Optimizing Election Forecasts: Integrating Ensemble Intelligence In Data Analytics

Submitted in partial fulfillment of the requirements for the degree of

Master of Science

In

Data Science

By

Ayushi Dey

22MDT0076

Under the guidance of:

Dr. Sri Rama Vara Prasad Bhuvanagiri

School of Advanced Sciences

VIT, Vellore.

May, 2024.

DECLARATION

I hereby declare that the thesis entitled "**Optimizing Election Forecasts: Integrating Ensemble Intelligence in Data Analytics**" submitted by me, for the award of the degree of *Master of Science* in *Data Science* to VIT is a record of bonafide work carried out by me under the supervision of **Dr. Sri Rama Vara Prasad Bhuvanagiri**.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date: 07.05.2024

Signature of Student

Ayushi Dey

CERTIFICATE

This is to certify that the thesis entitled "**Optimizing Election Forecasts: Integrating Ensemble Intelligence in Data Analytics**" submitted by **Ayushi Dey (22MDT0076)**, **School of Advanced Sciences**, VIT, for the award of the degree of *Master of Science* in *Data Science*, is a record of bonafide work carried out by her under my supervision during the period, **03.01.2024** to **07.05.2024**, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Vellore

Date: 07.05.2024

Signature of Guide
Department of Mathematics
SAS, VIT-Vellore

Signature of External Examiner

Head, Department of Mathematics
Dr. JAGADEESH KUMAR M.S.
SAS, VIT-Vellore

ACKNOWLEDGEMENT

With immense pleasure and deep sense of gratitude, I wish to express my sincere thanks to my guide, **Dr. Sri Rama Vara Prasad Bhuvanagiri** (Internal), **School of Advanced Sciences**, VIT, Vellore. Without his constant motivation and continuous encouragement, this research would not have been successfully completed.

I am grateful to the Chancellor of VIT, Vellore, **Dr. G. Viswanathan**, the Vice Presidents and the Vice Chancellor for motivating me to carryout research in the Vellore Institute of Technology, Vellore and for providing me with infrastructural facilities and many other resources that were needed for my research.

I express my sincere thanks to **Dr. Arunai Nambi Raj N**, Dean, School of Advanced Sciences, VIT, Vellore, for her kind words of support and encouragement. I would like to acknowledge the support rendered by my classmates in several ways throughout my research work.

I wish to thank **Dr. Jagadeesh Kumar M.S.**, Head of the Department of Mathematics, School of Advanced Sciences, VIT, Vellore for his encouragement and support.

I wish to extend my profound sense of gratitude to my parents and friends for all the support they extended during my research and also providing me with encouragement whenever required.

Signature of Student

Ayushi Dey

ABSTRACT

The modern political landscape is dynamic, with major countries significantly influencing the global economy through their policies. With elections looming in key nations like Canada, the United States, and India, understanding public sentiment towards political parties is crucial. This study focuses on predicting election outcomes by analyzing public sentiment towards political parties using tweets, specifically data related to Indian elections. The methodology involves collecting and organizing tweets, preprocessing it for feature extraction, and performing sentiment analysis to assess public opinion. Support Vector Machine, Logistic Regression and Naïve Bayes models have been trained on sentiment-labeled data for prediction, with ensemble techniques applied to improve accuracy. The study aims to identify trends in public sentiment towards political parties and evaluate the effectiveness of machine learning algorithms in predicting election results. Furthermore, the research aims to develop a machine learning framework to enhance election result prediction using sentiment analysis. The findings are expected to advance political forecasting and sentiment analysis, providing valuable insights for policymakers and analysts into the complex relationship between public opinion and election outcomes, using machine learning and social media data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
ANOVA	Analysis Of Variance
BJP	Bharatiya Janata Party
EDA	Exploratory Data Analysis
GBDT	Gradient-Boosted Decision Trees
GBM	Gradient Boosting Machine
HTML	HyperText Markup Language
MAE	Mean Absolute Error
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OvA	One-vs-All
OvR	One-vs-Rest
RFE	Recursive Feature Elimination
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
UK	United Kingdom
US	United States
VADER	Valence Aware Dictionary for Sentiment Reasoning

1. INTRODUCTION

1.1 OBJECTIVE

The project aims to improve the accuracy of election forecasting by going beyond traditional methods. While conventional approaches often rely on limited data sources like polls and surveys, this study leverages social media data, particularly tweets related to Indian elections, to gain a more comprehensive understanding of public sentiment. By analyzing this vast dataset using advanced machine learning models such as Logistic Regression, Support Vector Machine, and Naïve Bayes, along with ensemble techniques, the project seeks to improve the accuracy of election predictions.

The creation of an efficient machine learning framework for social media sentiment analysis-based election outcome prediction is another important goal of the research. In order to develop a model that can accurately forecast election results, the study will examine developments in public opinion toward political parties. This methodology will help forecast elections more accurately while also giving analysts and policymakers insightful information that will help them make better judgments.

Additionally, the project aims to deepen our understanding of the link between public sentiment on social media and actual election results. By exploring this relationship, the study seeks to uncover insights into how public opinion shapes political outcomes. This understanding could have significant implications for political forecasting and sentiment analysis methodologies, ultimately benefiting policymakers, analysts, and the public alike.

Overall, the project represents a significant step towards advancing our understanding of the complex relationship between public sentiment, social media, and election outcomes. By improving the accuracy of election forecasting, creating a robust machine learning framework, and gaining deeper insights into the link between public sentiment and election results, the study aims to contribute valuable knowledge to the field of political science.

1.2 MOTIVATION

The motivation behind this study stems from the dynamic nature of the modern political landscape, where major countries significantly influence the global economy through their policies. There is a growing need to understand public sentiment towards political parties and this understanding is crucial for policymakers, analysts, and stakeholders to make informed decisions.

Traditional methods of election forecasting often rely on limited data sources and may not capture the nuances of public opinion. By leveraging social media data, particularly Twitter, which provides real-time, unfiltered insights into public sentiment, this study seeks to improve the accuracy of election forecasting.

Furthermore, the study seeks to explore the link between public sentiment on social media and actual election results. Understanding this link can provide valuable insights into how public opinion shapes political outcomes, leading to more informed decision-making in the political arena.

Overall, the motivation behind this study is to advance our understanding of the complex relationship between public sentiment, social media, and election outcomes. By doing so, the study aims to contribute to more accurate and insightful election forecasting, ultimately benefiting policymakers, analysts, and the public alike.

1.3 BACKGROUND

The contemporary political arena is marked by its dynamic nature, where major nations wield substantial influence over the global economy through their policies. Understanding the public's sentiment towards political parties is crucial for predicting election results and guiding policy decisions.

Traditional methods of election prediction often rely on limited data sources like polls and surveys, which may not fully capture the breadth of public opinion. However, the emergence of social media, particularly platforms like Twitter, has provided a vast amount of real-time, unfiltered data that can offer valuable insights into public sentiment.

Twitter, with its expansive user base and the capacity to swiftly disseminate information, has become an invaluable asset for researchers and analysts examining public opinion. Through the analysis of tweets concerning political parties and election-related issues, researchers can assess real-time public sentiment and uncover trends and patterns that could impact election outcomes.

1.4 LITERATURE REVIEW

[1] Yavari et al. proposed a method to predict events using Twitter messages, focusing on changes in tweet volume over time. They preprocess and categorize tweets using techniques like non-negative matrix factorization and distance-dependent Chinese Restaurant Process incremental clustering. The categorized tweets are then analyzed for sudden increases or decreases in volume, which are used to predict the occurrence of specific events. The study shows the

effectiveness of their approach in predicting various events, such as political protests, natural disasters, and financial market fluctuations. The authors stress the importance of real-time analysis and temporal dynamics in capturing the evolving nature of events on social media. Their research highlights the potential of clustering and categorization techniques in extracting meaningful insights from large volumes of tweets. Yavari et al.'s work adds to the growing research on using social media for event prediction, offering a flexible and scalable framework adaptable to different events and data sources. The study underscores the value of temporal analysis and clustering techniques in extracting meaningful signals from social media data.

[2] Sharma and Moh (2016) explores the potential of sentiment analysis on Hindi tweets to predict election outcomes in India. The work builds upon the growing field of sentiment analysis, which uses computational techniques to extract emotional sentiment from text data. It highlights the increasing use of social media like Twitter as a valuable source of real-time public opinion, particularly relevant for understanding political sentiment. The research addresses the challenge of sentiment analysis in Hindi, a language with unique linguistic features and limited resources compared to English. It acknowledges existing work on Hindi sentiment analysis but emphasizes the need for further development and adaptation for capturing political nuances. The development of a domain-specific lexicon tailored to Indian politics is a valuable contribution to this area. The authors collected 42,235 Hindi tweets referencing five national political parties during the 2016 Indian general election campaign period. They employed both supervised and unsupervised approaches, utilizing sentiment analysis techniques like Dictionary-Based, Naive Bayes, and SVM algorithms. The analysis focused on identifying the sentiment of Twitter users towards each political party. All three techniques (Dictionary-Based, Naive Bayes, and SVM) identified BJP (Bhartiya Janta Party) as the party with the most positive sentiment on Twitter. Notably, the SVM model accurately predicted a 78.4% chance of BJP winning the most elections, aligning with the actual outcome where BJP secured 60 out of 126 constituencies, significantly outperforming other parties.

[3] Budiharto, et al. (2018) explores the use of sentiment analysis on tweets to predict the outcome of the 2019 Indonesian Presidential election. The authors propose a framework to collect, analyze, and classify Twitter opinions in order to predict the election results. They gathered tweets from the candidates' accounts and relevant hashtags, then used sentiment analysis to determine the positive, negative, or neutral sentiment of each tweet. Their findings suggest that Jokowi, the incumbent president, had a more positive sentiment than his challenger,

Prabowo. This aligns with the results of four survey institutes in Indonesia, suggesting that the method the authors used may be reliable. The study contributes to the field of social media analysis by demonstrating the potential of tweets for predicting election outcomes. It also highlights the importance of sentiment analysis as a tool for understanding public opinion. However, the study is limited by its focus on a single election and its use of a relatively small sample of tweets. Future research could explore the use of sentiment analysis on tweets to predict the outcome of elections in other countries and to study the impact of social media on public opinion more broadly.

[4] Alvi et al. (2023) reviews the use of tweets and sentiment analysis for election prediction. Sentiment analysis, a technique for gauging public opinion from text, is applied to social media data to forecast election outcomes. The authors examine the strengths and weaknesses of current methods, highlighting the potential for bias in tweets compared to traditional polling methods. They conclude that sentiment analysis on tweets can be a promising tool, but further research is needed to improve its accuracy for election prediction.

[5] Bansal and Srivastava (2018) explore the potential of Twitter sentiment analysis for predicting election outcomes. They propose a hybrid topic-based approach that analyzes both the sentiment and thematic content of tweets. This approach aims to address limitations of sentiment analysis alone, which might miss nuances in public opinion. The authors acknowledge the growing use of social media data for gauging public opinion in elections. Sentiment analysis of tweets offers a real-time and cost-effective method for election monitoring. However, they recognize that sentiment analysis alone might be insufficient due to the complexity of political discourse on social media. Their solution is a hybrid approach that combines sentiment analysis with topic modeling. Topic modeling helps identify underlying themes within the tweets, providing a more comprehensive understanding of public discussions. This combined analysis offers a potentially more accurate prediction of election outcomes.

[6] Coletto et al. (2015) explore a machine learning approach to focus on the potential of using social media data, specifically Twitter, for electoral predictions. A well-established approach in political science is analyzing public opinion through polls and surveys. However, social media offers a vast amount of real-time data that might provide new insights. Twitter, with its focus on short messages and public nature, presents a valuable resource for understanding public sentiment. Coletto et al.'s study (2015) is one example of leveraging this potential. By analyzing the content of tweets, they aim to develop a machine learning model

that can predict election outcomes. This approach highlights the growing interest in utilizing social media data for various research areas, including political science.

[7] Unankard et al. (2014) explore the potential of social media analysis for predicting election outcomes. Their approach leverages sub-event detection and sentiment analysis to capture public opinion on social media platforms. Sub-event detection helps identify specific moments within the election cycle, such as debates or policy announcements that might generate significant online discussions. Sentiment analysis then extracts the emotional tone (positive, negative, or neutral) surrounding these events. This combined approach offers a more nuanced understanding of public opinion compared to traditional sentiment analysis of the entire election period.

[8] In the realm of public opinion analysis, O'Connor et al. (2010) explored the connection between social media sentiment and public opinion time series. Their study, presented at the International Conference on Weblogs and Social Media (ICWSM), investigated the potential of using text sentiment extracted from tweets to understand fluctuations in public opinion. This research lays the groundwork for utilizing social media data as a tool for gauging public sentiment and its correlation with traditional methods like opinion polls.

[9] Tumasjan et al. (2010) examines Twitter's role in political discourse during the German federal election, investigating if Twitter reflects offline political sentiment. Using LIWC text analysis software, over 100,000 tweets mentioning political parties or politicians were analyzed. Results indicate that Twitter is used for political deliberation, with tweet volume mirroring election results and mentions of party alliances aligning with real-world coalitions. Sentiment analysis suggests tweets reflect the political positions of parties and politicians. The study suggests micro-blogging content can be a valid indicator of political sentiment, offering insights for future research.

[10] Maurya et al. (2021) aims to predict the probability of winning for political parties using both labeled and unlabeled data. Labeled data can be obtained through polling methods, but this may not provide high accuracy. Therefore, it's essential to use live data for more accurate election result predictions. Twitter, a microblogging site, allows users to post quick, real-time updates. By utilizing hashtags, we can easily generate and utilize the necessary data. We used the Python library "Tweepy" to access the Twitter API and fetch live data. We collected 350 tweets for each political party using keywords. Using the

"TextBlob" library in Python, sentiments were applied to each tweet. Based on the number of positive tweets for a particular party, the winning party is declared. Additionally, popular text classification algorithms like Naive Bayes, SVM, and Random Forest were used to train the model using labeled data. The accuracy of the predicted result is calculated and declared. Finally, the result is represented in the form of a bar graph for labeled data, showing the number of votes for each political party, and for unlabeled data, a pie chart is used to represent positive, negative, and neutral sentiments for each political party.

[11] Vendeville et al. (2021) introduces a new method for forecasting election results, distinct from the prevalent approach of using tweets for sentiment analysis due to concerns over its reliability. Instead, the authors base their model on a theoretical analysis of the voter model with stubbornness on strongly-connected graphs, building on their previous work. They use this model to predict the popular vote percentages for the Conservative and Labour parties in the UK, as well as the Republican and Democratic parties in the US, relying solely on official past election results. The authors' method achieves a Mean Absolute Error (MAE) of 4.74%, with errors ranging from 0.04% to 14%. However, this MAE is considered high for election prediction models, as previous works aimed for MAEs below 1 or 2%. For comparison, a simple baseline method that predicts the exact result of the previous election yielded an average error of 5.03%. The authors note that the choice to discard the first few election results was subjective, based on their observation of the model's behavior. Despite the relatively high MAE, the authors view their approach as a novel direction that provides insights into the political landscape. It solely relies on official data and also estimates the proportion of stubborn voters, those who always or never vote for a specific party, which adds depth to the analysis. To improve accuracy, the authors suggest incorporating in-between election polls and studying past election trends, such as landslide victories and incumbency reelection. They also propose combining their method with tweets-based estimations to enhance accuracy further. While their current method falls short of the desired MAE, the authors believe it offers a valuable starting point for future research in election forecasting.

[12] Rao et al. examined the use of Twitter in predicting election outcomes in India. They collected tweets from major political parties during the 2022 General State election and analyzed sentiment using VADER sentiment analyzer and machine learning. Their Random Forest model achieved 77.59% accuracy in predicting party popularity. The study suggests this approach could aid political parties in campaign strategy and provide long-term sentiment analysis for analysts.

[13] Singh et al. employ social media analytics to extract relevant information from tweets during the 2017 Punjab assembly elections in India. In addition to proposing a novel technique for seat forecasting—which predicts the number of seats a political party is likely to win in the elections—the authors used machine learning algorithms for polarity analysis, which aims to ascertain the sentiment of tweets. The study's findings suggested that the Indian National Congress would probably win, and when the election results were announced, that is exactly what happened. This study adds to the increasing corpus of research that uses social media data for political forecasting and analysis. By applying social media analytics techniques and machine learning algorithms to tweets, the study contributes to a deeper understanding of how social media can be leveraged to predict election outcomes. The findings highlight the potential of using big data from social media platforms for political analysis and forecasting, demonstrating the relevance and impact of social media in contemporary politics.

3. PROJECT DESCRIPTION & GOALS

PROJECT DESCRIPTION

This project aims to predict election outcomes in key nations like Canada, the United States, and India by analyzing public sentiment towards political parties using tweets. The study focuses on leveraging tweets related to Indian elections to improve the accuracy of election forecasting compared to traditional methods.

The process involves collecting and organizing tweets related to Indian elections, preprocessing it for feature extraction, and performing sentiment analysis to assess public opinion towards political parties. Support Vector Machine, Logistic Regression and Naïve Bayes models will be trained on sentiment-labeled data and will be used for prediction. Ensemble techniques will be applied to improve prediction accuracy.

The study will explore the link between public sentiment on social media and actual election results, aiming to provide deeper insights into the complex relationship between public opinion, social media, and election outcomes. By analyzing trends in public sentiment towards political parties, the study seeks to offer valuable insights for policymakers and analysts.

By integrating social media data into the study, the project will advance sentiment analysis and political forecasting techniques. Decision-makers and political stakeholders will gain from this more timely and nuanced understanding of public sentiment.

The overall goal of this research project is to advance knowledge of the intricate connections between social media, public opinion, and election results. This will help analysts and politicians make more informed decisions and increase the precision of election forecasts.

PROJECT GOALS

- Enhance Election Forecasting Accuracy: Improve the accuracy of election forecasting beyond traditional methods by leveraging social media data, particularly Twitter, related to Indian elections.
- Develop a Machine Learning Framework: Build a solid machine learning framework to use sentiment analysis in social media to forecast election outcomes. Support vector machine, logistic regression, and Naïve Bayes models will all be used in this framework, along with ensemble methods, to increase prediction accuracy.

- Explore the Link Between Public Sentiment and Election Outcomes: Gain deeper insights into the relationship between public sentiment on social media and actual election results. By analyzing trends in public sentiment towards political parties, the study aims to provide valuable insights for policymakers and analysts.
- Advance Political Forecasting and Sentiment Analysis: Contribute to the advancement of political forecasting and sentiment analysis methodologies by incorporating social media data into the analysis. This will provide a more nuanced and timely understanding of public opinion.
- Provide Insights for Decision-Makers: Offer valuable insights for policymakers, analysts, and stakeholders to make informed decisions based on the analysis of public sentiment towards political parties.
- Contribute to Academic and Research Communities: Contribute to academic and research communities by providing a detailed analysis of the complex relationship between public sentiment, social media, and election outcomes.

4. TECHNICAL SPECIFICATIONS

4.1 HARDWARE SPECIFICATIONS

- Processor : Intel® Core™ i3-6100U CPU
- Hard Disk : 400GB
- Memory : 4GB RAM

4.2 SOFTWARE SPECIFICATIONS

- Operating System : Windows 10
- IDE : Anaconda installed with Jupyter
- Frontend : Python
- Backend : NLTK, Scikit-learn and TextBlob installed

5. DESIGN APPROACH & DETAILS

5.1 MATERIALS, APPROACH & METHODS

MATERIALS:

- Python: Python is widely used for data analysis due to its simplicity and powerful libraries like Pandas, NumPy, and Matplotlib. These libraries offer tools for data manipulation, numerical computations, and visualization, making Python a go-to choice for tasks such as cleaning, processing, and analyzing data to extract valuable insights.
  - NumPy: NumPy is a vital Python package for scientific computing. Large arrays and matrices can be created and manipulated with its help, and a variety of mathematical functions are available to facilitate work with these data structures. NumPy was used in the project to carry out mathematical operations.
  - Pandas: NumPy's capabilities are expanded by the powerful Python data manipulation and analysis module Pandas. It provides flexible data structures that are perfect for managing structured data effectively, such as DataFrames and Series. Pandas is well-liked among data professionals because it makes activities like data cleansing, transformation, and analysis simpler. Pandas was used in the project to import and alter datasets.
  - NLTK: Natural Language Toolkit, or NLTK is an essential Python package for creating applications that work with data related to human language. It provides tokenization, parsing, categorization, stemming, tagging, and semantic reasoning, among other text processing functions. Because of NLTK's versatility, which allows it to be used for everything from simple research projects to complex programming exercises, it is a priceless tool for Python text data manipulation.
  - TextBlob: TextBlob is a Python toolkit for working with text and has an easy-to-use API for common natural language processing (NLP) operations. It provides features including sentiment analysis, noun phrase extraction, tokenization, part-of-speech tagging, classification, translation, and more. It was used in the project to preprocess the text data.
  - Matplotlib: Matplotlib makes it possible to create interactive, animated, and static visualizations. Plot types supported by it include scatter plots, bar charts, histograms, and line graphs, among others. Matplotlib was employed in the project to facilitate data visualization.

- Seaborn: Built upon Matplotlib, Seaborn is a Python framework that provides a high-level interface for creating statistical visualizations that are both aesthetically pleasing and educational. It requires little coding and makes it easier to create complex visualizations such as pair plots, violin plots, heatmaps, and others. For the aim of data visualization, Seaborn was used in the project.
- Scikit-learn: Scikit-learn, also referred to as sklearn, is a popular Python machine learning library. It provides a number of tools for various tasks such model selection, dimensionality reduction, clustering, regression, and classification. Machine learning models were developed throughout the project using sklearn.

- Twitter: Twitter is a popular social networking site where users may interact and exchange brief messages known as tweets. Users can interact with tweets by like, retweeting, and replying to other accounts, and they can follow accounts to read their tweets in a chronological timeline. Furthermore, Twitter provides academics, analysts, and marketers with useful data sources by providing insightful information about news events, trends, and public opinion.
  - User ID: A user ID on Twitter is a unique identifier assigned to each user account, allowing Twitter to distinguish between individual users. User IDs are used internally by Twitter to manage accounts and track user activity but are not typically displayed publicly.
  - Tweets: Tweets are brief messages, limited to 280 characters, that users post on Twitter to express thoughts, share information, news, and opinions. These messages often include hashtags to categorize them and make them easier for other users to discover.
  - Favourites & Retweets: Favourites on Twitter are a way for users to show appreciation or agreement with a tweet without retweeting it. Retweets, on the other hand, involve sharing another user's tweet with one's own followers, often to spread information or endorse a message.
- Original Dataset: The dataset consists of tweets gathered before the 2019 General Elections in India, centering on the prominent political parties BJP and Congress. These tweets were collected from December 2018 to March 2019. The dataset consists of the following:
  - User ID
  - Original Tweet
  - Favourite Count
  - Retweet Count

APPROACH:

- Flow Diagram

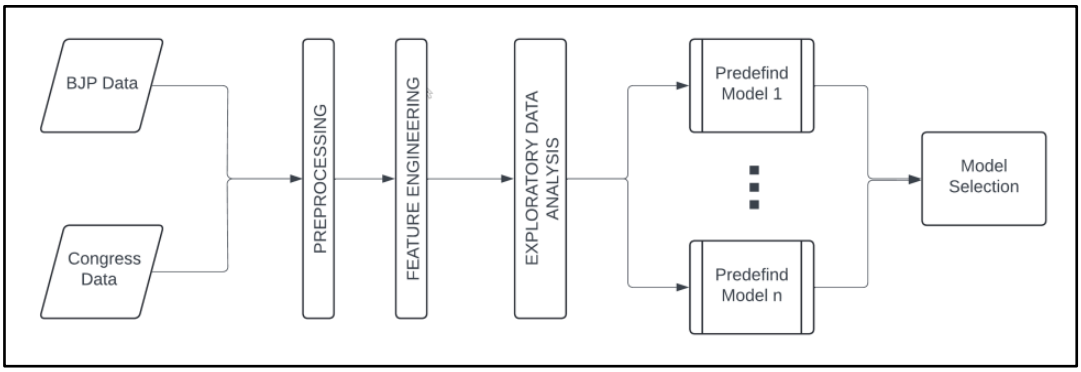


Fig. 1a: Flow Diagram 1

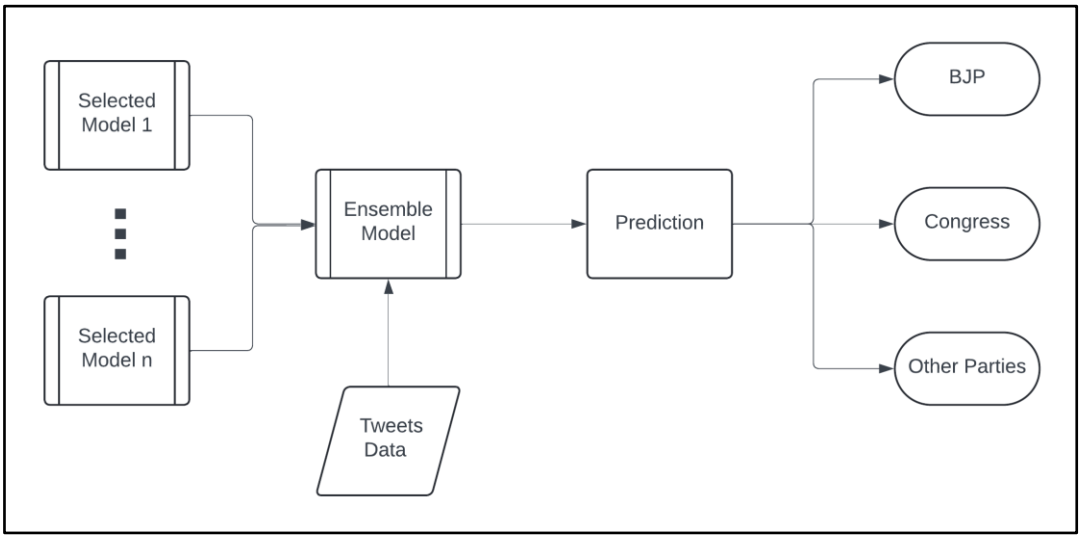


Fig. 1b: Flow Diagram 2

- Architecture Diagram

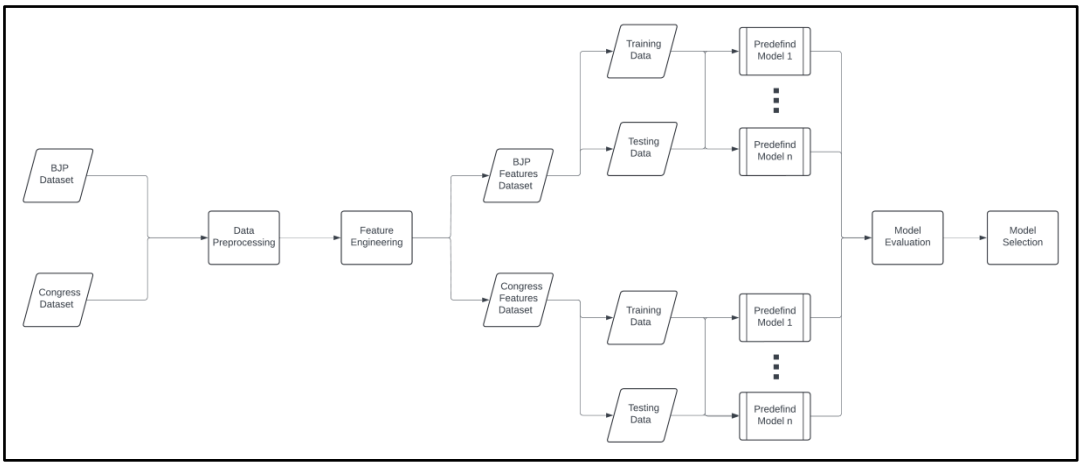


Fig. 2a: Architecture Diagram 1

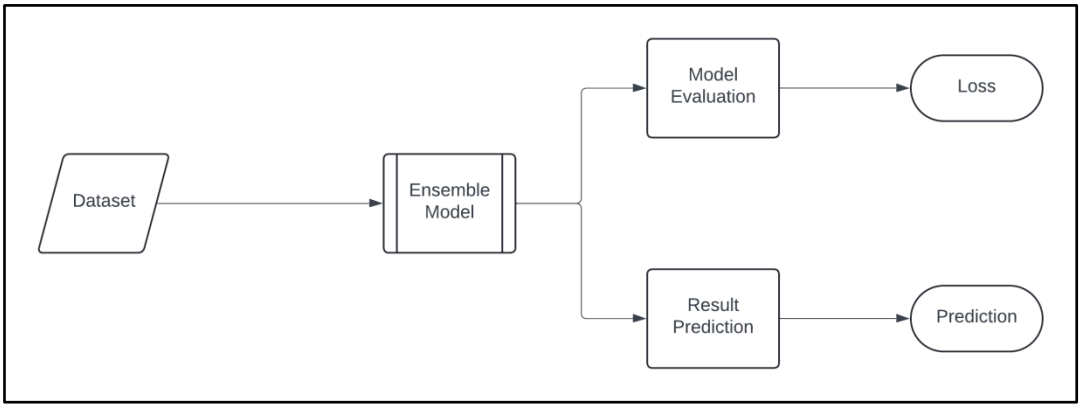


Fig. 2b: Architecture Diagram 2

- Flow Chart For Election Prediction

  The proposed model architecture for election prediction has been developed with the hypothesis that negative emotions towards one political party might be perceived as positive emotions towards another. As a result, the final ensemble model combines two individual models, one for each political party. This project aims to create a generalized approach to election prediction, regardless of the location, type of election, or number of political parties involved.

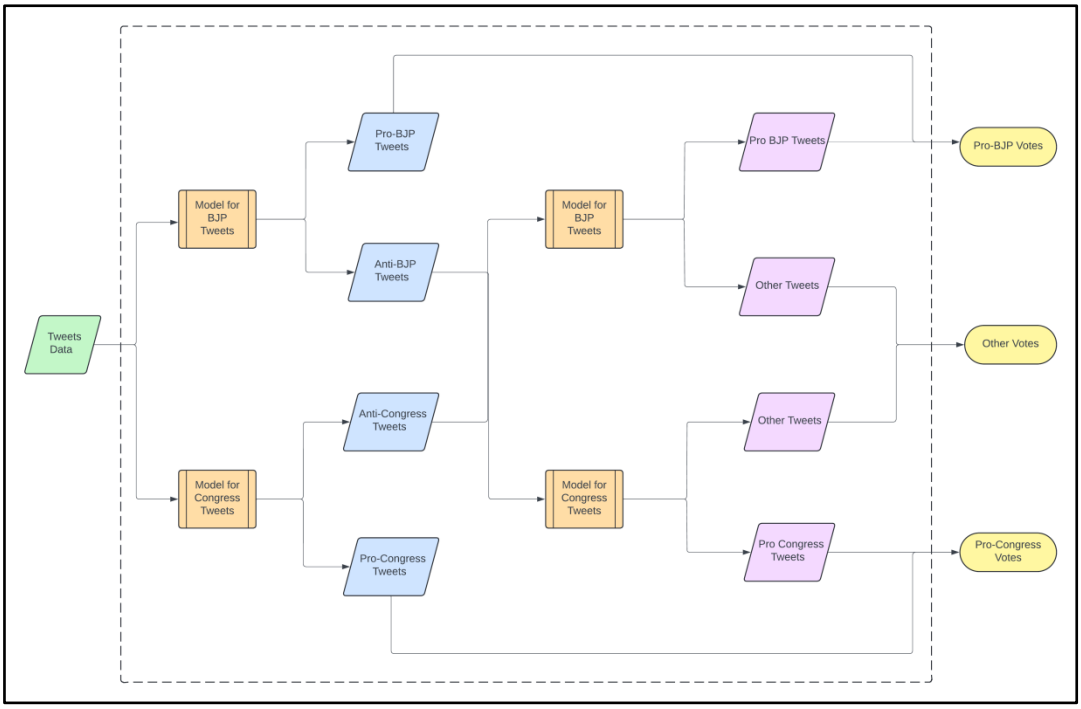


Fig. 3: Flow Chart for Election Prediction

METHODS

- Data Preprocessing: Data preprocessing is a crucial step in data analysis and machine learning, involving cleaning, transforming, and preparing raw data to make it suitable for further analysis. It includes tasks like handling missing values, encoding categorical variables, scaling features, and removing outliers, ensuring data quality and model performance.
  - Data Cleaning: Data cleaning in NLP involves removing noise, such as HTML tags and punctuation, and handling issues like spelling errors and tokenization. Stopwords and rare words may be removed, and text may be normalized through stemming or lemmatization. These processes enhance the quality of text data for analysis and modeling in NLP tasks.
  - Polarity: Polarity in sentiment analysis refers to the emotional tone of a text, indicating whether the expressed sentiment is positive, negative, or neutral. It helps gauge the overall sentiment of a piece.
  - Subjectivity: Subjectivity in text analysis refers to the extent to which a statement is influenced by personal opinions, feelings, or beliefs rather than factual information. It indicates the degree of bias or neutrality in a text. A greater value suggests a higher degree of personal opinion or bias in the text and a lower value indicates otherwise.
  - Data Labeling: Data labeling in natural language processing (NLP) involves annotating text to indicate the sentiment or opinion expressed. This process often requires human annotators to categorize text as positive, negative, or neutral. Supervised learning models use these labeled datasets to learn patterns and make predictions about the sentiment or opinion of unseen text. For this project, data labeling has been performed considering both polarity and subjectivity. Labeling tweets with both polarity and subjectivity allows for a more nuanced understanding of sentiment, improving the model's ability to capture subtle variations in opinion. This approach enhances prediction accuracy by considering the context and strength of sentiment expressed in tweets, leading to more robust election outcome predictions.

The data has been labeled (categorized as 'opinion' in the dataset) as follows:

Polarity	Subjectivity	Label
≥ 0.6	> 0.6	2
≥ 0.2	> 0.1	1
≥ -0.2	≤ 0.1	0
> -6	> 0.1	-1
≤ -6	> 0.6	-2

Table 1: Label Assignment

- Data Balancing: Data balancing in machine learning is a method used to tackle imbalances in class distributions within a dataset. It includes techniques like oversampling the minority class, undersampling the majority class, or utilizing synthetic data generation. Balancing data is crucial for enhancing the performance of machine learning models, particularly in classification tasks.
- Tokenization: Tokenization in natural language processing (NLP) refers to the process of dividing text into smaller units called tokens, which can be words, phrases, or symbols. This process is essential for text processing, as it allows for analysis at the word level. Tokenization plays a crucial role in various NLP tasks such as text classification, named entity recognition, and machine translation, as it provides a structured representation of text data.
- Lemmatization: Lemmatization in natural language processing (NLP) is the process of reducing words to their base or root form, known as a lemma. It employs vocabulary and morphological analysis to accurately identify the lemma of a word. Lemmatization enhances text analysis and understanding by treating different forms of a word as the same.

- Feature Engineering: Feature engineering involves creating new features from text data to improve machine learning model performance. This can include extracting n-grams, part-of-speech tags, sentiment scores, and other linguistic features. Techniques like TF-IDF, word embeddings, and syntactic parsing are also used to represent text in a way that is suitable for modeling.
  - TF-IDF Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is a method used to transform text documents into numerical forms. It assesses the significance of a word in a document compared to a corpus of documents. Words that are frequent in a document but infrequent in the corpus receive higher weights, aiding in capturing the distinctiveness of each document.

- Text Length: Text length in NLP refers to the number of words or characters in a piece of text, which can impact analysis.
- Part-of-Speech (POS) Tagging: Part-of-Speech (POS) tagging in natural language processing (NLP) involves assigning grammatical tags to words in a text according to their function in a sentence (e.g., noun, verb, adjective). This process aids in understanding sentence structure and is vital for various NLP tasks such as parsing, information extraction, and machine translation.

- Exploratory Data Analysis: Exploratory Data Analysis (EDA) involves examining and visualizing text data to understand its characteristics and patterns. This includes analyzing word frequencies, sentence lengths, and vocabulary usage. EDA helps identify data cleaning and preprocessing steps needed for effective model building and understanding the underlying structure of the text data.
  - Descriptive Statistics: Descriptive statistics in exploratory data analysis (EDA) for text data entail summarizing and describing important characteristics of the dataset. This includes metrics such as word frequencies, average sentence lengths, and vocabulary richness. Descriptive statistics aid in comprehending the fundamental properties of the text data, such as its distribution, central tendency, and variability. These statistics offer insights into the structure and content of the text, informing subsequent analysis and preprocessing steps in natural language processing (NLP) tasks.
  - Bar Plot: A bar plot is a visual representation of categorical data, showing the frequency or count of each category as bars. It helps in understanding the distribution of categorical variables.
  - Distribution Plot: A distribution plot visualizes the distribution of numerical data, showing the frequency or density of values along the range of the data. It helps in understanding data spread and skewness.
  - Pair Plot: A pair plot is a grid of scatterplots that display the relationship between pairs of variables in a dataset. It is useful for visualizing correlations and distributions among variables.
  - Correlation Matrix: A correlation matrix is a tabular representation displaying the correlation coefficients between pairs of variables in a dataset. It aids in identifying relationships and dependencies among variables.
  - Word Cloud: A word cloud is a graphical depiction of word frequencies in a text, with the size of each word indicating its frequency or significance. It assists in identifying key terms.
- Feature Selection: Feature selection involves choosing a subset of important features for model building, aiming to enhance model performance by reducing

the number of input variables and selecting the most informative features while minimizing overfitting. Techniques include:

- Filter Methods: These techniques choose features based on their statistical characteristics, like correlation, variance, or information gain. Examples include the Chi-square test, ANOVA, and correlation coefficient.
- Wrapper Methods: These approaches assess subsets of features using a predictive model and choose the subset that offers the best performance. Examples include Recursive Feature Elimination (RFE) and Forward Selection.
- Embedded Methods: These techniques integrate feature selection into the model construction process. Examples include Lasso regression and decision tree-based methods such as Random Forests, which inherently conduct feature selection.

For this project, Chi-square test and RFE have been used to conduct feature selection.

- Hyperparameter Tuning: Hyperparameter tuning involves finding the best hyperparameters for a machine learning model. Hyperparameters are set before the learning process, like the learning rate or the number of hidden units in a neural network. Tuning searches for the optimal hyperparameter combination using methods like grid search, random search, or Bayesian optimization to enhance model performance. In this project, random search was utilized for hyperparameter tuning.
- Logistic Regression: Logistic Regression is a statistical technique employed for classification tasks, particularly when the outcome variable is categorical. It estimates the probability of an input belonging to a specific category using a logistic function. For logistic regression models with multiple independent variables and multiple potential outcomes, multinomial logistic regression or softmax regression can be utilized. In this approach, the model estimates the probability of an observation belonging to each potential outcome using the softmax function. The softmax function is defined as:

$$P(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\beta_{2k}X_2+\cdots+\beta_{pk}X_p}}{\sum_{j=1}^{K} e^{\beta_{0j}+\beta_{1j}X_1+\beta_{2j}X_2+\cdots+\beta_{pj}X_p}}$$

Where,

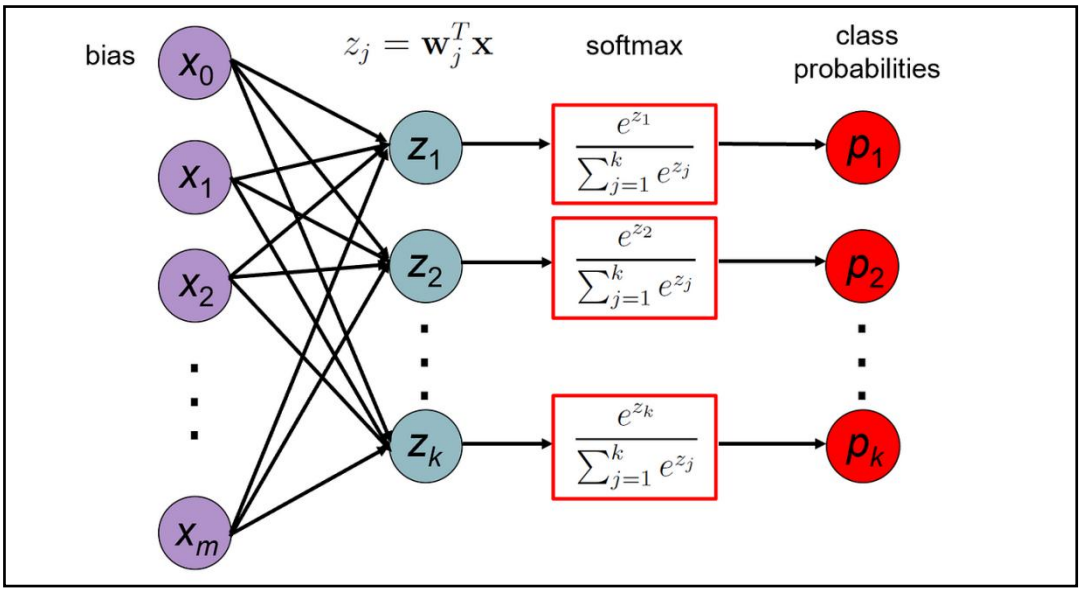
- $P(Y = k|X)$ is the probability that the outcome variable Y is category k given input X.
- K is the number of possible outcomes.
- $X_1, X_2, \dots, X_p$ are the independent variables.
- $\beta_{0k}, \beta_{1k}, \dots, \beta_{pk}$ are the coefficients for each independent variable for category k.

The model is trained to find the optimal values of β that minimize the logistic loss function:

$$L(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} 1(y_i = k)\, log\left(\frac{e^{\beta_k^T X_i}}{\sum_{j=1}^{K} e^{\beta_j^T X_i}}\right)$$

Where,

- $L(\beta)$ is the multinomial logistic loss function.
- N is the number of samples.
- y_i is the true class label for sample i.
- X_{ij} is the value of feature j for sample i.
- $1(y_i = k)$ is an indicator function that returns 1 if $y_i = k$ and 0 otherwise.

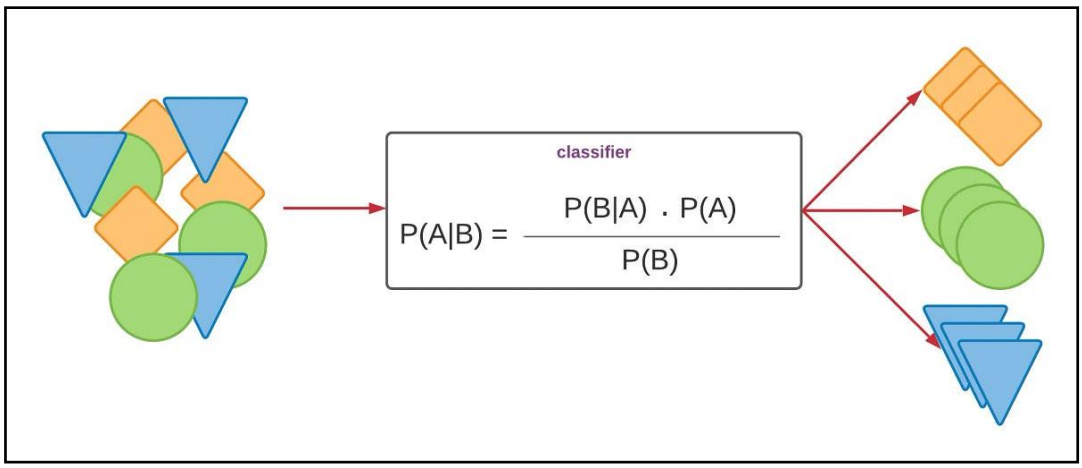


[27] Fig. 4: Multinomial Logistic Regression

- Naïve Bayes: Naive Bayes is a widely used probabilistic classification algorithm that relies on Bayes' theorem, assuming independence between features. It is highly effective for text classification applications such as spam filtering or sentiment analysis. Mathematically, Naive Bayes calculates the probability of a class given some features using Bayes' theorem:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y).P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)}$$

The "naive" assumption in Naive Bayes is that the presence of a particular feature in a class is independent of the presence of any other feature, given the class. This simplifies the calculation of $P(x_1, x_2, \dots, x_n)$ to $P(x_1|y).P(x_2|y) \dots . P(x_n|y)$.



[28] Fig. 5: Naïve Bayes Model

- Support Vector Machines: Support Vector Machines (SVM) is a supervised machine learning algorithm utilized for classification and regression. It identifies the hyperplane that most effectively separates distinct classes in the feature space. For multiclass classification, the One-vs-Rest (OvR) or One-vs-All (OvA) strategy is often employed with SVMs. This strategy entails training several binary classifiers, each dedicated to distinguishing one class from all others. The mathematical formulation of OvR SVM is as follows:
  Suppose we have K classes $\{1,2,\dots,K\}$. For each class i, we train a binary classifier $f_i(x)$ to distinguish class i from the rest. The decision function of each binary classifier is typically defined as:

$$f_i(x) = sign(w_i.x + b_i)$$

  Where,
  - w_i represents the weight vector
  - b_i is the bias term
  - x denotes the input feature vector

  During training, the parameters w_i and b_i are optimized to maximize the margin between the positive instances of class i and the negative instances of all other classes. Mathematically, for class i, the optimization problem can be formulated as:

$$min_{w_i,b_i} \frac{1}{2}\|w_i\|^2$$

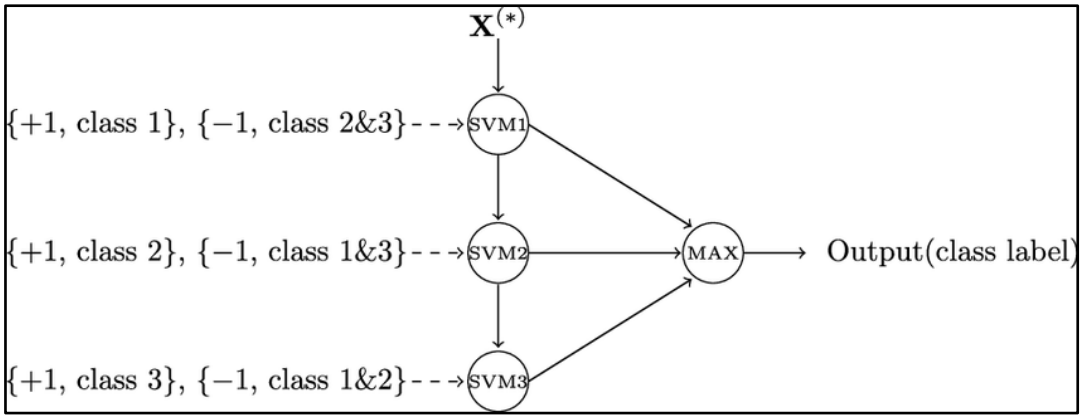
  Subject to:

$$y_j(w_i.x_j + b_i) \geq 1 \text{ for } j \neq i$$

  Where y_j is the class label (either +1 or -1) indicating whether x_j belongs to class j.
  During prediction, we compute the decision function for each binary classifier $f_i(x)$ and choose the class with the highest confidence:

$$\text{Predicted class} = \underset{i}{^{argmax}} \; f_i(x)$$

OvR SVM trains K binary classifiers, each specialized in distinguishing one class from the rest, making it a versatile approach for multiclass classification tasks.



[29] Fig. 6: Three Class One-vs.-All Support Vector Machine

- Bagging (Bootstrap Aggregating): Bagging, short for Bootstrap Aggregating, is an ensemble technique in machine learning used to enhance the accuracy and stability of models. It works by training multiple models, often of the same type, on different subsets of the training data and then combining their predictions. It works as follows:
  - Bootstrap Sampling: Multiple subsets of the training data are generated through random sampling with replacement, known as bootstrap sampling. Each subset is then utilized to train a distinct model.
  - Model Training: A base model, such as Logistic Regression or Support Vector Machine, is trained on each bootstrap sample. This process results in multiple models, each trained on slightly different datasets.
  - Voting or Averaging: In regression tasks, the predictions of the models are averaged to produce the final prediction. In classification tasks, the predictions may be combined using voting methods like majority voting to determine the final class label.

  Bagging helps reduce overfitting by reducing the variance of the model. It is particularly effective when using unstable models that are sensitive to the training data, such as decision trees. Random Forest, a popular ensemble method, uses bagging with decision trees to create a robust and accurate model.

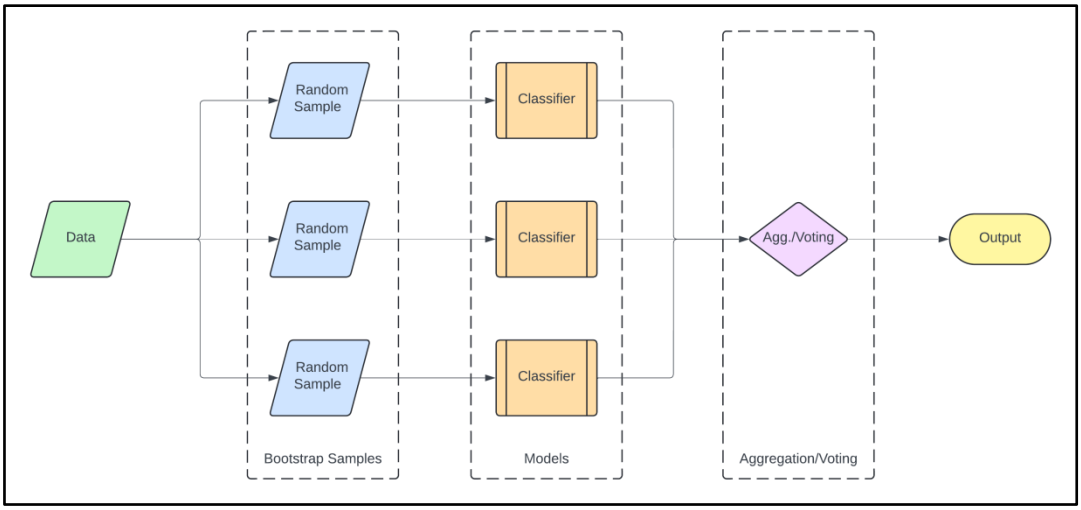


Fig. 7: Bagging Ensemble Method

- Boosting: Boosting is an ensemble technique in machine learning that seeks to enhance the overall predictive performance by combining the predictions of multiple base estimators, typically weak learners. The core concept of boosting involves training the base learners sequentially, with each new learner correcting the errors made by the preceding ones. Popular boosting algorithms include:
  - AdaBoost (Adaptive Boosting): In AdaBoost, each base learner pays more attention to the instances that the previous learners misclassified. It assigns weights to the training instances and adjusts them at each iteration to focus on difficult-to-classify instances.
  - Gradient Boosting: This refers to a general boosting algorithm in which a new base learner is trained to predict the residuals, which are the differences between the predicted value and the actual value, of the current ensemble's prediction. The final prediction is obtained by combining the predictions of all base learners.
    - XGBoost (Extreme Gradient Boosting): XGBoost is an optimized implementation of gradient boosting that prioritizes speed and performance. It incorporates regularization techniques to avoid overfitting and is capable of parallel processing.
    - LightGBM: LightGBM is a gradient boosting framework that introduces a unique method called Gradient-based One-Side Sampling (GOSS) to filter out data instances with small gradients. This technique helps to reduce training time and enhance efficiency.
    - CatBoost: CatBoost is another gradient boosting library that handles categorical features seamlessly by converting them into numerical values using various encoding techniques.

- Boosting Trees: Algorithms like GBM (Gradient Boosting Machine) and GBDT (Gradient Boosting Decision Tree) are implementations of boosting with decision trees as the base learners.

Boosting algorithms are powerful and widely used in various machine learning tasks, especially in classification and regression problems, due to their ability to improve model performance and handle complex relationships in the data.

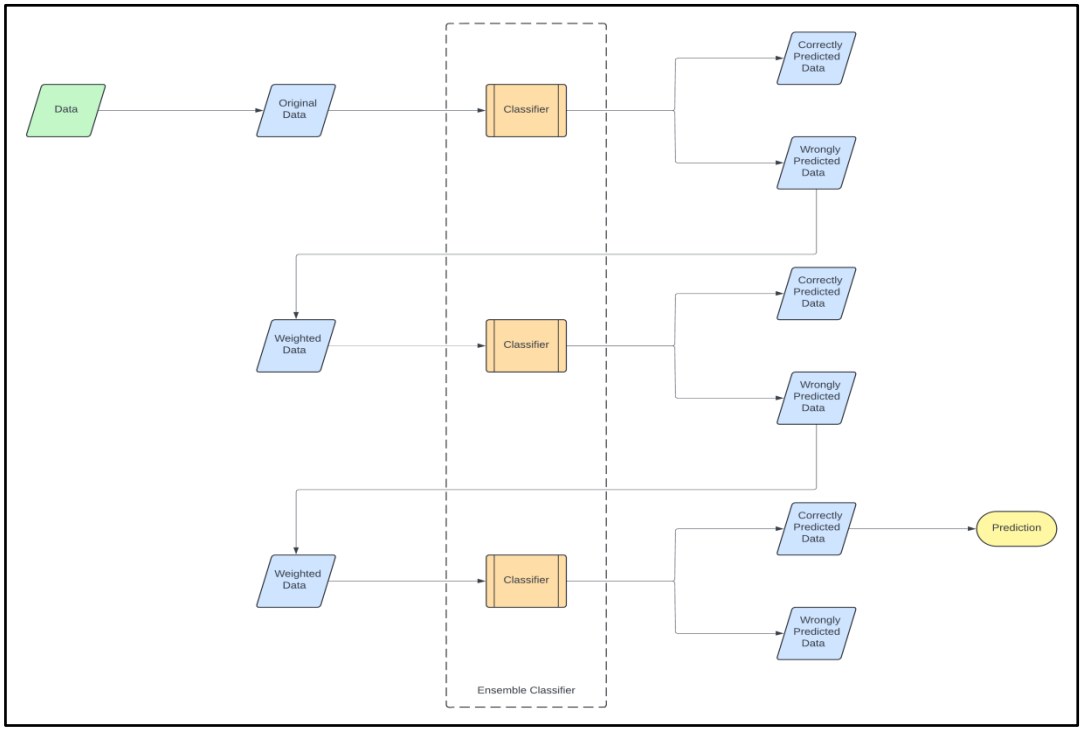


Fig. 8: Boosting Ensemble Method

- Performance Metrics: Performance metrics are utilized to assess the effectiveness of machine learning models. Some frequently used performance metrics are as follows:
  - Accuracy: Accuracy evaluates the ratio of correctly classified instances to the total instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

  - Precision: Precision assesses the ratio of true positive predictions to all instances predicted as positive.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

  - Recall: Recall evaluates the ratio of true positive predictions to all actual positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- F1-Score: The F1-score combines precision and recall into a single value, using the harmonic mean. This provides a balanced assessment of both measures.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6. Schedule, Tasks & Milestones

SCHEDULE

Sl. No.	Date	Task
1	03.01.2024 - 15.02.2024	Literature Review
2	16.02.2024 - 05.03.2024	Data Collection
3	06.03.2024 - 08.03.2024	Data Preprocessing
4	09.03.2024 - 12.03.2024	Feature Engineering
5	12.03.2024 - 16.03.2024	Exploratory Data Analysis
6	13.03.2024 - 30.04.2024	Model Building
7	01.05.2024 - 08.05.2024	Final Paper Write-up

Table 2: Schedule

TASKS & MILESTONES

Sl. No.	Task & Corresponding Milestone	Description
1	Data Collection	Collect tweets related to Indian elections, including tweets mentioning political parties and election-related topics.
2	Data Preprocessing	Clean and preprocess the collected data for feature extraction, including removing duplicates, handling missing values, and tokenizing text.
3	Feature Extraction	Extract features from the preprocessed data, such as sentiment scores, polarity, subjectivity, etc.
4	Sentiment Analysis	Perform sentiment analysis on the extracted features to gauge public opinion towards political parties.
5	Model Selection	Select appropriate machine learning models for predicting election results based on the sentiment-labeled data.

Sl. No.	Task & Corresponding Milestone	Description
6	Model Training	Train the selected models on the sentiment-labeled data to learn the patterns in public sentiment.
7	Ensemble Model	Plan & execute an ensemble model.
8	Evaluation	Use appropriate metrics to evaluate the efficacy of the trained models, such as F1-score, accuracy, recall, and precision.
9	Analysis of Results	Examine the outcomes to detect patterns in public sentiment towards political parties and evaluate the machine learning models' effectiveness in forecasting election results.
10	Insights Generation	Generate insights from the analysis to provide valuable information for policymakers, analysts, and stakeholders.
11	Documentation	Document the entire process, including data collection, preprocessing, model selection, training, evaluation, and analysis, for future reference and replication.
12	Report Writing	Write a comprehensive report summarizing the project aims, process, methods, results, and insights gained from the analysis.

Table 3: Tasks and Milestones

7. Project Outputs

DATA PREPROCESSING

- BJP tweets Dataset Before Preprocessing

id	original_text	favourite_count	retweet_count
advosushildixit	@anjanaomkashyap I am seeing you as future #bj...	67	63
UttarrPradesh	Which of the following should be top priority ...	84	56
ShaileshWrites	@RenukaJain6\nI still remember your video mass...	74	62
iamljp	#bjp @BJP4India @INCIndia @INCKarnataka how mu...	68	56
TheShobhitAzad	#AzadPrediction\n#LokSabhaElections2019 \n\nBJ...	73	67

Table 4: BJP tweets Dataset Before Preprocessing

- BJP tweets Dataset After Preprocessing

id	original_text	favourite_count	retweet_count	clean_text	sentiment	polarity	subjectivity	opinion	tokens	lemmas	lemmatized_text
advosushildixit	@anjanaomkashyap I am seeing you as future #bj...	67	63	@ anjanaomkashyap i seeing future # bjp spokes...	Sentiment(polarity=0.35, subjectivity=0.362500...	0.35	0.3625	1	[@, anjanaomkashyap, seeing, future, #, bjp, s...	[@, anjanaomkashyap, seeing, future, #, bjp, s...	@ anjanaomkashyap seeing future # bjp spokespe...
UttarrPradesh	Which of the following should be top priority ...	84	56	which following top priority modi government #...	Sentiment(polarity=0.25, subjectivity=0.3)	0.25	0.3	0	[following, top, priority, modi, government, #...	[following, top, priority, modi, government, #...	following top priority modi government # loksa...
ShaileshWrites	@RenukaJain6\nI still remember your video mass...	74	62	@ renukajain6 i still remember video massage	Sentiment(polarity=-0.5, subjectivity=0.425)	-0.5	0.425	-1	[@, renukajain6, still, remember, video, massa...	[@, renukajain6, still, remember, video, massa...	@ renukajain6 still remember video massage . d...
iamljp	#bjp @BJP4India @INCIndia @INCKarnataka how mu...	68	56	# bjp @ bjp4india @ incindia @ inckarnataka mu...	Sentiment(polarity=-0.3428571428571429, subjec...	-0.342857	0.757143	-1	[#, bjp, @, bjp4india, @, incindia, @, inckarn...	[#, bjp, @, bjp4india, @, incindia, @, inckarn...	# bjp @ bjp4india @ incindia @ inckarnataka mu...
TheShobhitAzad	#AzadPrediction\n#LokSabhaElections2019 \n\nBJ...	73	67	# azadprediction # loksabhaelections2019 bjp :-...	Sentiment(polarity=0.0, subjectivity=0.0)	0.0	0.0	-2	[#, azadprediction, #, loksabhaelections2019, ...	[#, azadprediction, #, loksabhaelections2019, ...	# azadprediction # loksabhaelections2019 bjp :-...

Table 5: BJP tweets Dataset After Preprocessing

- Congress tweets Dataset Before Preprocessing

id	original_text	favourite_count	retweet_count
Sunnysweet16	Wonder why no academic or journalist asks INC ...	22	11
drnitinchaube	Congrats for the change #australiavotes2019 an...	16	12
mrvivek07	Peopel Say “Govt Ne 70 Years Kya kiya”.\nUnse ...	6	4
JosephPravinP	@ajaymaken @RahulGandhi And as a final touch, ...	13	7
VandanaMegastar	#LokSabhaElections2019 Anyone not having mass ...	24	17

Table 6: Congress tweets Dataset Before Preprocessing

- Congress tweets Dataset After Preprocessing

id	original_text	favourite_count	retweet_count	clean_text	sentiment	polarity	subjectivity	opinion	tokens	lemmas	lemmatized_text
Sunnysweet16	Wonder why no academic or journalist asks INC ...	22	11	wonder academic journalist asks inc india rahu...	Sentiment(polarity=0.125, subjectivity=0.36904...	0.125	0.369048	0	[wonder, academic, journalist, asks, inc, indi...	[wonder, academic, journalist, asks, inc, indi...	wonder academic journalist asks inc india rahu...
drnitinchaube	Congrats for the change #australiavotes2019 an...	16	12	congrats change # australiavotes2019 # scottmo...	Sentiment(polarity=0.0, subjectivity=0.0)	0.0	0.0	-2	[congrats, change, #, australiavotes2019, #, s...	[congrats, change, #, australiavotes2019, #, s...	congrats change # australiavotes2019 # scottmo...
mrvivek07	Peopel Say "Govt Ne 70 Years Kya kiya",\nUnse ...	6	4	peopel say govt ne 70 years kya kiya , unse pu...	Sentiment(polarity=0.0, subjectivity=0.0)	0.0	0.0	-2	[peopel, say, govt, ne, 70, years, kya, kiya, ...	[peopel, say, govt, ne, 70, year, kya, kiya, ...	peopel say govt ne 70 year kya kiya , unse puc...
JosephPravinP	@ajaymaken @RahulGandhi And as a final touch, ...	13	7	@ ajaymaken @ rahulgandhi and final touch , # ...	Sentiment(polarity=0.32, subjectivity=0.565)	0.32	0.565	1	[@, ajaymaken, @, rahulgandhi, final, touch, ,...	[@, ajaymaken, @, rahulgandhi, final, touch, ,...	@ ajaymaken @ rahulgandhi final touch , # modi...
VandanaMegastar	#LokSabhaElections2019 Anyone not having mass ...	24	17	# loksabhaelections2019 anyone mass backing ca...	Sentiment(polarity=-0.16666666666666666, subje...	-0.166667	0.383333	0	[#, loksabhaelections2019, anyone, mass, backi...	[#, loksabhaelections2019, anyone, mass, backi...	# loksabhaelections2019 anyone mass backing ca...

Table 7: Congress tweets Dataset After Preprocessing

EXPLORATORY DATA ANALYSIS

- Descriptive Statistics of Features of BJP Related tweets

	favourite_count	retweet_count	subjectivity	polarity	opinion	text_length	nouns	verbs	adjectives	adverbs
count	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000
mean	58.114867	16.831800	0.435628	0.093230	-0.315533	182.270267	12.893600	2.916267	5.015400	1.136467
std	33.281517	15.562979	0.299447	0.337420	1.277268	73.996675	7.667572	1.962697	2.519812	1.149637
min	14.000000	1.000000	0.000000	-1.000000	-2.000000	22.000000	0.000000	0.000000	0.000000	0.000000
25%	34.000000	7.000000	0.200000	-0.071429	-2.000000	130.750000	9.000000	1.000000	3.000000	0.000000
50%	51.000000	13.000000	0.454226	0.000000	0.000000	185.000000	12.000000	3.000000	5.000000	1.000000
75%	70.000000	22.000000	0.650000	0.275000	1.000000	225.000000	16.000000	4.000000	7.000000	2.000000
max	182.000000	91.000000	1.000000	1.000000	2.000000	991.000000	110.000000	15.000000	22.000000	19.000000

Table 8: Descriptive Statistics of Features of BJP Related tweets

- Additional Statistics of Features of BJP Related tweets

	favourite_count	retweet_count	subjectivity	polarity	opinion	text_length	nouns	verbs	adjectives	adverbs
median	52.000000	13.000000	0.450000	0.000000	0.000000	185.000000	12.000000	3.000000	5.000000	1.000000
var	1058.928642	210.631179	0.087803	0.110738	1.606192	5183.734443	54.409460	3.829723	6.297404	1.315419
skew	1.410573	2.017794	0.028480	0.317258	0.040455	2.451700	4.729438	0.776656	0.596682	1.172307
kurt	2.020465	4.909882	-0.916303	0.476577	-1.096788	20.633844	45.397268	0.838843	0.405392	1.578861
25th percentile	35.000000	7.000000	0.207143	-0.071429	-2.000000	130.000000	9.000000	1.000000	3.000000	0.000000
75th percentile	70.000000	22.000000	0.641667	0.266667	1.000000	224.000000	16.000000	4.000000	7.000000	2.000000

Table 9: Additional Statistics of Features of BJP Related tweets

- Descriptive Statistics of Features of Congress Related tweets

	favourite_count	retweet_count	subjectivity	polarity	opinion	text_length	nouns	verbs	adjectives	adverbs
count	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.00000	15000.000000
mean	60.538467	14.978333	0.375848	0.079709	-0.568400	178.913467	12.991733	2.850867	4.60940	0.976667
std	40.490262	11.317656	0.310025	0.309488	1.291033	68.510350	6.924704	1.910725	2.47158	1.088149
min	1.000000	1.000000	0.000000	-1.000000	-2.000000	24.000000	1.000000	0.000000	0.00000	0.000000
25%	21.000000	6.000000	0.000000	0.000000	-2.000000	127.000000	9.000000	1.000000	3.00000	0.000000
50%	57.000000	12.000000	0.400000	0.000000	0.000000	180.000000	12.000000	3.000000	4.00000	1.000000
75%	88.000000	24.000000	0.600000	0.250000	1.000000	224.000000	16.000000	4.000000	6.00000	2.000000
max	160.000000	49.000000	1.000000	1.000000	2.000000	950.000000	110.000000	15.000000	18.00000	10.000000

Table 10: Descriptive Statistics of Features of Congress Related tweets

- Additional Statistics of Features of Congress Related tweets

	favourite_count	retweet_count	subjectivity	polarity	opinion	text_length	nouns	verbs	adjectives	adverbs
median	52.000000	13.000000	0.450000	0.000000	0.000000	185.000000	12.000000	3.000000	5.000000	1.000000
var	1058.928642	210.631179	0.087803	0.110738	1.606192	5183.734443	54.409460	3.829723	6.297404	1.315419
skew	1.410573	2.017794	0.028480	0.317258	0.040455	2.451700	4.729438	0.776656	0.596682	1.172307
kurt	2.020465	4.909882	-0.916303	0.476577	-1.096788	20.633844	45.397268	0.838843	0.405392	1.578861
25th percentile	35.000000	7.000000	0.207143	-0.071429	-2.000000	130.000000	9.000000	1.000000	3.000000	0.000000
75th percentile	70.000000	22.000000	0.641667	0.266667	1.000000	224.000000	16.000000	4.000000	7.000000	2.000000

Table 11: Additional Statistics of Features of Congress Related tweets

FEATURE ENGINEERING

- Features of BJP tweets

	id	original_text	0	1	2	3	4	5	6	7	...	997	998	999	polarity	subjectivity	text_length	nouns	verbs	adjectives	adverbs
0	advosushildixit	@anjanaomkashyap I am seeing you as future #bj...	0.0	0.0	0.294089	0.0	0.0	0.0	0.0	0.0	...	0.294089	0.000000	0.0	0.350000	0.362500	162.0	13.0	3.0	6.0	0.0
1	UttarrPradesh	Which of the following should be top priority ...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.250000	0.300000	62.0	3.0	1.0	2.0	0.0
2	ShaileshWrites	@RenukaJain6\nI still remember your video mass...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.259063	0.0	-0.500000	0.425000	187.0	10.0	3.0	8.0	2.0
3	iamljp	#bjp @BJP4India @INCIndia @INCKarnataka how mu...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	-0.342857	0.757143	177.0	15.0	6.0	1.0	2.0
4	Mdsr20351488	All Pakistanis had serious doubts on concept o...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.222222	0.655556	234.0	21.0	4.0	3.0	2.0

Table 12: Features of BJP tweets

- Features of Congress tweets

	id	original_text	0	1	2	3	4	5	6	7	...	997	998	999	polarity	subjectivity	text_length	nouns	verbs	adjectives	adverbs
	Sunnysweet16	Wonder why no academic or journalist asks INC ...	0.0	0.0	0.294089	0.0	0.0	0.0	0.0	0.0	...	0.294089	0.000000	0.0	0.125000	0.369048	166.0	10.0	4.0	5.0	0.0
	drnitinchaube	Congrats for the change #australiavotes2019 an...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.000000	0.000000	218.0	21.0	2.0	2.0	1.0
	mrvivek07	Peopel Say "Govt Ne 70 Years Kya kiya".\nUnse ...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.259063	0.0	0.000000	0.000000	254.0	27.0	4.0	3.0	2.0
	JosephPravinP	@ajaymaken @RahulGandhi And as a final touch, ...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.320000	0.565000	192.0	17.0	2.0	6.0	2.0
	VandanaMegastar	#LokSabhaElections2019 Anyone not having mass ...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	-0.222222	0.344444	199.0	12.0	10.0	0.0	3.0

Table 13: Features of Congress tweets

DATA VISUALIZATION

- Comparative Bar Chart to Show Comparison of Opinions between BJP and Congress

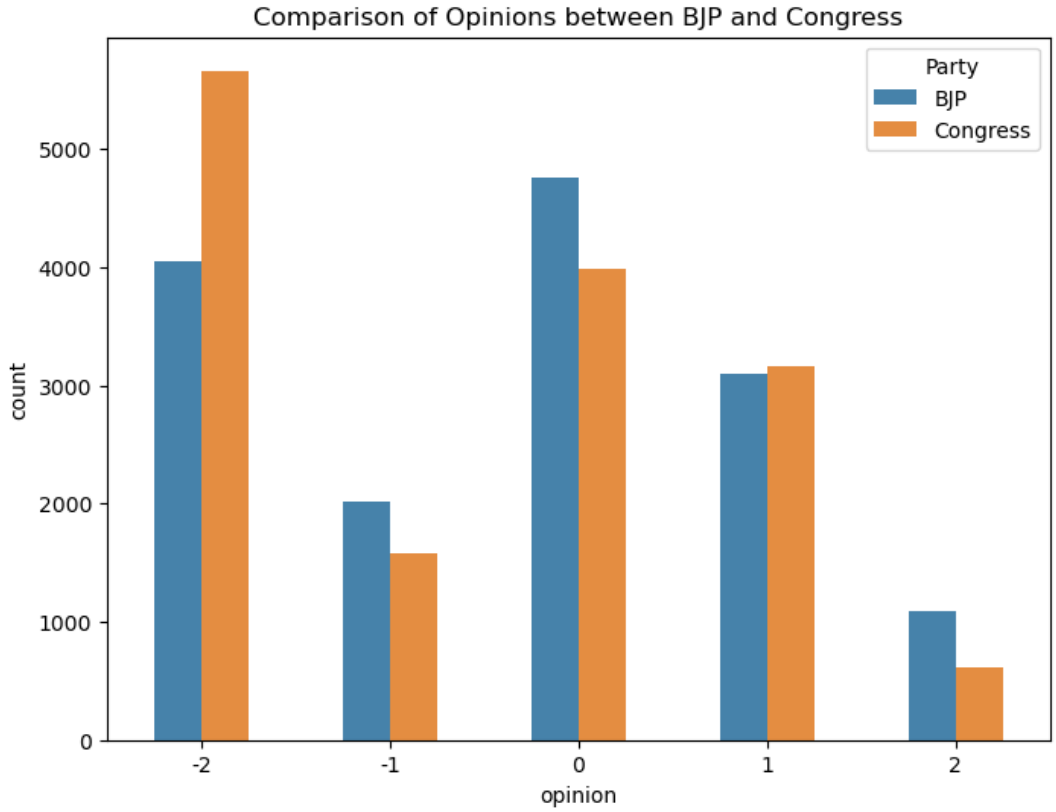


Fig. 9: Comparison of Sentiments between BJP and Congress

- Distribution Plots to Show Distribution of "Favourite Tweet" Count for BJP and Congress

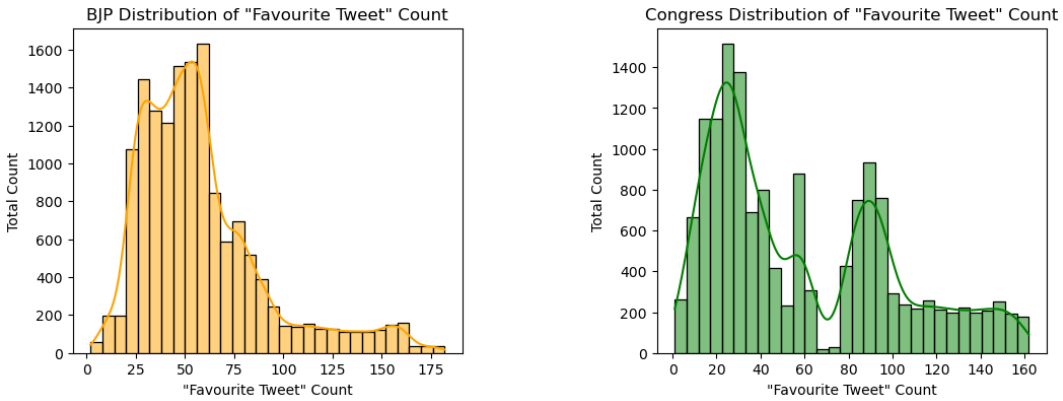


Fig. 10: Distribution of "Favourite Tweet" Count for BJP and Congress

- Distribution Plots to Show Distribution of “Retweet” Count for BJP and Congress

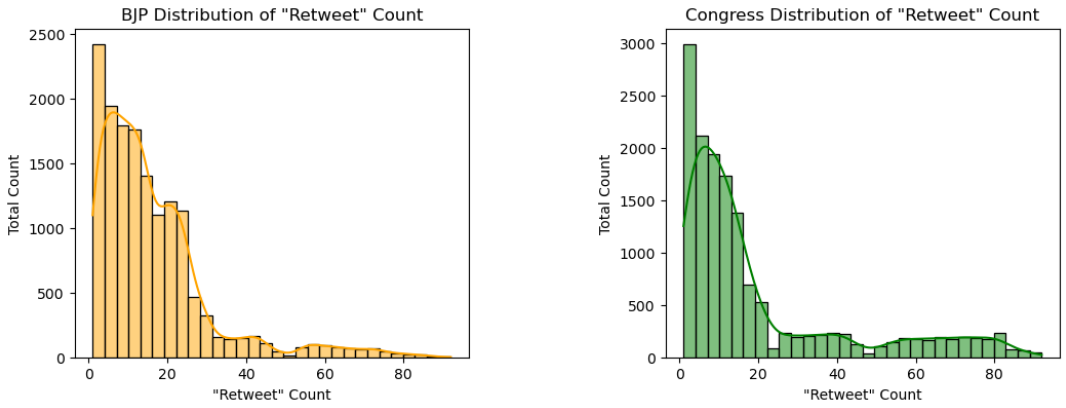


Fig. 11: Distribution of “Retweet” Count for BJP and Congress

- Distribution Plots to Show Distribution of Tweet Subjectivity for BJP and Congress

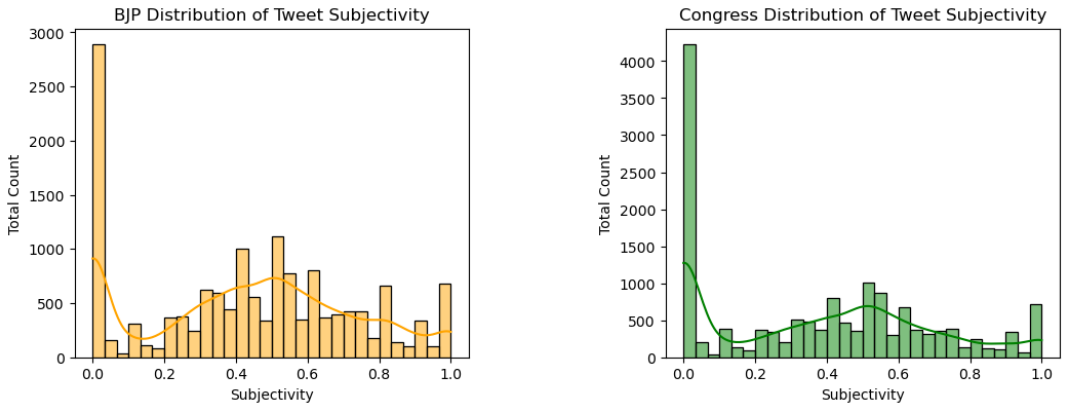


Fig. 12: Distribution of Tweet Subjectivity for BJP and Congress

- Distribution Plots to Show Distribution of Tweet Polarity for BJP and Congress

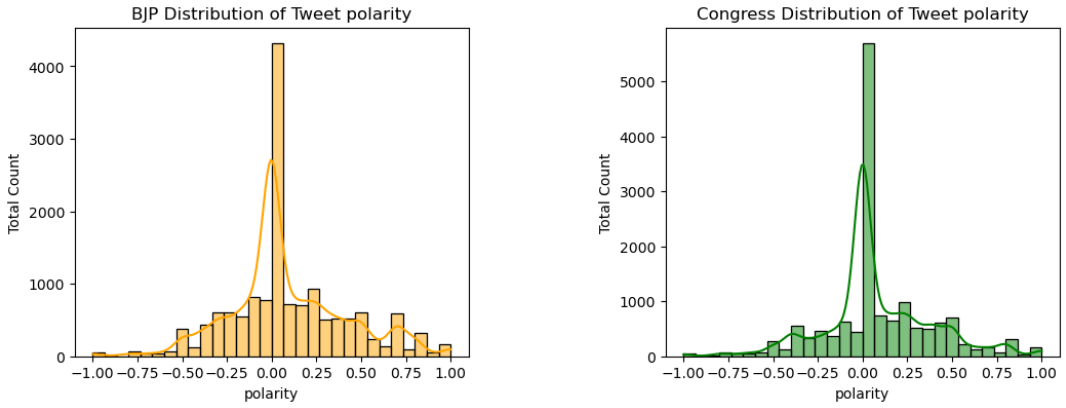


Fig. 13: Distribution of Tweet Polarity for BJP and Congress

- Pair Plot for BJP Features

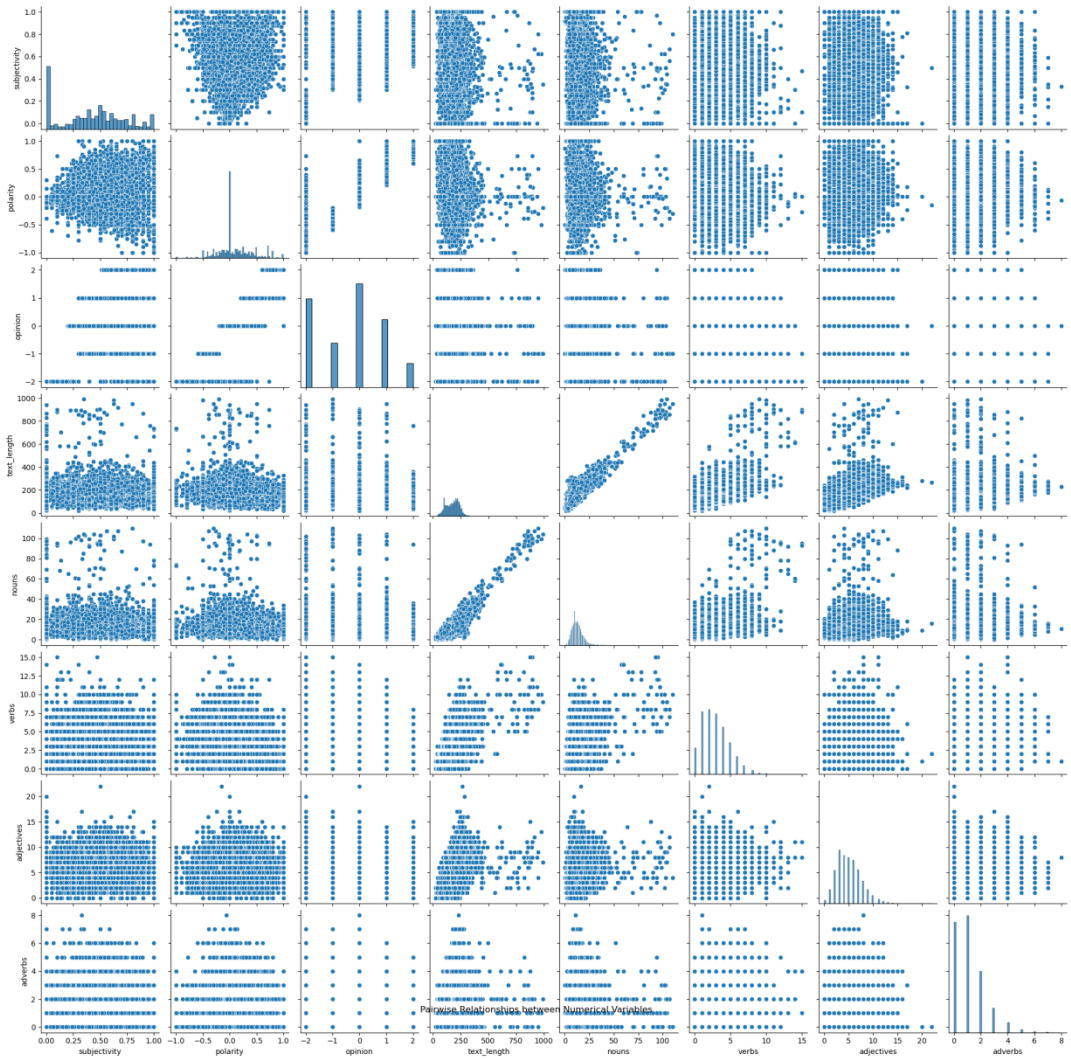


Fig. 14: Pair Plot for BJP Features

- Pair Plot for Congress Features

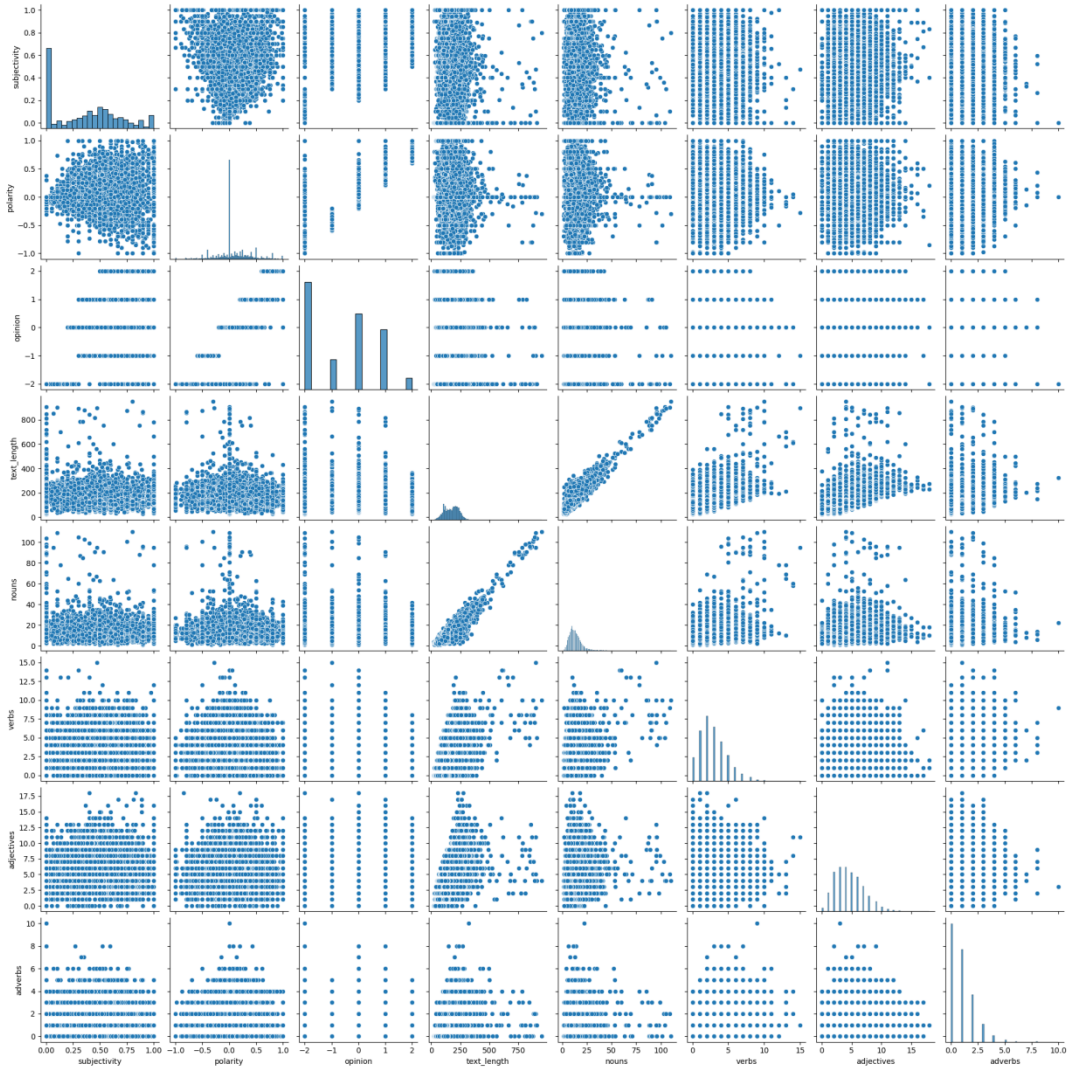


Fig. 15: Pair Plot for Congress Features

- Correlation Matrix for BJP Features

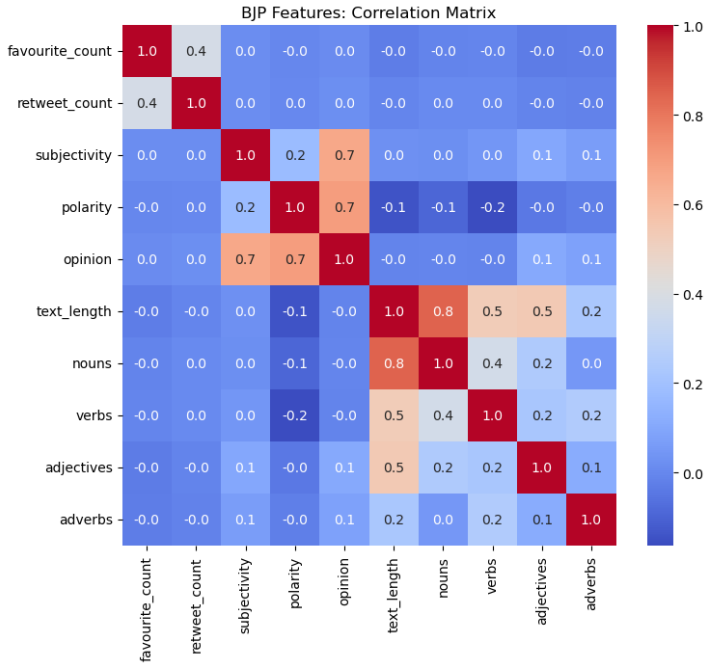
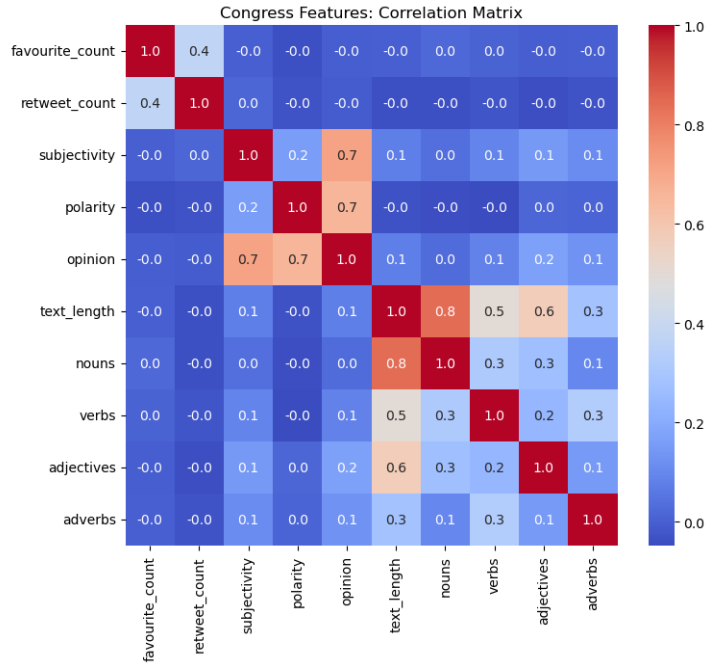


Fig. 16: Correlation Matrix for BJP Features

- Correlation Matrix for BJP and Congress Features



- Word Cloud of BJP tweets

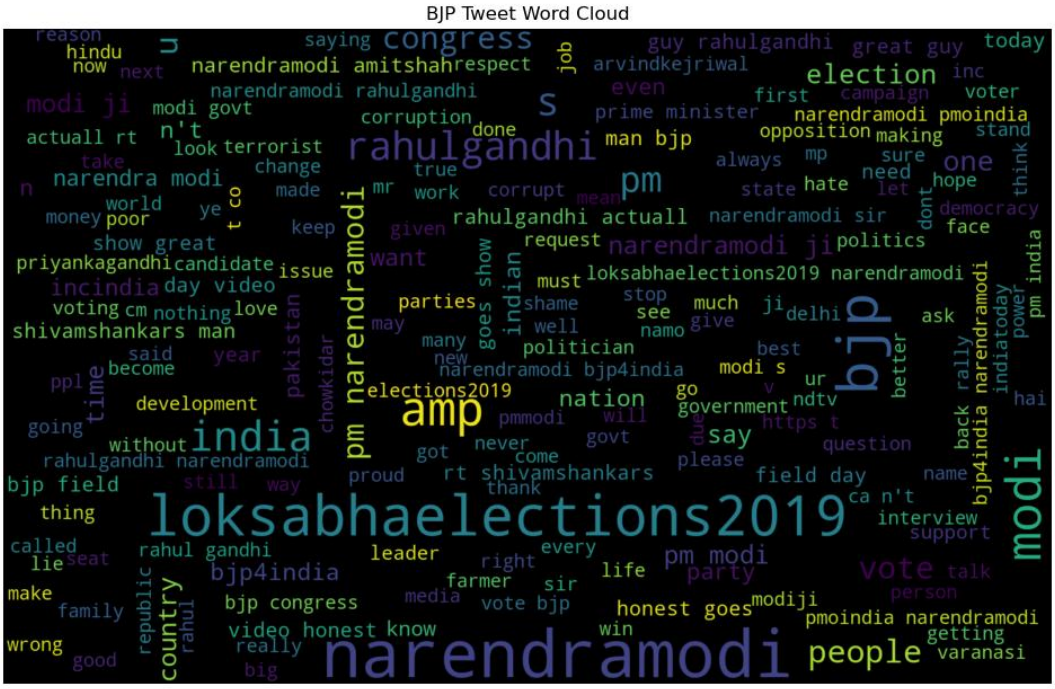


Fig. 17: Word Cloud of BJP tweets

- Word Cloud of Congress tweets

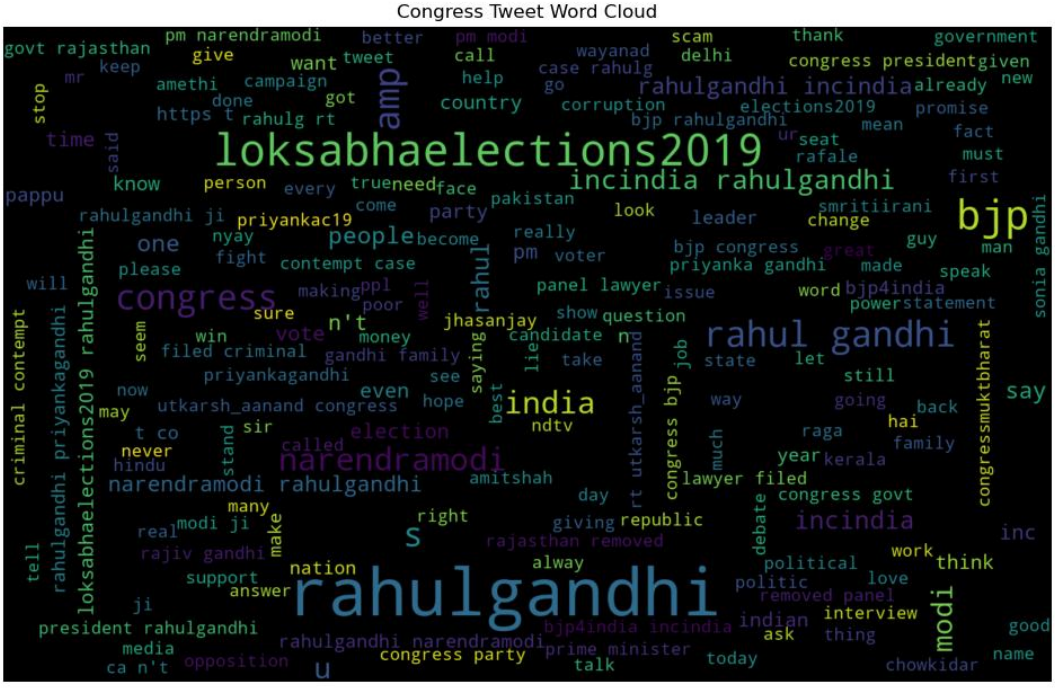


Fig. 18: Word Cloud of Congress tweets

8. Results & Discussions

In this project, we utilize three different datasets:

- Dataset 1: This dataset comprises over 15,000 tweets collected for both the BJP and Congress political parties, aiding in training models for election prediction
- Dataset 2: This dataset includes 7000 tweets each, taken into consideration for data balancing purposes.
- Dataset 3: This dataset consists of a total of 9000 tweets, used to predict the number of positive tweets for both the BJP and Congress, as well as neutral tweets.

The proposed model architecture for election prediction has been developed based on the hypothesis that negative emotions towards one political party might be perceived as positive emotions towards another. Consequently, the final ensemble model combines two individual models, one for each political party. This project aims to establish a generalized approach to election prediction, irrespective of the location, type of election, or number of political parties involved.

First, the data was preprocessed to clean the tweets by removing stopwords, punctuation, and other characters with lower relevance to the prediction. Next, the data was labeled based on polarity and subjectivity, and then divided into five categories as mentioned in table 1. The dataset was found to be imbalanced, and to address this, it was balanced using the second dataset. The final count of tweets corresponding to each label is as follows:

- BJP Dataset:

Label	Number of tweets
2	2983
1	3005
0	3010
−1	2995
−2	3007

Table 14: Tweet Opinion Tally for BJP Dataset

- Congress Dataset:

Label	Number of tweets
2	2979
1	2997
0	3018
−1	2993
−2	3013

Table 15: Tweet Opinion Tally for Congress Dataset

For this study, basic models such as Support Vector Machine, Logistic Regression and Naïve Bayes were considered. When the models were trained and tested for prediction, the performance metrics for both the BJP and Congress datasets were as follows:

- BJP Dataset

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	86.87%	0.8697	0.8687	0.8685
Naïve Bayes Model	87.77%	0.8803	0.8777	0.8778
Support Vector Machine	86.5%	0.8656	0.8650	0.8644

Table 16: Model Performance Metrics for BJP Dataset

- Congress Dataset

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85.7%	0.8579	0.857	0.8558
Naïve Bayes Model	59.14%	0.5964	0.5914	0.5584
Support Vector Machine	83.37%	0.8479	0.8337	0.8251

Table 17: Model Performance Metrics for Congress Dataset

Based on the charts above, we can conclude that the Naïve Bayes model exhibited the highest performance metrics for the BJP dataset, whereas the Logistic Regression model demonstrated the highest performance metrics for the Congress dataset.

Additionally, Bagging and Boosting ensemble methods, including AdaBoost, XGBoost, and Gradient Boost, were employed with each of the three base models for

each dataset. Among these, the bagging method yielded the best performance. The performance metrics were as follows:

- BJP Dataset

Base Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	86.94%	0.8702	0.8694	0.8692
Naïve Bayes Model	88%	0.8823	0.88	0.8801
Support Vector Machine	88.04%	0.8851	0.8804	0.8803

Table 18: Performance Metrics with Bagging Ensemble for BJP Dataset

- Congress Dataset

Base Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85.07%	0.8498	0.8057	0.8494
Naïve Bayes Model	55.56%	0.4685	0.5556	0.4577
Support Vector Machine	83.3%	0.847	0.833	0.824

Table 19: Performance Metrics with Bagging Ensemble for Congress Dataset

Therefore, we can conclude that for the BJP dataset, the Bagging Ensemble with SVM as the base performed the best, closely followed by the Bagging Ensemble with the Naïve Bayes Model as the base. Moreover, for all three models, an improvement in performance was observed for the BJP dataset when the Bagging Ensemble was utilized. All three models exhibited significantly good performance even without the ensemble method.

On the other hand, for the Congress dataset, no model showed any improvement when the ensemble method was applied. Overall, Logistic Regression demonstrated the best performance for the Congress dataset.

Finally, for the prediction framework, we will use the Bagging Ensemble with SVM as the base for BJP tweets and Logistic Regression for Congress tweets.

The tally of opinions in the original test dataset (Dataset 3) is as follows:

Opinion	Count	
	BJP	**Congress**
Positive	1491	1602
Neutral	546	544
Negative	963	854

Table 20: Tally of Opinions in Test Dataset

When this dataset was fed into the proposed framework, the output was obtained as follows:

- Pro BJP Votes: 2022
- Pro-Congress Votes: 1973
- Others: 1425

The proposed framework predicted some positive tweets for one party from the other party's negative tweets. This shows the possibility of existence of a relationship where negative opinion about one political party might be an indication of positive opinion about another political party, which might be valuable and worthwhile in future researches.

9. Limitations

The study faces several limitations that impact the accuracy and generalizability of its findings. Firstly, there are constraints related to time, funding, and the size and diversity of the dataset. These factors can limit the scope of the analysis and the ability to draw robust conclusions. Secondly, the quality of the data poses challenges. Tweets often contain noise such as typos, slang, and abbreviations, which can affect the accuracy of sentiment analysis. Additionally, automated sentiment analysis tools may struggle to capture the nuanced meanings in tweets. Furthermore, there are limitations in contextual understanding. Sentiment analysis may not always capture the context in which tweets are made, leading to potential misinterpretation. For example, a tweet expressing negative sentiment towards a political party may actually be using sarcasm or irony.

Temporal dynamics also play a role, as public sentiment on social media can change rapidly. A single-point analysis may not capture these fluctuations, necessitating longitudinal analysis to understand how sentiment evolves over time. External factors, such as news events or political scandals, can also influence public sentiment but may not be reflected in the tweets alone. Furthermore, while models like Logistic Regression, Support Vector Machine, and Naïve Bayes are commonly used for sentiment analysis, their performance can vary depending on the dataset and features used. Ensemble techniques can improve performance, but optimal results are not guaranteed.

Generalizability is another concern, as the study focuses solely on Indian elections. To achieve broader generalizability, the study would need to include tweets from a wider range of countries and contexts. Additionally, the study's findings may be influenced by the specific period of data collection, as sentiment on social media can be influenced by current events and trends. Therefore, caution should be exercised when extrapolating the findings to other contexts or time periods.

10. Conclusion

This study highlights the dynamic nature of the modern political landscape and the significant influence of major countries on the global economy through their policies. Understanding public sentiment towards political parties is crucial for predicting election outcomes and informing policy decisions. Traditional methods of election forecasting often rely on limited data sources, such as polls and surveys, which may not capture the full spectrum of public opinion. However, with the advent of social media, particularly platforms like Twitter, there is a wealth of real-time, unfiltered data that can provide valuable insights into public sentiment.

The study focused on predicting election outcomes by analyzing public sentiment towards political parties using tweets. The methodology involved collecting and organizing tweets related to Indian elections, preprocessing it for feature extraction, and performing sentiment analysis to assess public opinion towards political parties. Logistic Regression, Support Vector Machine, and Naïve Bayes models were trained on sentiment-labeled data and were used for prediction. Ensemble techniques were applied to improve prediction accuracy.

The findings of this research suggest that negative sentiment towards one political party may be indicative of positive sentiment towards another, highlighting a nuanced relationship that warrants further investigation in future research. The developed machine learning framework is expected to provide valuable insights into studying public sentiment towards political parties to predict election results.

The creation of a machine learning framework for social media sentiment analysis-based election outcome prediction is one of the primary outcomes of this research. Because they offer a more comprehensive knowledge of the intricate interaction between public mood, social media, and election outcomes, the research's findings have significance for policymakers, analysts, and stakeholders in the political arena.

It is crucial to recognize this study's limitations, though. The use of tweets, which might not accurately reflect the range of viewpoints held by voters, is one drawback. Time, money, and dataset size limitations are a few of these that may affect the analysis's breadth and resilience. The accuracy of automated sentiment analysis algorithms may be hampered by the noise and subtle meanings present in tweets, raising concerns about the quality of the data. While ensemble techniques may not always guarantee optimal results, models like Logistic Regression, Support Vector Machine, and Naïve Bayes are frequently utilized. However, their performance can vary. These limitations might be addressed in the future by broadening the study's scope to encompass a wider variety of political circumstances.

Overall, this research contributes to the growing body of knowledge on leveraging social media data for election forecasting and highlights the importance of understanding public sentiment in shaping political outcomes. Future research should continue to explore innovative approaches and consider a broader range of factors that influence public opinion and election results. By doing so, we can further advance our understanding of the complex relationship between public sentiment, social media, and election outcomes, ultimately improving the accuracy of election forecasting and providing valuable insights for policymakers, analysts, and the public alike.

11. References

[1] Yavari A, Hassanpour H, Rahimpour Cami B, Mahdavi M. Event prediction in social network through Twitter messages analysis. Soc Netw Anal Min. 2022;12(1):78. doi: 10.1007/s13278-022-00911-x. Epub 2022 Jul 9. PMID: 36618491; PMCID: PMC9807096.

[2] 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

[3] Budiharto, W., Meiliana, M. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. J Big Data 5, 51 (2018). https://doi.org/10.1186/s40537-018-0164-1

[4] Alvi Q, Ali SF, Ahmed SB, Khan NA, Javed M, Nobanee H. On the frontiers of tweets and sentiment analysis in election prediction: a review. PeerJ Comput Sci. 2023 Aug 21;9:e1517. doi: 10.7717/peerj-cs.1517. PMID: 37705657; PMCID: PMC10495957.

[5] Barkha Bansal, Sangeet Srivastava, On predicting elections with hybrid topic based sentiment analysis of tweets, Procedia Computer Science, Volume 135, 2018, Pages 346-353, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.08.183.

[6] Coletto, Mauro & Lucchese, Claudio & Orlando, Salvatore & Perego, Raffaele. (2015). Electoral predictions with Twitter: A machine-learning approach. CEUR Workshop Proceedings. 1404.

[7] Unankard, S., Li, X., Sharaf, M., Zhong, J., Li, X. (2014). Predicting Elections from Social Networks Based on Sub-event Detection and Sentiment Analysis. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds) Web Information Systems Engineering – WISE 2014. WISE 2014. Lecture Notes in Computer Science, vol 8787. Springer, Cham. https://doi.org/10.1007/978-3-319-11746-1_1

[8] O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: ICWSM, pp. 122–129 (2010)

[9] Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: ICWSM, pp. 178–185 (2010)

[10] Maurya, Ankit & Lodh, Satish & Joshi, Mayur & Singh, Prof. (2021). Election Result Prediction Using Sentiment Analysis. International Journal of Advanced Research in Science, Communication and Technology. 118-122. 10.48175/IJARSCT-V4-I3-018.

[11] Vendeville, A., Guedj, B. & Zhou, S. Forecasting elections results via the voter model with stubborn nodes. *Appl Netw Sci* **6**, 1 (2021). https://doi.org/10.1007/s41109-020-00342-7

[12] Rao, Kanade, Motarwar & Girme (2022), Election Result Prediction Using Twitter Analysis. International Research Journal of Engineering and Technology.
[13] Singh, Prabhsimran, Yogesh K. Dwivedi, Karanjeet Singh Kahlon, Annie Pathania, and Ravinder Singh Sawhney. "Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections." Government Information Quarterly 37, no. 2 (2020): 101444.
[14] K. Myilvahanan, Y. P, S. Pasha, M. Ismail and V. Tharun, "A Study on Election Prediction using Machine Learning Techniques," *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2023, pp. 1518-1520, doi: 10.1109/ICAIS56108.2023.10073693.
[15] Moawi, Hazim, Predicting Voting Behaviors and Election Results Using Digital Trace Data and Twitter (March 28, 2023). Available at SSRN: https://ssrn.com/abstract=4464047 or http://dx.doi.org/10.2139/ssrn.4464047
[16] Arnab Bhattacharyya, Palash Dey, Predicting winner and estimating margin of victory in elections using sampling, Artificial Intelligence, Volume 296, 2021, 103476, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2021.103476.
[17] Ram prasad, T. ., Tripathi, M. K. ., Moorthy, C. V. K. N. S. N. ., Nemade, V. A. ., Barpute, J. V. ., & Angadi, S. K. . (2024). Mathematical Modelling and Implementation of NLP for Prediction of Election Results based on social media Twitter Engagement and Polls.
[18] Georgiadou, Elena, Spyros Angelopoulos, and Helen Drake. "Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes." International Journal of Information Management 51 (2020): 102048.
[19] Alvi, Quratulain, Syed Farooq Ali, Sheikh Bilal Ahmed, Nadeem Ahmad Khan, Mazhar Javed, and Haitham Nobanee. "On the frontiers of tweets and sentiment analysis in election prediction: a review." PeerJ Computer Science 9 (2023): e1517.
[20] Rizk R, Rizk D, Rizk F, Hsu S. 280 characters to the White House: predicting 2020 U.S. presidential elections from tweets. Comput Math Organ Theory. 2023 Mar 28:1-28. doi: 10.1007/s10588-023-09376-5. Epub ahead of print. PMID: 37360912; PMCID: PMC10042672.
[21] P. Khurana Batra, A. Saxena, Shruti and C. Goel, "Election Result Prediction Using Twitter Sentiments Analysis," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2020, pp. 182-185
[22] Zuloaga-Rotta L, Borja-Rosales R, Rodríguez Mallma MJ, Mauricio D, Maculan N. Method to Forecast the Presidential Election Results Based on Simulation and Machine Learning. *Computation*. 2024; 12(3):38. https://doi.org/10.3390/computation12030038
[23] Ali, Haider & Farman, Haleem & Yar, Hikmat & Khan, Zahid & Habib, Shabana & Ammar, Adel. (2022). Deep learning-based election results

prediction using Twitter activity. Soft Computing. 26. 10.1007/s00500-021-06569-5.

[24] 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.

[25] Sang, E.T.K., Bos, J.: Predicting the 2011 dutch senate election results with twitter. In: EACL Workshop on Semantic Analysis in Social Media, pp. 53–60 (2012)

[26] Jungherr, A., Jurgens, P., Schoen, H.: Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, t. o., sander, p. g., & welpe, i. m. "predicting elections with twitter: What 140 characters reveal about political sentiment". Soc. Sci. Comput. Rev. 30(2), 229–234 (2012)

[27] https://towardsdatascience.com/deep-dive-into-softmax-regression-62deea103cb8

[28] https://medium.com/@priyankgupta2003/analyzing-rotten-tomatoes-reviews-with-naive-bayes-algorithm-b9dc4e1a714e

[29] https://www.researchgate.net/figure/Three-class-SVM-OvA_fig7_309200152