SENTIMENTAL ANALYSIS ON ELECTION CAMPAIGN DATA BASED ON ENSEMBLE LEARNING

Mr. Birendra Kumar
Associate Professor
Ajay Kumar Garg Engineering College
Ghaziabad, India

Rohit Kumar Singh
Department of IT
Ajay Kumar Garg Engineering College
Ghaziabad, India

Sejal Kulshrestha
Department of IT
Ajay Kumar Garg Engineering College
Ghaziabad, India

Shivi Tyagi
Department of IT
Ajay Kumar Garg Engineering College
Ghaziabad, India

Shourya Singh Sengar
Department of IT
Ajay Kumar Garg Engineering College
Ghaziabad, India

Abstract

Elections campaigns play a crucial role in governance, allowing people to select their representatives. Traditionally, communication involved in the election campaigns are spoken words and later transitioned to written formats. With the rise of social media, especially on platforms like Twitter (referred to as X), people now express their opinions and reviews about political parties openly. Analysing this wealth of data from election campaigns can unveil valuable insights into the election results. In response to this evolving landscape, this work has been undertaken to develop an artificial intelligence (AI) model. This model aims to predict whether campaign expressing opinions are positive, negative or neutral. To enhance the accuracy of this prediction, the project uses a stacking approach to make a hybridized model of an machine learning classifiers.

The significance of this project lies in its potential to aid in election predictions. By understanding the nature of opinions shared on social media, the developed AI model can contribute valuable insights into the dynamics of public sentiment during election campaigns. To make this information accessible and actionable, a web application has been created. This application integrates real-time election data, offering users a live feed of election trends and analyses based on the predictions made by the AI model. In essence, this project strives to leverage the power of AI and real-time data to provide a dynamic and informed perspective on election dynamics.

Keywords: Election data analysis, Sentiment Analysis, Twitter data analysis, social media reviews, Social media networks

---

I. INTRODUCTION

The rapid expansion of social media in recent times has empowered users with a robust platform to express their viewpoints. Popular platforms such as Facebook, Twitter, and Google+ are actively utilized for sharing ratings, reviews, and recommendations. The rapid evolution of social media has become a defining feature of contemporary times. Today, numerous social media platforms wield significant influence, leading to notable shifts in culture, ethics, norms, and fostering a more discerning mindset in responding to prevailing circumstances. Accessibility to

social media has expanded, reaching diverse social, political, and entertainment groups.

Functioning as an online medium, social media allows users from various backgrounds to effortlessly engage by following, sharing, and creating content, encompassing a wide range of formats such as photos, videos, blogs, social networks, and virtual experiences. Elections play a pivotal role in the democratic fabric of a nation. In the Indian parliamentary system, citizens exercise their right to determine the governing body for the next five years. Notably, from February 22 to March 22, elections were scheduled in five states, with Uttar Pradesh holding particular significance due to its contribution of the largest number of Members of Parliament to the national assembly. Over the years, market researchers have traditionally relied on standard methodologies like polls to gauge the beliefs and intentions of population segments. However, these approaches come with notable drawbacks, including the significant human effort required and the potential for being both costly and time-consuming. As technology advances, there is an increasing exploration of alternative and more efficient methods to obtain insights into public opinions and sentiments.

Sentiment Analysis (SA), also known as opinion mining, is the methodology that determines whether a word, sentence, or document conveys a positive, negative, or neutral sentiment. This innovation is widely employed to discern the sentiments of individuals towards a particular subject. The applications of sentiment analysis are diverse, extending to areas such as customer reviews, survey responses, and assessments of competitors. Its utility is widespread in business analytics, especially in scenarios involving the analysis of textual data. Sentiment Analysis seeks to unveil opinions, discern sentiments, and subsequently categorize these sentiments into various groups. A well-defined and accurate system for predicting sentiments could empower us to extract sentiments from the internet, forecast social behaviour, discern political trends, and identify emerging political factions within a specific geographical location.

However, when conducting sentiment analysis, several challenges arise because individuals do not consistently express their feelings in the same manner. A particular sentence may appear positive in one context but negative from a different perspective. Moreover, the presence of misspellings, intensifiers, and spams can introduce confusion during analysis. The myriad ways in which sentences can be constructed and the treatment of negations present additional complexities. With the rapid increase in subjective text on the internet, people turn to online platforms to obtain more nuanced, realistic, and subjective opinions on companies and products.

1.1 Scope of the work

The scope of the proposed system is to develop an AI based election campaign analysis application. Some applications of this works are,

1. The project focuses on analysing sentiments expressed on social media platforms, particularly Twitter, during election campaigns.

2. Political parties and candidates can use the insights provided by the AI model to understand public sentiment and adjust their campaign strategies accordingly.
3. Government agencies, political analysts, and public relations professionals can utilize the AI model to monitor media coverage and public opinion trends surrounding political events and figures.
4. Citizens can benefit from the web application by gaining access to real-time election data and analyses.

II. PROBLEM DEFINITION

The objective of this project is to classify the sentiment of a given election campaign as either positive or negative. Subsequently, utilizing these sentiment classifications, the goal is to conduct election result predictions based on the sentiment rates obtained from the analysis of the election campaign data. The overall flow of the suggested system is illustrated in the figure 1.

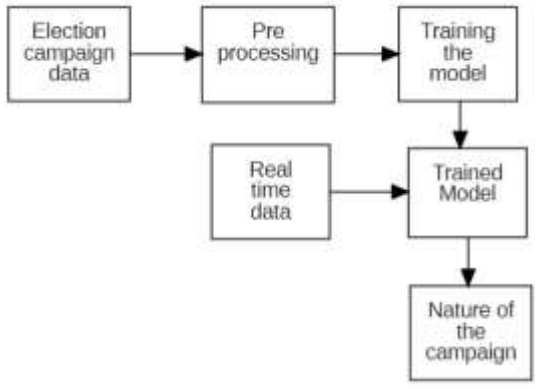

Fig 1. Work Flow Diagram

III. LITERATURE REVIEW

Many researches have evolved in the context of the election campaign sentiment study, election result prediction and some other election-oriented tasks. This literature review explores the existing works done in the election data analysis tasks.

Pradyumna et al[1] are engaged in experimental research with the objective of predicting the outcomes of the 2024 Indian General Elections, specifically the Lok Sabha Election featuring 545 Lok Sabha Seats. The research employs Data Analysis to parametrize computations related to coalitions and swings across all seats. Swing parameters are calculated using diverse machine learning techniques, including Linear Regression, Naive Bayes, Random Forest, and Time Series, utilizing prior election data specific to relevant constituencies. In addition to quantitative analysis, the researchers aim to understand the emotional context surrounding the elections. They plan to leverage a substantial corpus of recent publications and Twitter tweets to gauge sentiments. By applying swings to the vote shares of each political party in every constituency, the research intends to generate projections with the necessary biases derived from subjective data. The comprehensive analysis involves consideration of various parameters to gain insights into voter mindsets, including their needs, reasons for potential switches, satisfaction with current candidates, and whether their conditions are met. The research seeks to produce accurate projections by thoroughly analyzing data and capturing the diverse factors influencing voter decisions.

Krykun[2] introduces a novel approach for detecting election fraud. The proposed method relies on calculating the ratio of two standard normal random variables, followed by estimating parameters from the obtained sample. A crucial step involves comparing these estimates with the known theoretical values of parameters. The paper provides an illustrative example of the application of this method, demonstrating its practical utility in identifying potential instances of election fraud.

Chakraborty et. al[3] explores the structure and dynamics of a tweet-reply network centered around state assembly elections, specifically focusing on the context of India. The data, generated by Twitter users across the country over a 6-week period, is analyzed to understand the flow of Twitter activity during the West Bengal assembly elections. The study includes the identification of hashtags used by the three primary political contenders, enabling the recognition of cluster-level dominance in the Twitter network over the specified timeframe. Remarkably, the paper finds that this cluster dominance information correlates with the actual election outcomes, suggesting its potential as an effective forecasting tool. Additionally, the collected tweets are leveraged for lexicon-based emotion detection and subsequent analysis, adding a nuanced layer to the understanding of public sentiment during the electoral process.

In Ankita et. al[4] proposed work, Twitter serves as a valuable source of opinionated data, with the collection of tweets facilitated through Twitter APIs. The research utilizes the programming language R for tasks such as data acquisition, pre-processing, and analysis of the gathered tweets. The focus then shifts to sentiment analysis, employing various approaches to gauge public opinion. The timeframe for tweet collection spans from January 2019 to March 2019, aligning with the period leading up to the general elections in India. The study centers around two candidates, referred to as Candidate-1 and Candidate-2. The sentiment analysis results indicate that Candidate-1 is more favorably regarded and enjoys greater popularity compared to Candidate-2. Importantly, the paper concludes that these findings are consistent with the actual election results obtained in May 2019, showcasing the potential of Twitter sentiment analysis as a predictive tool for assessing public sentiments towards political candidates.

Widodo et. al[5] proposed an election data analysis using the Indonesia election data. The authors developed an algorithm and methodology to process significant data, identify top words, train the model, and predict sentiment polarity. Using the R language, experimental results indicate that Jokowi is leading in the current election prediction. Importantly, this prediction aligns with findings from four survey institutes in Indonesia, affirming the reliability of the proposed methodology.

Vasudevan et. al[6]., seeks to introduce an alternative approach to the current election system by emphasizing inherent loopholes. The existing electoral process primarily relies on the number of votes garnered by a candidate as the singular parameter for determining the winner. The proposed alternative aims to underscore the disparity between the candidate deemed most deserving to win an

election and the candidate projected to win based solely on popularity. The proposal advocates for a comprehensive evaluation of candidates, considering factors such as criminal records, educational qualifications, past social work, and previous term records. This multi-dimensional assessment aims to shift the focus away from a popularity contest, striving for a more nuanced and unbiased selection of candidates. The overarching goal is to enhance and refine the election process, promoting fairness and objectivity.

Tsai et. al[7]., introduces a machine learning-driven strategy for analyzing Twitter data to predict the outcomes of various local elections. The effectiveness of this strategy is validated by applying it to analyze Twitter data from the 2018 midterm elections in the United States. The findings indicate that the predicted results closely align with the actual election outcomes, suggesting the reliability and accuracy of the proposed machine learning approach in forecasting election results based on Twitter data.

Kellyton et.al[8] introduces SoMEN, a framework named Social Media for Election Nowcasting. SoMEN consists of a process and a machine learning (ML) model crafted for predicting election results in real-time. The framework utilizes social media (SM) performance metrics as input features, employing offline polls as labeled data. Furthermore, the paper delineates SoMEN-DC, an operational strategy for SoMEN that facilitates continuous predictions throughout the election campaign (DC). This innovative framework aims to improve the precision and immediacy of election result predictions by incorporating social media dynamics and leveraging real-time data.

Sangeetha et. al[9] presented the crucial aspect lies in selecting the appropriate methodology for election prediction. Numerous methodologies exist for prediction and sentiment analysis, but only a handful prove to be highly effective. Notably, Natural Language Processing (NLP), Machine Learning Techniques (MLT), and Decision Tree Learning (DTL) are among the prominent methodologies. However, the primary focus revolves around the widely utilized techniques of NLP and MLT for their efficiency in achieving accurate and insightful election predictions

Previous research in election analysis has primarily focused on utilizing classification algorithms to predict election outcomes by analyzing election data. However, there is a recognized need to identify the most effective approach for analyzing election campaign data. This research seeks to explore and establish a comprehensive pipeline for election campaign analysis, aiming to determine the most suitable methodology for this task..

IV. PROPOSED WORK

The proposed system employs the stacking, boosting and bagging approaches for the election campaign data classification and comparing the performance of the each boosting model in the context of election campaign data. Stacking and one of the boosting approaches which combines the multiple predictive models to improve the overall performance. Various the boosting improve the model performance by subsequently. Bagging . improves the performance of the model by reduce

variance within a noisy dataset. The proposed system holds immense significance in providing a technologically advanced and accessible tool for predicting election outcomes based on real-time sentiment analysis. By harnessing the power of AI, this system seeks to offer a nuanced understanding of public sentiment, contributing to a more informed perspective on election dynamics.

V. ALGORITHM USED

There are many traditional machine learning algorithms are playing their roles in different sector. It is crucial to identify which machine learning algorithm is works good in classifying the election campaign data. So that the popular algorithms namely Support vector machine, Gradient boosting, Logistic regression and KNN algorithms are used to train the model.

The generalized algorithm of the classification phase is;

i. Load the dataset
ii. Preprocess the data
iii. Split the data into training and testing data
iv. Initialize the classification algorithm
v. Fit the data into the classification algorithm
vi. Performance evaluation

VI. BLOCK DIAGRAM

Block diagram is the one which describes the entire system overall structure by block by block visually. The suggested system has five important process which is Collecting the election campaign data., pre processing the data, classification, performance evaluation and deployment which is illustrated in the figure 2

VII. IMPLEMENTATION

The proposed framework containing three important phases namely

i) Data collection and pre-processing
ii) Model development
iii) Analysis and deployment.

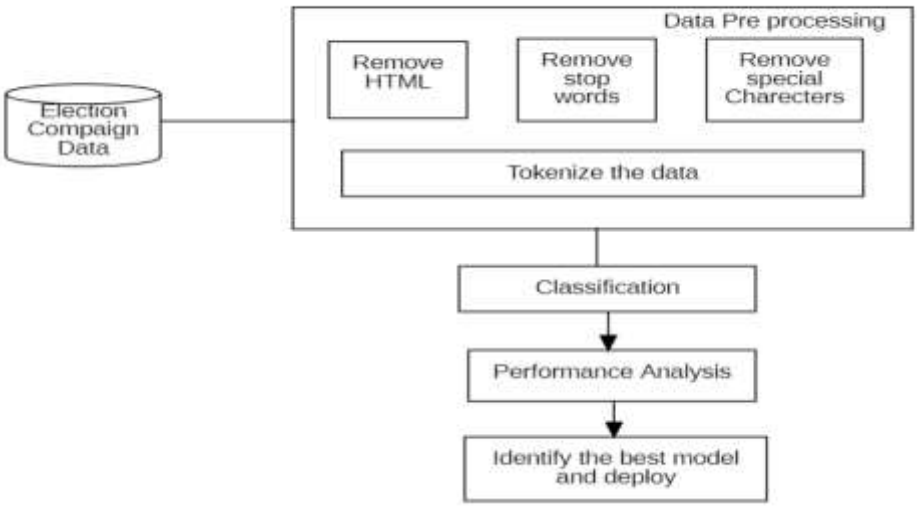

Fig 2. Block Diagram

Data Collection

For this project, we gathered data from the Kaggle data portal. The dataset includes election campaign tweets from two major political parties: PJP, with 49,477 tweets, and Congress, with 30,252 tweets. These datasets provide valuable insights into the online discussions surrounding these well-known political entities in Indian politics.

Data Preprocessing

Data preprocessing involves the essential task of formatting and tokenizing the text to transform it into a machine-readable format.

Labelling using TextBlob

The sentiment analysis script utili zes the TextBlob library to assess the senti ment of tweets. It begins with the importin g of the TextBlob class from the textblob m odule. The core of the analysis is found in t he analyze_sentiment function. This functi on accepts a tweet as input, conducts senti ment analysis using TextBlob, and then cat egorizes the sentiment as either positive, ne gative, or neutral.

Within the analyze_sentiment function, a T extBlob object called analysis is created for the input tweet. The sentiment property of t he TextBlob object is then utilized to obtai n the polarity score of the tweet. This polar ity score ranges from -1 (most negative) to 1 (most positive), with 0 indicating neutral sentiment.

Based on the polarity score obtained from T extBlob, the sentiment of the tweet is categ orized as follows: if the polarity is greater t han 0, the tweet is labeled as 'positive'; if th e polarity is less than 0, the tweet is labeled as 'negative'; and if the polarity is exactly 0 , the tweet is labeled as 'neutral'.

This sentiment analysis process is applied t o each tweet in the dataset using the apply f unction, resulting in the creation of a new c olumn named 'sentiment' in the DataFrame. This column contains the sentiment categor y ('positive', 'negative', or 'neutral') assigne d to each tweet based on its textual content. Overall, this approach offers a straightforw ard method for categorizing the sentiment e xpressed in tweets, though it's important to acknowledge that TextBlob's sentiment ana lysis is rule-based and may not capture all n uances of sentiment in natural language.

Remove HTML

The HTML and XML data forma ts are removed using the beutifulsoup pyth on libraray. Beautiful Soup (bs4) is a pyth on library primarily used to extract data fr om HTML, XML, and other markup lang uages. It’s one of the most used libraries f or Web Scraping.

Remove Special Characters and Conver t to Lowercase

The special characters are remov ed from the section by using the regular ex pression patterns. The text is converted to lowercase.

Removing Stop Words

Stop words are a set of commonl y used words in a language. Examples of s top words in English are “a,” “the,” “is,” “ are,” etc. Stop words are commonly used i n Text Mining and Natural Language Proc essing (NLP) to eliminate words that are s

o widely used that they carry very little us eful information.

Classification

After completing the preprocessing phase, the machine learning model is trained using a variety of algorithms, including Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, and Gradient Boosting. Once each model is trained with same data the best model identified based on the accuracies gained over the training and testing. Further each weak learners are used to construct the ensemble models such as stacking, boosting and bagging.

Stacking

Stacking involves training multiple base models, each analyzing sentiments from election-related social media posts using different machine learning algorithms like Linear SVC, K Nearest Neighbors Classifier, MLP Classifier , and Gradient Boosting Classifier. These base models independently predict whether the sentiments expressed in the posts are positive, negative or neutral. The predictions from these base models are then used as features to train a meta-learner, such as a Random Forest Classifier, which learns how to combine these predictions effectively. The figure 3 is illustrates the stacking of the proposed system.

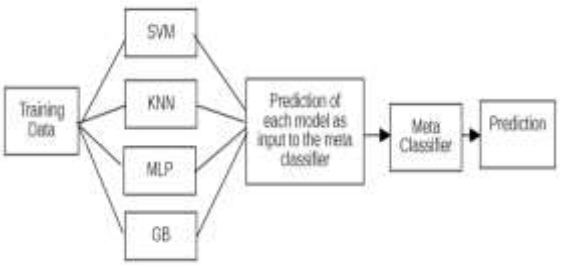

Fig 3. Stacking Diagram

Boosting

Boosting involves iteratively training weak learners (simple sentiment analysis models) to classify election-related social media posts. Each weak learner focuses on the instances that were misclassified by the previous ones, gradually improving the overall prediction accuracy. Boosting adapts to the complexities of sentiment expressions on social media by iteratively refining the model based on the mistakes made in earlier classifications.

Bagging

Bagging involves training multiple sentiment analysis models independently on different subsets of the election-related social media data, sampled with replacement (bootstrapping). Each model analyzes a subset of the data and predicts the sentiment of the posts. The predictions from these models are then aggregated, typically by averaging or voting, to make the final sentiment prediction for each post. Bagging helps reduce overfitting and variance in sentiment analysis by leveraging diverse models trained on different subsets of the social media data. By combining the predictions from multiple models, bagging provides a more robust and reliable sentiment analysis of election campaign expressions on social media.

Algorithm

1. Start
2. Read the raw data
3. Preprocess the data
   - Remove HTML
   - Remove special characters

- Remove stop words

4. Split the data into training and testing data
5. Vectorize the data using TfidfVectorizer
6. Train the machine learning classifiers using the same data
   - Support Vector Machine Classification

     LinearSVC()

     svr_lin.fit(features, target)
   - k-Nearest Neighbors Classification

     KNeighborsClassifier(n_neighbors=5)

     knn.fit(features, target)
   - MLP Classification

     lr = MLP()

     lr.fit(features, target)
   - Gradient Boosting Classification
7. Stacking the weak learners
8. Boosting the weak learners
9. Bagging the weak learners
10. Performance evaluation using the test data
11. Compare the accuracy of each algorithm
12. Extract the model which given good accuracy
13. Deploy the model

Performance Analysis

The performance evaluation has been done using the accuracy metric. Accuracy is defined as the ratio of correctly predicted examples by the total examples. Accuracy of the classification is calculated using

Accuracy = TP + TN / (TP + TN + FP + FN)

Where, TP – True positive, TN – True Negative, FP – False positive, FN – False Negative

For the given dataset the considered machine learning models were given different performance. The accuracy gained by each classifier are given in the table

Algorithm	Accuracy
SVM	76%
KNN	52%
MLP	84%
GB	81%
Stacking	88%
Boosting	86%
Bagging	86%

Table 1. Accuracy

From the tabulated performance gained by each algorithm., the stacking approach. By gaining the 88% of accuracy the stacking methodology proven that it is best approach in the context of sentimental analysis when compared with other ensemble or weak learners.

The next metric is called precision is used to evaluate the machine learning model which is the accuracy of positive prediction made by the model. Precision calculated as,

Precision = TP / (TP + FP)

Algorithm	Precision
SVM	81%
KNN	53%
MLP	71%
GB	81%
Stacking	88%
Boosting	87%
Bagging	86%

Table 2. Precision

From the table 2 the precision score of the the SVM, MLP and Stacking methodology given significant precision scores.

Recall is another method used to evaluate the model. Recall is a score of how many positive and negative samples are correctly identified. Recall calculated as,

$$Recall = TP / (TP + FN)$$

Algorithm	Recall
SVM	71%
KNN	50%
MLP	77%
GB	73%
Stacking	87%
Boosting	83%
Bagging	84%

Table 3. Recall

From the table 3 it comes to know that the stacking has a high recall score which means the stacking has a accurately detect the samples as it as.

Overall, the stacking ensemble method demonstrates strong capability in sentiment analysis for election campaigns as it achieves high accuracy, precision, and recall scores.

VIII. CONCLUSION

In conclusion, this project explores that how to predict election outcomes by employing artificial intelligence and real-time social media analysis. By focusing on sentiment analysis using advanced machine learning techniques, it is aim to offer a dynamic and insightful information's on public opinions during election campaigns. The comparison of classifiers enhances the accuracy of sentiment predictions. The results says that implementing the ensemble learning methodologies allows to get accurate prediction when comparing with the weak learners. The stacking methodology proven that it has good accuracy, precision and recall scores when compared with the weak learners. And the developed user-friendly web application for election campaign sentiment analysis makes this information accessible to a wider audience, providing real-time updates on election trends.

IX. FUTURE WORK

Looking in the future, the next work in advancing election prediction systems involves delving into multimedia analysis. In today's digital landscape, data isn't limited to just text; it encompasses images and videos, especially on social media. Therefore, a logical step for future work is to extend sentiment analysis to include these visual elements. This could entail understanding emotions conveyed in images and videos during election campaigns, extracting meaningful features from visual content, and integrating the analysis across various social media platforms. Moreover, exploring techniques that combine information from different forms of media can provide a more holistic grasp of public sentiment.

REFERENCES

[1] P. Parida, S. Sinha, A. P. Agrawal and R. Singh Yadav, "Predicting The General Election 2024 Using ML And Data Analytics," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-9, doi: 10.1109/INCET57972.2023.10170638.

[2] Krykun, Ivan. (2022). A new approach to Statistical analysis of election results.

[3] Chakraborty, A., Mukherjee, N. Analysis and mining of an election-based network using large-scale twitter data: a retrospective study. Soc. Netw. Anal. Min. 13, 74 (2023). https://doi.org/10.1007/s13278-023-01081-0

[4] Ankita Sharma, Udayan Ghose,Sentimental Analysis of Twitter Data with respect to General Elections in India, Procedia Computer Science,Volume 173,2020,Pages 325-334,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2020.

[5] Budiharto, W., Meiliana, M. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. Big Data 5, 51 (2018). https://doi.org/10.1186/s40537-018-0164

[6] Vasudevan,A.P.V, .Raghavendra, "Election Analysis and Pedagogy Forecast Using Big Data Analytics", 2017 IJEDR | Volume 5, Issue 4 | ISSN: 2321-9939

[7] M. -H. Tsai, Y. Wang, M. Kwak and N. Rigole, "A Machine Learning Based Strategy for Election Result Prediction," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2019, pp. 1408-1410, doi: 10.1109/CSCI49370.2019.00263.

[8] Kellyton Brito, Paulo Jorge Leitão Adeodato, Machine learning for predicting elections in Latin America based on social media engagement and polls, Government Information Quarterly, Volume 40, Issue 1, 2023. https://doi.org/10.1016/j.giq.2022.101782.

[9] Sangeeta Alagi, Mr. Tejas Kolambe, Mr. Vishal Bibe, Mr. Karan Gite, Mr. Sanket Chachar, "Survey on Election Prediction Using Machine Learning Technique"., Vol-9 Issue-2 2023.

[10] Ruben L. Bach, Christoph Kern, Ashley Amaya, "Predicting Voting Behavior Using Digital Trace Data", Social Science Computer Review2021, Vol. 39(5) 862-883.

[11] Zuloaga-Rotta, L.; Borja-Rosales, R.; Rodríguez Mallma, M.J.; Mauricio, D.; Maculan, N. Method to Forecast the Presidential Election Results Based on Simulation and Machine Learning. Computation 2024, 12, 38. https://doi.org/10.3390/computation12030038

[12] Abdul Ahad Abro, Mir Sajjad Hussain Talpur, Awais Khan Jumani, Waqas Ahmed Siddique, Erkan Yaşar, "Voting Combinations Based Ensemble: A Hybrid Approach", Celal Bayar University Journal of Science, Volume 18, Issue 3, 2022, p 257-263

[13] Kellyton Brito, Paulo Jorge Leitão Adeodato, Machine learning for predicting elections in Latin America based on social media engagement and polls, Government Information Quarterly,Volume 40, Issue 1, 2023.