IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

Unveiling the Power of Twitter: Sentiment Analysis for Election Prediction in India using Hybrid Model

OPEN ACCESS

Manuscript ID:

AG-2023-3002

Volume: 2

Issue: 3

Month: July

Year: 2023

ISSN: 2583-7117

Published: 13.07.2023

Citation:

Ambuj Kumar1, Dr. Pankaj Richhariya2. "Unveiling the Power of Twitter: Sentiment Analysis for Election Prediction in India using Hybrid Model" International Journal of Innovations in Science Engineering and Management, vol. 2, no. 3, 2023, pp. 12–18.

Ambuj Kumar[1], Dr. Pankaj Richhariya[2]

[1]*Research Scholar, Department of Computer Science, BITS, Bhopal*
[2]*Head of Department, Department of Computer Science, BITS, Bhopal*

Abstract

Every year, elections are held in India to choose new government officials. In the modern era, it's common practice to try to guess the results of elections months in advance. In the past few years, social media sites like Twitter, WhatsApp, & Facebook have grown to be major information resources. These sites provide an environment where anybody having internet access may voice their thoughts and ideas. Predicting the results of the Indian general election using sentiment analysis is the focus of this study. The purpose of this study is to analyze the efficacy of many machine learning algorithms for forecasting the results of elections through the analysis of the sentiments of tweets about political organizations. A number of different machine learning techniques, such as "Logistic Regression (LR)", "Random Forests", "Gradient Boosting (GB)", "SVM", "Multinomial Naive Bayes", and a suggested hybrid theory, were applied. The effectiveness of the algorithms was measured by their ability to correctly predict the results of the elections. The assessment results indicated that the suggested hybrid model had the highest accuracy (88.93%) compared to the other methods.

Keyword: election, election prediction, machine learning, twitter, hybrid model

I. INTRODUCTION

In a democratic society, the results of elections are crucial because it determines who will serve as leaders and enact policy on the public's behalf. Anticipating the results of elections properly may help political organizations improve their campaign plans and provide light on the public's voting behaviors. As social media has grown in popularity, political parties have begun using sentiment analysis to better understand how voters feel about them and their opponents. The number of political campaigns that make the most of social media has grown steadily. Twitter in particular has emerged as a powerful tool for political campaigns that can affect public opinion and win over voters. Twitter is a micro-blogging service where people may talk about whatever they want. The general mood of tweets about political parties may be analyzed using Twitter data, providing a window into the public's thinking. In a great deal of earlier study in this field, words and phrases have been categorized according to whether they formerly had positive or negative mood. [1] The purpose of sentiment analysis is to ascertain if a given text is meant to evoke a favorable, unfavorable or neutral reaction from the reader. This method may be used to assess the tone of tweets about political parties and make predictions about their performance in elections. Political parties may benefit from sentiment analysis by learning about the subjects that are most important to voters so that they can focus their campaigns on those areas.

A. Importance of social media in politics

The impact of social media on politics is growing rapidly across the globe, especially in India. The use of social media sites like Instagram, YouTube, Twitter, & Facebook to reach voters, interact with supporters, and activate the party's base has been revolutionary. Many observers have suggested that the internet, in general, and social media, in particular play a key role in amplifying existing economic, political, and cultural grievances and have their own independent effect on politics.[2] There are a number of reasons for the significance of social networks in politics, such as:

Wider reach: Social media allows political organizations the opportunity to reach more people than traditional media outlets. People in the realm of politics may bypass mediators by using social media to interact with the public and generate support.

Real-time engagement: Social media platforms enable political parties to interact with people in real-time as they address their questions, comments, and issues. As a result, political parties are better able to connect with their constituents and win their support.

Targeted messaging: With the use of social media platforms, political parties may send specific messages to voters based on their statistics and areas of interest. This facilitates communication between parties & people who share similar principles and ideas, ultimately leading to increased support for the party's platform.

Cost-effective campaigning: Campaigning on social media may save money for political organizations since it is more efficient than traditional methods. Smaller parties may compete with bigger ones since social media initiatives are frequently less expensive than traditional media campaigns.

Data analytics: Social media offers an enormous amount of information on user habits, preferences, and views that may be leveraged to create better campaign plans. Political parties may learn more about voters' priorities, voting patterns, and the various factors that impact their votes by studying data collected from social media platforms.

A. Sentiment analysis techniques

Sentiment analysis seems to be a common approach in natural language processing regarding acquiring qualitative insights from large amounts of text. Sentiment analysis applies to news stories, political disputes, stock markets, and Twitter data in addition to reviews and Twitter data [3, 4]. The following are examples of sentiment analysis methods that may be applied to texts data:

Rule-based techniques: In order to determine the sentiment of a piece of texts, rule-based methods use dictionaries and previously established rules. These dictionaries and rule sets are usually developed by humans who are specialists in the relevant field or languages

Machine learning techniques: Text sentiment prediction is made possible through the application of machine learning methods, which use algorithms to recognize patterns and correlations in text data.

Hybrid techniques: Hybrid methods enhance the precision of sentiment analysis by combining rules-based & machine-learning approaches. The machine learning approach, for instance, can be utilized to predict the sentiments of a text whereas a rule-based method can be used to discover negations and intensifiers in the texts.

The aim of this study is to utilize Twitter sentiment analysis to anticipate the results of the next general election in India. The research will examine Twitter data to determine the general public's opinion on the two largest political organizations in India, the Bharatiya Janata Party (BJP) & the Indian National Congress (Congress). The findings of this research may be utilized by both political parties and lawmakers to better understand public perception regarding political parties and produce more successful campaign strategies.

II. RELATED WORK

Joseph & Ferdin Joe John [5] explore a strategy for applying Twitter sentiment analysis for predicting the results of the 2019 Indian general election. Using a classifier based on decision trees for training and evaluating data, we find that the projected result is consistent with the actual result and the majority of the pre-poll analyses conducted thus far. This study exclusively reports tests done on English-language tweets that have received the most retweets from their original authors' followers. The results of several polling phases may be mapped in a timely manner using this strategy.

The Lok Sabha, or lower house, elections in India are scheduled for 2019, and this research Khare, Arpit, et al [6] analyzes Twitter data to examine popular opinion throughout the campaign. Due to the unregulated nature of this task, the researchers constructed a Transfer Learning-based automatic tweet analyser. The "Machine Learning Model" developed by the author uses "Linear Support Vector Classifiers" and the "Term Frequency Inverse Documents Fre-quency (TF-IDF)" method to analyze the text included inside tweets. The authors have also improved the model's ability to handle vitriol in tweets, something that has been overlooked by earlier studies.

[7] The use of Twitter by Indonesian politicians during the run-up to the 2019 Indonesian presidential elections has elicited a variety of reactions and opinions from the general population. The purpose of this research is to find the most

effective neural network algorithm for sentiment classification based on twitter data from the 2019 Indonesian presidential elections. The research team used many deep learning techniques to train their dataset. These included "Convolutional Neural Networks (CNNs)", "Long Short-Term Memories (LSTMs)", "CNN-LSTMs", "Gated Recurrent Units (GRUs)", & "Bidirectional LSTMs". Additionally, author trains their dataset using several conventional machine learning techniques, like "Support Vector Machines (SVM)", "Logistic Regression (LR)", & "Multinomial Nave Bayes (MNB)", to compare with our deep learning models. In the study, Bidirectional LSTM performed best, with an accuracy of 84.60 %.

The use of social media is becoming more important in political campaigns. To better understand the dynamics of social media debate, the author Alashri, Saud, et al. [8] of this research examines a dataset consisting of over 22,000 Facebook postings by candidates with more than 48 million responses. In this research, we focus on the 2016 U.S. presidential election and how candidates communicated with voters. We provide an innovative approach to categorizing comments as either strongly supported, supporters, dissenters, and strongly dissidents. Sentiment analysis is used to examine what each group has said about certain policies. Finally, we conclude with a discussion of potential future research directions to exploring the relationship between social media and election campaigns.

In recent years, a method has arisen that uses data from social media platforms, primarily Twitter, to foretell the results of elections. Burnap, Pete, et al. [9] develops a 'baseline' model for utilizing Twitter as a tool for election predictions and applies it to the 2015 UK General Elections. The research adds to the current literature on the topic by expanding the utilization of Twitter as a prediction tool to the UK setting and outlining its limits, especially in a multi-party society with spatial concentration of authority for small parties.

The author ianqiang [10] of this study introduces a word embedding approach derived from unsupervised training on large Twitter corpora, one that makes use of latent contextual semantic connections & co-occurrence statistical properties among words in tweets. Twitter's sentiment characteristics are constructed using word embedding's, n-grams, & word sentiment polarity scores. A deep convolutional neural network is fed the feature sets in order to train & predict the categorization of sentiment labels. The author conducts experimental comparison of our model's performance against a word n-grams method on "five Twitter data sets", & finds that our model outperforms the baseline approach in terms of accuracy & F1-measure for Twitter sentiments categorization.

III. METHODOLGY

A. Dataset

This research employed a dataset which includes 36,225 tweets on different political groups, such as the BJP & Congress, which was retrieved from GitHub. However, due to time and computational limitations, only 10,000 tweets were used for the analysis. The dataset used in this study consisted of tweets related to several political parties, like the BJP & the INC, and a corresponding compound score. The compound score was generated using a sentiment analysis tool that assigned a score between -1 & 1 to every tweet, with -1 signifying extremely negative sentiment, 0 signifying neutral sentiment, & 1 signifying extremely positive sentiment. The compound score is a useful metric for sentiment analysis as it takes into account the overall sentiment expressed in a tweet, including the presence of negation and sarcasm.

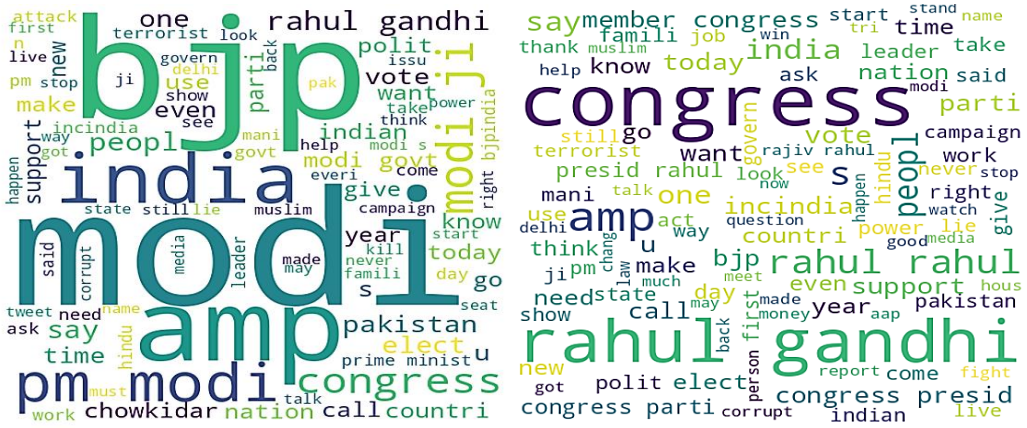

Figure 1 most words in BJP & Congress Tweets

To ensure the accuracy and reliability of the data, we conducted a thorough quality check before including it in our study.

B. Data Pre Processing

Data pre-processing is an important step in any data analysis project, and is especially critical for natural language processing (NLP) tasks such as sentiment analysis on Twitter. The quality of the classification is directly impacted by the pre-processing operations carried out. [11] Our data pre-processing steps allowed us to prepare the collected tweets for analysis, and to extract relevant features that we could use to build and evaluate models. The data cleaning process for this study involves a series of steps to prepare the collected data for analysis and model development. These steps may include:

Removing duplicates: To ensure that the data is unique and representative, we need to remove duplicated tweets from the dataset. This is done using panda's library.

Removing user mentions and hashtags: To focus on the content of the tweets rather than the users or hashtags, we need to remove user mentions and hashtags from the data. This is done using regular expressions.

Stemming and lemmatization: To standardize the text of the tweets, we need to apply stemming or lemmatization techniques to reduce words to their base form. This can help to improve the efficiency and effectiveness of subsequent processing steps. This is done using NLTK library.

Removing stop words: To focus on the more meaningful words in the text, we choose to remove stop words from the data. However, we carefully consider which stop words to remove and how to handle any potential edge cases or exceptions. This process was done using NLTK library.

Once the tweets were pre-processed, the next step was to convert them into a format that could be used for machine learning. This involved converting the text data into numerical vectors using word vectorization techniques. In this study, the tweets were vectorised using the Bag-of-Words (BoW) approach, which involves counting the frequency of each word in a document and representing the document as a vector of word frequencies.

	0	1	2	3	4	5	6	7	8	9	...	1490	1491	1492	1493	1494	1495	1496	1497	1498	1499
2011	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
445	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2313	0	0	0	0	2	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3862	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1917	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
405	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
368	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4935	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2083	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
232	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Figure 2 Data sample after Vectorization

C. Proposed Model

In this study, various machine learning algorithms such as "decision tree", "random forest", "multinomial naive Bayes", "gradient boosting", "logistic regression", and "support vector machine" were tested with two word vectorization techniques: Count Vectorise and Tfidf Vectorizer. However, none of these models provided satisfactory results in predicting the election outcomes. To improve the accuracy of the prediction models, hyper parameters of the algorithms were tuned using Grid Search CV. However, even after hyper parameter tuning, the results were not satisfactory. Therefore, a hybrid model was developed, consisting of a combination of Random Forest & Logistic Regression, using Stacking Classifier.

Several measures, including accuracy, recall, F1 score, precision, as well as confusion matrix, were used to assess the hybrid model's performance after it had been developed on the data set that had been pre-processed. The dataset used for training was divided 85:15 between testing and training sessions.

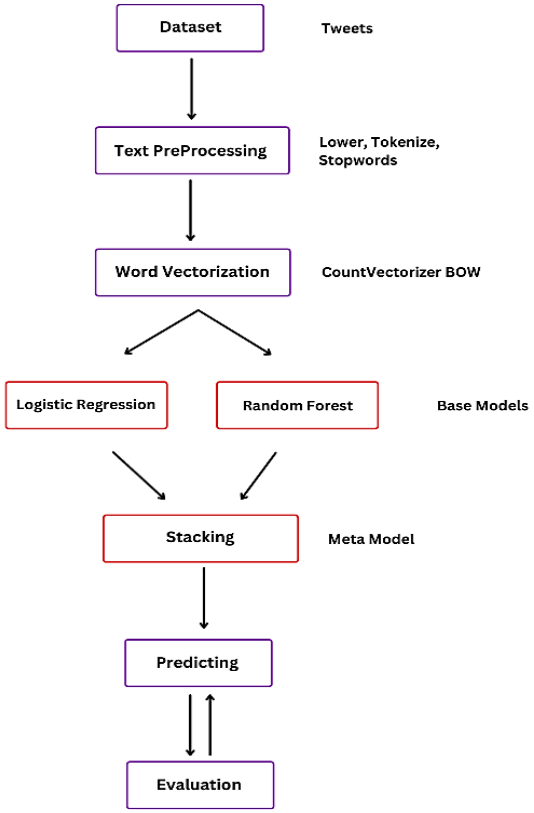

Figure 3 Proposed Model

The resulting test-dataset predictions were based on the training model. To determine how well the model performed, it was used to compare its predictions regarding actual election results. The technique successfully anticipated the result of elections for both the BJP & the Congress, as shown by the results. The model is trained, to evaluate the performance of the model it has to be test on unseen data. We already kept testing data aside for evaluation of the model after training. The algorithm was put to the test by fitting it with the test data so that it could make predictions. We compared many criteria, including accuracy scores & confusion matrix, to determine how well each one performed.

Accuracy Score: Using the proportion of right predictions represents the simplest straightforward metric by which to assess the efficacy of any given Classification algorithms. This is the exact reasoning for the Accuracy metric.

The following represents the mathematical formula.

$$Accuracy = \frac{Ture\ Negatives + True\ Positive}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

IV. RESULTS AND DISCUSSION

Six machine learning algorithms were used for sentiment analysis: logistic regression, random forest, gradient boosting, SVM, multinomial naive Bayes, and the proposed hybrid model. These algorithms were trained on the dataset using both count vectorizer and TF-IDF vectorizer techniques for word vectorization. The accuracy of each model was evaluated using the test dataset. The logistic regression model achieved an accuracy of 85.20%, the random forest model achieved an accuracy of 88.00%, the gradient boosting model achieved an accuracy of 79.60%, the SVM model achieved an accuracy of 80.00%, and the multinomial naive Bayes model achieved an accuracy of 81.06%.

However, the proposed hybrid model, which used a stacking classifier to combine the predictions of a random forest and logistic regression model, achieved the highest accuracy of 88.93%. This indicates that the hybrid model was the most effective in accurately predicting the sentiment of tweets related to political parties during the election. It is important to note that these results were obtained using the count vectorizer technique for word vectorization. When the TF-IDF vectorizer technique was used, the accuracy of the models was lower, with the hybrid model achieving an accuracy of 85.60%. This suggests that the choice of word vectorization technique can have a significant impact on the accuracy of the sentiment analysis models.

Below is the summary of the results in tabular form:

Table 1 Model Comparison

Models	Accuracy
Logistic Regression	85.20%
Random Forest	88.00%
Gradient Boosting	79.60%
SVM	80.00%
Multinomial Naive Bayes	81.06%
Decision Tree [12]	86.30%
Random Forest + Logistic Regression Hybrid Model (Proposed Model)	**88.93%**

The performance of our proposed hybrid model was compared with the existing work in the field of sentiment analysis for predicting election outcomes using Twitter data. The existing work reported an accuracy of 86.30% using a Decision tree algorithm. In comparison, our proposed hybrid model achieved an accuracy of 88.93%, which outperforms the existing work. This improvement in accuracy may be attributed to the use of multiple algorithms in the hybrid model, as well as the use of advanced techniques such as Count Vectorizer for feature extraction.

Here are visual results of proposed mode –

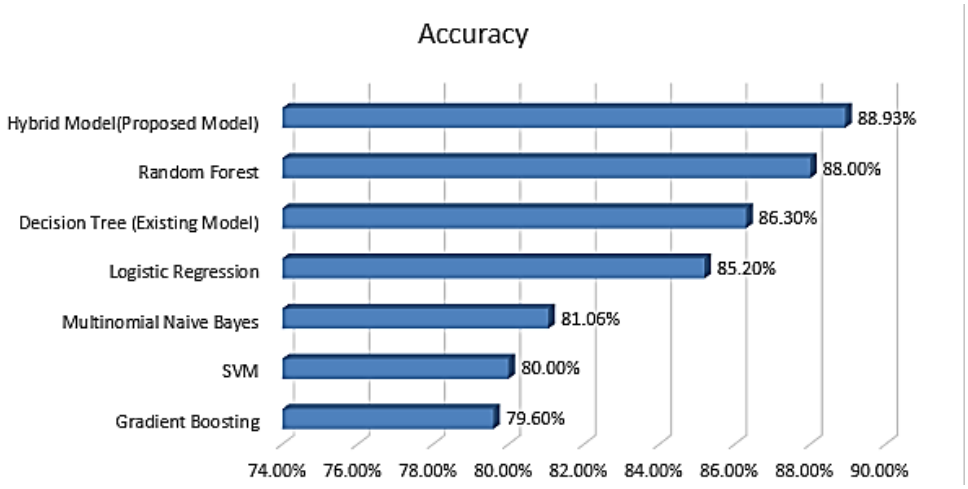

Figure 4 Result Comparison

V. CONCLUSION

The study aimed to develop an accurate and reliable model that could be used for real-time analysis of sentiment towards political parties and their potential impact on election outcomes. The methodology involved data collection, we used a pre-existing dataset consisting of 10,000 tweets related to political parties, including the Bharatiya Janata Party (BJP) and the Indian National Congress (INC). We used a sentiment analysis tool to generate a compound score for each tweet, which was used as a feature for building machine learning models. Six machine learning algorithms, including Logistic Regression, Random Forest, Gradient Boosting, SVM, Multinomial Naive Bayes, and a proposed hybrid model, were used for model development and training. We used Count Vectorizer technique to convert text data into numerical form, which was then fed into the models for training and testing. Our hybrid model achieved the highest accuracy of 88.93%, outperforming all other algorithms.

The results show that sentiment analysis based on Twitter data can be a useful tool for predicting election outcomes. The accuracy achieved by the models developed in this study is competitive with the existing literature, indicating the effectiveness of our proposed approach. Additionally, our proposed hybrid model performed better than individual models, suggesting the potential of combining multiple models for improving accuracy. There are several potential directions for future work building upon the findings and limitations of this study. Some suggestions would be Experiment with different pre-processing techniques and Incorporate more recent data.

REFERANCE

[1] Hatzivassiloglou, Vasileios, and Kathleen McKeown. "Predicting the semantic orientation of adjectives." 35th annual meeting of the association for computational linguistics and 8th conference of the european se of the association for computational linguistics. 1997.

[2] Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. "Political effects of the internet and social media." Annual review of economics 12 (2020): 415-438.

[3] Yu, Liang-Chih, et al. "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news." Knowledge-Based Systems 41 (2013): 89-97.

[4] Hagenau, Michael, Michael Liebmann, and Dirk Neumann. "Automated news reading: Stock price prediction based on financial news using context-

capturing features." Decision Support Systems 55.3 (2013): 685-697.

[5] Joseph, Ferdin Joe John. "Twitter based outcome predictions of 2019 Indian general elections using decision tree." 2019 4th International Conference on Information Technology (InCIT). IEEE, 2019.

[6] Khare, Arpit, et al. "Sentiment analysis and sarcasm detection of indian general election tweets." arXiv preprint arXiv:2201.02127 (2022).

[7] Hidayatullah, Ahmad Fathan, Siwi Cahyaningtyas, and Anisa Miladya Hakim. "Sentiment analysis on twitter using neural network: Indonesian presidential election 2019 dataset." IOP Conference Series: Materials Science and Engineering. Vol. 1077. No. 1. IOP Publishing, 2021.

[8] Alashri, Saud, et al. "The 2016 US Presidential Election on Facebook: an exploratory analysis of sentiments." (2018).

[9] Burnap, Pete, et al. "140 characters to victory?: Using Twitter to predict the UK 2015 General Election." Electoral Studies 41 (2016): 230-233.

[10] Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for twitter sentiment analysis." IEEE access 6 (2018): 23253-23260.

[11] Krouska, Akrivi, Christos Troussas, and Maria Virvou. "The effect of preprocessing techniques on Twitter sentiment analysis." 2016 7th international conference on information, intelligence, systems & applications (IISA). IEEE, 2016.

[12] Batra, Payal Khurana, Aditi Saxena, and Chaitanya Goel. "Election result prediction using twitter sentiments analysis." 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2020.