Prediction of Indian Election Using Sentiment Analysis of Twitter

Narsingh Pal Yadav[1], Bhramari Anand[2], Dr Bal Saraswat[3]
Department of Computer Engineering (Sharda University) Greater Noida, India

***Abstract*—social media is a collection of many online platforms that enables people to share contents with others. Social media has become big and massive platform to gather public opinions and sentiments in about any topics. Our study shows the potential of social media in the prediction of election outcomes. By analyzing public opinion, we will be able to identify voter behavior. Using machine learning techniques, we categorize social media post's comment section to identify sentiment keys such as- positive, negative and neutral. By using ML algorithms, we will be able to evaluates the accuracy of sentiment analysis in election results. It will provide insights in to public moods and its growing importance in today's time. This is evaluated using Naïve Bayes, SVM, and LSTM.**

***Index Terms*—Sentiment Analysis; Twitter; Indian Elections; Naive Bayes; Support Vector Machine; Long Short-Term Memory.**

I. INTRODUCTION

Natural Language Processing (NLP) can be classified into opinion mining and text mining. It is used in segregating the views of people's postings with respect to different social media applications like Facebook, Twitter, etc. Text or Sentiment mining is also helpful in different situations such as analyzing people's feelings about a movie, product, song, etc. and to differentiate between positive, neutral and negative reviews. It can be used in places like the stock market, e- commerce websites, song recommendations, etc. for better predictions and recommendations. There has been much research already conducted on Sentiment Analysis in the English language. Almatrafi et al [1] collected tweets using the Twitter data that considered only two major CandidateModi, Rahul gandhi and labelled them as negative, neutral and positive.

The aim of the paper was to analyze trends in the Indian General Election 2019 using location as a filter. They employed a supervised approach by applying a Naïve Bayes classifier and LSTM.

The problem statement: Is it probable to predict the popularity of any political Candidate and therefore extrapolate their chances of winning the election by utilizing sentiment analysis of Twitter data? To answer this question, it is imperative to analyze Twitter tweets to learn and study the sentiments of people in terms of positive polarity, neutral polarity and negative polarity. To analyze the problem statement, the authors obtained tweets, filtering the Data and then applied sentiment mining and prediction operations.

This takes us to certain research queries, for example, how to anticipate and break down what strategy is being accomplished? What steps are suitable for the task of election prediction? Using tweets, we can analyze the positive or negative feelings or opinions posted by people on social media. Furthermore, the preprocessing techniques, such as removal of emoticons, repeated words, Twitter mentions etc. are applied to dataset (tweets) and then classification models are applied for predicting the results.

II. LITERATURE REVIEW

This part of the paper is used to explain related study of opinion mining of different candidates, related techniques, micro- blogging system tasks and algorithms to fulfill those tasks. Furthermore, it talks about certain significant categories that emerged from this study. It involves the analysis of Indian Twitter Tweets to predict the results of the upcoming general elections.

Sentiment Analysis in Elections

Sentiment analysis has become an essential tool for understanding public opinion during elections. Almatrafi et al.

[1] utilized location-based sentiment analysis on Twitter data to identify trends during the 2014 Indian General Elections. Their study underscored the significance of geospatial filtering for refining sentiment insights. Similarly, Wang et al. [16] analyzed real-time Twitter data during the 2012 US Presidential elections, offering a robust framework for monitoring public sentiments on political events.

Sentiment Analysis in Indian Languages

The complexity of analyzing Indian languages has driven the development of unique linguistic resources. Das and Bandyopadhyay [2, 3] introduced SentiWordNet for Bangla and proposed fine-grained sentiment tagging for Bengali blogs [4]. For Hindi, Joshi et al. [5] proposed a fallback strategy for

sentiment classification, addressing the challenges posed by limited language resources. Bakliwal et al. [6, 8, 10] made significant contributions by developing Hindi subjective lexicons and employing graph traversal techniques to enhance sentiment analysis accuracy.

Comparative Analysis of Classifiers

Various studies have compared machine learning models for sentiment classification. Pak and Paroubek [13] demonstrated that Support Vector Machines (SVM) outperformed other classifiers like Naive Bayes (NB) and Long Short-Term Memory (LSTM) in terms of precision and accuracy. Similarly, Bermingham and Smeaton [15] found that SVM excelled in classifying sentiment in microblogs, even with short text data. These findings align with Mukherjee and Bhattacharyya's [7] discourse analysis approach and Mittal et al.'s [12] work on negation handling, which further improved sentiment classification performance.

Twitter Sentiment Analysis Approaches

Twitter's brevity and informal language present unique challenges. Go et al. [14] and Pak and Paroubek [13] used emoticons as distant supervision labels to classify tweets as positive or negative. This approach proved effective in generating large-scale training datasets. Sharma et al. [17] applied similar methodologies to Hindi tweets, integrating negation and discourse relation handling for more precise sentiment classification.

Sentiment Analysis Using Twitter

Recent research based on sentiment analysis says that the analysis of opinion utilizes simultaneous learning. Pak and Paroubek in [13] utilized tweets which end with emoticons like ":)" ":-)" as positive, and ":(" ":-(" as negative.

They accumulated models including LSTM, Support Vector Machines (SVM) and Naive Bayes and concluded that SVM performed the best amongst various others, attaining more precision which lead SVM to be the best performer of all the classifiers. They recorded that all distinctive models were beaten by the unigram model. To gather subjective information, they compile the tweets ending with emoticons comparatively as Go et al. In [14].

Birmingham and Smeaton [15] tested two distinct strategies, Multinomial Naïve Baye's (MNB) and SVM for web pages and scale blog. They found that MNB methodology outperforms SVM on scaled scale areas with short substance. Wang and Can et al. [16] build a reliable structure for the 2012 US races to recuperate political suppositions at work using Twitter. In the present systems, they are considering real time tweets, keeping location as filter and then analyzing people's sentiments.

Applications and Impact

Research has highlighted the value of sentiment analysis for political forecasting. By integrating real-time tweet analysis and advanced classifiers, such as SVM and LSTM, researchers have successfully captured public opinion trends, offering predictive insights into electoral outcomes. Studies like those by Almatrafi et al. [1] and Wang et al. [16] have shown the importance of combining linguistic and computational techniques to enhance the reliability of sentiment analysis in multilingual and diverse political contexts.

III. METHODOLOGY

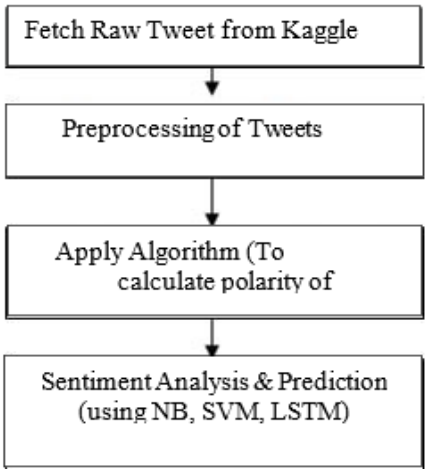

Fig. 1. Steps and techniques used in Experiment.

A. Data Collection

The data set for this study was sourced from Kaggle, which included tweets related to the Indian General Elections 2019. For each candidate, tweets were filtered based on the usage of specific hashtags associated with the candidates. For instance, hashtags like #Modi, #Rahul, and others were used to collect tweets supporting or opposing Narendra Modi and Rahul Gandhi, respectively.

The data set consists of English-language tweets collected between January 1, 2019, and May 23, 2019, covering the election campaign period. The data underwent a rigorous preprocessing stage to remove duplicates, retweets, and non-relevant content while ensuring the quality and relevance of sentiment analysis.

B. Preprocessing

The data extracted from twitter contains lot of special characters and unnecessary data which we not require [1]. If data is not processed beforehand, it could affect the accuracy as well as performance of the network down the lane. So it is very important to

process this data before training. We need to get rid of all the links, URLs and @ tags. Pre-processing also includes removal of stop words from the text to make analysis easier

C. Negation Handling

In English language there are certain words like "No", "Not" which can revert the meaning of the sentence. So, these words also help in finding the polarity of tweets.

D. Algorithms Used

We used a supervised approach such as classification algorithms Naive Bayes, Support Vector Machine and Long short-term memory. We took tweets with the names of Indian political Candidate such as #Modi, #Rahul. We collected a total of 23,998 tweets relevant to these hashtags.

a) Long Short-Term Memory

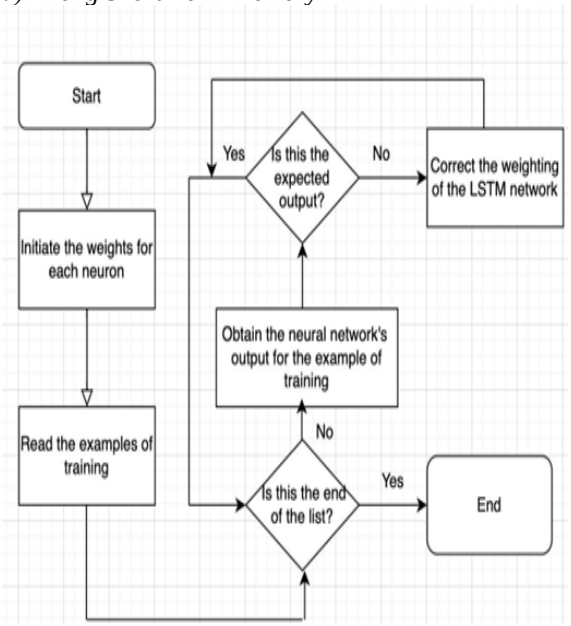

Fig. 2. Flow diagram of using Long Short-Term Memory

Long Short-Term Memory networks are an extension for recurrent neural networks, which extends their memory. It is very necessary to learn from important experiences that have very long-time gaps in between. They are capable of learning long-term dependencies. LSTM's work extremely well on a large variety of problems, and therefore are now widely used. They are specially designed to avoid the long-term dependency problem. To remember information for long periods of time is practically their default behavior.

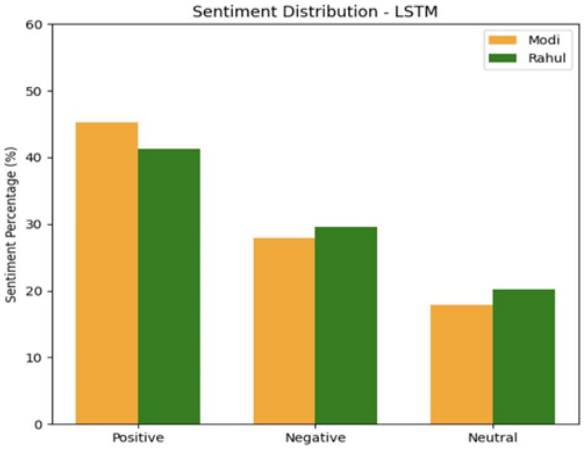

Fig. 3. Percentage of number of positive, negative and neutral tweets using Long Short Term Memory classifier

The above chart shows the political Candidate and their positive, neutral, and negative percentages. The algorithm gave an accuracy of 91%. According to this algorithm and the above figure, Modi with 45.22% positive tweets had a greater likelihood of winning the elections.

a) Naïve Bayes Classifier

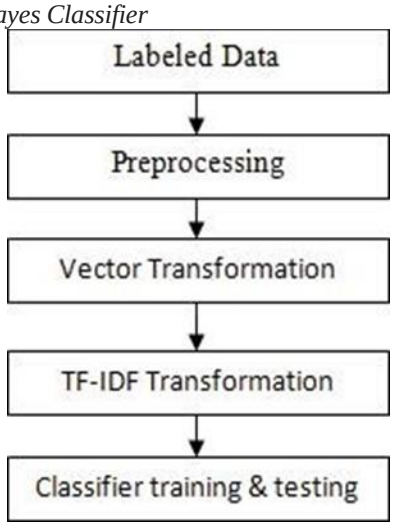

Fig. 4. Steps and techniques used in sentiment classification.

It is a simple probabilistic classifier based on the Baye's theorem. It assumes every feature is independent of each other. To assign labels for every input vector feature is utilized using the formula below.

$$P(\text{label} \mid \text{features}) = \frac{\{P(\text{label}) * P(\text{features} \mid \text{label})\}}{P(\text{features})}$$

Label in the above equation shows the polarity or sentimenti.e. positive, neutral and negative, and features are the words which have been extracted from the tweets.

Candidate	Total Tweets
Modi	53090
Rahul	30987

TABLE I. Total number of tweets for each Candidate.

We fetched a total of 153,000 tweets. After preprocessing, we were left with 84,077 tweets. We classified them using NB classifier. For further calculations, we manually labelled the dataset of 84,077 tweets and then performed 5-fold cross validation. In the cross-validation method, the process of training and testing was repeated 5 times utilizing 80% of the dataset as training data and the remaining 20% as testing data. After that, the average accuracy of classifiers was obtained.

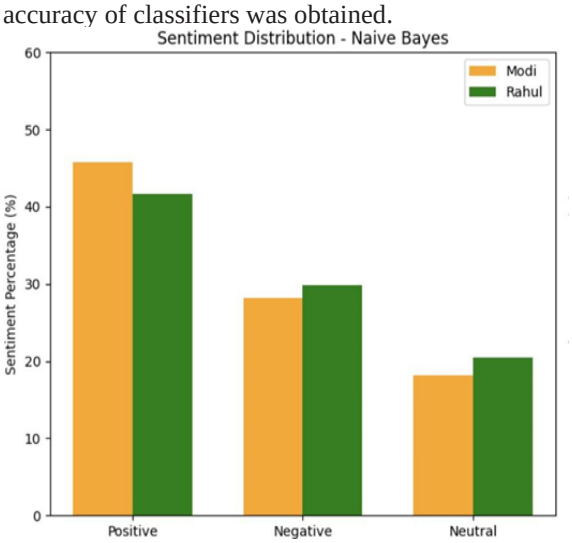

Fig. 5. Percentage of number of positive, negative and neutral tweets using NB classifier.

The above chart shows the political Candidate and their positive, neutral, and negative percentages. The algorithm gave an accuracy of 92%. According to this algorithm and the above figure, Modi with 45.71% positive tweets had a greater likelihood of winning the elections.

b) Support Vector Machine

It is a learning system utilizes hypothesis space in high dimensional feature space. It is more considered where the number of samples is smaller in number from the number of dimensions.

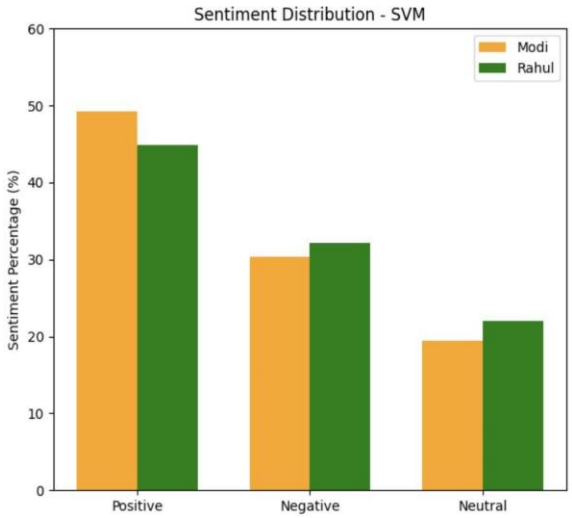

Fig. 6. Percentage of number of positive, negative and neutral tweets using the SVM classifier.

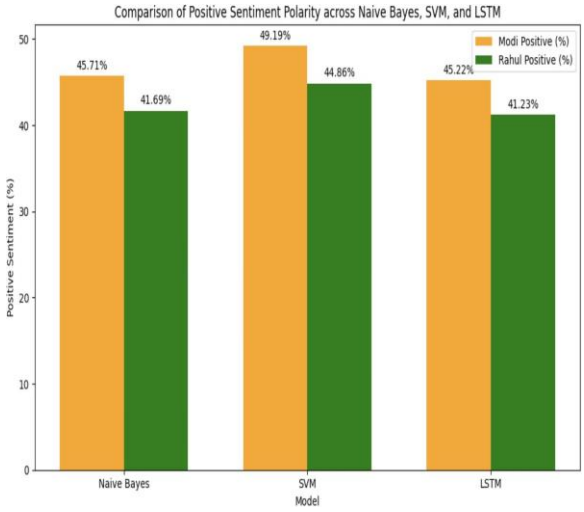

Fig. 7. Comparison of Positive polarity of all the three algorithms.

The above chart shows the political Candidate and their positive, neutral, and negative percentages. The algorithm gave an accuracy of 99%. According to this algorithm and the above figure, Modi with 49.19% positive tweets had a greater likelihood of winning the elections.

IV. RESULT AND DISCUSSIONS

As it is very difficult to predict the results of elections using other methods, including public opinion polls, and with the growing prevalence of social media, such as Facebook and Twitter, the authors decided to utilize sentiment analysis of Twitter tweets to predict the results of the Indian general election.

.

TABLE II. Accuracy of algorithm.

Algorithm	Accuracy
Naive Bayes	92%
SVM	99%
LSTM	91%

As shown in the above table the accuracy of the Naive Baye's algorithm was 92% and the accuracy of Support Vector Machine was 99% and LSTM 91%. We made our final prediction utilizing SVM, since the accuracy of the algorithm is higher. We predicted that the Candidate that had a better chance of winning the 2019 general election is Modi.

TABLE III. Precision Recall of algorithm

Algorithm	Precision	Recall
Naive Bayes	.83	.82
SVM	.85	.85
LSTM	.87	.86

We also calculated the precision and recall as shown in the above table. The result of the Naive Baye's algorithm was .83 and .82. For Support Vector Machine we obtained .85 as precision and .85 as recall. and LSTM obtained .87 precision and .86 recall.

V. LIMITATION

The limitation of our research is that we did not consider the emoticons which are also a relevant aspect when defining the polarity of a tweet. We can retrieve more tweets and classify them after the data was manually labeled because the number, 84,077, was insufficient to produce more accurate findings. In the future.

VI. FUTURE WORK

There could be many other prospective areas to conduct this research in, including the data from other big social media sites like Facebook to increase the size of the data set. We have more space to work with the training dataset such as considering the sample dataset in which the certain number of features of an algorithm is already defined. More machine learning algorithms such as Regression, Random Forest can also be considered for classification and further prediction.

VII. CONCLUSION

This research demonstrates the effectiveness of sentiment analysis in predicting electoral outcomes by leveraging Twitter data. Among the three models employed—Naive Bayes, SVM, and LSTM—the SVM model achieved the highest accuracy (99%), establishing it as the most reliable for our predictions. The analysis highlighted the superiority of machine learning models in capturing public sentiment and identifying trends, with the results suggesting Narendra Modi as the candidate with a greater likelihood of winning the 2019 Indian General Election. Despite certain limitations, such as excluding emoticons and relying on a relatively small dataset, this study provides a foundation for further exploration in social media-based sentiment analysis and electoral predictions. Future research can address these limitations by incorporating diverse datasets and advanced methodologies to enhance predictive accuracy.

REFERENCES

[1] O. Almatrafi, S. Parack, and B. Chavan, “Application of Location-Based Sentiment Analysis Using Twitter for Identifying Trends Towards Indian General Elections 2014,” Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, Article No. 41, Jan. 2015.

[2] A. Das and S. Bandyopadhyay, “SentiWordNet for Indian languages,” Proceedings of the 8th Workshop on Asian Language Resources, pp. 56–63, Aug. 2010.

[3] A. Das and S. Bandyopadhyay, “SentiWordNet for Bangla,” Knowledge Sharing Event-4: Task, Volume 2, 2010.

[4] D. Das and S. Bandyopadhyay, “Labeling emotion in Bengali blog corpus - a fine-grained tagging at sentence level,” Proceedings of the 8th Workshop on Asian Language Resources, pp. 47–55, Aug. 2010.

[5] A. Joshi, B. A. R, and P. Bhattacharyya, “A fall-back strategy for sentiment analysis in Hindi: a case study,” Proceedings of ICON 2010: 8th International Conference on Natural Language Processing, Dec. 2010.

[6] A. Bakliwal, P. Arora, and V. Varma, “Hindi subjective lexicon: A lexical resource for hindi polarity classification,” Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp. 1189–1196, May 2012.

[7] S. Mukherjee and P. Bhattacharyya, “Sentiment analysis in twitter with lightweight discourse analysis,” Proceedings of the 24th International Conference on Computational Linguistics (COLING), pp. 1847–1864, Dec. 2012.

[8] A. Bakliwal, P. Arora, A. Patil, and V. Varma, "Towards enhanced opinion classification using

NLP techniques," Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP, pp, 101– 107, Nov. 2011.

[9] B. R. Ambati, S. Husain, S. Jain, D. M. Sharma, and R. Sangal, "Two methods to incorporate local morphosyntactic features in Hindi dependency parsing," Proceedings of the NAACL HLT 1st Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL), pp. 22–30, June 2010.

[10] P. Arora, A. Bakliwal and V. Varma, "Hindi Subjective Lexicon Generation using WordNet Graph Traversal," International Journal of Computational Linguistics and Applications, Vol. 3, No. 1, pp. 25–39, Jan-Jun 2012.

[11] H. Gune, M. Bapat, M. M. Khapra and P. Bhattacharyya, "Verbs are where all the action lies: Experiences of shallow parsing of a morphologically rich language", Proceedings of the 23rd International Conference on Computational Linguistics, pp. 347–355, Aug. 2010.

[12] N. Mittal, B. Agarwal, G. Chouhan, N. Bania, and P. Pareek, "Sentiment Analysis of Hindi Review based on Negation and Discourse Relation," Proceedings of International Joint Conference on Natural Language Processing, pp. 45–50, Oct. 2013.

[13] A. Pak, and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), pp. 1320–1326, May 2010.

[14] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford University, pp. 1–12, 2009.

[15] A. Bermingham, and A. F. Smeaton, "Classifying sentiment in microblogs: Is brevity an advantage?" Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1833–1836, Oct. 2010.

[16] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp 115–120, July 2012.

[17] Y. Sharma, V. Mangat and M. Kaur, "A Practical Approach to Sentiment Analysis of Hindi Tweets," Preceedings of the 1st International Conference on Next Generations Computing Technologies (NGCT), Sept. 2015.