Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 6
May 2023

Analysis of Tweet Data for Predicting Election Outcomes with Machine Learning

Ravi kumar Kuchipudi
research scholar, ANU, Computerscience and Engineering ravikumary4@gmail.com
Dr. V.V. Jaya Rama Krishnaiah
Associate Professor, Department of Computer Science and Engineering KLE Vaddeswaram, India,
jkvemula@gmail.co

Abstract - The vast majority of Indians have always paid close attention to elections because of their significance. The current explosion of user-generated content on social media has given consumers a robust forum for their views. Twitter is one such site that regularly updates its users on current political happenings via hashtags and trends. People express themselves through their responses to these kinds of political occurrences. Our strategy is to compile a database of tweets from major political parties running for office in the General State election of 2022, and then calculate a sentiment score based on those tweets. The dataset includes both well-known and up-to-the-minute tweets about various topics.
party politics. To collect tweets about a particular political party, we utilise search terms like "BJP elections 2022," "#UPelections BJP," "#Punjabelections BJP," etc. Methods that we used includedUsing the test data, we created a classifier using VADER Sentiment Analyzer and traditional machine learning methods like Random Forest Classifier, Support Vector Machines, etc. Therefore, this work uses sentiment analysis to forecast election outcomes based on collected tweets.
.**Key Words:** Sentimental Analysis, Twitter, Supervised learning, Natural Language Processing, Machine Learning

1. Introduction

Opinions can now be widely disseminated via social media. Facebook, Twitter, and Google+ are just a few examples of the many social media outlets available for rating and reviewing content. There are millions of people who follow the official Twitter accounts of the world's main political parties and their representatives. They see it as a way to reach out to the youth vote and build support among that demographic. There has been a dramatic increase in the number of Indian Twitter users since the pandemic began, and individuals are using the platform to express their opinions on recent political decisions.

Sentimental Analysis [1] is a technique for teaching a computer to recognise and respond to human emotions in written material. A text could be anything from a short review to a lengthy social statement, a tweet, or a message. The public and political parties alike can utilise Twitter Sentiment Analysis of tweets on elections to gain insight into how voters feel about various candidates and ultimately make more informed voting decisions.

Elections are a crucial part of any functioning democracy. The people of India have the power to choose their leaders for the next five years under the country's parliamentary system. There will be five state elections from February 22 to March 22, the most significant of which is in Uttar Pradesh, which sends the most members of parliament to the national legislature. The major national political parties that are running for office are 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).

2. Literature Survey

The outcomes of the 2016 Indian general election were predicted using Hindi tweets by Parul S and Teng-Sheng Moh [2].

Thelanguageaccuracy.Over the course of a month, we compiled a dataset containing 42,235,62.1% tweets using the text mining method NaveBayes. Three ML methods were used in this study.

precision of the Support system. The percentage for Vector Machine was 78.4%. Because of its

superior accuracy, SVM was used for the final forecast.

In order to train and test their algorithm, Dr. D. Rajeswara Rao and his colleagues [3] collected a dataset of more than 500,000 tweets. They made an educated guess as to which political party was more popular online. Advised a system that took more than 2 days to train the dataset and provide a classifier. Experimental results demonstrated that SVM, with an accuracy of 80%, was the most accurate model developed.

The 2019 Indian General Elections were predicted using a decision tree by Ferdin Joe and John Joseph [4]. The outcomes of the proposed strategy for forecasting the outcomes of the Indian election showed great potential.

Middle Georgia State University's Meng-Hsiu Tsai and colleagues [5] recently proposed a machine learning technique for analysing Twitter data in order to forecast the outcomes of US local elections. extremely positive, positive, neutral, negative, and extremely negative were the five categories they used to classify their findings. To quantify how they felt about something, they applied the RNTN model.

The outcomes of the 2019 Lok-Sabha elections were predicted by Payal Khurana Batra and her colleagues [6]. After cleaning up the data, they separated tweets associated with the BJP and those associated with the Congress. In order to train their model, they use**d**

There are five distinct ML algorithms. Above 80% accuracy was achieved with both the decision tree and XGBoost.

3. Proposed Methodlogy

There are five steps to enacting the proposed methodology. The first two steps are repeated on a regular basis, and then the prediction step is implemented.

3.1 Data Collection

The tweets used to train the dataset were gathered in the months leading up to the 'Punjabelection' in December of 2021.Nearly 12,000 tweets were collected in support of 'Yogi Adityanath' for the BJP in the upcoming 2022 election.Numerous 'INCPunjabhashtagselections' were made, along with other options such as 'like count,''retweet count,' 'useretc.', and 'date created,' 'date used,' and 'useretc.'So, it's finally time to domaintweet'It was a spwerecificalsocorporaincludedwasin.

-set of data. The model was put through its paces using data collected every 5 days from January to February, covering the leading political parties in each Indian state (running in elections-2022). Table-1 displays the top three political parties in each state as reported by OneIndia [7] opinion polls.

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.data. We collected information throughout time by following various hashtags. It often causes tweets from the same user or users using the same hashtags to be repeated. In the process of gathering information, we made sure no duplicates were made. The subsequent procedures for text data preparation are as follows:

1. Use of regular expressions:

Change "@handles" to "handles," "#hashtags" to "hashtags," and "regular" to "multiple" and "multiple" to "regular" to "remove" the single space from the website.WeURLs, also omitted accents and other punctuation.

2. Commonly used terms that are removed during the opword isofast process include the, a, an, in, as, etc. These words serve no purpose other than to pad out the sentence. These terms are used extremely frequently. We'd rather avoid having to allocate extra storage space for, and spend more time processing, these terms.

3. Lemmatization:

All terms in the corpus should be in their lemmatized, or dictionary, form. In no way do winning, our prototype "winner," for which we consider "wins"There are two distinct words here: two arewords, all converted having sameto contextualtheir root formmea 'win'as. Case in point: a few words

3.2 Labelling the dataset

The dataset has to be labelled after preprocessing is complete. The VADER(Value Aware Dictionary and sEntiment Reasoner) [8] package helped us categorise the tweets in the dataset as good, negative, and neutral. It labels the data using a combination of lexical analysis and rules. To determine the tone of a tweet, we multiply its polarity score by its component parts.

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

3.3 Data Preprocessing

The preparation of texts is a crucial step in the machine learning process. The process essentially boils down to cleaning the data such that only relevant information is retained. Making ensuring there are no duplicate or useless records is the first step in data cleaning.

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 word prominence in the text as measured by its frequency, expressed as a numerical value. Rare words have a higher tfidf value and are valued in model development because of this.More so, a classification model is constructed using supervised machine learning methods. The predicted result was calculated using a combination of Logistic Regression, Support Vector Machine, and Random Forest Classifier. We also used ensemble voting approaches to combine some of the aforementioned algorithms. When training our model, we built a pipeline combining tfidf and ML methods.

- **Logistic Regression:**

It is typically employed in situations when the expected outcome is a scalar. It's an extension of linear regression where an S-shaped function is fitted that maximises the range from 0 to 1. It provides the likelihood of an expected outcome class..

- **Support Vector Machine:**

It can be used when a decision boundary, in this case a hyperplane, is needed to categorise datapoints in an n-dimensional space with many classes. The size is determined by the total number of characteristics.

- **Random Forest Classifier:**

A classification model is constructed using a mixture of n decision trees. The results of n trees' forecasts are averaged out. Because it is an ensemble method, it can often outperform individual supervised algorithms.

- **Voting Classifier:**

To create a voting classification model, we utilised a few of the aforementioned methods. We constructed two distinct ensemble models using hard voting (selects a model using majority voting technique) and soft voting (selects a model by computing the probability of each class and average it).

Among the models presented in table-3, we discovered that Random Forest Classifier produced the highest accuracy. We put it to use by making predictions on an additional, 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

made individual databases for every state and political party. Each of these datasets was run through the algorithm to determine underlying sentiment in the tweets. We calculated the popularity score, provided by, to gain a political party's support and predict their success in State elections.,

Popularity_score = (sum(tweets with positive sentiment) - sum(tweets with negative sentiments) / (Total tweets over

In addition to its more common name, "Effective positive Rate," thetheperiabove))*100score has several names. We used this score to rank the political parties and create a graphic representation of them. We also determined the share of supportive, critical, and neutral tweets for each candidate. We also included a timeline of tweets that expressed opinions about each party during the campaign season (January and February 2022).

The Effective Results

Charts 1 and 2 compare the top three political tweets from Uttar Pradesh and Punjab in terms of the percentage of favourable, negative, and neutral tweets.

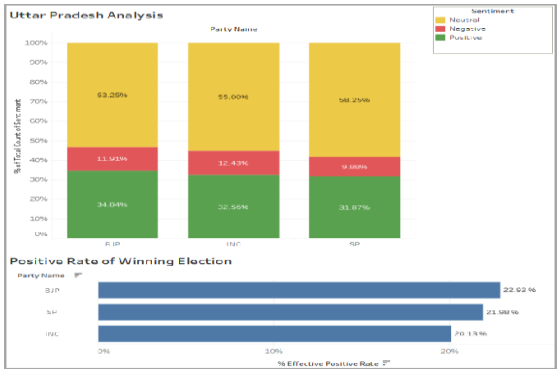

Chart -1: Uttar Pradesh Analysis

here is a greater likelihood of a BJP victory in the upcoming UP state elections given the party's 34.84% favourable tweets and popularity score of 22.93%. However, an effective positive rating of 22.37% places AAP as the likely victor in the upcoming Punjab elections.

In-depth study of the BJP in Uttar Pradesh is provided in Chart 3, which displays a timeline of tweets expressing various emotions regarding the party. Since most tweets are about news stories and political events associated with a specific party, the data above suggests that most tweets have a neutral tone. Twitter popularity rankings are calculated in the same way for all five states.

As compared to other states, Manipur has a relatively low rate of internet and social media penetration, which explains why the state's elections have received so little attention on Twitter. Thus, the suggested research demonstrates that Twitter may be used efficiently as an election result indicator for most Indian states.

Table -4: Actual and Predicted Winner

State Name	Predicted winner	Actual winner
Punjab Analysis	Party Name	Sentiment: Neutral, Negative
Uttar Pradesh	BJP	BJP
Punjab	AAP	AAP
Uttarakhand	BJP	BJP
Goa	BJP	BJP
Manipur	INC	BJP
[redacted]	[redacted]	[redacted]

% Effective Positive Rate

Chart -2: Punjab Analysis

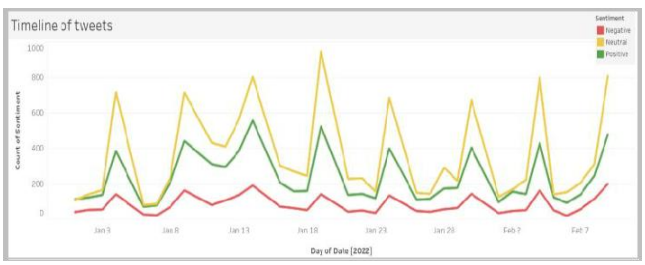

Chart -3: Uttar Pradesh-BJP: Timeline of tweets

Table 4 shows that the BJP has been the dominant political party in most of the states. OneIndia.com[10] provides the exact results. All states, with the exception of Manipur, were accurately predicted.

5. FUTURE WORK AND LIMITATIONS

Predicting election outcomes in the proposed approach does not take into account users' locations, as Twitter is not believed to provide sufficient information about the blogosphere. To further enhance precision, this study can be expanded to include tweets written in languages spoken by residents of Indian states other than English. Twitter currently supports the regional languages of Hindi (Gujrati), Marathi (Math), Urdu (Urdu), Tamil (Tamil), Bengali (Bengali), and Kannada (Kannada).

Some examples of sarcasm were missed because the semantics were misused, and not everyone has access to social media where they may stand out and show their support for one another.

6. CONCLUSION

The popularity of a political party can be predicted using a Random Forest classification model that achieved an accuracy of 77.59%. Political parties can use the proposed mechanism to enhance their election season campaigns. As part of social media analytics, it can be used to look at how other political groups are faring. By viewing the historical trends of political parties, users are better able to make educated voting decisions. A political party's long-term strategy can benefit from this approach, which can also be used by political scientists to analyse public opinion over extended time periods. Based on research on the rising popularity of social media, this study investigated Twitter as a potential tool for political campaigns.

REFERENCES

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

[2] [Online]. Available: https://www.datarobot.com/blog/introduction-to-sentiment-analysis-what-is-sentiment-analysis/.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] "Uttar Pradesh Assembly Election 2022 Opinion Poll,"

[9] Oneindia, 2022. [Online]. Available:

[10] https://www.oneindia.com/uttar-pradesh-election-

[11] 2022-opinion-poll-and-exit-poll/.

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

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

[14] "Indian Elections 2022," Oneindia, 25 March 2022.

[15] [Online]. Available: https://www.oneindia.com/elections/.

[16] [11] Opinion Mining And Snark Analysis In General Election Tweets

[17] Dr. V.V. Jaya Rama Krishnaiah, Ravi Kumar Kuchipudi

[18] Journal Of Data Acquisition And Processing, 2023, 38 (2): 4770-4785