A STUDY: SENTIMENTAL ANALYSIS FOR ELECTION RESULTS BY USING TWITTER DATA

Nuthanakanti Bhaskar¹, Pudi Sandeep², Nethi Amarnath³, Joel Thankachan⁴
¹Associate Professor, ²,³,⁴Student
¹,²,³,⁴CSE Department, CMR Technical Campus, Hyderabad, Telangana, India.
¹bhaskar4n@gmail.com, ²177r1a05a1@gmail.com, ³177r1a0596@gmail.com,
⁴177r1a0584@gmail.com

ABSTRACT

The entire world is changing at a breakneck pace, and technology is no exception. User-generated data is abundant on social networking platforms like Twitter. Users from all around the world offer their thoughts, opinions, ideas, and feelings about a variety of topics, including products, movies, and politics. Manual sentiment analysis is a time-consuming task. Opinion mining has recently gained popularity as a result of the large volume of opinionated data available on social networking sites such as Twitter. In this paper, we used a chronological approach to data collection, data pre-processing, emotional analysis, and machine learning analysis to forecast the outcome of the US 2020 presidential election using Twitter emotional analysis. We used a Random Forest classifier after completing a literature review and comparing all supervised ensemble machine learning algorithms to determine which one was the best. The proposed technique was tested on Twitter data, and it outperformed existing approaches.

Keywords: Twitter Sentiment Analysis, Twython, Random Forest Classifier, Polarity Detection, Social Network, Feature Engineering.

I. INTRODUCTION

Sentiment analysis, also referred to as "opinion mining" or "emotion Artificial Intelligence," refers to the systematic recognition, extraction, evaluation, and examination of emotional states and subjective information using natural language processing (NLP), text mining, computational linguistics, and bio measurements. Sentiment analysis is concerned with the voice in customer materials, such as online surveys and evaluations and web-based social networks [3].

Twitter, also known as the Online Social Network (OSN), is a rapidly growing online platform where users may create, post, update, and read brief text messages known as tweets. Users can contribute their opinions, beliefs, and thoughts on a certain topic via tweets. Sentiment Analysis (SA) is a method of recognizing and categorizing polarity in general. This method is employed in a variety of industries, including e-commerce, health care, entertainment, politics, and so on [4]. The ability to draw conclusions from such unstructured and uniform data is extremely beneficial in a variety of industries. The extremely unstructured format of opinion data available on the internet, however, makes mining difficult [11].

Opinion mining (OM) and sentiment analysis (SA) are two new fields aimed at assisting users in finding opinionated content and detecting sentiment polarity. The terms OM and SA are frequently interchanged to indicate the same concept. Some researchers, on the other hand, claim to be attempting to solve two distinct challenges. The focus of OM is on determining whether a piece of text contains opinion, a problem known as subjectivity analysis, whereas the focus of SA is on sentiment polarity detection, which assigns a positive or negative sentiment to the opinion of the investigated text. OM or SA is the “in-text computational analysis of opinions, feelings, and subjectivity” [12].

In this work, we used 3-way sentimental polarity classification (i.e., Twitter sentiment analysis) they are Positive, negative and neutral: a. Positive and negative tweets are far less common than neutral tweets. This is in contrast
to other sentiment analysis domains (such as product reviews), which are primarily positive or negative; b. There are issues with language representation, such as feature engineering; c. Tweets are often short and contain little emotional indications [10].

Social media is now used by the general population to react to political parties. These crucial public comments are not processed to produce relevant information that can be used to construct a better image of the public voice in the political environment due to the enormous number and diversity of posts. Obama's campaign elevated social media and the field of opinion mining to new heights, resulting in election predictions that were only 2.5 percent off the mark. The impact extended far beyond forecasting election winners [13].

Sentimental analysis is used in a variety of domains and characteristics. We took it to the next level in this work by using polarity to forecast the winner among rivals. However, based on tweets, this isn't a final decision. We aim to convey that we care about the national interest or the voter's choice before they cast their vote.

We utilised Natural language processing because most tweets are written in natural language (NLP). According to the opinions mentioned [23], NLP plays a significant role and can be utilised. Because inconsistency can influence accuracy, feature extraction or pre-processing is critical in prediction.

Random Forest, Support Vector Machine, Nave Bayes, Maximum Entropy, and Decision Tree are just a few of the base classifiers used to forecast accuracy. We chose the Random Forest classifier out of all the supervised learning techniques because it achieves the highest level of accuracy across all classifiers, can handle enormous data with ease, and time consistency is a crucial factor.

II. LITERATURE SURVEY

Budiharto et al. [8] employed sentiment analysis to forecast the Indonesian Presidential election result using tweets from President Candidates Jokowi and Prabowo, as well as tweets from pertinent hashtags, collected from March to July 2018. The authors used an algorithm and approach to count relevant data, top words, train the model, and forecast sentiment polarity. According to the results of the experiment, Jokowi is now leading the current election projection. This forecast result corresponds to four Indonesian survey institutes that confirmed the system delivered accurate forecast results.

Gull et al. [13] proposed an approach to make opinion mining easier and more efficient by combining linguistic analysis with opinion classifiers to identify polarity for Pakistan's political parties. A method is offered that preprocesses Twitter's raw data and compares two classification approaches, such as Naive Bayes and Support Vector Machines, to classify the data and achieve the necessary accuracy.

Chauhan et al. [27] discussed and evaluated the effectiveness of several volumetric, emotion, and social network techniques in predicting key decisions from online social media platforms. In this regard, election result forecasting is an application of sentiment analysis aiming at predicting the outcomes of a current election by evaluating popular sentiment via social media. They looked at sentiment analysis tools and tried to explain how the researchers may use social media material to predict election results. This research aids in determining the political views of online users who utilise social media sites like Facebook and Twitter.

Hasan et al. [15] analysed public perceptions of a product using Twitter data. To begin, they created a preprocessed data architecture based on natural language processing (NLP) to filter tweets. Second, sentiment was analysed using the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) model concepts. This is a project that combines BoW and TFIDF to precisely classify positive and negative tweets. Finally, utilising the NLP approach, the Model achieved 85.25 percent accuracy in sentiment analysis.

Ahmad et al. [2] used three approaches: Lexicon-based techniques, Machine Learning-based techniques, and hybrid techniques, and applied them to the unigram, bigram, and trigram, respectively, to display different scores. Random Forest, Naive Bayes, Support Vector Machine, and other Machine Learning Tools were employed.

With the goal of increasing the performance and accuracy of the sentiment classification methodology, Ankit et al. [4] created an ensemble classifier that combines the base learning classifier to produce a single classifier. The suggested ensemble classifier outperforms stand-alone classifiers and the widely used majority vote ensemble classifier, according to the results.
Ankita Sharma et al. [5] developed an approach that combines text mining with sentiment analysis, and this study combines the two. The strategy was sound and methodical. It featured a text analysis of tweets prior to undertaking sentiment analysis. The tools used in this project were chosen after a thorough examination of the literature and feature engineering.

In comparison to typical machine learning algorithms, Asgarnezhad et al. [6] proposed a new strategy to analysing natural language and the frequency of terms from one sentiment class, resulting in a better sentiment classification. The authors compare the results achieved with typical machine learning techniques using RapidMiner with entirely polarity centred strategies utilising the Python programming language and the NLTK module in this study.

Bouazizi et al. [7] introduced a novel methodology that, in addition to the aforementioned binary and ternary classification tasks, delves deeper into the categorization of texts obtained from Twitter, requiring a set of tweets to be classified into seven different sentiment classes. The proposed method is scalable, and it may be used to categorise texts into several categories. The authors begin by introducing SENTA, a customised tool designed to assist users in selecting the attributes that are most appropriate for their application from a large number of options, and then running the classification through a simple graphical user interface. In comparison to the present system, their work obtained a greater level of accuracy.

eili et al. [9] worked on a variety of Datasets and concentrated primarily on polarity classification. They pre-processed each dataset's data and using chosen Machine Learning to see how accurate the prediction was, and they got the results they wanted.

Da Silva et al. [10] proposed a method for automatically classifying tweet sentiment utilizing classifier ensembles made up of Multinomial Naive Bayes, Support Vector Machines, Random Forests, and Logistic Regression, as well as lexicons. In relation to a query keyword, tweets are categorized as positive or negative. This method can help customers and businesses get to know one another better.

Giachanou et al. [12] investigated and briefly described the methods proposed for sentiment analysis in Twitter to provide an overview of the field. They explored topics such as Twitter opinion retrieval, sentiment monitoring over time, irony recognition, emotion recognition, and tweet sentiment quantification, as well as other topics relating to sentiment analysis on Twitter. Naive Bayes, Maximum Entropy, Support Vector Machines, Multinomial Naive Bayes, Logistic Regression, Random Forest, and Conditional Random Field are some of the supervised machine learning techniques employed.

Pollacci et al. [21] presented a method for automatically extending the dictionary used in lexicon-based sentiment analysis based on epidemic spreading. The resulting lexicon is proven to have valences that correspond well with human-annotated sentiment and to provide tweet sentiment classifications that are comparable to the original lexicon, with the added benefit of being able to tag more tweets than the original dictionary. SVM is used to categorise the train and test sets to cross-validate, and the method is easily adaptable to multiple languages and suitable to vast volumes of data.

Desai et al. [11] first described a complete approach for performing sentiment analysis on extremely unstructured Twitter data and categorizing it into true or false categories. Second, they explored several strategies for performing sentiment analysis on Twitter data, including knowledge-based and machine learning techniques such as SVM, Naive Bayes, and maximum entropy. Furthermore, they are given a parametric comparison of the described strategies based on the parameters.

Hasan et al. [14] used a hybrid strategy that included both a sentiment analyser and machine learning. They also used supervised machine-learning algorithms like Naive Bayes and support vector machines to compare sentiment analysis techniques in the analysis of political viewpoints (SVM).

Abbas et al. [1] On the Twitter processed dataset, they proposed a method that combined four classifiers, each using a different technique—naive Bayes, SVM decision trees, multilayer perceptron, and logistic regression—and they show the contrast between individual classifier score and the proposed method, achieving higher accuracy than the individual classifiers.
Kolchyna et al. [17] proposed a system that incorporates both lexicon-based and machine learning approaches to sentiment analysis. They talked about how to create features and how to choose features for machine learning sentiment categorization. They run these algorithms using a benchmark Twitter dataset from the SemEval-2013 competition and task 2-B to measure the performance of the main sentiment analysis methods over Twitter. The results suggest that the SVM and Naive Bayes classifiers-based machine learning approach beats the lexical technique.

Machine learning, ensemble approaches, and dictionary (lexicon) based approaches were all discussed by Alsaeedi et al. [3]. In addition, hybrid and ensemble approach for analyzing Twitter sentiment was investigated. Maximum Entropy and Naive Bayes are used in the Support Vector Machine to determine how accurate the polarity classification prediction is on the data.

Ismail et al. [16] tested multiple sentiment analysis classifiers in different experimental setups to see how effective they were at sentiment mining Twitter data. The Stanford Testing Sentiment dataset (STS) was used for this research. Multinomial Naive Bayes outperforms other classifiers in Twitter sentiment analysis, and is less impacted by data sparsity, according to the findings of this study.

Krouska et al. [18] explained how to gather the necessary data to pre-process reviews in order to identify sentiment and determine whether it is favourable or unfavourable. A detailed comparison of sentiment polarity classification algorithms for Twitter text is presented, as well as the role of text pre-processing in sentiment analysis. Possible combinations of approaches and reports on their efficacy were included in the set of tests, which were conducted using manually annotated Twitter datasets. Finally, it has been demonstrated that feature selection and representation can have a positive impact on classification performance.

Machine learning, ensemble approaches, and dictionary (lexicon) based approaches were all discussed by Alsaeedi et al. [3]. In addition, hybrid and ensemble approach for analyzing Twitter sentiment was investigated. Maximum Entropy and Naive Bayes are used in the Support Vector Machine to determine how accurate the polarity classification prediction is on the data.

In the framework of document categorization, Martinčić-Ipšić et al. [20] conducted a comparative examination of 3 models for a feature representation of text documents. To categorize the document, they employed the most commonly utilized family of bag-of-words models and the Random Forest Classifier. Finally, superior performance was demonstrated in the job of classifying huge documents.

A sentiment classification model for Twitter tweets was proposed by Sevim et al. [22]. Short communications are first expanded with BabelNet, a concept network, in this architecture. The messages are then enlarged and the original form is included in an ensemble learning model. As a result, when the ensemble model was compared to traditional classification techniques, the F-measure value was shown to be higher NB, DT, Sequential Minimal Optimization.

Singh et al. [23] discuss how pre-processing Twitter data might improve sentiment categorization, particularly in terms of slang phrases. To check the relevance and sentiment translation of the slang term, the suggested pre-processing approach depends on the bindings of slang terms on other coexisting words. They employed n-gram to locate the bindings and conditional random fields to determine the slang word's relevance. Experiments were carried out to see how the proposed method affected sentiment categorization, and the results clearly show that the accuracy of classification has improved.

Yadollahi et al. [24] presented state-of-the-art methods such as i. a taxonomy of sentiment analysis; ii. a survey on polarity classification methods and resources, particularly those related to emotion mining; iii. a comprehensive survey on emotion theories and emotion-mining research; and iv. some useful resources, such as lexicons and datasets.

Zhao et al. [25] investigated the impact of text pre-processing on sentiment classification performance in two types of classification tasks, and compared the classification results of six pre-processing approaches using two feature models and four classifiers on five Twitter datasets. The results reveal that utilising the pre-processing methods of expanding acronyms and replacing negation improves the accuracy and F1-measure of the Twitter
sentiment classification classifier, but that deleting URLs, numerals, or stop words little impacts the accuracy and F1-measure. When various pre-processing approaches were used, the Naïve Bayes and Random Forest classifiers were found to be more sensitive than Logistic Regression and support vector machine classifiers.

Jaiswal et al. [26] employed Python, Natural Language Toolkit, Natural Language Processing, and Data Collection to study sentiment analysis. In addition, polarity is represented by a plotting shape.

Every effort connected to Twitter sentiment analysis, according to our survey, is polarity detection oriented, which is how opinion mining works. We devised our proposed methodology, illustrated in Fig 1, to forecast and achieve the requisite accuracy of prediction from individual tweets.

![Proposed Method](image)

**Fig 1. Proposed Method**

To analyse sentiments, we followed six steps in this proposed method: a. Data collection. b. Pre-Processing of data. c. Polarity. d. Subjectivity and sentiment extraction. e. Data visualisation and f. Applying machine learning Algorithm on pre-processed data to achieve desired accuracy.

### III. METHODOLOGY

In this paper, we offer a Random Forest Classifier-based technique for forecasting election results on Twitter sentiment. The use of a Random Forest Classifier is intended to improve accuracy. The suggested method entails data collecting, pre-processing, sentiment analysis, and machine learning analysis, with the findings visualised in a bar graph, as shown in Fig 1.

![Architecture](image)

**Fig 2. Architecture**

#### 3.1 Data Collection

Because we chose Twitter as our real-time data source, Python provides a library called Twython for retrieving data from Twitter. Twython makes it simple to use the Twitter API without leaving the current window. To use twython to retrieve tweets, we'll need 4 keys: the Consumer Key, Consumer Secret, Access Token, Access Token Secret, and Authentication Keys. We created a Twython instance with the application keys and the users' OAuth tokens using these. The search function is displayed in Fig 3 after the tag has been set up.
3.2 Data Pre-processing

Data mining on Twitter is a difficult task. Raw tweets, often known as noisy data, are what we get from Twitter. For further machine learning analysis, this must be pre-processed, also known as feature extraction. Feature engineering is the name given to the process of extracting features. Because the accuracy of raw tweets can be affected. Lower case conversion, punctuation, removal of hashtags and other Twitter notations (@ RT), URLs, stop words, Tokenization, and stemming are all part of the pre-processing activity.

In this work we used following 7 steps for Data Pre-processing:

a. Lower case conversion: The system is unable to distinguish between uppercase and lowercase letters. Lower casing can be accomplished in Python by utilising the lower () built-in function.

b. Remove Punctuations: regular expressions are used all punctuation marks should be replaced with blanks.

c. Remove hashtags, @RT and URL: Replace all hashtag, @, and https preceding content with blanks using regular expression.

d. Remove stop words: Import stop-words from the Natural Language Toolkit (NLTK) dataset and delete them during feature extraction.

e. Stemming of words: WordNet is an English dictionary based on Natural Language Processing (NLP) that lemmatizes words to the root word. It may be downloaded from the NLTK library.

f. Tokenize of words: To tokenize the statement, the NLTK library download punkt uses an unsupervised tokenizer algorithm.

g. Remove missing values: First, see if there are any missing values in the fetched data, and if so, eliminate them.

3.3 Sentiment Analysis

A frequent NLP task is sentiment analysis, often known as opinion mining. The most important part of sentiment analysis is analysing a body of text in order to comprehend the sentiment communicated by the content. TextBlob is a direct library for sentiment analysis in Python that yields two properties: polarity score (PS) and subjectivity score (SS). Each tweet's polarity and subjectivity were retrieved using TextBlob. We categorised the tweets as favourable, negative, or neutral using the polarity score, as shown in Fig 4.
3.4 Random Forest Classifier

A random forest classifier is an ensemble learning classification method that works by generating a large number of decision trees during training and then outputs the classification class. This means that each decision tree should collect the results and output the most important ones.

The classifier data must be handled before moving on to Random Forest (RF). Bag of Words (BOW) is a feature extraction method that may be applied to the training of a Random Forest classification machine. It generates vectors after creating a vocabulary of all the unique terms found in all of the papers in the training set. A large document with a large produced vocabulary may produce a vector with many 0 values. This is referred to as a sparse vector. When modelling using sparse vectors, more memory and computing resources are required. To circumvent this, the countvectorizer import from the sci-kit learn package is used to convert a given text into a vector 0 based on the frequency of each word in a tweet. Then, as illustrated in Fig 5, fit the result to a sparse matrix that can handle big data.

For splitting data arrays into two subsets, training set and testing set, the machine uses the sci-kit learn import train test split module. The set test was set to 30%. It's now completely prepared for machine learning analysis. Import random forest classifier from sci-kit learn sub module ensemble learning, fit the training data into the RF classifier, then predict the test data. Finally, determine the accuracy of the test and training set.

Because we retrieved real-time tweets, the number of tweets varies from time to time. We worked on Donald Trump and Joe Biden's tweets because they are the front-runners in the US presidential election in 2020. The accuracy of the model we developed and trained on two tweets ranges from 81% to 89% for Trump tweets and 83% to 85% for Biden tweets.

IV. RESULTS AND DISCUSSION

This section delves into the details of the outcomes. To begin, emotion classification of tweets using TextBlob is referred to as polarity detection (i.e., 0: neutral, 1: negative, 2: positive). Figure 6 depicts the understood polarity of Trump and Biden.
In addition, we diversified the polarity and contrasted both candidates in Fig 7. The accuracy score for Trump tweets is 83.22 percent, while Biden tweets are 85.73 percent, according to the below Instance polarity and forecast by Random Forest accuracy score.

![Fig 7. Individual Polarity Visualisation of Trump and Biden](image)

The purpose of a word cloud visualisation is to examine how tweets are distributed, with the size of each word indicating the frequency. A word cloud can be used to emphasise similar data points. Data from numerous social networking websites is frequently analysed using word clouds.

![Fig 8. Trump Tweets Word Cloud](image)

Out of tweets, Fig. 8 shows a word cloud visualisation of Trump, while Fig. 9 shows a word cloud visualisation of Biden.

![Fig 9. Biden Tweets Word Cloud](image)

V. CONCLUSION

In this paper, we show how social media such as Twitter can be used to predict future outcomes such as elections, specifically to extract the sentiment or views of people who are likely to vote in the general election or have an influence on those who will vote, and to classify their sentiment using Sentiment Analysis. First, we demonstrated sentimental analysis; second, we addressed knowledge extraction pre-processing in order to be precise; and finally, we talked machine learning prediction.
REFERENCES


5. Ankita Sharma, Udayan Ghose, Sentimental Analysis of Twitter Data with respect to General Elections in India, Procedia Computer Science, Volume 173, 2020, Pages 325-334, ISSN 1877-0509, httpsdoi.org10.1016j.procs.2020.06.


28. www.twitter.com

