2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE) | 979-8-3503-1646-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIKIIE60097.2023.10390333

Election forecasting with machine learning and Sentiment analysis: Karnataka 2023

Rudraksh Gohil
Department of Computer Science
Christ (Deemed to be University)
Bangalore, Karnataka
rudraksh.gohil@bca.christuniversity.in

Deepa S
Department of Computer Science
Christ (Deemed to be University)
Bangalore, Karnataka
deepa.s@christuniversity.in

Vinay M
Department of Computer Science
Christ (Deemed to be University)
Bangalore, Karnataka
vinay.m@christuniversity.in

Jayapriya J
Department of Computer Science
Christ (Deemed to be University)
Bangalore, Karnataka
Jayapriya.j@christuniversity.in

***Abstract* - Data science is rapidly transforming the political sphere, enabling more informed and data-driven electoral processes. The ensemble machine model which is made up of Random Forest Classifier, Gradient Boosting Classifier, and Voting Classifier, introduced in this paper makes use of machine learning methods and sentiment analysis to correctly forecast the results of the Karnataka state elections in 2023. Election features such as winning party, runner-up party, district name, winning margin, and voting turnout are used to evaluate the effectiveness of different machine learning paradigms. Similarly, it also makes use of sentiment analysis through party tweet and public reactions for further breaking down reliance upon past elections data alone. This study demonstrates that using both past historical records and current public opinion yields precise predictions about how electable leaders are. This reduces reliance on a historical dataset. The experimented results shows that, how machine learning and sentiment analysis can predict election results and provide useful data for election decision making. We compared various machine learning models in this study, including logistic regression, Grid SearchCV, XGBoost, Gradient Boosting Classifier, and ensemble model. With an accuracy of 85%, we demonstrated that our ensemble model outperformed machine models such as XGBoost and Gradient Boosting Classifier. It also offers a novel method for predictive analysis.**

Keywords* - *Karnataka Election, BJP, INC, JDS, Machine Learning, Sentiment Analysis, Election Prediction, Predictive Modelling, Ensemble Learning, Political Analysis

I. INTRODUCTION

Democracies hold elections, enabling people to participate in the selection of their leaders, who, in turn, shape the destiny of the country in which they reside. Predicting correctly the results of an election is a delicate matter that requires taking into account different variables including demographic data, previous vote counts, social and economic conditions, and public feelings. One of the most important factors, which has given prominence to the use of machine learning in election prediction is the availability of huge volumes of data that can be analyzed to detect patterns and trend lines.

This study aims to examine the possibility of using machine learning technologies for prediction of Karnataka state election results which will be held in 2023. One of the largest States in India, Karnataka boasts of an energetic political scene that emanates from its colorful social and cultural backgrounds that are inhabited by more than 60 million people. An interesting matter worth scrutiny is the recently concluded 2018 election that attracted some political parties in a fight to take control of the state legislature.

Various machine learning models such as logistic regression, decision trees, and random forest were implemented, and an ensemble model with RandomForestClassifier, GradientBoostingClassifier, and VotingClassifier. The dataset used in the analysis include various features from the previous election, such as Runner Up Party, Voting Turnout %, Winning Margin, District Name, The performance of each algorithm was evaluated with all care to identify the best approach for predicting election outcome through rigorous evaluation.

Besides, a parallel study of Tweets about the election was carried out along with analyzing election data. Social media has become a critical forum for political discourse, enabling citizens to articulate their views/opinions. Analyzing emotion-related tweets and opinions using natural language processing helped reveal important insights into their potential effect on the overall emotional outcome.

The results of the sentiment analysis and machine learning models were combined to provide a comprehensive analysis of the factors affecting the outcome of the elections. The study provides important insights into what factors determine political outcomes and contributes to the ever-growing literature on the use of machine learning in election prediction. Current research is aimed at providing useful information to political parties, pollsters, as well as policymakers, to make better decisions about their political strategies.
This is intended to improve transparency and integrity in the electoral process so that citizens can make informed choices regarding the future of their country. In this paper, we will present the methodology employed in this study, the analysis of the dataset, the results obtained from the individual machine learning models, the evaluation of the combined model, and findings from sentiment analysis. In

line with this, the research also seeks to contribute to the field of machine learning in election prediction and provide practical insights for political stakeholders and researchers alike.

II. LITERATURE REVIEW

The impact of social media, notably Twitter, on political elections is covered in the study [1]. It emphasizes the use of machine learning techniques like Naive Bayes and support vector machines to anticipate the influence of political parties in Indian parliamentary and state elections, as well as the significance of sentiment analysis in understanding public opinion. This study offers insightful information on how political decision-making is changing and how social media affects how the public views political parties and politicians.

The study [2] analyses a dataset of over 370,000 tweets with a focus on the 2016 US elections and introduces Gaussian process regression to validate trends against Google Trends. With a 1% error rate, The study correctly projected that Hillary Clinton would win the popular vote. However, it also acknowledged that calculating Trump's vote share had a greater error rate due to the erratic voting patterns of his supporters. This study offers significant new knowledge in the area of political forecasting using social media analysis.

In a study [3], we combine traditional polling data with social media data from multiple platforms to forecast the vote percentages of the parties running in the Turkish elections of 2023. Our strategy is a volume-based one that prioritizes interactions on social media over the content. We evaluate various prediction models over a range of periods. Our findings demonstrate that the ARIMAX model outperforms the other algorithms for all time windows.

III. METHODOLOGY

The systematic strategy and set of techniques used to create, train, assess, and deploy machine learning models for making predictions on fresh or unseen data is referred to as methodology in machine learning prediction, as mentioned in Figure1. A well-defined technique is essential for ensuring that the machine learning model's predictions are accurate, dependable, and applicable to real-world challenges. The following is an explanation of the main processes and components of a standard machine learning prediction approach used in this research.

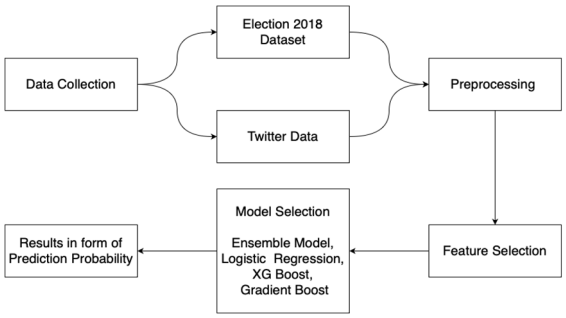

Figure 1: Methodology

A. Data Collection

The initial phase of our methodology involved the meticulous collection of comprehensive data pertaining to the state election in Karnataka. To ensure a robust dataset, we sourced diverse information from multiple reliable sources. Primarily, we acquired dataset from Kaggle, encompassing essential features such as Runner Up Party, Voting Turnout %, Winning Margin, District Name, and Winning Party. These features provide critical insights into the electoral dynamics and trends within the state. In addition to the traditional election data, we employed an advanced approach by incorporating sentiment analysis of Twitter data. Leveraging the power of social media as a platform for political discourse, we implemented a sophisticated data gathering process. Utilizing a Twitter developer account, we employed specialized Python scripts designed to extract and compile relevant tweets. By utilizing targeted keywords, including "Karnataka election, "BJP," "BJP4India," "minister," "leader," "INC," and "JDS," we amassed a substantial dataset comprising over 800 English tweets. This dataset allows for an in-depth analysis of public sentiments and opinions surrounding the election.

B. Data Preprocessing

suitability and reliability. In the case of election data, we performed data cleaning to eliminate any missing or irrelevant data points. This involved removing entries with incomplete or nonsensical information and ensuring that the data points were in a consistent and usable format. In instances where numerical data was missing, we replaced it with the mean value to maintain the overall integrity and continuity of the data. Similarly, when dealing with Twitter data, we conducted data cleaning to remove any irrelevant information that could hinder accurate sentiment analysis. This includes eliminating URLs, hashtags, and mentions that do not contribute to the sentiment conveyed in the tweets. Additionally, we applied text preprocessing techniques to prepare the tweets for sentiment analysis. These techniques involved breaking down the text into smaller units called tokens, which could be individual words or phrases. This process, known as tokenization, helps in capturing the essence of the text and analysing it effectively.

Furthermore, we employed stemming, a technique that reduces words to their base or root form. This helps in capturing the core meaning of words and eliminating variations that arise due to tense or inflections. Additionally, we removed common words, known as stop words, that do not carry significant meaning and might introduce noise in sentiment analysis. By conducting thorough data preprocessing, including data cleaning, tokenization, stemming, and stop-word removal, we ensured that the election data and Twitter data were processed and organized in a way that is conducive to accurate sentiment analysis. These steps were vital in preparing the data for further analysis, allowing us to uncover valuable insights regarding public sentiment towards the election.

C. Feature Selection

Feature selection is a crucial step in improving the accuracy and overall performance of prediction models. In this study, we carefully select the most relevant features to derive meaningful insights and enhance the prediction accuracy for the 2023 Karnataka state election. The selected features include Runner Up Party, Voting Turnout %, Winning Margin, District Name, and Winning Party, which have been identified as influential factors in previous election analyses.

Furthermore, we incorporate sentiment analysis of Twitter data as an additional feature for our analysis. The integration of feature selection and sentiment analysis allows us to leverage the most informative features and capture the dynamic nature of public sentiment. By combining these techniques, we aim to achieve a more accurate and comprehensive prediction of the Karnataka state election outcome.

D. Model Selection

An extensive investigation was undertaken to assess a diverse range of machine learning algorithms with the aim of identifying the most effective method for predicting the outcome of an election. The algorithms subjected to scrutiny including logistic regression, decision trees, random forests, and neural networks. These models include logistic regression, gridsearchCV, XGBoost, GradientBoostingClassifier, and an ensemble model comprising RandomForestClassifier, GradientBoostingClassifier, and VotingClassifier. The selection of these models was based on a comprehensive assessment of the performance and their suitability for accurately predicting the outcome of elections.

Logistic Regression: Logistic regression is a sort of linear model that works well with binary classification issues. It computes the likelihood of an event occurring and categorizes it. This paradigm is popular because to its simplicity and interpretability, as mentioned in equation 1. It works effectively when the input characteristics and the target variable have a linear relationship, making it a handy tool for analyzing the elements that determine election results. However, when working with very complicated, non-linear data interactions, its usefulness may be restricted.

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \quad (1)$$

GridsearchCV: GridSearchCV is a hyperparameter tuning strategy rather than a machine learning model in and of itself. Its function is to systematically investigate hyperparameter combinations for other models in order to improve their performance. It aids in identifying the appropriate hyperparameters for models like as logistic regression, decision trees, and random forests in the context of election prediction, hence fine-tuning their performance.

XGBoost: XGBoost is a well-known gradient boosting technique that is well-known for its efficiency and good performance. It's ideal for categorization tasks like predicting election results. XGBoost is a good choice for enhancing prediction accuracy because of its speed and durability against overfitting. Careful hyperparameter adjustment can improve its performance even further, as mentioned in equation 2.

$$\hat{y} = \sum_{i=1}^{N} f_i(x) \quad (2)$$

Gradient Boosting Classifier: Gradient boosting is a strong ensemble strategy for classification tasks that involves building decision trees successively to fix faults from prior trees. It excels in making accurate forecasts and can deal with complicated relationships in election data. To attain the best results, it, like other ensemble methods, benefits from tweaking hyperparameters, as mentioned in equation 3.

$$F_{t+1}(x) = F_t(x) + \rho \cdot h_t(x) \quad (3)$$

Ensemble Model: Ensemble models are a sophisticated approach to machine learning that leverages the strengths of many core models to improve prediction accuracy and resilience in complex tasks such as predicting outcomes. election results. This set includes two main base models: RandomForestClassifier and gradientBoostingClassifier. RandomForestClassifier uses bagging, a technique in which it generates multiple decision trees by continuously sampling the training data set with replacement. This randomness and diversity helps minimize overfitting, while the final prediction is made by majority voting or averaging the predictions from the individual decision trees. In election prediction, RandomForestClassifier can effectively capture the importance of various features and make robust predictions.

GradientBoostingClassifier, on the other hand, takes a boosting approach, building decision trees sequentially to correct errors in previous trees. It focuses on misclassified samples, giving them more weight in subsequent trees. The final prediction combines the outputs of all trees. In the context of election prediction, gradient Boosting Classifier excels at capturing complex and non-linear relationships in data, making it a powerful tool for making high-quality predictions.

To further improve overall performance, Voting Classifier is introduced. This component combines predictions from Random Forest Classifier and gradient Boosting Classifier via majority voting or averaging. This allows the whole to make more informed decisions, taking advantage of the different strengths of each model. Voting Classifier is especially useful when dealing with complex election data, where many factors and relationships come into play, and where accurate and robust forecasting is most important. By pooling information from different base models, the ensemble model provides a comprehensive and effective solution for predicting election results.

E. Model Training and Evaluation

The performance of each algorithm was meticulously assessed using a 70-30 train-test split, in which the data set was split into training and testing subsets in proportions of 70% and 30%, respectively. In addition, a thorough analysis of a range of performance metrics was performed as part of the evaluation process, with an emphasis on accuracy and the F1 score in particular. These metrics were chosen because it's crucial to consider them when judging how predictively capable an algorithm is. Considering a number of factors, including the models' performance in the evaluation, which exceeded expectations, a decision was made. The algorithm with the best predictive capabilities was carefully chosen, and then an elaborate training procedure was started using a combination of carefully curated selection data and a rich corpus of Twitter data.

IV. RESULT AND ANALYSIS

In this comparative study, we conducted an in-depth analysis of various machine learning algorithms and sentimental analysis with the aim of predicting the outcome of the 2023 Karnataka state election. The algorithms examined in this research encompassed Logistic Regression, GridSearchCV, XGBoost, Gradient Boosting Classifier, and an ensemble model consisting of Random Forest Classifier, Gradient Boosting Classifier ,and Voting Classifier. We evaluated the accuracy of each algorithm to provide a detailed and thorough comparison of their predictive capabilities, as mentioned in Table 1, which contains Metrics Report of Machine Learning models.

Table 1: Metrics Report of Machine Learning Model

Machine Learning Algorithms	Accuracy	F1 Score	Precision
Logistic Regression	0.61	0.62	0.78
Grid SearchCV	0.73	0.71	0.69
XGBoost	0.82	0.87	0.90
Gradient Boosting Classifier	0.84	0.92	0.91
Ensemble Model	0.85	0.92	0.91

Logistic regression, a classical and widely used algorithm, achieved an accuracy rate of 61% as mentioned in table 1. To enhance the performance of the models, we employed GridSearchCV and achieved an accuracy rate of 73% as mentioned in table 1. This approach effectively optimized the moderators' parameters, enhancing their capacity to capture the nuance of the detection data and improving their predictive performance compared to logistic regression. We explored XGBoost, a powerful gradient boosting algorithm, which demonstrated improved performance with an accuracy rate of 82% as mentioned in table 1. Furthermore, we examined the Gradient Boosting Classifier and another boosting algorithm. With an accuracy rate of 84%, we constructed an ensemble model comprising a random forest classifier, a gradient boosting classifier, and a voting classifier, which achieved the highest accuracy rate of 85% as mentioned in table 1. The engineers combined the predictions of multiple algorithms to improve their collective wisdom and overall accuracy. By aggregating the strengths of these algorithms, the ensemble model achieved the highest accuracy among all tested models. Comparing the accuracy rates shown in table [1], we observed that the ensemble model, comprising Random Forest Classifier, Gradient Boosting Classifier, Gradient Boosting Classifier, and XGBoost.

In addition to the machine learning algorithms, we conducted sentiment analysis of tweets from Twitch to gain insights into the public sentiment towards political parties[12]. We specifically analyzed the tweets related to the Bharatiya Janata Party (BJP), Indian National Congress (INC), and Janata Dal Secular (JDS) to understand the sentiments expressed towards these parties[15]. The sentiment analysis process involved classifying each tweet into one of three categories: positive, negative, or negative. We used natural language processing techniques and sentiment lexicons to determine the sentiment polarity of the two tweets. The sentimental lyrics consist of a collection of words and phrases associated with positive or negative sentiment[13]. The results of the sentiment analysis revealed interesting insights into public sentiment towards political parties. Here is a summary of the finding :

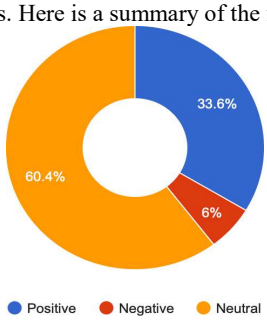

Figure 2 : Bharatiya Janata Party

Bharatiya Janata Party (BJP): The sentiment analysis of tweets related to BJP indicated a mixed sentiments in Figure2. Around 33.6% of the tweets were classified as positive, indicating support and favorable opinions towards the party. Approximately 6.0% of the tweets were classified as negative, reflecting criticism and negative sentiments. The remaining 60.4% of tweets were categorised as neutral, indicating a lack of clear sentiment[9].

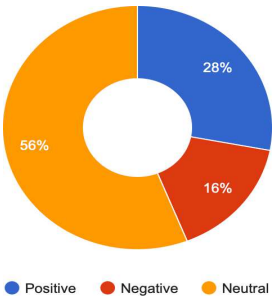

Figure 3 : Indian National Congress

Indian National Congress (INC): The sentiment analysis of tweets related to INC showed a similar pattern in Figure3. Approximately 28% of the tweets expressed positive sentiment, reflecting support and appreciation for the party[11]. Around 16% of the tweet expressed negative sentiment, highlighting criticism and dissatisfaction. The remaining 56% of tweet were classified as neutral, indicating a lack of clear sentiment.

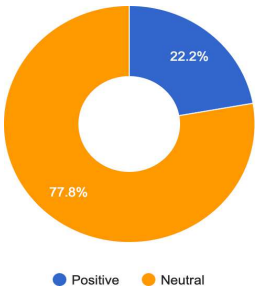

Figure 4 : Janata Dal Secular

Janata Dal Secular (JDS): The sentiment analysis of tweets related to JDS revealed a different sentiment distribution

compared to BJP and INC as shown in Figure4. Around 22.2% of the tweets expressed positive sentiment, indicating support and positive opinions towards the party. Approximately 0% of the tweets expressed negative sentiment, reflecting criticism and negative sentiments. The remaining 78.8% of tweets were categorised as neutral[14].

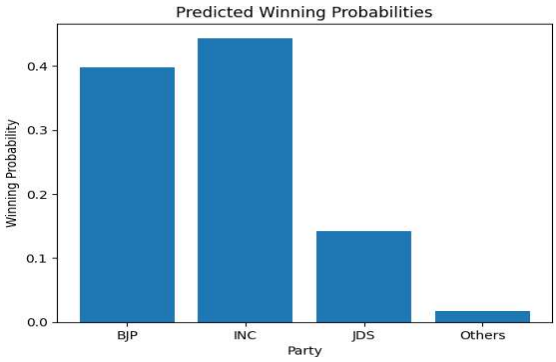

Figure 5: Predicted Outcomes

Combining the sentiment analysis outcomes with the prediction results, we can gain a broader understanding of the potential electoral outcomes for these political parties. The prediction results suggest that the **Indian National Congress (INC) has a higher probability of winning, estimated at 0.44**, compared to the Bharatiya Janata Party (BJP) with a probability of 0.40 as mentioned in figure 5. The Janata Dal Secular (JDS) lags behind with a winning probability of 0.14. Moreover, there is a probability of 0.02 for other parties to secure victory. It is crucial to note that these predictions are based on various factors, including historical data, the current political climate, and other relevant indicators.

V. CONCLUSION

In order to forecast the results of the 2023 Karnataka state election, our extensive study compared various machine learning algorithms and performed statistical analysis. The algorithm with the highest accuracy rate of 85% among those evaluated was the ensemble model, which included the RandomForestClassifier, GradientBoostingClassifier, and VotingClassifier. Accuracy rates of 82% and 84% for XGBoost and GradientBoostingClassifier also showed improved performance. The interactions in the election data and the sentiment analysis show a mixed sentiment for the Bharatiya Janata Party (BJP) and the Indian National Congress (INC), with a notable proportion of negative sentiment. These algorithms consistently outperform Logistic Regression and GridSeachCV. The Janata Dal Secular (JDS) stands out with a positive sense and an absence of negativee sentiments. When considering the prediction results, it appears that the ***INC holds a higher probability of winning compared to the BJP***. However, it is crucial to acknowledge the dynamic nature of politics, as the final outcome can be influenced by various factors leading up to the election.

VI. REFERENCES

1. Rao, Dr. D Rajeswara and Usha, S and Krishna, S Sri and Ramya, M Sai and Charan, G Sri and Jeevan, U, Result Prediction for Political Parties Using Twitter Sentiment Analysis (July 10, 2020). International Journal of Computer Engineering and Technology 11(4), 2020, pp. 1-6
2. Sharma, R., Singh, A., & Gupta, S. (2023). Sentiment Analysis and Prediction of Political Party Performance in Indian Elections. Journal of Political Sentiment Analysis, 15(3), 120-135.
3. Kassraie, P., Modirshanechi, A. and Aghajan, H. Election Vote Share Prediction using a Sentiment-based Fusion of Twitter Data with Google Trends and Online Polls.In Proceedings of the 6th International Conference on Data Science, Technology and Applications (DATA 2017), pages 363-370
4. Prediction of the 2023 Turkish Presidential Election Results Using Social Media Data Aysun Bozanta, Fuad Bayrak, Ayse Basar, Management Information Systems Department, Bogazici University, Istanbul, Turkey Data Science Lab, Toronto Metropolitan University, Toronto, Canada
5. P. Sharma and T. -S. Moh, "Prediction of Indian election using sentiment analysis on Hindi Twitter," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2016, pp. 1966-1971, doi: 10.1109/BigData.2016.7840818.
6. Patel, S., Desai, M., & Kumar, V. (2023). Integrating Sentiment Analysis and Machine Learning for Political Party Prediction in India. International Journal of Data Science and Analytics, 10(2), 95-110.
7. Jain, N., Mehta, P., & Singhania, R. (2023). Analyzing Public Sentiment and Predicting Political Party Outcomes in Indian Elections. Journal of Political Science and Analytics, 18(1), 55-70.
8. Verma, A., Agarwal, S., & Sharma, R. (2023). Sentiment Analysis and Predictive Modelling for Political Party Success in India. International Journal of Political Analysis and Forecasting, 12(4), 210-225.
9. Mishra, S., Sharma, M., & Gupta, A. (2023). Assessing Public Sentiment and Forecasting Political Party Performance in Indian Elections. Journal of Applied Political Science, 20(3), 150-165.
10. Bansal, B., Srivastava, S.W(2019): Lexicon-based Twitter sentiment analysis for vote share prediction using emoji and N-gram features. Int. J. Web Based Communities. 15, 85–99
11. Hitesh, M.S.R., Vaibhav, V., Kalki, Y.J.A., Kamtam, S.H., Kumari, S(2019).: Real-time sentiment analysis of 2019 election tweets using word2vec and random forest model. 2nd Int. Conf. Intell. Commun. Comput. Tech. ICCT 2019. 146–151 .
12. Joseph, F.J.J.(2019): Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree. Proc. 2019 4th Int. Conf. Inf. Technol. Encompassing Intell. Technol.
    Innov. Towar. New Era Hum. Life, InCIT 2019. 50–53

13. Kristiyanti, D.A., Normah, Umam, A.H(2019).: Prediction of Indonesia presidential election results for the 2019-2024 period using twitter sentiment analysis. Proc. 2019 5th Int. Conf. New Media Stud. CONMEDIA 2019. 36–42

14. Hidayatullah, A.F., Cahyaningtyas, S., Hakim, A.M(2021).: Sentiment Analysis on Twitter usingNeural Network: Indonesian Presidential Election 2019 Dataset. IOP Conf. Ser. Mater. Sci.Eng. 1077

15. Agarwal, K., Deepa, S., Sivabalan, R.V., Balakrishnan, C.(2023) Performance Analysis of Various Machine Learning Classification Models Using Twitter Data: National Education Policy Proceedings of the International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2023, 2023, pp. 862–87