International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

# A Systematic Review of Predicting Elections Based on Social Media Data

**[1] Nayana H M, [2]Roopa R**

[1]Sem MCA Student, Department of MCA, BIET, Davanagere
[2]Assistant Professor, Department of MCA, BIET, Davanagere

**Abstract -** *Elections in India has always been considered as an important event and has been keenly followed by majority of people. The rapid increase of social media in the recent past has provided end users a powerful platform to voice their opinions. Twitter, being one such platform, provides day-to-day updates on political events through different hashtags and trends. People provide their opinion by reacting on such political events. Our approach is to gather a collection of tweets of top political parties contesting within the General State election, 2022, then compute the sentiment score. Dataset contains mixture of both popular as well as recent tweets related to specific political party. Specific keywords are used to extract tweets for a party like 'BJP elections 2022', '#UPelections BJP', '#Punjabelections BJP', etc. We utilized a combination of VADER Sentiment Analyzer and classic machine learning algorithms like Random Forest Classifier, SVM, etc. to build our classifier and classify the test data as positive, negative and neutral tweets. Therefore, this work analyses tweets collected from twitter and predicts election results by performing sentimental analysis on them.*

***Key Words*: Sentimental Analysis, Twitter, Supervised learning, Natural Language Processing, Machine Learning**

## 1. INTRODUCTION

Social media has become a powerful tool for sharing opinions. There are social media platforms like Facebook, Twitter and Google+ to share opinions, reviews and ratings. All major political parties and their members all over the world have their official accounts on Twitter with millions of followers. They consider this platform as a medium to connect with young people who might vote them. With significant rise of Indian users on Twitter during the pandemic, people have been more vocal to criticize or appreciate a political decision.

Sentimental Analysis is a method to teach a machine to extract emotion from a given text [1]. A text can be anything, a simple review, a social statement, tweets or messages. Twitter Sentiment Analysis of tweets regarding elections can be used by the general public as well as the political parties to understand the positive or negative views of people regarding a particular political party, thus, helping to predict the election results during that period.

Elections play an important role in a democratic country. Indian parliamentary system gives its people the right to decide who will govern them for the next five years. During the tenure of Feb 22 to March 22, five state elections are lined up, with the important one being at Uttar Pradesh, which sends the largest number of MPs to parliament. The major national political parties contesting in the elections are 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).

## 2. LITERATURE SURVEY

Parul S and Teng-Sheng Moh [2] predicted the results of 2016 Indian general elections using tweets in Hindi language. They performed text mining on 42,235 tweets collected over a month. They applied three ML algorithms. The accuracy of the Naïve Bayes' algorithm was 62.1% and the accuracy of Support. Vector Machine was 78.4%. Final prediction was done by utilizing SVM, since the accuracy was higher.

Dr D. Rajeswara Rao and team [3] gathered a dataset of more than 500,000 tweets out of which 80% was used for training and rest for testing. They predicted which political party had more influence on social media. Proposed a system that trained the dataset for more than 2 days and a classifier was built. Experiments proved that SVM was the most accurate model built with an accuracy of 80%.

Ferdin Joe and John Joseph [4] used decision tree to predict 2019 Indian General Elections. The results obtained in the proposed methodology showed it had a promising future in predicting Indian election results.

Meng-Hsiu Tsai and his team [5] at Middle Georgia State University presented a machine learning strategy to analyse Twitter data for predicting the results of local elections in US. They categorized their results into 5 classes namely very positive, positive, neutral, negative and very negative. They used RNTN model to calculate weighted sentiment scores.

Payal Khurana Batra and her team [6] predicted election results of Lok-Sabha 2019. After pre-processing, they split the data into two parts containing BJP and Congress tweets in separate sets. They trained their model using

five different ML algorithms. Decision tree and XGBoost gave higher accuracy above 80%.

## 3. PROPOSED METHODLOGY

The proposed methodology can be implemented in five phases. The first two phases are done periodically, followed by prediction phase.

### 3.1 Data Collection

Tweets used for training the dataset were collected during the period of November-December, 2021. A total of nearly 12000 tweets were collected. Different hashtags and phrases like 'UPelection', 'Punjabelection', 'Yogi Adityanath', 'BJP elections 2022', 'INC Punjab elections', etc. were used. Thus, a domain-specific corpora was created and other features like 'like count', 'retweet count', 'user name' and date-time of tweet' were also included in the dataset. Datasets of top political parties for every Indian state(contesting in elections-2022) were collected every 5 days during Jan-Feb period for testing the model. Top three political parties considered for each state, according to OneIndia [7] opinion polls, is shown in table-1.

**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 |
| Goa | BJP, INC, AAP |
| Manipur | BJP, INC, NPF |

Two important libraries used were:

1. Tweepy : It is provided by Twitter. A collection of latest as well as popular tweets of a particular hashtag were collected and combined together.

2. Snscrape : As tweepy has a restriction on the amount of tweets to be extracted and tweets older than 7 days cannot be extracted, snscrape was used to overcome these limitations.

### 3.2 Data Preprocessing

Text preprocessing is an important phase in ML workflow. It essentially involves cleaning of the data to extract only the meaningful information. First step towards data cleaning is to ensure there is no duplicate or irrelevant data. We acquired data using different hashtags over different periods of time. It can often lead to redundant tweets by the same or different users having multiple common hashtags. We eliminated duplicity in data collection phase itself. Next steps for preprocessing done on text data 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 merely used 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 two words, having same contextual meaning, as two different words. For example, the words 'winning', 'winner', 'wins' are all converted to their root form 'win'.

### 3.3 Labelling the dataset

After preprocessing, it is important to label the dataset. We used VADER(Valence Aware Dictionary and sEntiment Reasoner) [8] library to label the dataset into positive, negative and neutral tweets. It uses lexical and rule-based analysis to label the dataset. We use the compound value of the polarity score to get the sentiment behind the tweet. 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 |

### 3.4 Model Training

For the proposed work, the data was split into training(0.75) and test data(0.25) and the feature extraction technique used was tf-idf. The tfidf technique[9] multiplies term frequency and inverse document

frequency, which provides a numerical value, denoting the weight of a particular word in the document. The rare words have higher tfidf value and considered important to model training.

Further, supervised machine learning algorithms are used to build a classification model. We used Logistic Regression, Support Vector Machine and Random Forest Classifier to predict the final output. We also used combination of the above algorithms using ensemble voting techniques. . We created a pipeline of tfidf along with ML algorithms while training our model.

- **Logistic Regression:**

  It is mainly used when output is a categorical variable. It is an extension of Linear Regression, but instead of fitting the regression line, we fit an S-shaped function which saturates all the values between 0 and 1. It gives the probability of the predicted output class.

- **Support Vector Machine:**

  It is applicable when we want to classify an n-dimensional space of datapoints having multiple classes using a decision boundary called hyperplane. The dimension depends upon the number of features.

- **Random Forest Classifier:**

  It uses a combination of n decision trees to build a classification model. It averages the accuracies of all predictions of n trees. Since, it is an ensemble technique, it has the ability to outweigh other supervised algorithms in most of the cases.

- **Voting Classifier:**

  We used a combination of the above algorithms to build an voting classification model. We used both, hard voting(selects a model using majority voting technique) as well as soft voting(selects a model by calculating the probability of each class and averaging it), to build two different ensemble models.

By comparing the accuracies of the above models shown in table-3, we found out Random Forest Classifier provided better results than others. Thus, we used it to make further predictions on unlabeled dataset.

**Table -3:** Model Analysis

| 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% |

### 3.5 Model Predictions

We created separate datasets for each political party for each state. We applied the model on each of these datasets to get the sentiment behind the tweet. In order to get the popularity of a political party and extrapolate their chances to win the State elections, we calculated the popularity score, given by,

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

The above score can also be called as 'Effective positive Rate'. We ranked and visualized the political parties by the above score. We also calculated the percentage of positive, negative and neutral tweets for all the parties. To provide further analysis, we provided a timeline of tweets, providing sentiments, for every party during the campaigning period(Jan-Feb 2022) of election.

## 4. 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.

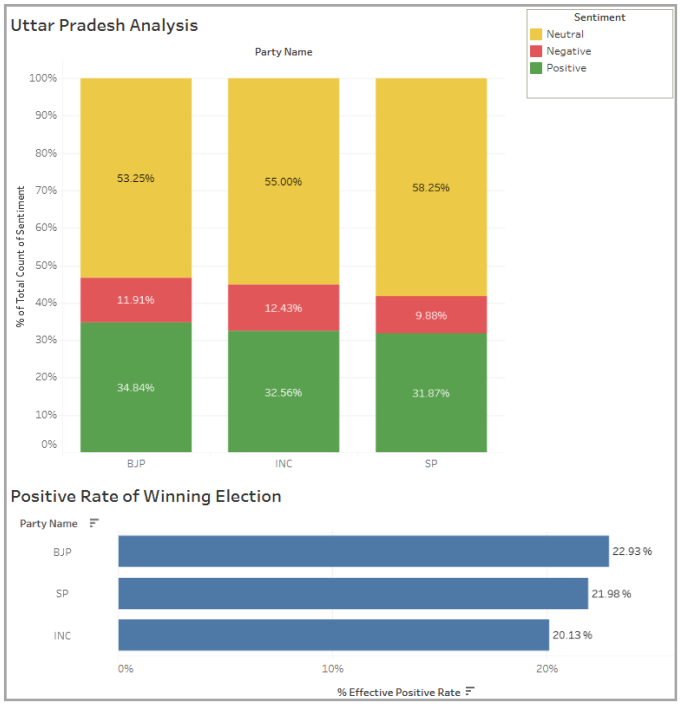


**Chart -1**: Uttar Pradesh Analysis

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.

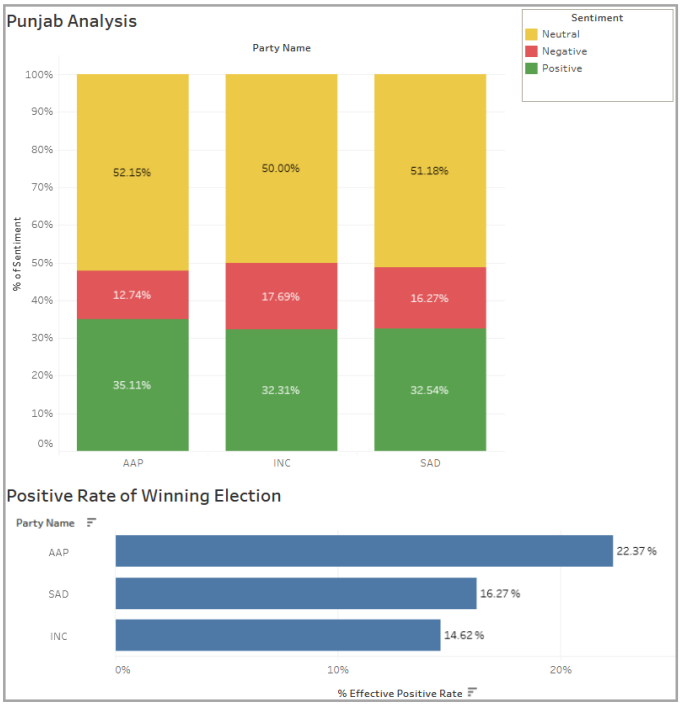


**Chart -2**: Punjab Analysis

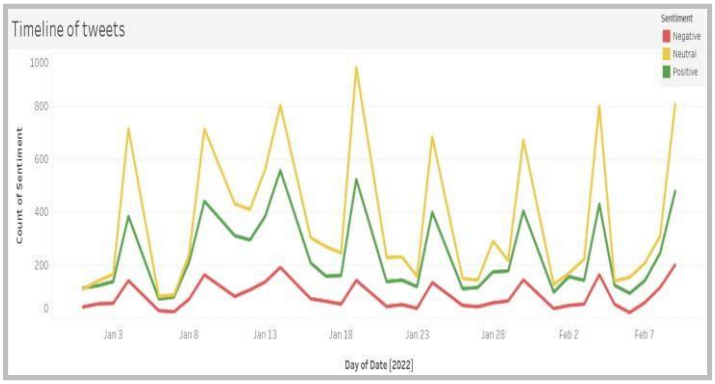


**Chart -3**: Uttar Pradesh-BJP: Timeline of tweets

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 correct predictions for all states, except Manipur.

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 |
| Goa | BJP | BJP |
| Manipur | INC | BJP |

## 5. FUTURE WORK AND LIMITATIONS

The proposed system does not consider geographical location of the tweet as a filter for state elections, as Twitter does not provide adequate information about user's location, thus, a general mood of entire blogosphere is considered to predict the election results. This work can further be extended on tweets of different regional languages of Indian states other than English to improve accuracy. Currently, regional languages supported by Twitter are Hindi, Gujrati, Marathi, Urdu, Tamil, Bengali, and Kannada.

Sarcasm was not detected in some sentences due to misuse of the semantics and not everyone has access to social media where they can stand out from the crowd and express their support for one another, are some of the limitations of the proposed model.

## 6. CONCLUSION

An effective Random Forest classification model, with an accuracy of 77.59%, was built to predict the popularity of political party. The proposed system can be used by political parties to improve their campaigning strategies during the election period. It can be used by them as a part of social media analytics to study the trends of other political parties as well. User can make informed decision in voting by seeing the current trends of political parties. Political analyst and strategist can use this methodology, as application, as a long term plan for a political party to study the sentiments of people over a long time period. Observing the expanded use of social media platforms, this project concentrated on exploring of social platform (Twitter) as the chase for elections' campaign.

## 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/.