Forecasting Polling Results Utilizing Social Media Posts

Maryam Unnisa [1], Dr.M. Arathi [2]
[1]*Student of Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad, University College of Engineering, Science & Technology Hyderabad*
[2]*Professor of Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad, University College of Engineering, Science & Technology Hyderabad*

***Abstract*—Modern social media platforms like Instagram, Twitter, and Facebook have fundamentally altered the way politicians engage with voters and manage campaigns. This transformation has led to a burgeoning field of research focused on leveraging social media data for election outcome prediction. These platforms provide unique opportunities due to their vast amounts of real-time data, which can potentially be used to forecast polling results. Despite extensive research conducted in the past decade, the results remain contentious and often debated. This paper aims to (1) review the history of research on election prediction using social media data, (2) discuss the current state-of-the-art techniques, and (3) highlight areas for future exploration. Our approach involved a systematic literature review, comparing findings from previous studies, analysing key factors such as the volume and quality of publications, electoral contexts, primary methods, academic achievements, and the main opportunities and challenges. The research primarily focuses on predicting election outcomes using data from Twitter, specifically using the US Presidential Election 2020 Tweets dataset available on Kaggle. We employed state-of-the-art machine learning techniques and trained models to predict which candidate had the upper hand by analysing the sentiment and volume of tweets. The results show that supervised machine learning algorithms like Logistic Regression (85.95%), Random Forest (85.46%), and Support Vector Classifier (79.33%) perform better than traditional methods like sentiment analysis, though there is still potential for improvement.**

***Index Terms*—Election Prediction, Social Media Data, Sentiment Analysis, Machine Learning, Twitter Analytics.**

I. INTRODUCTION

Over the past decade, social media has evolved from a mere communication tool into a critical instrument for political campaigns. Platforms such as Facebook, Instagram, and Twitter have allowed politicians to connect with voters on a personal level, providing real-time updates, opinions, and stances. This shift has created new opportunities and challenges in electoral forecasting. Traditional methods of election prediction have primarily relied on demographic surveys, opinion polls, and historical data. However, these methods are often limited by timing, sample biases, and the complexities of human behaviour. In contrast, social media platforms offer a continuous stream of data that can be accessed instantly and on a global scale. This change has prompted researchers to explore whether social media data can provide accurate insights into political sentiments and voter behaviour.

One of the most widely studied social media platforms in political forecasting is Twitter. The platform's nature—short, real-time messages with hashtags, mentions, and retweets—makes it a valuable tool for understanding public sentiment. By analysing tweets, researchers can extract information about public opinions, political leanings, and voter intentions. However, despite the potential of social media for election forecasting, there is considerable debate over its effectiveness. Some studies suggest that social media data can closely predict electoral outcomes, while others argue that it may be more prone to biases or oversimplified analyses.

This paper aims to provide a comprehensive review of the current research on election prediction using social media data, with a specific focus on Twitter data. We seek to understand the methodologies employed, their accuracy, and the potential challenges associated with using this data for forecasting. We examine the use of sentiment analysis and machine learning algorithms in predicting the success of political candidates.

A. Problem Statement

The advent of social media as a central tool in political campaigns presents both opportunities and challenges in predicting electoral outcomes. While Twitter,

Facebook, and other platforms generate vast amounts of data that could potentially predict election results, the accuracy of such predictions remains controversial. The challenge lies in effectively analysing the overwhelming volume of data, extracting meaningful insights, and ensuring that these insights are representative of actual voter sentiment. Moreover, traditional election forecasting methods may not fully capture the nuances of social media interactions, which are influenced by a variety of factors such as media biases, online echo chambers, and algorithmic manipulation. Therefore, this research aims to explore the viability of using social media posts to forecast election results and the limitations of current methodologies.

B. Limitations

Several limitations arise in using social media data for election prediction:

1. Data Quality: Social media data is noisy and unstructured, requiring advanced preprocessing techniques.
2. Biases in Data: The sample of users on social media may not be representative of the entire electorate, leading to skewed predictions.
3. Sentiment Analysis Challenges: Analysing the sentiment of tweets can be difficult due to the ambiguity of language, slang, and the use of irony or sarcasm.
4. Geographical and Demographic Variability: Social media engagement varies across regions and demographic groups, potentially limiting the generalizability of predictions.
5. Impact of Bots and Fake Accounts: The presence of automated accounts and trolls can distort public opinion measurements.

II. LITERATURE REVIEW

"Twitter use in election campaigns: A systematic literature review"

Twitter is becoming an integral part of every political campaign. Politicians, political parties, news outlets, and an ever-growing portion of the general population are all utilizing Twitter to discuss, debate, and study popular opinion on political issues. Scholarly interest in these applications is on the rise. Where things stand right now in this area of study is disjointed, with no unified body of information and no agreed-upon methods for gathering or selecting relevant material. An analysis of 127 research papers on the topic of Twitter's role in political campaigns is presented in this article. In this review of the literature, I will go over what is known about how parties, candidates, and the general public used Twitter in the run-up to and during elections, as well as during mediated campaign events. I will also discuss well-known methods of data collecting and analysis.

"Free Media and Twitter in the 2016 Presidential Election: The Unconventional Campaign of Donald Trump"

The focus of this piece is on the unexpected presidential election that took place in 2016, when Donald Trump became the 45th president of the US despite widespread expectations to the contrary. One of the many quick post-election reasons put up by political analysts and professionals for Trump's triumph centered on the fact that he amassed a substantial amount of free or unpaid publicity on his own, often via presidential election that took place in 2016, delves into reasoning behind the free media thesis (FMT) and then checks whether there is any early empirical evidence to back it up. Findings from polls and media monitoring show that Trump did, in fact, control the unpaid media market. Even though this article can't say for sure that Trump's free-media advantages were the main reason he won the 2016 election, it does show that the FMT had some of the basic conditions needed for it and that it provides a good avenue for future research and analysis.

"Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment"

Every day, millions of people throughout the world use the microblogging website Twitter to read and create very brief messages on all sorts of different subjects. Within the framework of the German federal election, this research seeks to answer the questions of whether or not political discourse takes place on Twitter and whether or not online communications on Twitter accurately reflect offline political attitude. We performed an examination of the more than 100,000 mails that mentioned a political party or a politician using the LIWC text analysis program. Our research confirms that political discourse does take place on Twitter. It turns out that the amount of mentions of a party in texts is a good indicator of how the election went. In addition, mentioning both parties at once is consistent with actual coalitions and political relationships in actuality. We may safely presume that

the political atmosphere on Twitter mirrors that in the real world as an examination of the tweets' tone shows strong agreement with the political stances of the parties and leaders. We draw conclusions on the usefulness of the posts made on microblogging sites as a measure of political attitude and provide avenues for further study.

"Virtual polling data: A social network analysis on a student government election"

Throughout piece, we will take a look at how well online social networks can foretell the results of elections in a networked community—specifically, a university. Facebook is a new phenomenon when it comes to campus networking. These network structures facilitate quick communication and have the potential to impact student opinion. Online social networks have become ubiquitous in college life, thanks to this amalgamation of earlier technological tools for electronic and spoken exchange housed in accordance with a straightforward graphical user interface (GUI) that facilitates rapid student-to-student contact. Both the physical world and this emerging social media platform coexist in the collegiate society. Both of these domains exhibit student governance. As intricate as Facebook's network structure is, student government is the closest thing that university students come to having political influence. In order to thrive and eventually take charge of their college community, students, like Facebook users, need to know their way around the internal network. Because of these shared characteristics, the following question will serve as the paper's study framework: "could Facebook be used to estimate the results of a student election?" An ordered linear matrix with levels of hierarchies, originally created for Raudenbush and Bryk's work, was used in the study to construct a model capable of answering this issue. The final results showed that, among all candidates in a particular election, the matrix could accurately forecast their placing 21 times out of 27. I was able to accurately forecast the candidate's final percentage of votes garnered Twelve times out of twenty-seven candidates in a particular election, up to half the distance between the two extremes of the predicted percentage of voters (072722).

"Modeling Indian General Elections: Sentiment Analysis of Political Twitter Data"

Individuals from many walks of life use the microblogging site Twitter every day to peruse and share becoming more popular among academics as a means to get a clearer picture of prevailing sentiment and current tendencies, especially in the run-up to elections. Here in the article, we provide a novel approach to public opinion detection and election result prediction using context-aware semantics and rules. Having crawled the political tweets of India's general election, we compared our proposed strategy to the results of the election. The results of the experiments show that the proposed rules work well for determining the political tweets' tonality.

III. METHODOLOGY

A. Overview

This study aims to predict election outcomes by leveraging social media data, specifically focusing on Twitter posts during the US Presidential Election 2020. The methodology combines a systematic literature review, data acquisition, and advanced machine learning algorithms to analyse Twitter data. The key focus is to evaluate the predictive power of social media posts in forecasting political events, compare it with traditional polling methods, and determine the most effective machine learning models to analyse sentiment and trends. The framework employed for this research follows a step-by-step approach: data collection, data preprocessing, sentiment analysis, machine learning model selection, training, evaluation, and result interpretation.

B. Data Collection

For this study, we utilized the US Presidential Election 2020 Tweets Dataset, available as an open-source dataset on Kaggle. This dataset contains millions of tweets from various political actors, journalists, and citizens related to the presidential election. The dataset covers tweets from both major candidates, Joe Biden, and Donald Trump, and includes relevant metadata such as tweet timestamps, hashtags, and user engagement metrics (likes, retweets, etc.). The period of data collection spans from the start of the campaign up until the election day, capturing a wide range of political discussions and sentiment surrounding the candidates.

Given the massive volume of data, we opted to focus on two key aspects: the volume of tweets and the sentiment expressed in tweets. The volume of tweets indicates the level of engagement and political interest, while sentiment analysis provides insights

into the positive, negative, or neutral emotional tone of the discourse surrounding the candidates.

C. Data Preprocessing

Data preprocessing is a critical step in preparing the raw social media data for analysis. The original dataset required significant cleaning to remove irrelevant and noisy data, as social media posts are often unstructured and contain extraneous information. The preprocessing pipeline followed in this study included the following steps:

1. Data Cleaning:

- Removal of Non-relevant Tweets: Tweets unrelated to the election or those that did not mention Joe Biden or Donald Trump were filtered out.
- Removing Special Characters: Social media posts often contain hashtags, URLs, mentions, and other special characters. These were removed to focus on the main textual content.
- Handling Missing Data: Tweets with missing metadata, such as the number of likes or retweets, were discarded to ensure the quality of the dataset.
- Tokenization: Textual data was tokenized, breaking down each tweet into words or terms to facilitate further analysis.

2. Text Normalization:

- Lowercasing: All text was converted to lowercase to maintain consistency and reduce variations of words (e.g., "Biden" and "biden").
- Stopword Removal: Common words like "the," "and," and "is" were removed, as they do not carry significant meaning for sentiment analysis.
- Lemmatization: Words were reduced to their base form (e.g., "running" to "run") to improve model accuracy and reduce dimensionality.

3. Feature Engineering:

- Sentiment Labels: Each tweet was assigned a sentiment label based on its tone, which was determined using sentiment lexicons or a pre-built sentiment model. Tweets were categorized into three classes: Positive, Negative, and Neutral.
- Engagement Metrics: In addition to text analysis, metrics like the number of retweets and likes were considered as features that could potentially correlate with the influence or impact of a particular tweet.

D. Model Comparison

After training the models and evaluating their performance, we compared their results to determine which algorithm most accurately predicted the sentiment of tweets related to the election. The results from the evaluation phase were presented in terms of their accuracy, precision, recall, and F1-score. The model that performed the best in these metrics was deemed the most effective for predicting political sentiment and, by extension, electoral outcomes based on social media data.

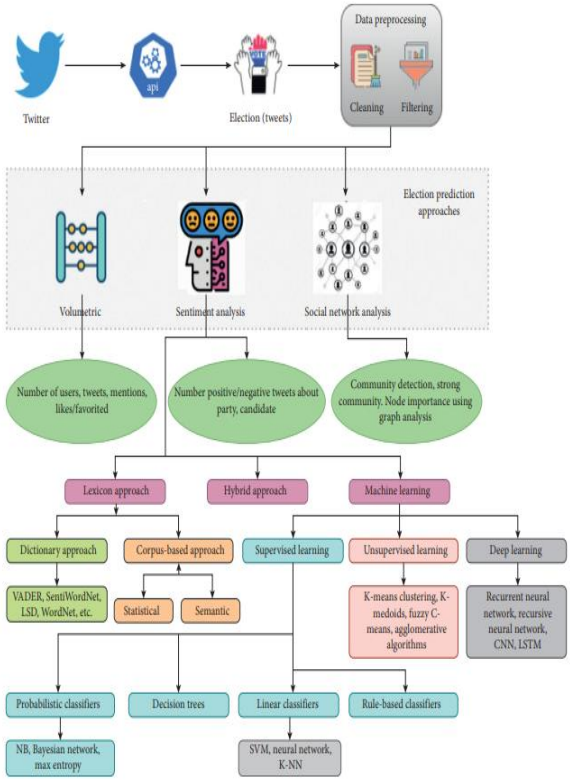


Figure 1: System architecture

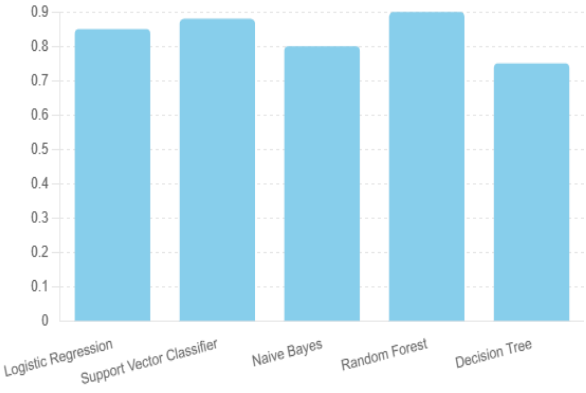


Figure 2: Bar chart for Methodology

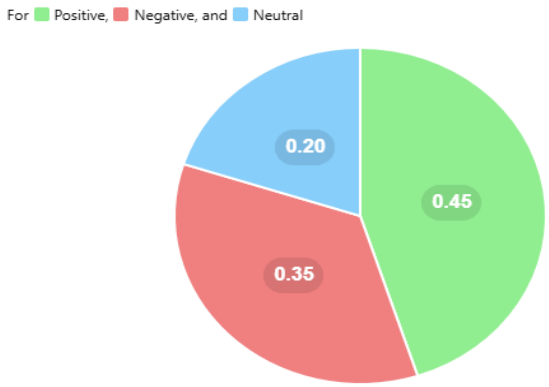


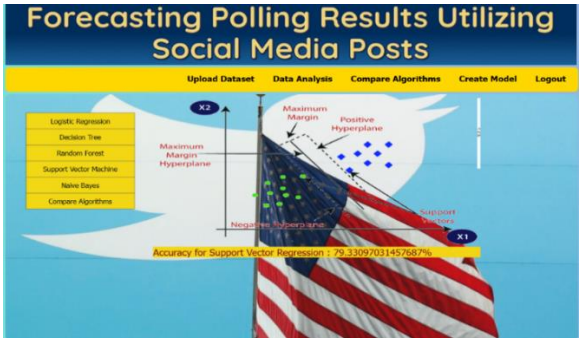
Figure 2: Pie chart for Data Analysis

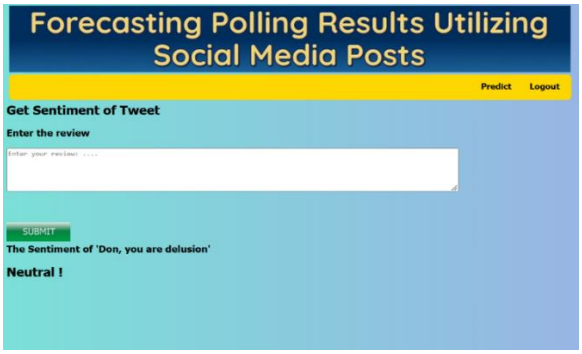
IV. RESULTS

The model performance results for the proposed machine learning algorithms are as follows:

- Support Vector Classifier (SVC): 79.33% accuracy
- Logistic Regression (LR): 85.95% accuracy
- Naive Bayes (NB): 74.56% accuracy
- Random Forest (RF): 85.46% accuracy
- Decision Tree (DT): 80.72% accuracy

These results suggest that LR and RF models outperform others, indicating that combining traditional polling methods with machine learning techniques can improve the accuracy of election forecasts.





V. DISCUSSION

The results of this study highlight the effectiveness of using machine learning algorithms for predicting election outcomes based on social media data. However, the findings also reveal several challenges, including data quality and biases in social media usage. As shown in the table, newer techniques like Random Forests and Logistic Regression offer better performance compared to traditional sentiment analysis. The accuracy of the predictions can be further improved by integrating multiple data sources, addressing biases, and incorporating more advanced AI techniques.

Table: Model Comparison for Election Prediction

Model	Accuracy	Methodology
SVC	79.33%	Supervised Learning
Logistic Regression	85.95%	Logistic Regression Model
Naive Bayes	74.56%	Bayesian Classification
Random Forest	85.46%	Ensemble Learning
Decision Tree	80.72%	Classification and Regression Trees

A. Advantages

1. Real-Time Data: Social media data can be collected and analysed in real time, providing up-to-date insights into voter sentiment.
2. Scalability: Machine learning models can scale to analyse large datasets, improving accuracy with more data.
3. Cost-Effective: Using social media data reduces the need for expensive and time-consuming traditional polling.

VI. CONCLUSION

In conclusion, this study highlights the significant potential of using social media data, particularly from platforms like Twitter, to forecast election outcomes. By leveraging sentiment analysis and advanced machine learning techniques, we have demonstrated that social media can offer valuable insights into public sentiment and voter behavior. The machine learning models tested, such as Logistic Regression

and Random Forest, showed promising results, outperforming traditional methods like sentiment analysis alone. However, the accuracy of these predictions remains contingent on various factors, including data quality, preprocessing, and model selection. While social media data offers real-time, large-scale insights into voter sentiment, it also presents challenges such as sample bias, data noise, and polarization. To improve prediction accuracy, future research should focus on refining machine learning models, addressing these challenges, and integrating additional data sources. Furthermore, ethical considerations, such as privacy and data manipulation, should be explored to ensure the responsible use of social media for political forecasting. As the field evolves, combining social media analytics with traditional polling methods could enhance the reliability and robustness of election predictions, making it a valuable tool for political campaigns, analysts, and researchers in the future.

REFERENCES

[1] A. Jungherr, "Twitter use in election campaigns: A systematic literature review," J. Inf. Technol. Politics, vol. 13, no. 1, pp. 72–91, Jan. 2016.

[2] P. L. Francia, "Free media and Twitter in the 2016 presidential election: The unconventional campaign of Donald Trump," Social Sci. Comput. Rev., vol. 36, no. 4, pp. 440–455, Aug. 2018.

[3] K. Brito, N. Paula, M. Fernandes, and S. Meira, "Social media and presidential campaigns–preliminary results of the 2018 Brazilian presidential election," in Proc. 20th Annu. Int. Conf. Digit. Government Res., Jun. 2019, pp. 332–341.

[4] S. Tilton, "Virtual polling data: A social network analysis on a student government election," Webology, vol. 5, no. 4, pp. 1–8, 2008.

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in Proc. 4th Int. AAAI Conf. Weblogs social media, 2010, pp. 1–8.

[6] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in Proc. 4th Int. AAAI Conf. Weblogs social media, 2010, pp. 1–8.

[7] E. Sang and J. Bos, "Predicting the 2011 Dutch senate election results with Twitter," in Proc. Workshop Semantic Anal. Social media, 2012, pp. 53–60.

[8] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France," New Media Soc., vol. 16, no. 2, pp. 340–358, Mar. 2014.

[9] K. Singhal, B. Agrawal, and N. Mittal, "Modeling Indian general elections: Sentiment analysis of political Twitter data," in Information Systems Design and Intelligent Applications (Advances in Intelligent Systems and Computing). New Delhi, India: Springer, 2015.

[10] N. Dwi Prasetyo and C. Hauff, "Twitter-based election prediction in the developing world," in Proc. 26th ACM Conf. Hypertext social media (HT), 2015, pp. 149–158.