. By examining the sentiments expressed in tweets towards candidates Donald Trump and Joe Biden, this project aims to uncover potential correlations between online sentiment and electoral outcomes. Additionally, incorporating social network analysis adds another dimension to understanding the dynamics of public opinion as expressed on social media.

Through the examination of retweets and user interactions, the project seeks to identify community leaders and influential nodes within the network of Twitter users discussing the election. This network analysis aims to shed light on the flow of information and the dissemination of sentiment across the platform. By leveraging centrality measures such as degree, betweenness, closeness, and eigenvector centrality, we can identify key actors who play pivotal roles in shaping the discourse surrounding the election.

While promising, it is crucial to recognize the limitations of using social media data, such as representation biases and the influence of non-human actors like bots. However, by employing a combination of sentiment analysis and network analysis, this project endeavors to provide valuable insights into the complexities of public opinion in the digital age. These insights not only contribute to the academic discourse on predictive analytics in political science but also highlight the ethical considerations and methodological challenges inherent in analyzing social media data.

Moving forward, the methodologies proposed herein could prove instrumental in enhancing the accuracy and reliability of political forecasting using digital platforms. By understanding the interplay between sentiment, network dynamics, and electoral outcomes, researchers and policymakers can better grasp the intricacies of contemporary political discourse and decision-making processes.