International Journal of Gender, Science and Technology
ISSN: 2040-0748
UGC Care Group I Journal
Vol-12 Issue-02 Dec 2023

Election Result Prediction using Twitter Analysis

Dr.A.Nagaraju[1].,Srikanth Vaddi[2].,Devadurga Gurunath[3].,Dharani Babu Dorai[4].

[1] Professor &HOD, [2,3,4]StudentsB.Tech-IT.
Malla Reddy Institute of Technology and Science.,Maisammaguda.,Medchal.,Telangana,India

[1]dranraju.mrits@gmail.com,[2] vsrikanth380@gmail.com,
devadurgagurunath@email.com,[4]doraidharani@gmail.com

Abstract -

The majority of Indians have always regarded elections as a momentous occasion worthy paying close attention. The exponential growth of social media in the last few years has given regular people a great platform to share their thoughts and feelings. One such site is Twitter, which informs users daily on political happenings using various hash tags and trends. In response to these kinds of political occurrences, people voice their opinions. Our strategy entails compiling a set of tweets from the leading political parties running in the 2022 General State election and then calculating the sentiment score. Tweets pertaining to a particular political party, both old and new, are included in the dataset. To find tweets on a certain party, we utilize phrases such as "BJP elections 2022," "#UP elections BJP," "#Punjab elections BJP," and so on. In order to construct our classifier and categorize the test data as either positive, negative, or neutral tweets, we combined VADER Sentiment Analyzer with traditional machine learning methods such as Random Forest Classifier, SVM, etc. Consequently, this study uses sentiment analysis on Twitter data acquired from the platform to forecast the outcome of the next election.

1. INTRODUCTION

The rise of social media has given people a voice like never before. You may express your thoughts, ideas, and ratings on social networking sites such as Google+, Face book, and Twitter. The official Twitter accounts of every major political party and its representatives throughout the globe have millions of followers. They want to reach out to young people who could cast ballots for them on this platform. People in India have been increasingly outspoken in their criticisms and praises of government decisions since the epidemic saw a dramatic increase in the number of Twitter users in the country. Using Sentimental Analysis, one may train a computer to identify and categorize emotions in text [1]. Anything from a little review to a lengthy social statement, tweets, or messages might be considered text. Anyone interested in the election results, including political parties, may utilize Twitter Sentiment Analysis to see how people feel about a certain party and whether that feeling is good or bad. In a democratic nation, elections are crucial. Under India's parliamentary system, citizens get to choose their leaders for the next five years. There will be five state elections between February 22 and March 22, the most significant of which will be in Uttar Pradesh, which has the highest number of members of parliament. The Bhartiya Janata Party (BJP), the Indian National Congress (INC), the Aam Aadmi Party (AAP), the Samajwadi Party (SP), the Shiromani Akali Dal (SAD), and the Naga People's Front (NPF) are the main national political parties that are running for office.

2. LITERATURE SURVEY

Using Hindi-language tweets, Parul S. and Teng-Sheng Moh [2] forecasted the outcomes of the 2016 Indian general elections. Using 42,235 tweets gathered over a month, they conducted text mining. In all, three ML algorithms were used. The Support Vector Machine achieved an accuracy of 78.4%, whereas the Naïve Bayes method achieved 62.1%. The final prediction was made using SVM because of its better accuracy. With a dataset of over 500,000 tweets, Dr. D. Rajeswara Rao and colleagues [3] trained their model using 80% of the data and tested it with the remaining 20%. They made an educated guess as to which political faction dominated online discourse. A classifier was constructed using the proposed approach, which trained the dataset for over two days. Among the models constructed, SVM achieved the highest level of accuracy (80%) in the experiments. For the 2019 Indian General Elections, decision trees were used by Ferdin Joe and John Joseph [4]. Predicting the outcomes of Indian elections using the suggested technique shows promise based on the results collected.

Using Twitter data for US local election prediction, Meng-Hsiu Tsai and colleagues [5] from Middle Georgia State University introduced a machine learning approach. Very favorable, positive, neutral, negative, and very negative were the five categories into which they placed their findings. Weighted sentiment ratings were computed using the RNTN model. The 2019 Lok Sabha election outcomes were forecasted by Payal Khurana Batra and colleagues [6]. Separate sets of tweets from the BJP and the Congress were added to the data after pre-processing. During model training, five distinct ML methods were used. With an accuracy of more than 80%, decision trees and XGBoost performed better.

3. PROPOSED METHODLOGY

There are five steps to follow while using the suggested technique. There is a periodicity to the first two steps, and then there is a prediction phase.

3.1 Collecting Data

The dataset was trained using tweets that were gathered during November and December of 2021. Approximately 12,000 tweets were collected. Yogi Adityanath, 'BJP elections 2022,' 'INC Punjab elections,' 'UPelection,' and many more hashtags and keywords were used. This led to the development of domain-specific corpora, which were supplemented with additional elements such as 'like count,"retweet count,' user name, and date-time of tweet'. In order to test the model, datasets representing the leading political parties in each of the Indian states that would be participating in the 2022 elections were gathered every five days from January to February. Table 1 displays the results of the top three political parties in each state as determined by the OneIndia [7] opinion surveys.

Rankings of the major political parties in each state

State Name	Top 3 political parties considered
Uttar Pradesh	BJP, SP, INC
Punjab	AAP, INC, SAD
Uttarakhand	BJP, INC, AAP
Goa	BJP, INC, AAP
Manipur	BJP, INC, NPF

The following two libraries were significant:

First, Tweepy is something that Twitter provides. The most recent and widely used tweets tagged with a certain hashtag were collected and analyzed.

2. Snscrape: To get over tweepy's constraints (can't extract tweets older than 7 days and limited quantity of tweets), we utilized inscape.

3.2 Preprocessing the Data

An integral part of any machine learning process is text preparation. Data cleansing is the process of removing irrelevant information from large datasets. Eliminating duplicate or unnecessary data is the first stage in data cleansing. Various hash tags were used during different time periods to gather data. The same or different people may end up tweeting the identical thing many times if they utilize the same hash tag. In the data collecting step itself, we removed any duplication. The following are the subsequent procedures for preparing text data:

1. Application of regular expressions: To eliminate URLs, we substituted '@handles' for 'handles,' '#hash tags' for 'hash tags,' and several spaces for a single space using regular expressions. Punctuation and special characters were also eliminated.

2. Eliminating stop words: Words like "the," "a," "an," "in," etc., that are often used but should not be retained. As filler, these words contribute nothing to advance the meaning of the phrase. There is a very high frequency of these terms. It is our hope that these words will not add unnecessary storage space to our database or lengthen the time it takes to process data. The third step is lemmatization, which ensures that all words in the corpus are represented as lemma, the basic or dictionary form. We would like that our model not treat as distinct terms any two words that have the same meaning in a given context. "Win" is the source of the English terms "winning," "winner," and "wins," among others.

3.3 Dataset Labeling

It is critical to label the dataset after preprocessing. The dataset was categorized into positive, negative, and neutral tweets using the VADER package, which stands for Value Aware Dictionary and Sentiment Reasoned [8]. It labels the dataset using lexical and rule-based analysis. To determine the tone of the tweet, we calculate the polarity score's compound value. Table 2 displays the sentiment-to-score mapping.

Vader compound score labeling (Table 2)

Compound score	Sentiment
>=0.05	Positive
>=-0.05 and <=0.05	Neutral
<=-0.05	Negative

3.4 Model Training

The suggested study used tf-idf as its feature extraction approach and divided the data into training (0.75) and test (0.25) sets. A numerical number that indicates the weight of a given word in the document is produced by multiplying the tfidf technique[9] with the inverse document frequency and term frequency. The model training relies on the uncommon terms because of their high tfidf value. In addition, a classification model is constructed using supervised machine learning methods. To forecast the end result, we used Logistic Regression, Support Vector Machine, and Random Forest Classifier. Additionally, we used ensemble voting approaches to combine the aforementioned algorithms. In order to train our model, we built a pipeline that includes tfidf and ML techniques.

• Regression based on logic:

For most cases, it's the go-to method for producing categorical data. Fitting an S-shaped function that saturates all the values between 0 and 1 is used instead of fitting the regression line in this extension of Linear Regression. The output class prediction probability is provided.

Vector Support System:

Using a decision boundary termed a hyper plane, it may be used to categorize data points in an n-dimensional space that have numerous classifications. As the number of characteristics increases, so does the dimension.

• Classifier based on random forests:

To construct a classification model, it makes use of n decision trees in combination. It takes an average of how accurate all n trees' forecasts are. Compared to other supervised algorithms, it often performs better since it is an ensemble approach.

Table 3: Analysis of the Model

Model	Accuracy Score
Random Forest Classifier	77.59%
Voting Classifier(Soft)	74.69%
Logistic Regression	74.22%
Voting Classifier(Hard)	73.86%
Support Vector Machine	73.28%

4. RESULTS AND DISCUSSIONS

The first graphic shows the "Effective Positive Rate," while the second compares the proportion of positive, negative, and neutral tweets from the top three political accounts in Punjab with those from Uttar Pradesh.

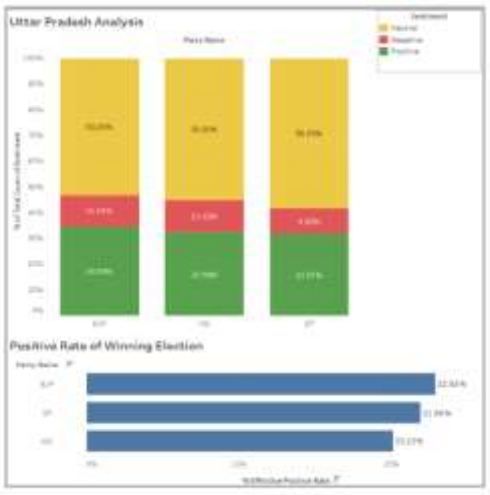


Chart -1: Uttar Pradesh Analysis

The BJP's chances of winning the UP state elections are greater with 34.84% favorable tweets and a popularity score of 22.93%. However, with an effective positive percentage of 22.37%, AAP is projected to emerge victorious in the Punjab elections.

Chart-3 shows the history of tweets of various feelings, providing an in-depth examination of the BJP in Uttar Pradesh. Since the majority of tweets are about news stories and political events involving a certain party, the data above suggest that most tweets lack a clear mood. Similarly, all five states are ranked according to their popularity on Twitter.

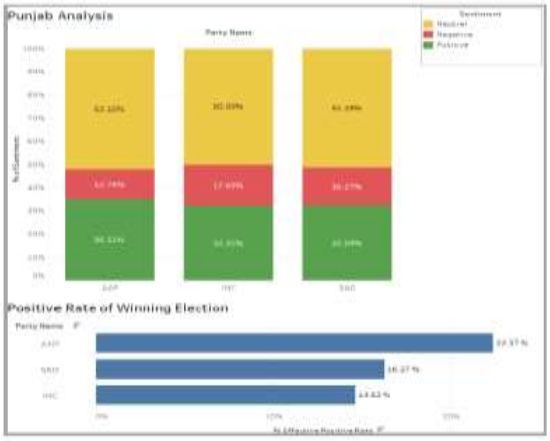
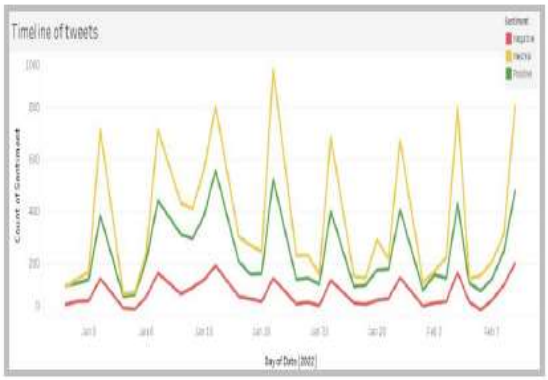


Chart -2: Punjab Analysis



Twitter timeline between the Uttar Pradesh BJP and the state

Table 4 shows that the BJP is the most popular political party in most states. Results from OneIndia.com[10] are used to get the exact results. With the exception of Manipur, all of the states' forecasts were confirmed by the observations. Twitter is unable to accurately forecast Manipur's election outcomes because to the state's very low internet and social media penetration compared to other states. Consequently, there have been little debates and buzz around the Manipur elections on Twitter. Consequently, the suggested article demonstrates that Twitter may be used as a medium to successfully indicate election results for the majority of Indian states.

Figure 4: Real and Expected Victory

State Name	Predicted winner	Actual winner
Uttar Pradesh	BJP	BJP
Punjab	AAP	AAP
Uttarakhand	BJP	BJP
Goa	BJP	BJP
Manipur	INC	BJP

5. FUTURE WORK AND LIMITATIONS

Since Twitter does not provide sufficient information about user's location, the suggested approach takes into account the overall mood of the whole blogosphere to forecast the outcomes of state elections rather than the tweet's actual location. The accuracy of this work may be enhanced by applying it to tweets in languages other than English, such as the regional languages of the Indian states. Among the regional languages that Twitter currently supports are Bengali, Kannada, Urdu, Tamil, Gujrati, and Marathi. The suggested model has a few flaws, such as the fact that it missed sarcasm in a few words because the semantics were misused and the fact that not everyone has access to social media, which allows people to stand out and show their solidarity.

6. CONCLUSION

In order to forecast which political party will be more popular, a powerful Random Forest classification model was developed, which achieved an accuracy of 77.59%. Political parties may enhance their campaign plans throughout the election time by using the suggested approach. As part of their social media analytics, they may utilize it to look at how other political parties are doing. With the most up-to-date party tendencies shown, users may cast more educated ballots. This approach, when used as a long-term strategy for a political party, may help analysts and strategists evaluate public mood over extended periods of time. This research focused on investigating Twitter as a hunting ground for political campaigns, in response to the growing use of social media.

REFERENCES

[1] DataRobot, "Introduction to Sentiment Analysis: What is Sentiment Analysis?," DataRobot, 26 March 2018. [Online]. Available:

https://www.datarobot.com/blog/introduction-to-sentiment-analysis-what-is-sentiment-analysis/.

[2] T.-S. M. Parul Sharma, "Prediction of Indian Election Using Sentiment Analysis," 2016 IEEE International Conference on Big Data (Big Data), pp. 1966-1971, 2016.

[3] S. U. S. K. M. S. R. G. C. U. J. Dr D Rajeswara Rao, "Result prediction for political parties using twitter sentiment analysis," International Journal of Computer Engineering and Technology, no. 11(4), 2020.

[4] F. J. J. Joseph, "Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree," 2019 4th International Conference on Information Technology (InCIT), Bangkok, THAILAND, pp. 50-53, 2019.

[5] Y. W. M. K. a. N. R. Meng-Hsiu Tsai, "A machine learning based strategy for election result prediction," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1408-1410, 2019.

[6] A. S. S. a. C. G. Payal Khurana Batra, "Election Result Prediction Using Twitter Sentiments Analysis," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 182-185, 2020.

[7] "Uttar Pradesh Assembly Election 2022 Opinion Poll," Oneindia, 2022. [Online]. Available: https://www.oneindia.com/uttar-pradesh-election-2022-opinion-poll-and-exit-poll/.

[8] GeeksforGeeks, 7 Oct 2021. [Online]. Available: https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/.

[9] Mamun, "Medium," 20 June 2019. [Online]. Available: https://medium.com/@imamun/creating-a-tf-idf-in-python-e43f05e4d424. [Accessed 20 May 2022].

[10] "Indian Elections 2022," Oneindia, 25 March 2022. [Online]. Available: https://www.oneindia.com/elections/.