Twitter Sentiment Analysis to predict the outcome of the 2020 US Presidential Elections

Abhishek Kumar Bais
Department of Computer Science
San Jose State University
San Jose, CA, USA
abhishek.bais@sjsu.edu

Wasae Qureshi
Department of Computer Science
San Jose State University
San Jose, CA, USA
wasae.qureshi@sjsu.edu

Samer Baslan
Department of Computer Science
San Jose State University
San Jose, CA, USA
samer.baslan@sjsu.edu

***Abstract*—Sentiment Analysis also known as opinion mining is a technique used to extract emotions from textual information using natural language processing and text analysis to accurately classify texts as carrying positive, neutral, or negative emotions. In recent years, Twitter's popularity as a modern medium of public engagement has grown manifold. Consequently, the volume of opinionated data available on twitter has also grown manifold. Thus, Sentiment Analysis on Twitter data has become a subject of research in a wide array of fields. In this paper, we will perform Sentiment Analysis on Twitter data referencing the two presidential candidates, Donald Trump of the Republican Party and Joe Biden of the Democrat Party, from the one month preceding the 2020 US presidential elections to predict the outcome of the elections.**

***Index Items* —Twitter, Sentiment Analysis, VADER, Natural Language Processing, Machine Learning, Bag of N-grams, TF-IDF, Supervised Learning, Logistic Regression, Random Forest, XGBoost.**

I. INTRODUCTION

In the USA, as of July 2020, there are 62.55 million active twitter users[1]. Every day more and more people are taking to twitter to share their opinion and engage in debate and discussion on a wide array of subjects. In the political space, it has been observed that twitter traffic goes up particularly in the run-up to the presidential elections. For example, in the six-month period between June 1, 2016 and November 8, 2016, a total of 171 million unique political tweets were made with the keywords "Trump" or "RealDonaldTrump" or "DonaldTrump" or "Hillary" or "Clinton" or "HillaryClinton"[2].

As the volume of political tweets increases, so does the likelihood of misinformation and propaganda, reaching the masses. Although misinformation has existed in the public space since ancient times[3] , widespread digital misinformation is the new elephant in the room. Many studies have investigated the phenomenon of misinformation on Twitter[4,5,6].These investigations have found a herd spread of sentiments, whether positive, neutral, or negative in polarity, gelled closely with human beliefs. They tend to create echo chambers and spread virally on social media platforms[7]. In this paper, using a dataset comprising tweets that reference the two presidential candidates, Donald Trump of the Republican Party and Joe Biden of the Democrat Party, from the immediate month preceding the 2020 US elections, we will gauge the extent of reach of such echo chambers by performing Sentiment Analysis and using the information gained to predict the outcome of the 2020 US presidential elections.

There are two popular techniques used to perform Sentiment Analysis on generic text data. These are Symbolic techniques and Machine learning techniques [8]. The symbolic techniques use a dictionary-based approach to draw emotions from tweets. The dictionary meaning associates a polarity with the word. By comparing it to the tweets, twitter messages can be annotated with appropriate polarity. Machine learning techniques on the other hand use algorithms to extract information from tweets and use it to annotate polarities. Twitter feed is uniquely different from more generic text data. Tweets are short and limited to 280 characters. Moreover, they often contain slang words and misspelling.

II. METHODOLOGY

Twitter Sentiment Analysis is a multi-step process. In the first step, the data is pre-processed by cleaning it to remove slang words and misspellings. Next, feature extraction is performed. This is done in two phases. In the first phase, twitter specific features such as hashtags, acronyms and emoticons are extracted.

In the second phase, hashtags, acronyms, and emoticons are removed to create regular text data on which Natural Language Processing techniques such as "*Bag of N-grams*" its "*TF-IDF*" are employed to extract text features. All the extracted features together define the feature vectors of the tweet. Sentiment Analysis is performed using the extracted features to classify the tweets by polarity.

Finally, supervised learning models such as *Logistic Regression, Random Forest* and *XGBoost* are often employed

to validate the accuracy of the sentiment classifications and predictions made should the accuracy be within acceptable range. An architecture framework for our system is shown in Fig 1.0 below.

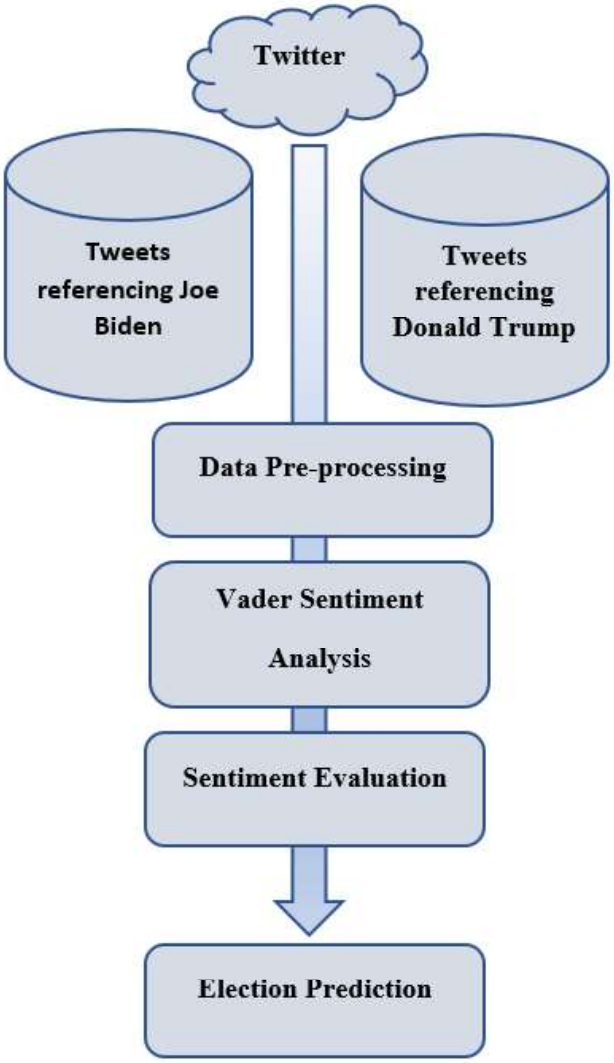

Fig 1.0. Architectural framework for system

In this paper, we classify tweets into "*positive*", "*neutral*" and "*negative*" emotions using Valence Aware Dictionary and Sentiment Reasoner (*VADER*), a lexicon widely used on social media texts to extract sentiment scores from opinions.

Supervised learning models such as *Logistic Regression, Random Forest* and *XGBoost* with "k-fold cross validation" are then used to measure the accuracy of the sentiment scores from *VADER*. In this process the data is partitioned into '*k*' random, equal size sub-samples. One sub-sample is set aside as test data, while the remaining k-1 sub-samples are used to train the model. Upon repeating the process k-times with each sub-sample used exactly once, the models test the accuracy of the sentiment score.

Lastly, to predict the outcome of the 2020 US presidential elections, we analyze the trends in "*positive*", "*neutral*" and "*negative*" sentiment tweets over time as the election approaches to draw conclusions.

Section III shows the data analysis performed and outlines the different preprocessing techniques used. Section IV describes the *VADER* sentiment analysis process. Sections V describes the sentiment evaluation performed on the *VADER* classification using *Logistic Regression, Random Forest* and *XGBoost* and multiple feature extraction techniques. Section VI presents the results of the sentiment analysis and evaluation. Finally, Section VII draws on the results to make the election prediction.

III. DATA ANALYSIS

A. Data Exploration

1,747,805 tweets with the keywords "RealDonaldTrump" or "JoeBiden" from October 2020 were extracted and formed the corpus for our dataset. These included 970,919 tweets referencing Donald Trump and 776,886 tweets referencing Joe Biden. The tweets are made from all over the world.

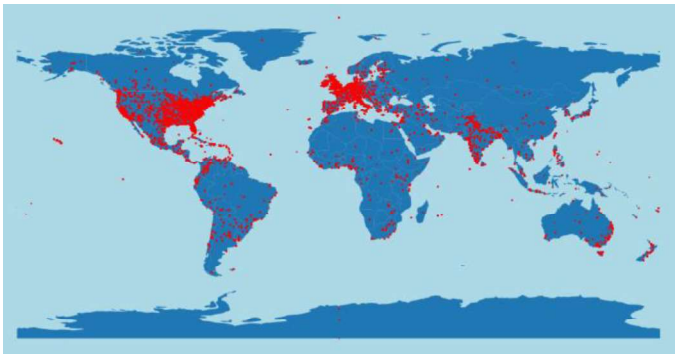

Fig 2.0. Tweets grouped by geographical location of origin

The top five countries of origin of most tweets for which geographic location were available are shown in Fig 3.0. These countries were USA, UK, Canada, and Germany and France. 337,079 tweets originated from the USA. The tweets came from multiple mediums such as Twitter web, Iphone, Android, Ipad, Tweet deck and Hootsuite. The vast majority of tweets, over 400,000 came from Twitter web.

In our paper, we performed sentiment analysis on all tweets regardless of the country of origin. We did this consciously because in today's globally connected world the viral spread of sentiments is not dependent on the location of the tweet origin. Once a tweet finds its audience and gains acceptance by a sizeable mass of people via likes and retweets, its reach creates echo chambers, capable of influencing the human mind and thereby influence election results.

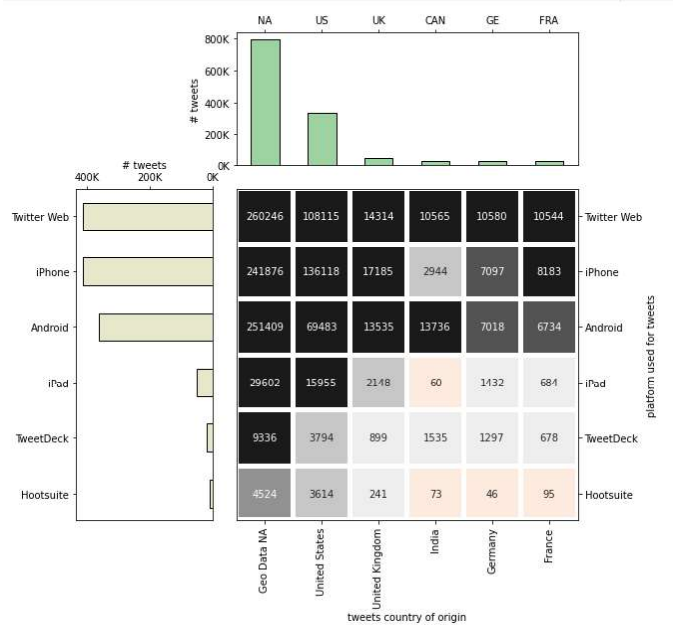

Fig 3.0. Tweets quantified by medium and country of origin

B. Data Preprocessing

Although the dataset contained both English and non-English tweets, as shown in Fig 4.0 below, over 60% of the tweets were made in the English language.

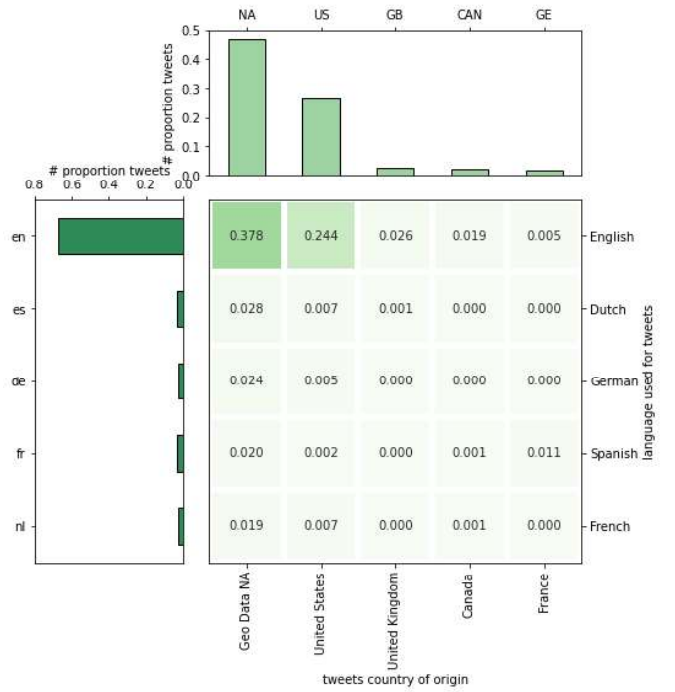

Fig 4.0. Tweets quantified by source language

The tweets contained emoticons, whitespace both trailing and leading, numbers, links, user information, single characters, repeat characters, erroneous punctuation, spelling mistakes, etc. These features increase dimensionality and require careful data processing to achieve accurate sentiment analysis results.

Pre-processing of twitter data involves several techniques, many of them are outlined by Angiani et al[13]. In our paper we employed the data processing techniques outlined in the Table 1.0 below.

High dimension data construct	Preprocessing Techniques
Emoticons	Replace with 'positive', 'negative' keywords
Trailing or leading white spaces	Remove
Numbers	Remove
Links/ URLs	Remove
User Information	Remove
Single character words	Remove
Repeat characters in words	Replace with single character
Erroneous punctuations	Remove
Word Contractions	Replace with expanded form
Unnecessary capitalizations	Replace with lowercase
Contentless words	Stop words removal
Inconsistent word form	Lemmatize

Table 1.0. Preprocessing techniques used

Emoticons like *"sad", "weep", "love", wink", "laugh"* and *"smiley"* were recognized and replaced by *"positive"* or *"negative"* keywords. Over 100-word contractions such as "aren't", "couldn't" were replaced with their expanded form words such as "are not", "could not". Stop words such as *"no", "not". "never", "can't", "don't"* present in the Natural Language Toolkit, commonly called '*nltk*' convey sentiments. To keep these words in the tweet corpus, a custom stop words list was created and read in. Several tweets contained words with repeat characters such as *"baaaaaaad"*. From a sentiment analysis perspective, *"baaaaaaad"* and *"bad"* relay the same emotion. Replacing repeat characters by a single character allowed differently spelled words, conveying the same meaning to be treated identically. In this way, using various pre-processing techniques we transformed raw twitter data to a more feature extraction friendly form, suitable for sentiment analysis and particularly subsequent model evaluation.

IV. VADER SENTIMENT ANALYSIS

Valence Aware Dictionary for Sentiment Reasoning *(VADER)* is a sentiment lexicon developed by Hutto and Gilbert[12] that enhanced widely accepted sentiment word-banks such as (LIWC, ANEW, and GI) by adding emoticons, acronyms and initialisms to extract over 9,000 lexical features from text data. They then used crowd sourcing to obtain sentiment scores for all feature on a scale ranging from "[–4] Extremely Negative" to "[4] Extremely Positive", with allowance for "[0] Neutral (or Neither, N/A)".

In our paper, we used *VADER* to classify twitter data into "*positive*", "*neutral*" and "*negative*" categories. Next, proportion of tweets referencing "RealDonaldTrump" and "JoeBiden" were plotted over time and analyzed. Mean sentiment score for both presidential candidates across all US

states and in particular swing states of Arizona, Florida, Georgia, Michigan, North Carolina and Pennsylvania are also calculated, visualized and analyzed over time for the one-month period immediately preceding the 2020 presidential elections. Trends in mean sentiment score for the swing states are particularly important in an electoral colleges US election system where voting patterns in swing states eventually determines the outcome of the presidential elections.

We chose *VADER* as our tool of choice for sentiment analysis as it has a well-evolved lexicon and does not require training data to generate sentiment scores. For our corpus of 1,747,805 tweets referencing "RealDonaldTrump" and "JoeBiden", *VADER* successfully generated sentiment scores in 7 minutes running time.

V. SENTIMENT EVALUATION

To validate the sentiment classification obtained from *VADER*, supervised learning models namely *Logistic Regression, Random Forest* and *XGBoost* from the scikit-learn python library with k-fold cross validation was used. This validation was essential as we relied on the *VADER* sentiment scores to make the prediction for the 2020 US presidential elections. Multiple models were used to obtain a higher confidence quotient in the prediction.

In order to facilitate robust and accurate model evaluation, we used the *Bag-of-Ngrams*, and its Term Frequency – Inverse Document Frequency *(TF-IDF)* variant for linguistic feature extraction.

A. Bag of N-grams

In this approach, the preprocessed tweet corpus is represented as an unordered set of *N-grams*. By associating weights to group of words that occur together, an accurate sentiment feature extraction of the corpus can be performed. In our paper we used N=2 and N=3 grams. Fig 5.0 shows the top ten, 2-gram, and 3-grams of the tweet corpus from the two presidential candidates.

B. Term Frequency-Inverse Document Frequency (TF-IDF)

In this approach, the relevance of words to determining the sentiment of the tweet is additionally accounted for. Mathematically, Term Frequency, *TF(d, t)* is defined as

$$TF(d,t) = \begin{cases} 0 & if\, freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & otherwise. \end{cases}$$

while, Inverse Document frequency is defined as,

$$IDF(t) = \log \frac{1 + |d|}{|d_t|},$$

where d, is the entire tweet corpus while d_t is the set of tweets containing term t. If $|d_t| << |d|$, the term t will have a large IDF scaling factor and vice-versa.

Words found uniquely in "*positive*", "*neutral*" or "*negative*" categorized tweets are given higher importance than words found across tweet categories. This allows for a more accurate

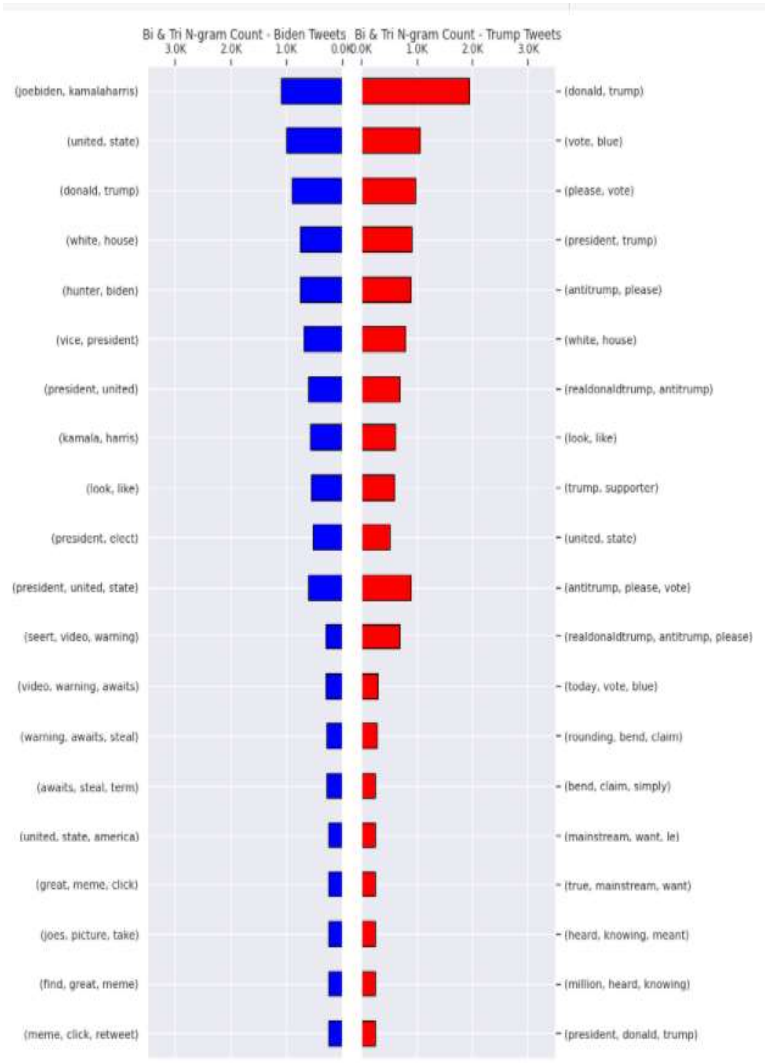

Fig 5.0. Top ten words in 2-gram, 3-gram of tweet corpus

sentiment feature extraction of the corpus. Based on this notion, the Term Frequency-Inverse Document Frequency is defined as

$$TF\text{-}IDF(d,t) = TF(d,t) \times IDF(t).$$

Once the sentiment features are extracted, the *Logistic Regression, Random Forest* and *XGBoost* models are used to evaluate the *VADER* classification, using the k-fold cross validation process, and evaluated on accuracy, precision and recall metrics, defined underneath

$$Accuracy = (TP + TN) / (TP + FP + TN + FN) \quad -1$$

$$Precision = TP/\ (TP+FP) \quad -2$$

$$Recall = TP/\ (TP+FN) \quad -3$$

where, TP is true positive (the occurrences of correctly identified sentiments) and FP is false positive (the occurrences of incorrectly identified sentiments).

VI. RESULTS AND DISCUSSION

An analysis of the *VADER* extracted sentiments over time showed that while the proportion of "*positive*" and "*neutral*" tweets increased for both presidential candidates Donald Trump and Joe Biden in the run-up to the 2020 presidential elections, the rate of increase of such tweets was higher for Joe Biden than Donald Trump by 25%.

At the same time, while the proportion of "*negative*" tweets decreased for both presidential candidates, the rate of decrease of such tweets was also higher for Joe Biden by 50%. These results are captured in Fig 6.0.

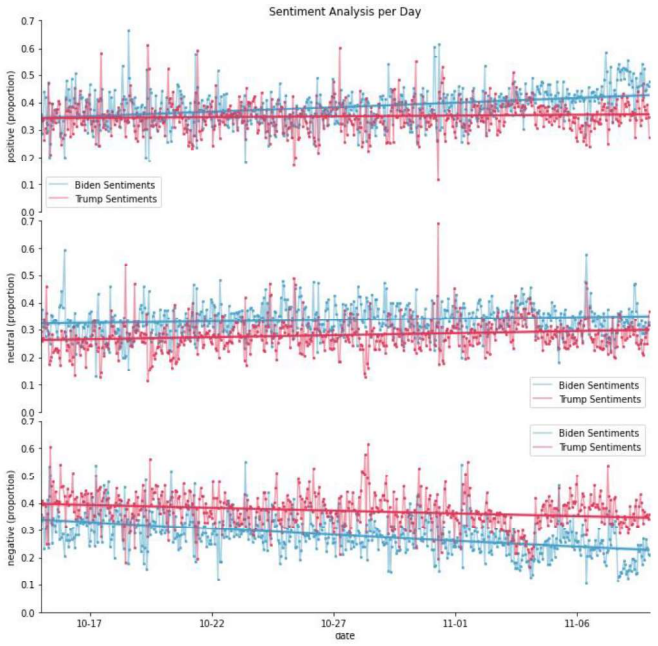

Fig 6.0. Proportion of positive, neutral, negative tweets over time

An analysis of the mean sentiment score over the one-month period immediately preceding the elections for all US states showed that while President Trump's mean sentiment scores lay mostly in a neutral [-0.5 -to +0.5] range, those for Joe Biden lay in a positive [+0.5 -to- +1.0] range. These results are captured in Fig 7.0 below.

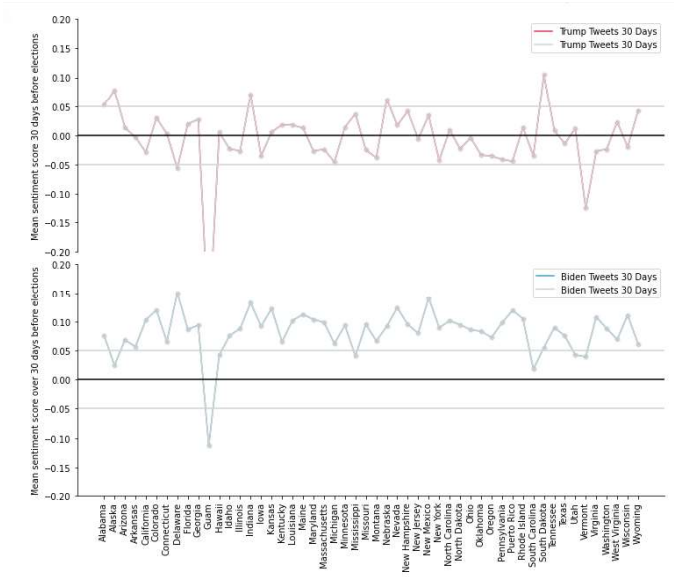

Fig 7.0. Mean sentiment score over time for all US states

An analysis of the mean sentiment score over the same one-month period immediately preceding the elections for only the swing states of Arizona, Florida, Georgia, Michigan, North Carolina and Pennsylvania showed an even more positive [0.10 -to- 0.15] range for Joe Biden while that for President Donald Trump continued to lay within the [-0.5 -to- +0.5] range. These results can be seen in the Fig.8.0 below.

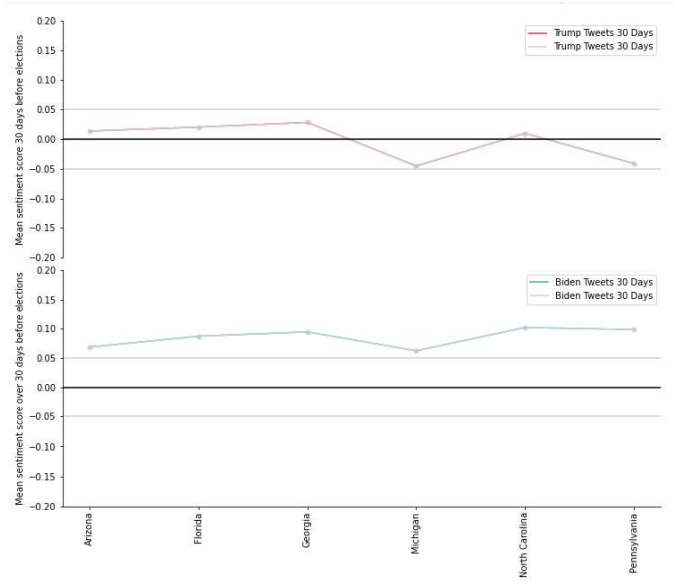

Fig 8.0. Mean sentiment score over time for all US swing states

An evaluation of the *VADER* sentiment categorization of tweets by sentiment polarity using different feature extraction techniques described previously in section V. (A-C) using the *Logistic Regression, Random Forest* and *XGBoost* models is captured in the Table 2.0 below.

Feature Extraction Technique	Classifier	Accuracy	Precision	Recall
Bag of n-grams	Random Forest	82.8%	86%	80.37%
Bag of n-grams + TF-IDF		82.3%	86%	78.89%
Bag of n-grams	Logistic Regression	85.4%	83%	80.3%
Bag of n-grams + TF-IDF		81%	84%	78.5%
Bag of n-grams	XGBoost	80.3%	80%	80%
Bag of n-grams + TF-IDF		80.3%	80%	80%

Table 2.0. Sentiment Evaluation

VII. CONCLUSION

This paper used the 9,000 features strong *VADER* lexicon to perform sentiment analysis on 1,747, 805 tweets from the one-month immediately preceding the US presidential elections, referencing "RealDonaldTrump" or "JoeBiden", originating from all around the world.

The sentiment scores were visualized and analyzed to obtain trends in proportion of "*positive*", "*neutral*", and "*negative*" sentiments over time. Mean sentiment scores were also calculated visualized and analyzed for all US states and swing states over time.

The *VADER* sentiment scores were measured over multiple metrics for accuracy using supervised learning models such as *Logistic Regression, Random Forest* and *XGBoost* and k-fold cross- validation using linguistic feature extraction techniques such as *Bag of N-grams* and its *TF-IDF* variants to ensure a higher confidence of prediction subsequently made.

Based upon the sentiment analysis results and their evaluated accuracy over multiple models, we predict with high degree of confidence that Joe Biden will emerge victorious in the 2020 US presidential elections.

VIII. REFERENCES

[1] J. Clement, "Leading countries based on number of twitter users as of July 2020", July 24, 2020

[2] Alexandre Bovet , Flaviano Morone & Hernán A. MakseJ, "Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump ", Scientific Reports, April 2017

[3] Jacky Mansky, "Age old problem fake news", Smithsonian magazine, May 7, 2018

[4] Fang Jin, Wei Wang, Liang Zhao, Edward Dougherty, Yang Cao, Chang-Tien Lu, Naren Ramakrishnan, ""Misinformation propagation in the age of twitter", IEEE, Volume:47, Issue 12, 2014

[5] Suchita Jain, Vanya Sharma, Rishabh Kaushal, "Towards automated real-time detection of misinformation on Twitter", 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)

[6] Mabrook S. Al-Rakhami ; Atif M. Al-Amri, "Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter", IEEE, Volume 8, 2020

[7] Michela Del Vicario, Sabrina Gaito, Walter Quattrociocchi, Matteo Zignani, Fabiana Zollo, "News Consumption during the Italian Referendum: A Cross-Platform Analysis on Facebook and Twitter", 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)

[8] Gang Li, Fei Liu, "Application of a clustering method on Sentiment Analysis", Journal of Information Science, Feb, 2012.

[9] Gleen. A. Dalaorao, Ariel. M. Sison, Ruji. P. Medina, "Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy", 2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA)

[10] Sahar A. El. Rahman, Feddah Alhumaidi AlOtaibi, Wejdan Abdullah AlShehri, "Sentiment Analysis of Twitter Data", 2019 International Conference on Computer and Information Sciences (ICCIS).

[11] Hay Mar Su Aung; Win Pa Pa , "Analysis of Word Vector Representation Techniques with Machine-Learning Classifiers for Sentiment Analysis of Public Facebook Page's Comments in Myanmar Text", 2020 IEEE Conference on Computer Applications(ICCA).

[12] C.J Hutto, Eric. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", 2015 Conference: Proceedings of the Eighth International AAAI Conference on Weblogs and Social MediaAt: Ann Arbor, MI.

[13] Giulio Angiani, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi "A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter", 2016 KDWeb.