Election Sentiment Analysis Through Natural Language Processing of Political Discourse

Kadir Hasan
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
kadir.hasan@g.bracu.ac.bd

Kaji Sajjad Hossain
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
kaji.sajjad.hossain@g.bracu.ac.bd

MD Zubairul Islam
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
md.zubairul.islam@g.bracu.ac.bd

Noshin Tabassum
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
noshin.tabassum2@g.bracu.ac.bd

Annajiat Alim Rasel
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat@gmail.com

Md Sabbir Hossain
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

***Abstract*—Sentiment analysis, often known as opinion mining, is a natural language processing (NLP) methodology that uses algorithms and computational methods to identify and extract sentiments, views, or attitudes conveyed in text. The purpose of sentiment analysis is to understand and categorize the subjective information given in the text as positive, negative, or neutral. Our project aims to use advanced Natural Language Processing (NLP) methods to understand and study the political speech that happens during elections; so that we can predict the winner of the election. The US election is a very hot issue around the world as it affects almost every aspect of the world. Moreover, it is the only election about which people from every country share and post their opinions and judgments on social media such as Facebook, Twitter, Youtube etc. Our paper focuses on deriving significant observations from textual data, including speeches, news and articles that are uploaded on Twitter in order to assess the dominant sentiments pertaining to electoral candidates and matters. The methods we used to carry out our analysis is data preprocessing, data mapping and sentiment analysis. We preprocessed the collected dataset in order to remove unnecessary texts that are not considered useful for sentiment analysis. For data mapping, we used various processes of data visualization such as pie-charts, bar charts in order to show the numerical data of tweets from various countries. Finally, for sentiment analysis, we used three different Natural Language Processing models namely Textblob, BERT and VADER.**

***Index Terms*—sentiment analysis, NLP, election, twitter, trump, biden**

I. Introduction

The progression of media coverage throughout history, encompassing print, television, and social media, underscores the dynamic nature of information distribution. The allusion to John F. Kennedy serves to emphasize the importance of campaign speeches in providing the electorate with information that enables them to form well-informed judgments regarding leadership, political ideology, and potential courses of action. In light of this particular context, the objective of our research is to conduct an analysis of political discourse that occurs throughout election campaigns.The 2020 United States Presidential Election witnessed a fiercely contested clash between the current President, Donald Trump, who represented the Republican Party, and Joe Biden, the Democratic Party's contender and former Vice President. Both candidates possessed discernible policy platforms and ideologies, with President Trump emphasizing his administration's economic record prior to the pandemic and conservative agenda, while Joe Biden concentrated on unity, healthcare reform, and tackling social justice concerns. In the end, Joe Biden successfully achieved victory by capturing crucial swing states and obtaining a clear majority in the Electoral College. The objective of our research might be to comprehend the sentiments conveyed in campaign speeches during US election 2020, with an investigation into the ways in which language is employed to sway public opinion and decision-making. The project may endeavor to discern patterns, sentiments, and trends within the extensive textual data produced throughout election campaigns by employing NLP techniques to political discourse. Our analysis would yield the future winner of an election by analyzing the sentiments of people during the election.

Through the ability to actively engage and communicate, social media platforms like Twitter have been able to bring people from all over the world closer together. Sentiment analysis is a technique that can analyze public opinion on a particular problem. Election-related political discourse is dynamic and provides an extensive amount of vocabulary that reflects the views, opinions, and sentiments of the population. To explore this complicated system, we are using sentiment analysis of Natural Language Processing (NLP) approaches to extract, analyze, and grasp the sentiment. We used NLP tools namely TextBlob, BERT and VADER to facilitate diverse texts and comments of people. TextBlob has an easy-to-use interface for tasks such as sentiment analysis, part-of-speech

tagging, and noun phrase extraction. BERT, a cutting-edge deep learning model, excels in comprehending context and semantics, delivering subtle insights into sentiment and language structures in more complicated texts. VADER, a sentiment analysis tool that uses a lexicon and rule-based approach, is specifically designed to evaluate the intensity and polarity of sentiment, finally analyzing comments on social media platforms.

Our study shows how engineering education can be used in many different ways. Our paper stresses how important it is to make decisions based on data, think about ethics when using AI, and communicate technical results clearly to a wider audience—all skills that engineering students need to learn. Along with that, our study shows how technological advances can be used for the greater good, showing how they can help us understand how people really feel and make participation in government more meaningful.

II. Dataset

A. Dataset Description

The dataset we used in our paper was fetched from the website Kaggle. The main context of this dataset was the election of US which happened on 3rd November, 2020. The impact of this election was a big deal to the world. This dataset was created by analysing data from the social media platform "Twitter.". Two API's named "status_lookup" and "snsscrape" were used to find the keywords. This dataset got several updates from October to November 2020. There are two different contestants "Donald Trump" and "Joe Biden" and almost 958580 tweets were fetched in the analysis of the paper. The noticeable features are, "tweet" which indicates full tweet text, "likes" which indicates number of likes, "source" that indicates utility used to post the tweets, "user_location" which indicates the location from where the tweet posted, "city" that indicates the city parsed from the user's location, "country" indicates the country name of the user which was parsed from "user_location".

The inspiration for this dataset was the form of correlation between sentiment of users on twitter and the eventual election result. The author of the dataset also focused on some points like, whether the tweets can manipulate the election or if we can predict the outcome, including each state.

B. Implementation of Data

We used a variety of features for our paper and pre-process them as a usable data variables. We also did data cleaning and also dropped useless columns like tweet_id', collected_at' , 'user_description', 'collected_at'. There were individual datasets for Trump and Biden and we mixed them together for usability. Then we visualized the mixed dataset using some visualization methods to find out the number of tweets in various circumstances for both contestants, Trump and Biden.

III. Literature Review

Finity et al. [1] approach to examine the 2020 US Presidential Election campaign speeches consisted of gathering speech transcripts from YouTube using the API, processing them, and doing an analysis. Speech-to-text transcripts that were automatically generated were utilized instead of official transcripts. Following the cleaning, tokenization, and preprocessing of the data, 172 speeches by Trump, Biden, Pence, and Harris were produced. Because the material was unstructured, the analysis mostly relied on Bag of Words (BOW) techniques. While PCA and HCA assisted in clustering utterances based on similarities, techniques like TF-IDF were utilized to find essential terms. Speech themes were found using Latent Dirichlet Allocation (LDA). The feelings elicited by each speech were also recorded using emotional intensity analysis utilizing the NRC Emotional Intensity Lexicon. A transparent analysis that is comprehensible to a broad audience was the aim, and potential political bias was to be avoided.

Gorro et al. [2] focused on analyzing student behavior on social media one week before the 2022 Philippines Presidential Election. 300 students from Cebu Technological University Carmen Campus contributed 5321 posts on Facebook in total. The analysis of these postings was done by the study using machine learning methods such as Support Vector Machine (SVM), Latent Dirichlet Allocation (LDA), and K-means Clustering. The final result was a 73% accuracy rate in identifying posts as neutral, positive, or negative marketing. Due to the study's small sample size of a particular student group and reliance on quantitative methodologies, generalizability may be limited. To mitigate this, future research must use a variety of platforms, mixed-methods, and diverse sample sizes to conduct thorough analyses.

Solitana et al. [3] aims to comprehend the dynamics of hate speech in the internet debate surrounding the Philippine election season. By making hate speech classifiers better, it hopes to lessen its prevalence in the world's cyberspace and steer online discussions in the direction of more constructive discourse. This study makes use of a dataset of tweets related to the 2016 Philippine election, which includes a significant number of tweets classified as hate or non-hate and linked to a variety of hate targets (such as race, physical characteristics, sex, handicap, religion, class, and quality). Deep learning classifiers, distributional semantics, and keyword-based techniques are some of the methods used to classify hate speech. Specifically, the study uses cluster analysis and linguistic comparisons to examine tweets that contain hate and those that do not. The study's application is limited to the dataset from the 2016 Philippine Presidential Election; hence, in order to assure wider generalizability, more research using other datasets is necessary.

Nugroho et al. [4] basically focuses on the lexicon based token analysis in order to analyze the sentiment of a sentence. Lexicon-based sentiment analysis evaluates each word's sentiment in the text according to the sentiment scores listed in the lexicon. Here, based on the dataset sentiment analysis is carried out on the tweet using the VADER model. This model is used to research lexicons and rule based sentiment that is specifically matched with social media statements.Using context-aware rules, ongoing updates, and domain-specific improvements can help overcome the shortcomings of the VADER model's lexicon-based sentiment analysis and increase accuracy.

Naiknaware et al. [5] focuses on evaluating the public sentiment towards Indian government schemes using Twitter data. The study employs sentiment analysis to classify opinions on schemes like GST, Demonetization, and Digital India. By analyzing Twitter datasets from 2016 to 2018, the research predicts the 2019 Indian election outcomes based on sentiment polarity. The findings suggest positive sentiments for GST, Digital India, Make in India, Startup India, Swacha Bharat, and Yoga Day, while negative sentiments surround Demonetization and Kashmir. The study underscores the efficacy of sentiment analysis in gauging public opinion. Diversifying sources (social media platforms, polls, interviews) can help reduce the study's possible bias from depending just on Twitter data, which lacks broad representation. This will allow for a more thorough examination of public attitude.

Kavitha et al. [6] explores sentiment analysis on Twitter data using NLP and machine learning. It focuses on tweets tagged in voting systems. Employing text mining and algorithms like Random Forest, Decision Tree, and Logistic Regression, the study achieves a 95% accuracy rate with Random Forest. This research offers insights into public sentiment, emphasizing the prominence of social media for expressing opinions. The study suggest potential applications for predicting user behavior based on sentiment analysis, highlighting avenues for future research in the field. Due to platform-specific biases resulting from the study's Twitter-centric focus, future research should diversify data sources to provide a more thorough understanding of public sentiment across a variety of online platforms.

IV. Methodology

In this section, we discussed the proposed methods we used for our sentiment analysis to get an overview of US Election 2020. In our paper, we generally focused on the analysis of the dataset by implementing different NLP sentiment analysis based models and find the best possible models to visualize different sentiment conditions like, positive, negative and neutral to predict the outcome of the election.

A. Data Pre-Processing

Data pre-processing is a crucial part of our analysis. Firstly, we retrieve data from two CSV files for Biden and Trump. Then we dropped some of the useless columns to mitigate the complexity of our dataset. Further on, we mixed two datasets for Trump and Biden to compare the results between these two contestants and sort them using the feature 'created_at'. We also provided some visualizations of the processed data for both contestants in terms of the number of likes and tweets. Later on, we used some pie charts to show the number of tweets posted from different countries and states. Furthermore, we created some class function for chart building to visualize the top 5 countries of the source of the tweets we collected from. We also visualize the top 10 sources of posting the tweets from. At the end, we summarized a visual bar chart of the top continents and cities where the tweets were posted from. We also did text cleaning by removing square brackets, hashtags, links, punctuations and words that contain numbers. The uppercase letters are also being lowercase in that process.

B. Sentiment Analysis

*1) **TextBlob**:* TextBlob is a Python package used in natural language processing (NLP) that offers pre-built models for activities like sentiment polarity evaluation, part-of-speech tagging, noun phrase extraction, and more. This makes text processing jobs like sentiment analysis easier. Text is read by TextBlob's sentiment analysis technology, which determines if the content is favorable, negative, or neutral. It is useful for deciphering sentiments in sentences since it accomplishes this by employing unique techniques and learning from linguistic patterns.

In our paper, we implement this model to figure out the sentiment analysis for both Trump and Biden. At first, we conduct the model on the dataset of Trump where duplicate rows with same user are dropped by keeping only the first occurrence. Then we dropped any rows with missing values and filter out the country column only with United States of America. Later on, we used a clean function by creating a new column. Three new columns were added 'subjectivity', 'polarity', and 'analysis', created by applying the functions getSubjectivity, getPolarity, and getAnalysis to the 'ClearTweet' column, respectively. The subjectivity shows how subjective the tweet is. The sentiment polarity is represents by the polarity and lastly the analysis explained three different values of positive, negative and neutral.

After that, we used counter class to count the occurrence of each unique value in the polarity for Trump and later on for Biden. In the case of analysis of textblob function, we created a sentiment score to verify the result. In that case, if the score is less than 0, it returns negative, if equal to 0, it returns neutral and if greater than 0, it returns positive.

*2) **BERT**:* Google created a state-of-the-art deep learning model called BERT (Bidirectional Encoder Representations from Transformers) that analyses words in text bidirectionally,

taking into account both the words that come before and after them to better understand their meanings.

In our paper, we firstly imported some libraries and initialize sentiment analysis using BERT model. The BERT analysis function takes a piece of text as input and analyze the sentiment. The result contains in a dictionary extracting using a label and result. Based on the label, the function returns negative, positive and neutral. Then we applied our result to the ClearTweet column which initially contains the cleaned text from our dataframe. The finalized results are also visualized later on.

3) ***VADER****:* With word analysis, VADER is a sentiment analysis tool that can identify emotions in text and provides a sentiment score that indicates whether the text is favourable, negative, or neutral. VADER, a lexicon-based sentiment analysis tool, gauges sentiments in text by analyzing words and punctuations to determine whether the expressed emotion is positive, negative, or neutral. Its strength lies in evaluating sentiments in social media content due to its built-in rules that comprehend emoticons, slang, and even capitalization.

In the paper, we firstly initiated necessary libraries for this model and initialize VADER intensity analyzer. Then we implement a function where it takes text inputs to capture compound sentiment scores. If the compound score is greater than equal to "0.05", it returns positive sentiment. If the score is less than equal to "-0.05" then it returns negative sentiment. Otherwise it will pass a neutral sentiment. After that, we created a new column to store the result fetched from VADER analysis. We also compare the result with ClearTweet to make it usable for further visualization.

V. Result Analysis

A. *TextBlob*

TextBlob employs a pre-trained sentiment analysis model and a sentiment lexicon to categorize text as positive, negative, or neutral. While applying Textblob on the processed data for Trump, we noticed that the maximum number of polarity scores imply Neutral. Our obtained percentages of positive, negative and neutral after application of Textblob on the texts relating to Trump are 38.90%, 20.00% and 41.10% respectively. Again, while applying Textblob on the processed data for Biden, we noticed that the maximum number of polarity scores are Positive. Our obtained percentages of positive, negative and neutral after application of Textblob on the texts relating to Biden are 44.70%, 18.50% and 36.80% respectively.

From the charts, it is noticeable that the percentage of positivity of Biden (44.70%) is higher than the percentage of positivity of Trump (38.90%). That means, maximum tweets are tweeted in support of Biden.

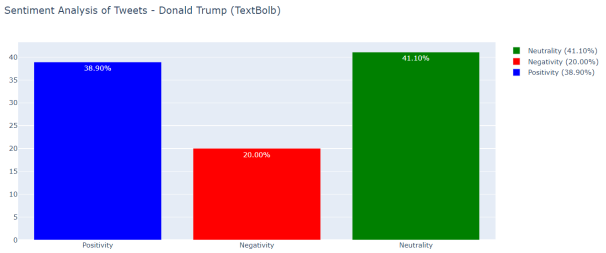

Fig. 1. Sentiment Analysis of Tweets - Donald Trump (TextBlob)

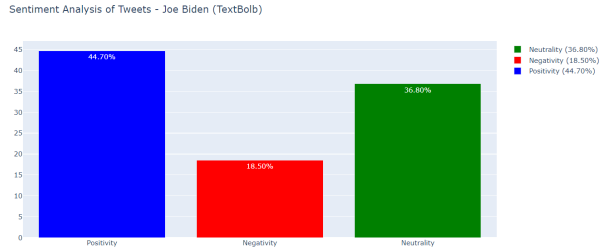

Fig. 2. Sentiment Analysis of Tweets - Joe Biden (TextBlob)

B. *BERT*

While applying BERT on the processed data for Trump, we noticed that the maximum number of polarity scores labeled as Negative; that means maximum tweets are against Trump. Our obtained percentages of positive and negative after application of BERT on the texts relating to Trump are 24.80% and 75.20% respectively. Again, while applying BERT on the processed data for Biden, we noticed that Biden's positivity of polarity scores are higher than Trump's positivity of polarity score. Our obtained percentages of positive and negative after application of BERT on the texts relating to Biden are 31.40% and 68.60% respectively. Moreover, the negativity of polarity score after application of BERT is higher in case of Trump, implying that maximum tweets that are related or mentioning him are against public interests.

From the charts, it is noticeable that BERT's prediction does not give any neutrality, it only gives an accurate analysis of the sentiments.

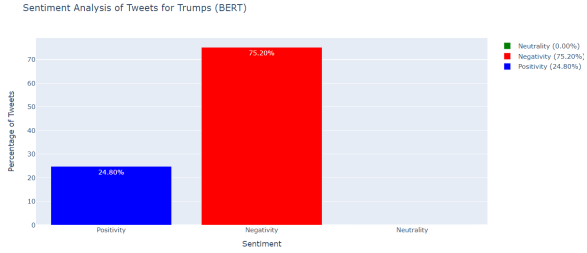

Fig. 3. Sentiment Analysis of Tweets - Donald Trump (BERT)

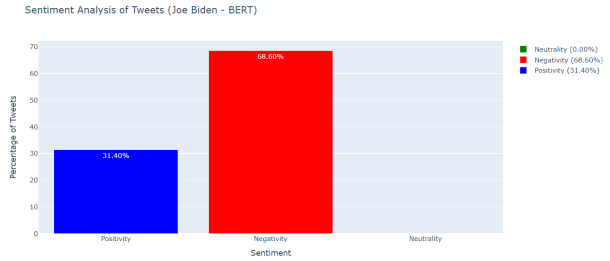

Fig. 4. Sentiment Analysis of Tweets - Joe Biden (BERT)

C. *VADER*

While applying VADER on the processed data for Trump, we noticed that the maximum number of polarity scores labeled as Negative; that means maximum tweets are against Trump. Our obtained percentages of positive, negative and neutral after application of VADER on the texts relating to Trump are 34.40%, 38.40% and 27.20% respectively. Again, while applying VADER on the processed data for Biden, we noticed that the maximum number of polarity scores labeled as Positive; that means maximum tweets are in support of Biden. Our obtained percentages of positive, negative and neutral after application of VADER on the texts relating to Biden are 39.70%, 30.40% and 29.90% respectively.

The charts clearly imply that the maximum number of people express a positive opinion about Biden than that of Trump.

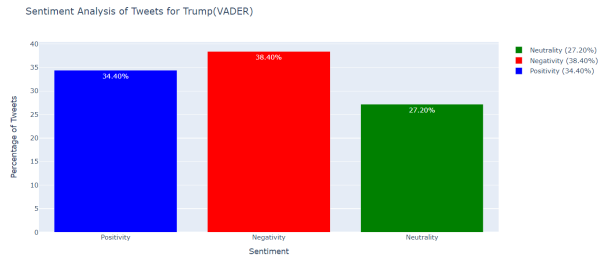

Fig. 5. Sentiment Analysis of Tweets - Donald Trump (VADER)

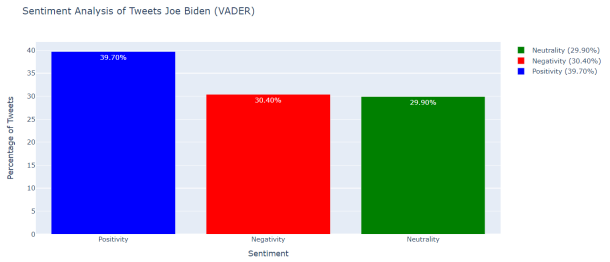

Fig. 6. Sentiment Analysis of Tweets - Joe Biden (VADER)

Final Analysis of the Result: The TextBlob model analyzes the winner of the election by comparing the percentage of Positive labels of Trump and Biden. Secondly, the BERT model analyzes the winner of the election by comparing the percentage of Negative labels of Trump and Biden. Finally, the VADER model analyzes the winner of the election by clearly indicating which trained dataset got the highest percentage label; after application of VADER model, we noticed Trump's trained dataset got the highest percentage for Negative label whereas Biden's trained dataset got the highest percentage for Positive label.

Label	Donald Trump			Joe Biden		
	TextBlob	BERT	VADER	TextBlob	BERT	VADER
Positive	38.90%	24.80%	34.40%	44.70%	31.40%	39.70%
Negative	20.00%	75.20%	39.40%	18.50%	68.60%	30.40%
Neutral	41.10%	-	27.20%	36.80%	-	29.90%

Fig. 7. Table of the Overview of Different Models

FUTURE WORKS

- **Cross Linguistic Analysis:** Expand analysis to include sentiment analysis in different languages or multilingual sentiment analysis to capture a diverse user base. In our paper, we only use English text language to analyze tweets from the internet.
- **User Behavior Analysis:** Different engagement metrics (retweets, replies and mass user influences) can be applied to understand the impact of influencers and community on different sentiments. The dataset can be used in different aspects to find the best models for sentiment analysis.
- **Multi-Platform Data Analysis:** In this paper, we only use twitter social media to extract our dataset. This can be multi-platform base like Facebook, Reddit, Instagram etc. Different social data extracting methods like, newspaper, speech and record can be also used as the datagram for this paper.

CONCLUSION

In conclusion, our election sentiment analysis research using advanced Natural Language Processing (NLP) models delivers valuable insights with extensive applicability. Our research assists political campaigns in improving strategy and tailoring messaging to resonate effectively with the voters by comprehensively assessing popular opinions in political debate. The sentiment analysis results are used to monitor public opinion in real time, assisting political entities, policymakers, and media outlets in remaining informed and altering narratives. Policymakers can use our findings to help influence decision-making and match policies with public preferences. Media outlets can improve their reporting and analysis, resulting in a more comprehensive view of election dynamics. Our study also adds to social and political research by serving as a resource for in-depth studies of electoral behavior and candidate perception. Furthermore, the capacity

to detect sentiment trends acts as an early warning system, allowing stakeholders to address issues as soon as they arise. Overall, our findings not only enrich academic understanding, but also provide practical insights that have the potential to improve decision-making, communication methods, and public involvement in the vital domain of electoral politics.

REFERENCES

[1] Kevin Finity, Ramit Garg, and Max McGaw. A text analysis of the 2020 u.s. presidential election campaign speeches. In *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6, 2021.

[2] Ken Gorro, Leodivino Lawas, Rosein A. Ancheta, and Anthony Ilano. Understanding social media behavior in philippines presidential election using natural language processing. In *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, pages 1–5, 2022.

[3] Nico T. Solitana and Charibeth K. Cheng. Analyses of hate and non-hate expressions during election using nlp. In *2021 International Conference on Asian Language Processing (IALP)*, pages 385–390, 2021.

[4] Deni Kurnianto Nugroho. Us presidential election 2020 prediction based on twitter data using lexicon-based sentiment analysis. In *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 136–141, 2021.

[5] Bharat R. Naiknaware and Seema S. Kawathekar. Prediction of 2019 indian election using sentiment analysis. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, pages 660–665, 2018.

[6] M. Kavitha, Bharat Bhushan Naib, Basetty Mallikarjuna, R. Kavitha, and R. Srinivasan. Sentiment analysis using nlp and machine learning techniques on social media data. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 112–115, 2022.