SEMANTIC SENTIMENTAL ANALYSIS OF TWEETS ON US ELECTION RESULTS 2020

INFO 5731- Section 002 #Group-3

COMPUTATIONAL METHODS FOR INFORMATION SYSTEMS

COLLEGE OF INFORMATION SCIENCE

Hari Prasad Laggoni – 11448940

Vamsi Krishna Bharghava Muktinutalapati-11443219

Lakshmi Naga Sai Veera Durga Uppala- 11447891

Sai Priyanka Mittapelly – 11436112

Prasanna Kumar Tumati– 11366427

UNT UNIVERSITY OF NORTH TEXAS

SEMANTIC SENTIMENTAL ANALYSIS OF TWEETS ON US ELECTION RESULTS 2020

INTRODUCTION

- Analysis of the relationship between social media data and electoral outcomes has started with the broad use of social media. Numerous studies have shown that social media has an impact on election outcomes in various countries and can be used effectively to forecast election result.
- Twitter is a social media application where many of the users use this across the globe to share their views on the events that occur.

- In this project, we have extracted tweets from twitter and estimates were made as per the tweets which are positive, negative and neutral as per the contesting candidates.

- We have also used extracted tweets as an input data for machine learning models in order to identify whether tweets are biased to Trump or Biden.
- We have also visualized the percentage of positive, negative, and neutral tweets for both candidates.
- From this research, we have acquired the sentiment of each tweet and analyzed each model's accuracy, precision, recall and F-1 score.

RESEARCH QUESTIONS

The research questions that are to be answered in this project are:

- After performing sentimental analysis, determine what is the polarity of tweets between Biden and Trump.
- What are the frequent words used in tweets in reference to the 2016 and 2020 US Elections?
- Upon training machine learning models, determine which tweets are biased towards Biden or Trump.

METHODOLOGY

Below are the steps involved in the process of this research:

- Tweets are extracted from Twitter using Tweepy and Kaggle data set has been used for the sentimental analysis of tweets.
- To make an extracted data set into a structured format, we have used data cleaning techniques like stemming, lemmatization, stop words removal, emoji removal, etc.
- For the above cleaning techniques, we have used Python libraries like 'Spacy', 'Text blob', 're' for cleaning the unstructured data.
- TF-IDF has been used for extracting features from the tweets.
- Different Machine learning models have been applied and evaluated on the terms of accuracy, precision, recall, F-1 score.
- Finally, comparison of accuracies of different models will provide the conclusion to our analysis.

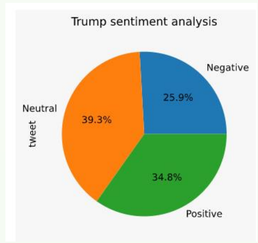
DATA COLLECTION

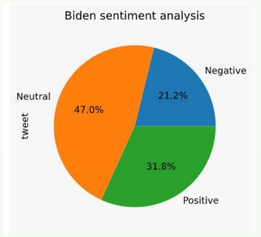
Below are the sources we collected the data from (Kaggle, Twitter)

- https://www.kaggle.com/manchunhui/us-election-2020-tweets
- https://www.kaggle.com/benhamner/clinton-trump-tweets
- Collected tweets from Twitter using Tweepy. Tweepy is a Python library used to extract Tweets.
- Collected data is stored in .csv format for complete analysis using panda's library in Python.
- We have collected 6000 tweets to create an accurate machine learning model for the election analysis.

DATA ANALYSIS

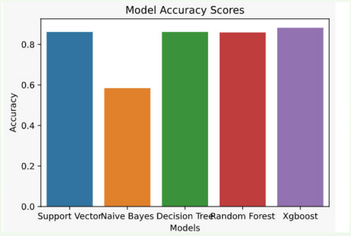
We conducted the data cleaning techniques and calculated the polarities for each tweet of Biden and Trump on the collected data of tweets from Twitter and Kaggle sites. From the sentiment analysis, visualization of sentiment of tweets has been shown in a pie chart. We have also visualized the most frequent words used by users in Twitter in the 2016 and 2020 elections using Word Cloud.





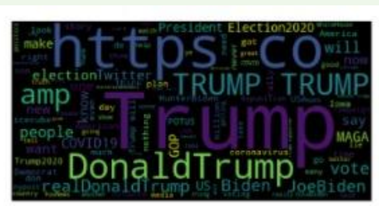
Results and Discussion

We can see that most of the tweets for Biden and Trump have a neutral sentiment. Meanwhile, Trump has more negative sentiment feedback from tweets in comparison to Biden. Below are the results for the different models. From this XG Boost is the best model to train the data followed by decision tree and support vector. Naïve Bayes is the model with poorest performance.



Frequent Words used in tweets by users in the elections 2016 and 2020

2016



2020

CONCLUSIONS

In the sentiment analysis that we have conducted, the results show that Biden received less negative feedback in comparison to Trump. This type of analysis could be used in real life when predicting election results as shown by the result of the Biden vs. Trump election; Biden won. However, through this analysis it also shows that most people had very neutral opinions with regards to this election. We ran different machine learning models for the collected data sets and tried to figure out which type of model is best for training the data. We concluded that the XG Boost model performed the best in training the data.

RESULTS

By performing all the machine learning models on the datasets, we get the below results:

Model Name	Accuracy	F-1 Score
XG Boost	0.882217	0.862534
Decision Tree	0.861432	0.846939
Random Forest	0.859122	0.836461
SVM	0.861432	0.835165
Naïve Bayes	0.584296	0.560976

Future Work:

- This model can be helpful for various psephologists and political parties to improve their performance in administration as per the sentimental analysis of people on the current affairs and their decisions.
- We can also perform sarcasm detection on these tweets which will be helpful for political parties to find the political trend before elections.
- Various other factors could be analyzed to better understand the polarity of the tweets.