e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

ELECTION OUTCOME PREDICTION USING SENTIMENT ANALYSIS ON TWITTER

Mahi Verma[*1], Prathamesh Suryawanshi[*2], Sampada Deore[*3], Priyanka Mundhe[*4], Prof. A.G. Phakatkar[*5]

[*1,2,3,4]Undergraduate Student, Dept. Computer Engineering, Pune Institute Of Computer Technology, Pune, India.

[*5]Professor, Dept. Computer Engineering, Pune Institute Of Computer Technology, Pune, India.

DOI : https://www.doi.org/10.56726/IRJMETS41569

ABSTRACT

Democracy is a system of governance wherein individuals possess the freedom to select their leaders. In our nation, India, regular elections are conducted to determine the political representatives. Presently, there is a growing inclination among people to anticipate the election outcomes in advance. They make predictions based on news reports, personal discussions, and online platforms. Social media websites such as Facebook, Twitter, and WhatsApp have gained significant popularity inrecent times. These platforms are easily accessible to anyone with an internet connection, and they serve as virtual spaces for individuals to express their opinions. Twitter, in particular, has emerged as a prominent platform during significant eventsin our country. Any news item quickly becomes a trending topic on Twitter, as users actively share their views, criticisms, and campaigns related to political parties, public figures, or leaders. This platform proves to be valuable in predicting election results even before the official exit poll results are announced. Hence, this study involves analyzing tweets collected from Twitter and employing sentiment analysis techniques to forecast election outcomes.

Keywords: Sentimental Analysis, Twitter, Supervised Learn- Ing, Natural Language Processing, Machine Learning.

I. INTRODUCTION

Social networking has evolved into a powerful tool for expressing opinions. Platforms like Facebook, Twitter, and Google+ have provided avenues for sharing viewpoints, re- views, and ratings. Official Twitter accounts of major politicalparties and their representatives worldwide attract millions of followers. These platforms are seen as a means to engage with the younger demographic who may potentially support them in elections. The surge in Indian Twitter users during the pandemic has led to increased outspokenness in expressing support or criticism of political decisions.

Sentiment analysis is a technique used to teach computersto identify emotions in text. Texts such as regular reviews, social statements, tweets, and messages can all be subjectedto this analysis. The general public and political parties can utilize Twitter sentiment analysis of election-related tweets to gauge the positive or negative sentiments of people towards a particular political party. This aids in anticipating the election results during that period.

Elections hold significant importance in a democratic nation. In the Indian parliamentary system, citizens have the rightto determine who will govern them for the next five years. From February 22 to March 22, five state elections were scheduled, with Uttar Pradesh being particularly crucial asit sends the largest number of Members of Parliament to the national assembly. The major national political parties competing in these elections are the Bhartiya Janata Party (BJP), Indian National Congress (INC), Aam Aadmi Party (AAP), Samajwadi Party (SP), Shiromani Akali Dal (SAD), and Naga People's Front (NPF).

II. LITERATURE SURVEY

Parul S and Teng-Sheng Moh [2] forecasted the outcomes of the 2016 Indian general elections by analyzing tweets written in Hindi. They conducted text mining on a dataset of 42,235 tweets collected over a one-month period. Three machine learning algorithms were employed, with Na¨ıve Bayes achiev-ing an accuracy of 62.1%

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

and Support Vector Machine (SVM)achieving an accuracy of 78.4%. Due to its higher accuracy, SVM was chosen for the prediction.

Dr. D. Rajeswara Rao and their team [3] assembled a datasetof over 500,000 tweets, with 80% of the data used for training and the remainder for testing. Their objective was to determinewhich political party had a greater influence on social media. They proposed a system that trained the dataset for morethan two days, resulting in the construction of a classifier. Through experiments, they demonstrated that SVM was the most accurate model, achieving an accuracy of 80%.

Ferdin Joe and John Joseph [4] employed a decision tree approach to predict the outcomes of the 2019 Indian General Elections. The results obtained from their proposed methodology indicated a promising future for predicting Indian electionresults.

Meng-Hsiu Tsai and their team [5] at Middle Georgia State University presented a machine learning strategy for analyzingTwitter data to predict the results of local elections in the US. They categorized their results into five classes: very positive, positive, neutral, negative, and very negative. They utilized theRNTN model to calculate weighted sentiment scores.

Payal Khurana Batra and their team [6] predicted the election results of the Lok Sabha in 2019. After preprocessing the data, they divided it into two sets: one containing tweets related to the BJP and the other containing tweets related to theCongress party. They trained their model using five different machine learning algorithms. Decision tree and XGBoost yielded higher accuracy above 80%.

III. PROPOSED METHODLOGY

The suggested approach can be implemented in five distinct phases. The initial two phases are executed periodically, and they are followed by the prediction phase.

A. Data Collection

The training dataset was compiled by collecting tweets during the period of November-December 2021, resulting in a total of approximately 12,000 tweets. Various hashtags and phrases such as 'UPelection', 'Punjabelection', 'Yogi Adityanath', 'BJP elections 2022', 'INC Punjab elections', andothers were utilized to gather relevant tweets. This approach led to the creation of a domain-specific corpus. Additionally, the dataset included additional features such as 'like count', 'retweet count', 'user name', and the date-time of the tweet.

For testing the model, datasets of the top political parties for each Indian state participating in the 2022 elections were collected every five days during the January-February period. These datasets were obtained based on the top three political parties considered for each state, as determined by the OneIn- dia opinion polls [7]. Please refer to Table-1 for the specific parties considered for each state.

Table 1: Top Political Parties Of Each State

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

Two important libraries used were:

- Tweepy: Tweepy is a library provided by Twitter that enables access to and utilization of the Twitter API. It serves as a valuable tool for collecting and working with tweets from various sources. With the help of Tweepy, the project was able to efficiently retrieve both recent and popular tweets related to specific hashtags. This allowed for the accumulation of a comprehensive dataset, which could then be analyzed and processed for further use in the project. By leveraging the functionalities provided by Tweepy, the project was able to streamline the process ofcollecting and consolidating tweets from Twitter.
- Snscrape: Tweepy has certain limitations, such as re- strictions on the number of tweets that can be

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

extracted and the inability to access tweets older than 7 days. To overcome these limitations, the project incorporated the use of Snscrape.Snscrape is an alternative library that provides additional capabilities for scraping tweets from Twitter. Unlike Tweepy, Snscrape allows for the extraction of tweets beyond the 7-day limitation, enablingaccess to historical tweet data. By leveraging Snscrape alongside Tweepy, the project was able to overcome the restrictions imposed by Tweepy and gather a more comprehensive dataset, including tweets that are older than 7 days. This combination of Tweepy and Snscrape enhanced the ability to collect a broader range of tweet data for analysis and further processing in the project.

B. Data Preprocessing

Text preprocessing is a crucial phase in the machine learningworkflow, as it involves cleaning and preparing the data to extract meaningful information. The initial step in datacleaning is to identify and remove duplicate or irrelevant data.In the case of acquiring data using different hashtags andover various time periods, there is a possibility of obtaining redundant tweets from the same or different users that share multiple common hashtags. To address this issue, duplicity was eliminated during the data collection phase itself.

The subsequent steps involved in preprocessing the textdata are as follows:

1) Use of regular expressions: We used regular expressions to remove website URLs, replace '@handles' with 'handles', 'hashtags' with 'hashtags', and multiple spaces with single space. We also removed special characters and punctuations.

2) Removal of stopwords: A stopword is a commonly used word such as 'the', 'a', 'an', 'in', etc. These words do not add any meaning to the sentence and merelyused as a filler. The frequency of these words are very high. We would not like these words to take more space in our database and increase data processing time.

3) Lemmatization: It is important that all words in the corpora are in their root or dictionary form, known as lemma. We do not want our model to consider twowords, having same contextual meaning, as two different words. For example, the words 'winning', 'winner', 'wins' are all converted to their root form 'win'.

C. Labelling the dataset

Once the dataset has undergone preprocessing, the next step is to label the data. In this project, the VADER (Valence AwareDictionary and Sentiment Reasoner) [8] library was employedfor labeling the dataset into three categories: positive, negative,and neutral tweets. VADER utilizes lexical and rule-basedanalysis to assign sentiment labels to the dataset.

The sentiment behind each tweet is determined using the compound value of the polarity score provided by VADER. The polarity score represents the intensity and direction of the sentiment expressed in the tweet. By mapping the compound score to sentiment categories, the dataset is labeled accordingly Table-2 shows the mapping of score to sentiment.

Table 2: Labelling The Vader Compound Score

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

D. Model Training

In the proposed work, the dataset was divided into training data (75

To build a classification model, supervised machine learn- ing algorithms were utilized. The following algorithms were employed:

1) Logistic Regression
2) Support Vector Machine (SVM)
3) Random Forest Classifier

Additionally, ensemble voting techniques were used to combine the predictions of the above algorithms for the final output. This approach helps improve the accuracy and robustness of the model.

During the model training phase, a pipeline was created to incorporate the tf-idf feature extraction technique

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

along with the machine learning algorithms. This pipeline ensures that thetext data is transformed using tf-idf and then passed through the specified algorithms for training and prediction.

By leveraging these techniques and algorithms, the proposed work aimed to develop a classification model capable of predicting sentiment or other relevant outcomes based on the provided dataset.

Table 3: Model Analysis

State Name	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%

E. Model Predictions

In order to analyze the sentiment and popularity of each political party in the State elections, separate datasets were created for each party in each state. The developed model was then applied to these datasets to determine the sentiment behind the tweets related to each political party.

To assess the popularity of a political party and make inferences about their chances of winning the State elections,a popularity score was calculated. The popularity score, also referred to as the 'Effective Positive Rate', was computedusing the following formula:

PopularityScore = (sum(tweets with positive sentiment)-sum(tweets with negative sentiments)/(Total tweets over the period))*100.

This score indicates the proportion of positive sentiment expressed in tweets related to a political party. By calculating the popularity score for each party, they were ranked and visu-alized to provide an understanding of their relative popularity.

Furthermore, the percentage of positive, negative, and neutral tweets was determined for all the parties, offering additional insights into public sentiment towards each party.

To provide further analysis, a timeline of tweets was gen- erated, displaying the sentiments expressed in tweets for each party during the campaigning period (January-February 2022) leading up to the elections. This timeline helped to understand the sentiment trends and fluctuations associated with different political parties during the election campaign.

IV. RESULTS AND DISCUSSIONS

The 'Effective Positive Rate' and comparison of percentage of positive, negative and neutral tweets of top three political tweets of Uttar Pradesh and Punjab are shown chart-1 and chart-2 respectively.

BJP having 34.84% positive tweets and popularity score of 22.93% has higher chances of winning the UP state elections. AAP, on the other hand, is predicted to be the winner of Punjab elections, with effective positive rate of 22.37%.

Chart-3 provides in-depth analysis of BJP in Uttar Pradesh, showing timeline of tweets of different sentiments. The above figures show most of the tweets have neutral sentiment, since most of the tweets are news articles and political events of a particular party. Similarly, ranking of popularity, over Twitter,is done for all the five states.

From table-4, it is clear that the political party, BJP has dominated in majority of states. The actual results are obtained from OneIndia.com[10] results. The observations shows cor- rect predictions for all states, except Manipur.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

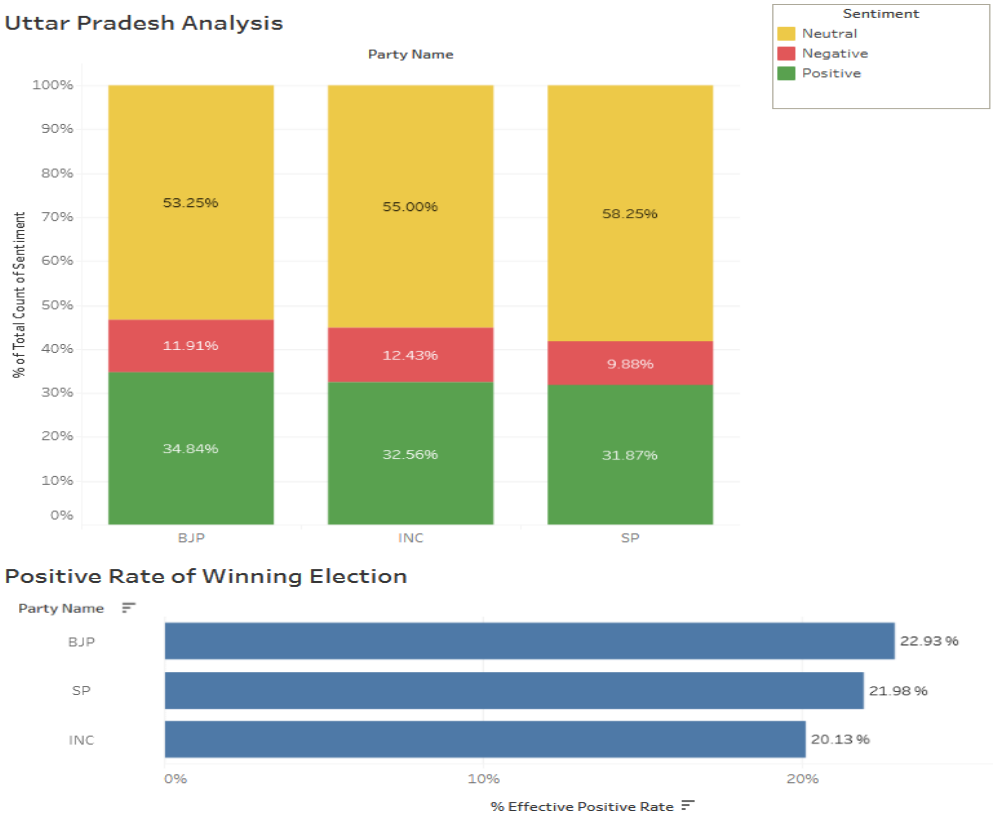

Fig. 1. Uttar Pradesh Analysis

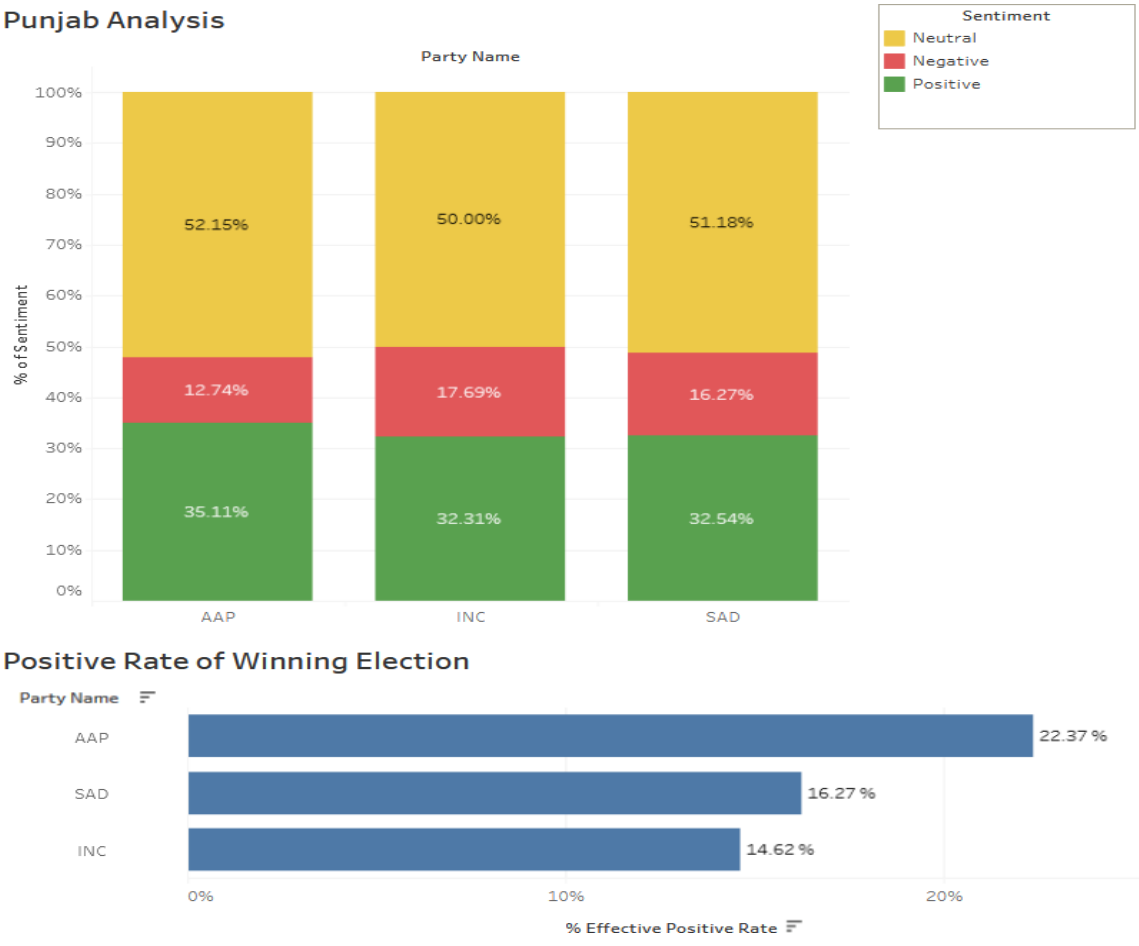

Fig. 2. Punjab Analysis

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

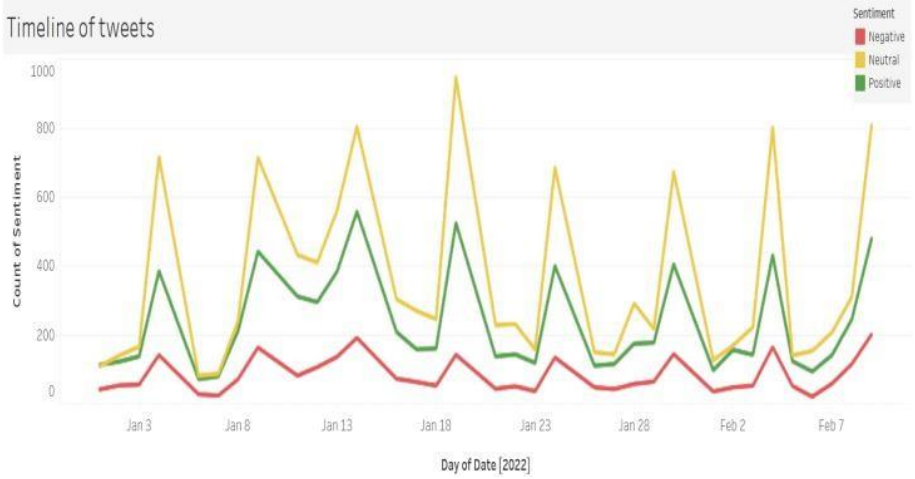

Fig. 3. Uttar Pradesh-BJP: Timeline of tweets

Since the penetration of social media and internet is low in Manipur among other states, thus leading to few discussions and low buzz of Manipur elections on Twitter, the results of this state cannot be rightly predicted by Twitter. Therefore, the proposed paper shows that Twitter as a platform can be effectively used as an election result indicator for majority of Indian states.

Table 4: Actual And Predicted Winner

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

V. FUTURE WORK

- This study can be extended to include tweets in various regional languages of Indian states in addition to English to improve accuracy. The regional languages currently supported by Twitter are Hindi, Gujarati, Marathi, Urdu, Tamil, Bengali and Kannada.
- The proposed system does not consider the location of tweets as filters for state elections, because Twitter does not provide enough information about the location of users and therefore the opinion of the entire blogosphere is considered a predictor of the elections.

VI. CONCLUSION

This project focuses on the exploration of social mediaplatforms, particularly Twitter, as a crucial component of election campaigns. Recognizing India as a highly socially connected country, with a significant portion of its youngpopulation below the age of 35, social platforms play a vital role in the lives of the youth. The goal of this project is to analyze the impact of social platforms on the political system, specifically in various states during elections.

The system employs an effective Random Forest classification model, achieving an accuracy of 77.59%. It serves as a valuable tool for political parties to enhance their campaigningstrategies during the election period. By utilizing social media analytics, political parties can gain insights into the trends of other parties and make informed decisions. Moreover, politicalanalysts and strategists can utilize this methodology as a long-term plan to study the sentiments of the population over an extended period. Overall, this project recognizes the widespread usage of social media platforms and specifically focuses on Twitter as a prominent platform for election campaigns. It emphasizes the importance of understanding and leveraging social platformsto better comprehend the perspectives and sentiments of the people during the election process.

VII. REFERENCES

[1] Random forest classifier. https://miro.medium.com/max/1400/ 0*f qQPFpdofWGLQqc.png, 2019. [Online; accessed 20-02-2022]..

[2] Javapoint.com. Support vector machine. https://static.javatpoint.com/tutorial/ machine learning/ images/support-vector-machine-algorithm.png. [Online; ac- cessed 20-02-2022.

[3] Parul Sharma and Teng-Sheng Moh. Prediction of indian election using senti- ment analysis on hindi twitter. In 2016 IEEE International Conference on Big Data (Big Data), pages 1966–1971. IEEE, 2016.

[4] Dr D Rajeswara Rao, S Usha, S Krishna, M Sai Ramya, G Charan,and U Jee- van. Result prediction for political parties using twitter sentiment analysis. International Journal of Computer Engineering and Technology, 11(4), 2020.

[5] Ferdin Joe John Joseph. Twitter based outcome predictions of 2019 indian general elections using decision tree. In 2019 4th International Conference on Information Technology (InCIT), pages 50–53. IEEE, 2019.

[6] Meng-Hsiu Tsai, Yingfeng Wang, Myungjae Kwak, and Neil Rigole. A machine learning based strategy for election result prediction. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI), pages 1408–1410. IEEE, 2019.

[7] Jyoti Ramteke, Samarth Shah, Darshan Godhia, and Aadil Shaikh. Election result prediction using twitter sentiment analysis. In 2016 Inter- national Conference on Inventive Computation Technologies (ICICT), volume 1, pages 1–5. IEEE, 2016.