Exploring the Non-Political Factors Behind Young Voter Enthusiasm: A Machine Learning Approach

by

Md Israk Hossain
20201150
Raya Subah
20201132
Mashrur Ahmed Utsho
20201072

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
February 2025

© 2025. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md Israk Hossain
20201150

Raya

Raya Subah
20201132

Mashrur Ahmed Utsho
20201072

Approval

The thesis/project titled “Exploring the Non-Political Factors Behind Young Voter Enthusiasm: A Machine Learning Approach” submitted by

1. Md Israk Hossain (20201150)
2. Raya Subah (20201132)
3. Mashrur Ahmed Utsho (20201072)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on February 7, 2025.

Examining Committee:

Supervisor:
(Member)

Md Saiful Islam
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Syed Zamil Hasan Shoumo
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

To build a democratic nation, a voting system provides the foundation and it represents the fundamental rights of citizens to voice their decisions. It also represents the responsibility of citizens for shaping their government. The participation of young voters is important as they form a significant portion of a country and are able to bring new perspectives regarding decision-making or policies revolving around a country. However, a lot of young people refrain from voting due to political violence, thinking it will not make a difference or due to various other reasons. They can be engaged through awareness on social media, political campaigns or discussions in classrooms. Their involvement can increase voter enthusiasm which is pivotal in determining higher voter turnout. With the rise of Artificial Intelligence, the effort of people has decreased significantly in the extraction of data and finding meaningful insights. Hence, we will utilize machine learning to make future decisions based on the primary data we have collected. The aim of this research is to propose a multi-modal agent that can help to infer voter turnout and understand the factors that influence the behavior of our youths when participating in voting using artificial intelligence. We aim to focus only on the non-political factors that affect the voting behavior of young people. We have carried out a survey on the students of BRAC university who were asked to fill out a questionnaire containing both multiple choice questions and opinion based questions. The answers to the multiple choice questions will be processed as tabular data and fed to an artificial neural network for inference. Similarly, the answers to the opinion based questions will be fed to an extreme gradient boosting model for sentiment analysis. The true label for both levels of inference will be whether a person would participate in voting or not. The proposed multi-modal agent will concatenate the outputs from the artificial neural network and the extreme gradient boosting model and provide a final level of prediction. Moreover, to assess the predictions of the artificial neural network, XAI models such as SHAP and LIME will be used to produce global and local level explanations. A further analysis of these explanations will be made to understand which features are affecting our model. Therefore, the proposed model can help us to gauge young voter turnout and shed a light into the influencing factors that steer their decisions.

Keywords: Machine Learning, Voting, Neural Networks, Multilayer Perceptron (MLP), Voter Enthusiasm, Non-political factors, Sentiment Analysis, Natural Language Processing, NLP on text data, Random Forest (RF), Support Vector Machine (SVM), Linear and Polynomial Kernel, Extreme Gradient Boosting (XGBoost), Bidirectional Encoder Representations from Transformers (BERT), BERT-base, Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Principal Component Analysis (PCA), Synthetic Minority Oversampling Technique (SMOTE), Correlation Test, University, Student, Voting Behavior, Young Voter Enthusiasm, Functional API, Bangladesh, K means clustering, Elbow Method.

Dedication

We sincerely dedicate this work to our late friend, *Sayed Hasan Emon*, who passed away tragically in a bike accident on 29th November, 2023. Emon was a hardworking and meritorious student who had dreams to see our work published. He has contributed significantly to this research and even if this paper does not end up getting published, we at least hope that we have made him proud. Our final message would be, live and let live; life is short and we never know when is our last day with someone.

Acknowledgement

All praises to the Almighty, it is Him who have made this work possible to be completed despite the challenges we have faced. We wholeheartedly thank our supervisor sir, Md. Saiful Islam and co-supervisor sir Syed Zamil Hasan Shoumo for their unwavering support, guidance and motivation which have made this journey enjoyable. We thank Professor Dr. Riaz Partha Khan for his unparalleled intellect and friendly demeanor. His knowledge in the field of politics have been really insightful and we have gained a new perspective on how social issues can combine with computational methods and technology, something which we had little to no knowledge about. Most importantly, it was Professor Dr. Mahbubul Alam Majumdar and Professor Dr. Wasiqur Rahman Khan who initiated this idea and through this research we were able to view machine learning through the lens of political science. Lastly, we express our gratitude to our family for being with us throughout. Whenever life gets difficult, it is them to whom we return for some solace.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ANN Artificial Neural Network

$BERT$ Bidirectional Encoder Representations from Transformers

$CART$ Classification And Regression Tree

$CHAID$ Chi-square Automation Interaction Decision

$GSCV$ Grid Search Cross Validation

$LIME$ Local Interpretable Model-agnostic Explanations

MLP Multilayer Perceptron

NLP Natural Language Processing

NN Neural Network

PCA Principal Component Analysis

$ReLU$ Rectified Linear Unit

$SHAP$ SHapley Additive exPlanations

SVM Support Vector Machine

SVR Support Vector Regression

$XGBoost$ Extreme Grading Boosting

Chapter 1

Introduction

1.1 Overview

Young voters play an important role in the future of a country, but their voter turnout is often lower than that of older age groups. There has been many studies which focus on the political reasons for such lower turnout rate, but non-political factors such as social influences, personal beliefs are less explored. Understanding these non-political factors can help identify the reasons that encourage or discourage the young voter from voting. Hence, the purpose of this research is to focus on identifying the non-political factors by using advanced machine learning methods to analyze both tabular data, such as demographic information, and textual data, such as opinions expressed in text.

1.2 Motivation Behind the Research

In recent years, a declining trend has been noticed in young voters' participation in democratic nations around the world. Analysis of data from numerous democratic countries shows a lower voter turnout among young adults (typically aged 18 to 29) compared to older age groups. This trend raises questions about the factors influencing their disinterest in voting. In 2012, the voter turnout in the US Presidential election was 55% of the total voters, and in 2014, only 36% of the total voters voted in the midterm election. Another study has shown young people under 18 are more likely to live at home with their families and still attend school, leading to potentially different socialization effects at the time of their first election [5]. There may be several other interconnected reasons linked to this. Some voters feel voicing their views and opinions will not impact the political landscape in any way, as an area might be dominated by a certain party, hence leading to voter indifference. Understanding politics intricately can be complex for a young voter due to the inadequacy of civic education in schools, making it tough to understand the dynamics of voting. There can also be a lack of access to unbiased and true information. Young voters' choices to vote can also be hampered by having a busy schedule amid work and studies, their families showing little to no concern for voting, and also reading fewer newspapers or articles to engage with the field of politics. Our aim is to reach out to a wide spectrum of the young audience at our university to determine the non-political factors that are affecting the voting behavior. The lack of proper research revolving around this topic was the reason there was an absence of primary dataset.

Hence for data collection, we have used Google Forms to perform an anonymous survey, making people feel completely safe to share their concerns regarding this matter.

1.3 Research Objectives

The main objective of this research is to understand the non-political factors that affect whether young people choose to vote. By combining ideas from political science and artificial intelligence, this study aims to use advanced machine learning techniques to provide more accurate and useful predictions.

First of all, our aim is to collect primary data by preparing a questionnaire by consulting and discussing with professors from different departments. After that, we aim to do a multi-modal analysis using different types of data. Our research takes a multi-modal approach, meaning we work with both tabular data and textual data to identify patterns that influence voting behavior. The tabular data, which includes information such as demographics, and other related factors, is analyzed using a Multilayer Perceptron (MLP) model to measure voter interest. Simultaneously, we use sentiment analysis on the textual data to study the opinions related to voting. This textual data is first processed into numerical form using TF-IDF transformation, and then machine learning models such as Random Forest, XGBoost, Support Vector Machines (linear and polynomial) and BERT are applied to predict the sentiments. We tune these models to get the best results by analyzing their accuracy and loss values and select the one that performs the best.

Besides, we aim to use the predictions from the MLP model and the sentiment analysis as inputs to a new ensemble model built with functional API. These inputs are merged in the model, which passes the combined data through the layers. The output layer provides the final prediction of whether a young individual is likely to vote or not. Using the predictions of MLP, we further aim to use explainable AI models to compare and find out the influential factors behind young voter enthusiasm.

1.4 Challenges Faced

We collected around 1000 responses, and after pre-processing and cleaning the data, the total responses dropped to around 850. This led to the dataset being small as originally we targeted the overall responses to be around 2000. Furthermore, there was an imbalance in our dataset, for which we had to apply SMOTE to reduce the biasness of the machine learning models. This led to synthetic data creation, leading to lower variation in them. A lot of students did not take the task seriously or understand the purpose of the research despite our explanation. Hence, they filled out the form with illogical remarks which hampered our time. Additionally, there was no prior work which is similar to our research. For this, we had to make a whole set of questionnaire by scheduling meetings with professors. This work was hectic and if there were any prior publication on this, we could have modified the existing work or gained some insight from the papers. Lastly, the July student movement put the work to a halt, as everything went to a shutdown. This slackened our progress a bit and it was sort of tough for us to get back to the activity after a long gap.

1.5 Thesis Outline

The main objective of this research was to find out the non-political factors which affect young voter enthusiasm. We have created a dataset by creating a questionnaire and surveying it. The participants were the students of BRAC University. The research was multi-modal where we obtained separate predictions from both tabular and textual data. These predictions were later inputted to a new model from which the final prediction was obtained. Explainable AI models were used to find out the possible influential factors.

Chapter 1 discusses the overview of the topic with the research objectives. It also gives an idea of the challenges that we faced while conducting the research.

Chapter 2 is about our study on research papers from where we have identified the possible machine learning models and used them for the sentiment analysis portion and the tabular classification task. This chapter also shows the papers from where we gained insights about the questions that could be used for the survey's questionnaire.

Chapter 3 shows the theoretical knowledge of the models we have used for the research. The models' workflows and structures are described with visual presentation.

Chapter 4 proposes the methodology and workflow of the research. Here, the data collection method is described with details of the data. This chapter also describes the data pre-processing steps and the dataset preparation for multimodal classification, ensuring the addressing of class imbalance.

Chapter 5 is about the model implementation phase where the models were evaluated.

Chapter 6 is the analysis of the results where all the results from the models are analyzed. Moreover, the most influential factors (features) which affect the voting behavior of young voters are identified from the explanations of explainable AI models and correlation tests through comparative analysis.

Chapter 7 provides the conclusion of the research and elaborates the possible ways in which future-works can be conducted to overcome the challenges faced in this research.

Chapter 2

Literature Review

2.1 Possible Questions For The Survey

According to Kim, Alvarez, and Ramirez (2020) [15], there are only a few demographic variables that actually affected voter turnout. Interestingly, the study found that age didn't play a role, which we think could be a limitation in the model. For our research, we want to include factors like place of residence, educational background, and occupation to get a more accurate picture.

In a study by Blais (2006) [2], the main reasons for voter turnout include compulsory voting, the electoral system, and whether there is a single legislative house. First, compulsory voting increases turnout, especially when penalties are in place for those who don't vote; however, this is more effective in well-established democracies. Second, turnout is higher in countries with large, competitive voting districts where parties put in more effort to win votes. However, in places like Latin America, the electoral system doesn't seem to increase turnout, as there's less connection between the voting system and turnout outside of Europe. Third, voter turnout tends to be higher in countries with a single legislative house, where power is more concentrated; in countries with multiple legislative houses, turnout is often lower since power is spread out across different elections. From this paper, we learned that non-political factors affecting voter enthusiasm can include the voting age and voting rules. Based on this, we included questions in our survey asking participants about their age, whether they voted in the most recent election, if they plan to vote in upcoming elections, and if they would vote if the election day were scheduled on a weekend.

Study of Fowler [6] found that age to be one of the key factors affecting voter turnout. In 2008, young voters (under 30) were less likely to vote if they had little interest in the election process. In 2013, rallies and events increased young voter participation. For middle-aged voters, factors like supporting certain candidates or parties, political interest, and involvement in politics played an important role. Other social factors, such as religion, gender, and employment status, had little impact on turnout. It was also observed that there was a certain proportionality between election outcomes and number of participation. If the sets of policies made do not satisfy a certain number of people, they will refrain from voting, thus deteriorating government performance. When people are unhappy with the policies, they might not vote, which can weaken government performance. In our research, we also include age as a non-political factor to predict interest in voting. We look at different age groups to see who is more or less interested in voting. We also ask survey questions to find out if students are politically aware, concerned about election violence, or if they have voted before.

According to Coffé and Voorpostel (2010) [4], their study tries to find out if young people tend to adopt parents' political party choice in Switzerland. The research collected data from 580 pairs of parents and 919 young people. The study analyzes the influence of parental voting behavior and attitudes on their children's support for the Swiss People's Party (SVP). The controlled variables are educational level, age, gender, language, occupational status and church attendance. The four models of the multilevel regression are model 1(Parental Influence), model 2 (Parental Attitudes), model 3 (Mother's Influence), model 4 (Gender Differences). After analyzing the results, it is seen that for model 1, parents' SVP support highly influences child's preference for SVP with mothers exerting stronger influence. For model 2, it is seen that parents' negative attitude towards the EU positively affected the child to support SVP. According to model 3, it shows that the influence of mothers remains significantly higher than fathers' and model 4 suggests that the influence of mothers is more on girls than boys. From the findings of this research, we can see that there can be influence of parents on their child which determines their voting behavior. Likewise, if parents are enthusiastic about voting then their children may be more likely to vote in elections. We can also see that factors like age, educational level, gender and occupational status can have an effect on voting behavior. For this reason, we have added questions regarding parents' voting enthusiasm, participant's occupation, gender and educational background of both participant's and parents'.

The study of Zhu (2021) [20], focuses on analyzing the factors which influence voter turnout among young individuals aged 18 to 29 in the United States. Key factors affecting voter turnout include race and ethnicity, geography, social media usage, education and income level which are the factors that are within the control of the young voters. The paper explains that the voting rate has dropped over time, and it also varies between different racial groups. When it comes to gender, women are more likely to vote than men and are often involved in politics in quieter ways. Social media is another big influence on young voters because it keeps them informed about political events and voting trends. Since young people spend a lot of time on social media, this plays an important role in boosting their interest in voting. Geography also matters a lot. Data from the Census Bureau shows that areas with larger populations have more young voters. Income is another factor that impacts voter turnout—people with higher incomes are more likely to vote. According to the Census Bureau, citizens with more income are more likely to vote than citizens with less income and also there is a positive correlation between voter participation and GDP per capita according to a research. Apart from this, as the level of education rises, there is a significant increase in the voter turnout. From the above study, we can conclude that these non political factors can affect young voters' enthusiasm. As a result, in our survey we have set questions about geography, educational background, social media usage time and occupation.

In this paper [7], Sandvoss examines the effects of blogging, the new age technology to express one's feelings or opinions towards a particular phenomenon. It is seen that blogging is becoming increasingly popular and a prominent area of public discussions which is drawing more and more people towards politics. It talks about how people's political engagement online can be compared to being fans of celebrities or sports teams. Just like fans are passionate about their favorite stars, some people are equally passionate about politicians and political issues. The study shows that supporters of different political parties may have different views which tend to divide them into two completely opposing groups. Besides, it also raises concerns about how this emotional attachment to politicians might affect democracy. Instead of addressing important issues and how to resolve them, people might become more biased towards their favorite politicians or parties, which could weaken the idea of democracy and invite dictatorship in the digital age. Through a popular political blog named Daily Kos, this research uses qualitative methods like observing and analyzing the posts and comments to find out the similarities between political blogging and fan cultures. Even though our focus was not to compare between the two, a similar approach was made to find out how social media presence and subscription to news channels and daily watching or reading news affects one's enthusiasm to vote. Due to the emergence of technology, we also consider online presence a means to influence a young adult's mind. Through our survey, we have asked the university students about their time spent on social media, if they consider it along with television, online news portals or websites, as a feasible and reliable source of information. If yes, we can assume the person is somewhat or to a greater extent aware of the political system of their country and also globally. This can either motivate them to vote or be interested in participating in the elections or simply be unwilling to vote. For better prediction, we also have added other non-political factors such as their support of the notion of democracy, previous voting experience, awareness of the parliamentary system, family members' influence, eligibility, etc to get a clearer idea if they choose to vote or refrain due to external factors or simply by their own choice.

This paper [3] looks at the reasons behind why young adults choose to vote or not, using a model called the information–motivation–behavioral skills (IMB) model. They conducted two studies to test how well the IMB model predicts voting intention and behavior. It is seen that the IMB model predicts voting behavior better compared to other models like Theory of Reasoned Action and the Theory of Planned Behavior. As we know, the United State is a democratic nation, the voter turnout is still low, with only 55% participation to vote. Mostly the young candidates are not voting. In the year 2004, the United state presidential election had only 47% of young voters compared to 55% and 72% in older age groups, according to the U.S. Census Bureau. We see age to be one of the most important factors so we included 'age' as an important factor for our study. When a large number of people, particularly young adults, choose not to vote, the democratic process suffers. Furthermore, research has shown that when people begin voting at an early age, they are more likely to be interested in politics later in life. A key reason young individuals avoid voting is that they have bad attitudes toward politics—they may lack trust in the government or believe their vote has no impact.. In this study, participants filled out a questionnaire about trust in the government and their sense of political influence, or "efficacy." Similarly, we included questions to understand if young people feel skeptical about voting due to past experiences or political beliefs about upcoming elections. We also asked about past voting experience, parents' education, and knowledge of the country's judicial system to get a clearer picture and make more accurate predictions in our study.

The study [16] shows that while there are many theories to explain why people vote, the unique nature of the 2016 election pushed researchers to consider factors outside of traditional frameworks. They found out that factors like voter registration status, age, political interest, voting history, and primary election participation are crucial for predicting turnout. Therefore, in our paper, we used similar factors like age, political interest, voting history, and election participation. We also added new questions, like whether violence on election day, parental influence, and encouragement to vote affect turnout.

2.2 Possible Models To Use

With so many factors in this large dataset, analyzing everything at once became difficult. According to Kim, Alvarez, and Ramirez (2020) [15] , machine learning can help with this problem by reducing the number of variables to the most important ones, making analysis easier and more insightful. One specific machine learning method used in the study was "fuzzy forests" which is an extension of random forests. Fuzzy forests work by identifying and selecting the most important variables, helping to narrow down the many interconnected pieces of data. The results showed that even when using only a small number of variables, fuzzy forests performed well. The study revealed three main insights about voter turnout in the 2016 election. First, it found that political issues like Obamacare, fiscal policy, and climate change helped indicate if registered voters participated, but these issues didn't strongly affect the overall turnout. Second, most demographic variables, except age, were not strongly linked to turnout. Finally, voting registration and administrative processes were found to be significant factors in turnout. The paper concludes that machine learning tools, like fuzzy forests, can help researchers better understand voter turnout. In comparison, we chose neural networks (NN) instead of fuzzy forests for our analysis. Neural networks are better at managing large, complex datasets and can handle unstructured data more effectively with the right hardware, like GPUs. They're also great for tasks like sentiment analysis. Because of this, we believe neural networks will provide more accurate results than fuzzy forests.

In the field of elections [18], Malaysia follows a democratic system, encouraging everyone's independence to choose and elect their own leader to shape the country's political landscape. Overall. Malaysia has undergone 14 general elections and its highest voter turnout is recorded to be 84.8 percent. However when factors such as eligibility to vote and age is considered, it is examined that voting participation is significantly low. Thus this study focuses primarily on Malaysian voters' turnout in the year 2008 and 2013 using decision tree algorithms. The three types of decision trees used are CHAID, CART AND C5.0. From all it is observed that CHAID gave the most accurate results in predicting voter turnout in Malaysian general elections. Apart from this, other algorithms such as Support Vector Machine, Random Forest and Boosting C5.0 are also considered reliable to predict the accuracy of output. While the earlier study used decision trees, we will use neural networks (NN) to capture complex patterns and relationships in data. Neural networks can handle complex data better than decision trees, which may give lower accuracy on complex datasets. Techniques like dropout and regularization in NN help prevent overfitting by reducing the model's dependency on certain features. NN's hidden layers also allow it to learn hidden relationships, improving prediction accuracy.

According to the paper [22], Lasso Regression is a technique that helps make predictions while keeping the model simple. They began by selecting important factors like public sentiment in June and any scandals involving the president. To find the most important factors, they used two techniques: forward regression, where they added one factor at a time until the prediction couldn't improve further, and backward regression, where they started with all factors and removed the least important ones. Both methods highlighted similar factors, such as public sentiment and campaign spending. They then adjusted a parameter, called lambda, to balance the weight of each factor. Testing their model with data from past elections, they found that it accurately predicted outcomes, showing a 41.63% vote share for the incumbent party, indicating that the party would likely lose. The dataset used was complex, covering elections from 1952 to 2016 and including many economic and non-economic factors. Despite the challenges, the researchers identified key factors and built a successful model. They suggested future improvements, like adding factors such as social media activity or voter demographics and exploring more advanced methods like Deep Learning for more precise predictions. In our research, instead of Lasso Regression, we use deep learning models to better understand voting behavior and achieve more accurate results. Since our focus is on a different topic, we changed our survey questions to suit our area of study. We also consulted with political science professors to refine the questions for young voters and used advanced models to predict their voting enthusiasm.

The study "Modeling and Forecasting US Presidential Elections Using Learning Algorithms" [11] aims to create accurate models for predicting US presidential election outcomes using artificial neural networks (ANN) and support vector regression (SVR). It emphasizes the role of data preprocessing, which improves model accuracy by 50%, and uses stepwise regression to find key variables, with the president's approval rate being the most important. SVR and ANN are selected for their ability to handle complex patterns and non-linear relationships. The study uses the tangent hyperbolic sigmoid (Tansig) function as the activation function in ANN models. Findings show that SVR models outperform ANNs and traditional linear regression in predicting the last three US presidential elections (2004, 2008, 2012), proving SVR's superior accuracy. The research highlights the importance of advanced preprocessing and proper variable selection to boost algorithm performance. It concludes that SVR is the most accurate for election forecasting and suggests future research should focus on state-level modeling and further parameter optimization. This study offers valuable insights for policymakers and researchers, showing the effectiveness of learning algorithms in political forecasting. For research on young voter enthusiasm (our research paper), using neural networks and advanced deep learning techniques are recommended to achieve the desired outcomes.

Machine learning is becoming more popular in political science research, helping to improve accuracy, refine measurements, address complex patterns in data, and analyze new types of information. Study [16] shows how machine learning can help us understand voter turnout by applying different methods to survey data from the 2016 election. Research on what affects voter turnout in U.S. elections covers a range of factors: demographic traits like age, race, education, and income; political attitudes and behaviors, such as interest in politics and voting history; institutional factors, like voter registration status; and campaign dynamics, such as the importance of issues. Models like Random Forest and Naive Bayes have been used to predict voter turnout with a focus on both accuracy and understanding the key predictors. These models show results which align with previous research. However, challenges remain, such as making sure models are interpretable, avoiding data biases, and understanding the limits of prediction for explaining cause and effect. Despite these hurdles, machine learning shows promise in enhancing our understanding of voting and informing future research and policy. As a result, for our research, we chose neural networks (NN) for our analysis, focusing on deep learning, which has become widely used due to its power in handling complex data. Furthermore, to improve the models' interpretability, we can use Explainable AI techniques.

According to Chandra and Saini (2021b) [17], social media has become an important platform for information during election campaigns. Platforms such as Twitter and Facebook enable users to express their opinions and share news regarding the election. Traditional methods like surveys and opinion polls can no longer capture voters' preferences in the highly polarized environments of modern elections as seen in the USA's two-party system. Also, traditional approaches can not handle the dynamic changes of the rapidly changing public sentiment. Due to this limitation of the traditional methods, researchers have explored that social media sentiment analysis using Natural Language Processing (NLP) methods would be an alternative way to forecasting election outcomes. These methods have previously been successfully applied to other fields such as consumer behavior, stock market prediction. Recent advances in NLP have enabled more sentiment analysis methods to be used as a predictive tool for election outcomes. Models such as Long Short-Term Memory (LSTM) have widely been used on social media data. Another model that brought substantial improvements to the NLP tasks is the BERT (Bidirectional Encoder Representations from Transformers) model, developed by Google, from the Transformer-based models. Researchers have applied both BERT and LSTM to analyze the social media sentiment during election campaigns and found that they performed very well. For instance, in the 2020 US presidential election, BERT based sentiment analysis predicted a potential Biden victory. Due to its bidirectional nature, BERT can capture contextual data well which increases its ability to predict correctly. Though sentiment analysis works well on social media data, there are limitations to it. For example, not all voters but only a sample of it share their sentiment on social media whereas there are others who prefer not to share their political views on social media. Moreover, due to the Covid 19 pandemic, the data collected in the 2020 election had biases because public sentiment was affected by economic instability and geographical tensions. From this paper, we understand that BERT is a very powerful tool to predict election outcomes by sentiment analysis.

Hence we decided to use it in our research where it will perform sentiment analysis on individuals' comments about the next election and based on the predictions we will get to know if an individual will actually vote or not.

BERT's bidirectional architecture enables it to outperform the results of other traditional NLP models which do text classification by relying on predefined linguistic rules. This is because, according to [21], BERT can work with a huge amount of unsupervised data which can be used to pre-train and can be tuned for specific tasks with minimal labeled data. Traditional approaches like TF-IDF (Term Frequency - Inverse Document Frequency) and n-gram models showed better performance on smaller datasets but can not handle language intricacies in large datasets. Previous studies have shown that classical approaches cannot handle small datasets well whereas BERT excels. For example, upon applying BERT on IMDB dataset, an accuracy of 0.9387 was obtained which was far better than the results obtained from Support Vector Classifiers and Logistic Regression. In comparative experiments, it has been found that BERT outperformed TF-IDF based classifiers in diverse text classification tasks. Apart from this, in the RealOrNot tweet experiment, BERT gave an accuracy of 0.8361 where the task was to differentiate between tweets about real disasters and general tweets. In the same task, H2O's AutoML gave an accuracy score of 0.77607 which showed the power of BERT in binary classification with noisy text data. When BERT was used in a Portuguese news experiment, it was seen that it achieved accuracy comparable to the results that was obtained from English text. BERT achieved an accuracy score of 0.9093 whereas Gradient Boosting Classifier scored 0.8480 during text classification using nine classes of Portuguese news articles. Similarly, BERT scored 0.9381 in the text classification using Chinese sentences which was obtained from Santiago's hotel review dataset. All these shows that BERT not only provides greater accuracy but also makes the training process simpler and can also perform well in multilingual contexts. In the future, Bayesian hyperparameter optimization can boost the accuracy of BERT even more. From this paper, we have got the idea that BERT can be used for the task of sentiment analysis in our research as it can outperform other traditional approaches even in multilingual contexts.

In this paper [12], a research has been conducted by using K-means clustering with elbow method to find the best cluster of customer profiles. Customer segmentation is a commonly used approach to understand the customer behavior and characteristics. Clustering techniques are used to find patterns between customer data which can be used to improve the marketing strategies and CRM (Customer Relationship Management). The K-means clustering algorithm is widely used for this task where different clusters are formed. An initial number of clusters (k) is taken and the data points are assigned to the nearest cluster centroids, iteratively updating the centroid until the cluster stabilizes. However, the initial choice of centroid can heavily impact the algorithm's performance and especially can result in inconsistent clustering in large datasets. To deal with this, the Elbow Method is used where for different values of k it plots the SSE (Sum of Squared Errors) and finds out the point where adding more clusters does not decrease the SSE substantially. This point is the elbow point from where the optimal number of clusters is found out. For customer segmentation, the optimal number of clusters was 3 as calculated by the elbow method. From this research, we figured out that for clustering our data points we can use the elbow method and K-means clustering so that we can get the optimal number of clusters. This will allow us to explain each point in a cluster easily using Explainable AI.

For text analysis, LIME (Local Interpretable Model-agnostic Explanations) has become a known technique for understanding complex, "black-box" models. As mentioned [19], LIME changes parts of the input data (like omitting some words or altering them in a sentence) and sees how these changes affect the model's predictions. This process helps to identify which specific words influence the model's decision the most. LIME calculates the "cosine distance" between two feature sets to measure how similar or different two samples are. LIME uses this distance to weigh the importance for different samples in the analysis. Samples with small distances are called close samples and are given more importance because they're considered more relevant to the input. An exponential function is used to find out this weight which makes nearby samples more significant. LIME uses a regularized linear regression model which includes a penalty to keep the explanation simple and avoid overfitting to noise. The bandwidth parameter controls how LIME reacts to feature changes. A higher bandwidth means more distant points influence the explanation, while a lower bandwidth keeps it focused on nearby points. The research finds important findings about the performance of LIME's surrogate model when a number of features are removed and the weighting method is used. When there are many features or the settings have low bandwidth, larger sample size is required, and it is observed that the model becomes stable with increasing sample size. Furthermore, the coefficients of the surrogate model are linear to the original model, making it easier to interpret across different models.
From this paper, we have got the idea that LIME can be a useful tool in understanding the non-political factors influencing young voter enthusiasm. LIME can alter the input data, observe their effects on a model's predictions, and can help us identify which factors play a significant role in voter enthusiasm.

The paper [13] examines how different ways of picking out important details from text can impact the accuracy of sentiment analysis, particularly when using Twitter data. It compares two common methods: TF-IDF and N-Grams. To get the tweets ready for analysis, the researchers cleaned and organized the data using several steps. These included breaking down sentences into words (tokenization), making everything lowercase (normalization), simplifying words to their base forms (stemming), removing common but irrelevant words (like "the" or "is"), and filtering out unnecessary symbols and noise. They used the SS-Tweet dataset, which includes tweets marked as positive, neutral, or negative. To determine the sentiment of these tweets, they tried out six machine learning models: Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), and Random Forest. They found that TF-IDF performed better than N-Grams, improving accuracy by 3-4%. Of all the models tested, Logistic Regression stood out as the top performer, scoring the highest on accuracy, precision, recall, and F-Score. The main motive of this study was to understand that TF-IDF is a better approach for this type of text analysis and Logistic Regression is the most dependable algorithm. It is very important to understand how crucial it is to choose features and models to get the best output. Therefore, from the findings of this paper, we decided to use TF-IDF to form the vector from the corpus for sentiment analysis.

Chapter 3

Background Analysis

3.1 Multilayer Perceptron

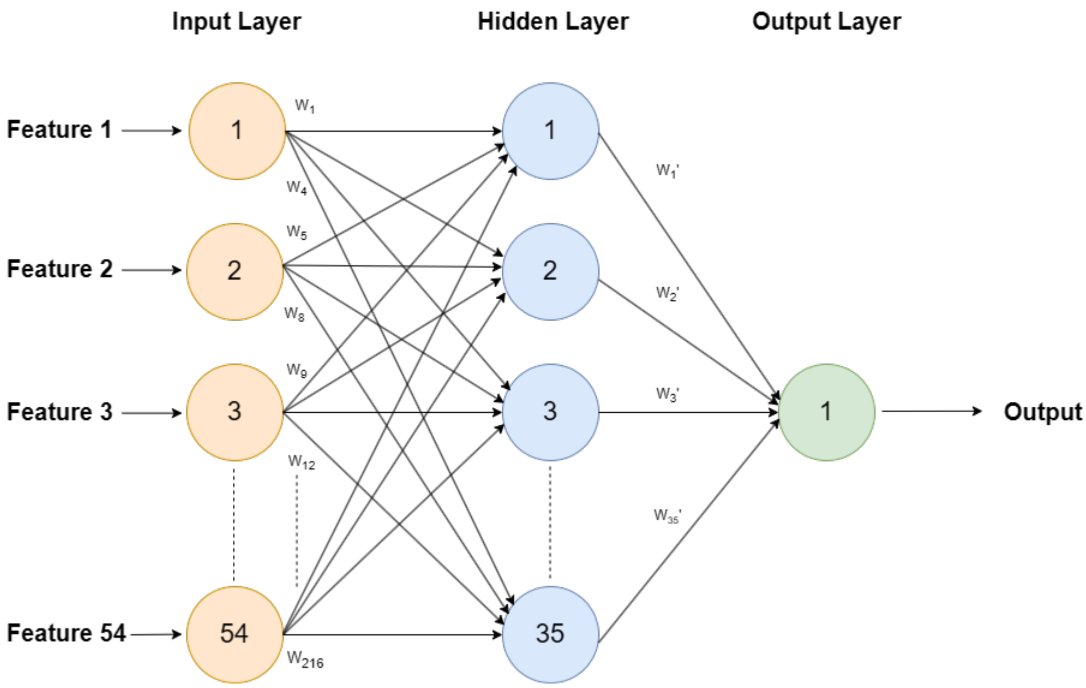

Figure 3.1: Structure of the Multilayer Perceptron Model

A Multilayer Perceptron (MLP) is a type of artificial neural network used for classification and regression tasks. As one part of our research is to classify if an individual will vote or not, we are using MLP for this classification task. The basic structure of MLP consists of an input layer, one or more hidden layers and an output layer. The input layers' neurons take input from features, add weight, bias and propagate them forward to the neurons in the hidden layer and in the same manner from hidden layer to output layer where certain activation functions are applied. During back propagation, the error, calculated from loss function of the output is reduced by updating weights and gradients using optimization algorithms. Forward and back propagation take place in an iterative way up to the epoch value that is being set.

3.2 Random Forest

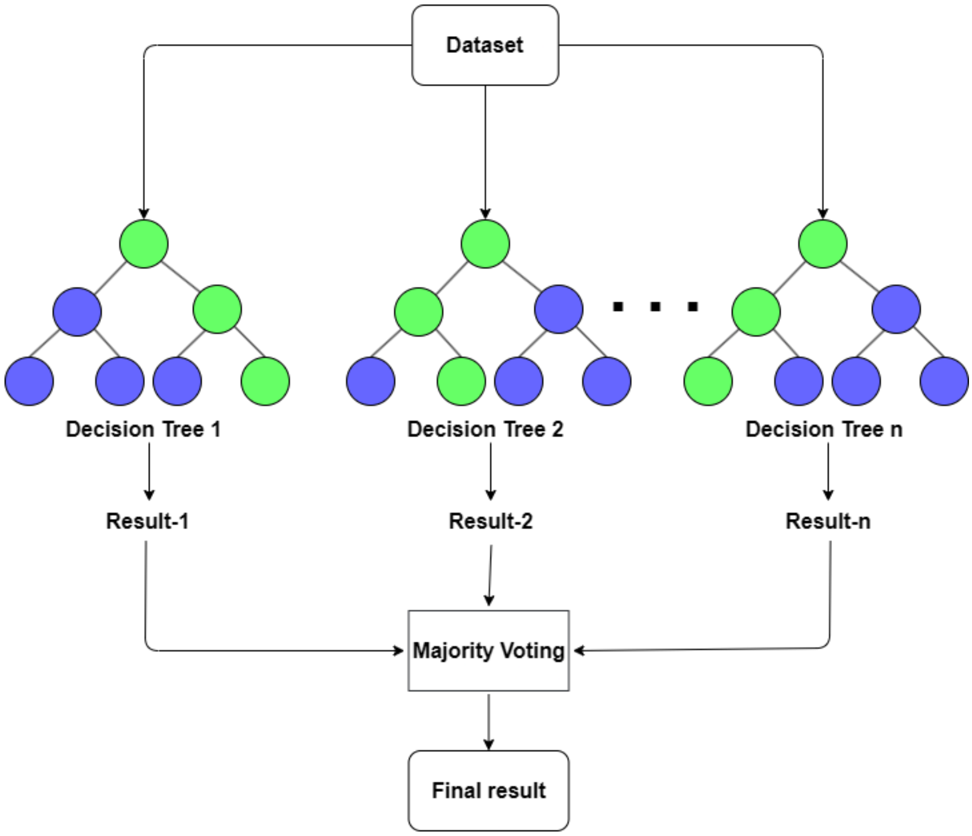

Figure 3.2: Random Forest

Random Forest is a machine learning algorithm that works by combining a collection of decision trees that ensembles the decisions into outputs to achieve better predictions and more accuracy compared to one single model. The multiple trees are trained on different sets of data to prevent overfitting, thus making the model more robust. By using "bootstrapping", random forest trains each tree on a specific subset of features. This ensures that each tree is unique and diverse. In cases of classification, the final output is based on the majority of votes and for regression, all the decision trees are averaged and that average prediction is the final outcome. Random forest is used to handle complex datasets and for its robustness.

3.3 Support Vector Machine

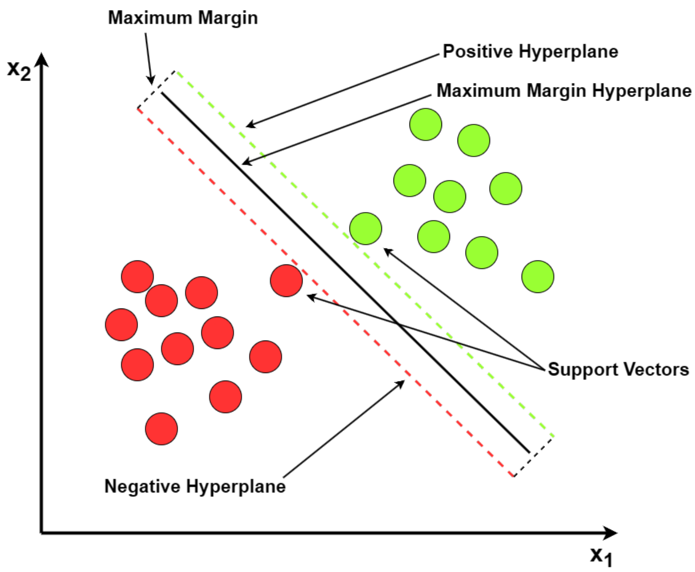

Figure 3.3: SVM with Linear Kernel

Support Vector Machine (SVM) is a supervised machine learning algorithm that is used for classification and regression tasks. It can effectively handle high-dimensional data and works by finding the hyperplane (optimal separating boundary) that differentiates between data of distinct classes. Out of all possible hyperplanes, SVM chooses the line that has the maximum distance between the boundary and the points that are nearest to each class. These closest points are called support vectors and are crucial when it comes to defining the boundary. For non-linear data, SVM uses the kernel trick (a technique which transforms the data into a higher dimensional space) [1] which makes the non linear data linearly separable in the new space without the need of calculations for higher dimensions.

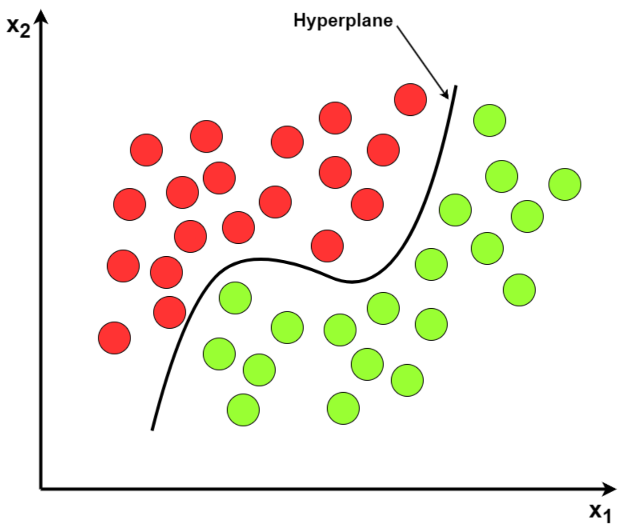

Figure 3.4: SVM with Polynomial Kernel

3.4 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a powerful and highly efficient machine learning algorithm which merges several weak models' predictions to get a stronger prediction. The algorithm sequentially builds an ensemble of decision trees, where each tree works to correct the errors made by the previous tree [8]. It minimizes the loss function iteratively where it calculates the errors (residuals) at each iteration and trains a new tree to minimize these residuals. In other words, weights are applied to all the independent features and then a decision tree predicts the result. The weights which are not correctly predicted by the first tree are fed into the next decision tree. The final prediction is made by summing up all the individual contributions of all the trees. A general formula for the final prediction can be expressed as:

$$\hat{y}_i = \sum_{t=1}^{n} \eta f_t(x_i) \tag{3.1}$$

Here,

- $\hat{y}_i$ represents the final prediction,
- η is the learning rate,
- n represents the total number of trees,
- $f_t(x_i)$ is the output of the t-th tree.

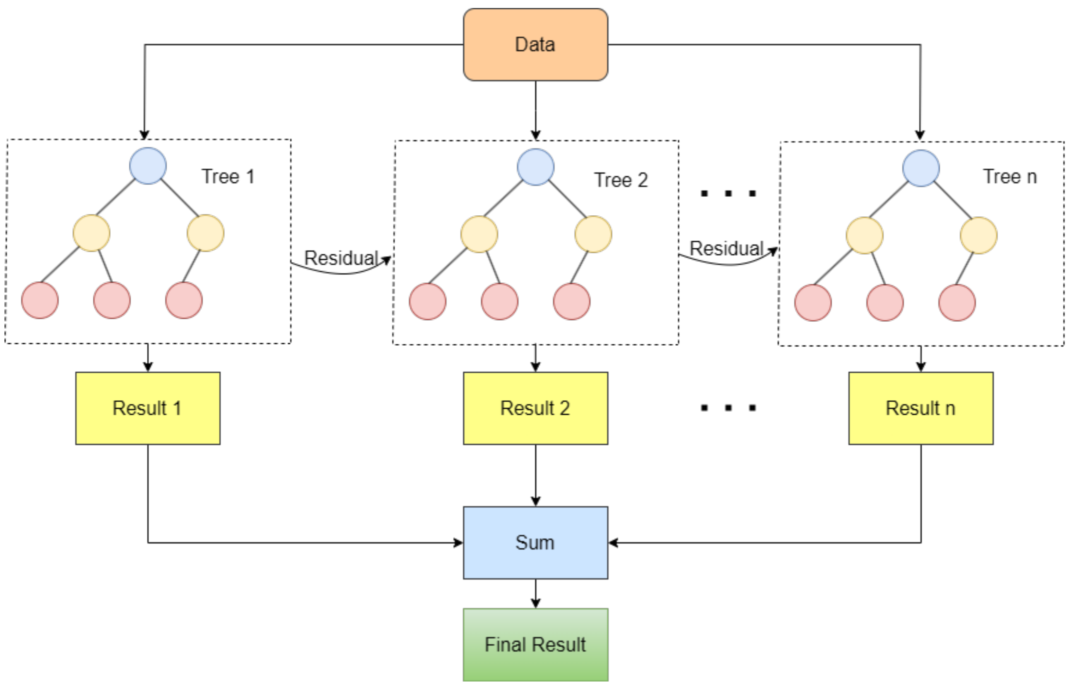

Figure 3.5: XGBoost Model

3.5 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a transformer based machine learning model. Developed by Google, BERT can perform NLP tasks such as sentiment analysis, classification of texts etc. Unlike traditional models, BERT can process text bidirectionally which means that when it finds the context of a target word, it simultaneously looks at the two words which are to the immediate left and right of the target word [14].

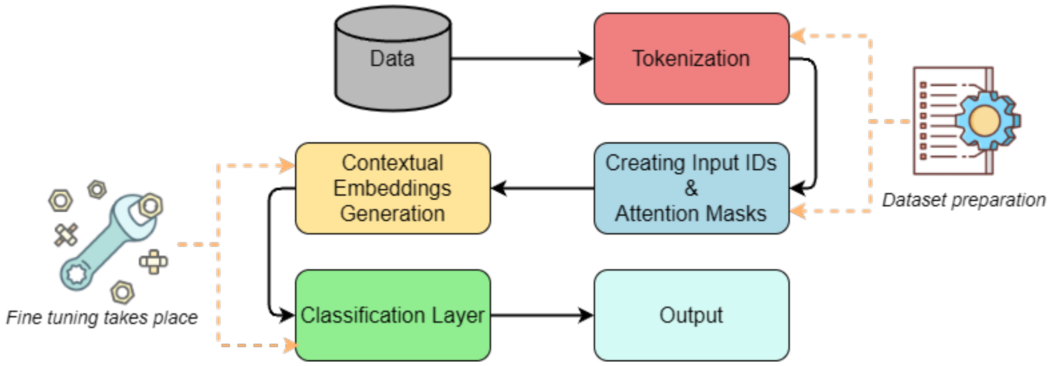

Figure 3.6: Workflow of BERT

The BERT base model, unlike BERT large, has 12 transformer layers, 768 hidden units, and 12 attention heads which enable BERT to efficiently encode the contextual information of the input text. These transformer layers process the input texts bidirectionally to understand the context. The text classification starts with tokenization using the Bert Tokenizer which converts textual data into numerical representation. In the process of tokenization, each text sample is tokenized to small tokens which are then converted into numerical token IDs and attention masks. After that, a dataset class is used which organizes the tokenized inputs and their corresponding labels making sure that the data is ready for training and testing. Then the tokenized inputs, along with the attention masks, are passed to the transformer layers to generate contextual embeddings. These embeddings are then passed through the classification layer to produce outputs. The *BertForSequenceClassification* class is used for the classification layer, which adds classification heads on top of the transformer layers. The heads help to predict the target labels based on the encoded input text. For fine-tuning the BERT base model, during training, the *Trainer*class is used which saves the parameters, like epochs, batch size etc., of the best-performing models. Finally, the fine-tuned model generates the outputs.

3.6 Creating the Ensemble Model

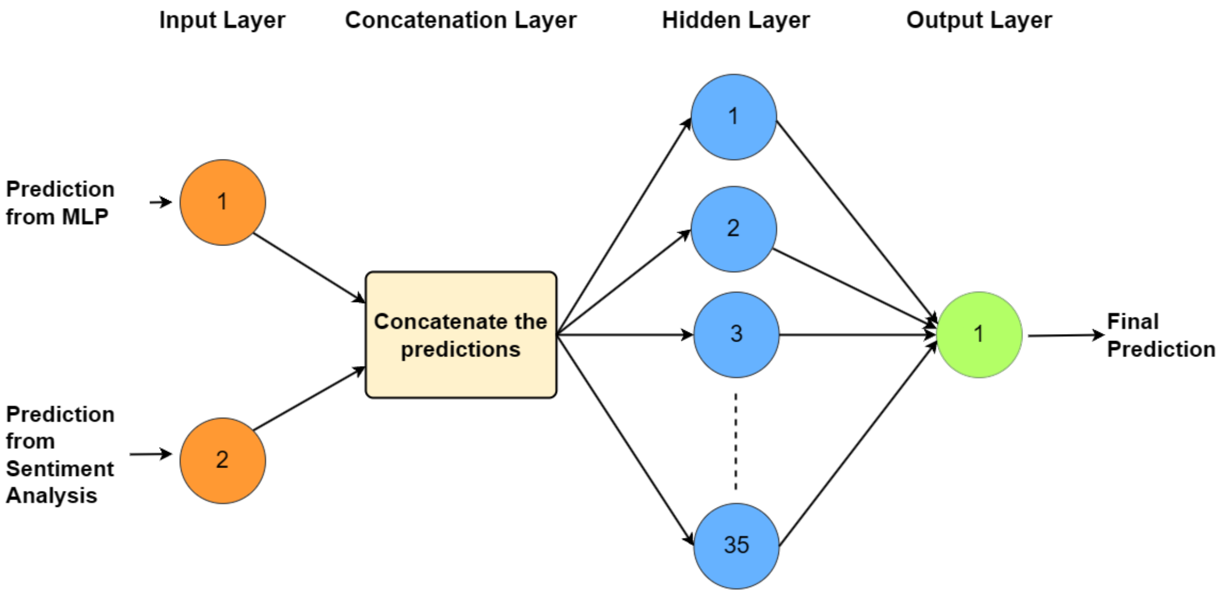

Figure 3.7: Structure of the Ensemble Model

Our main aim in designing the ensemble model is to combine the predictions of MLP and sentiment analysis model. The goal of this approach is to produce a more accurate final prediction. Firstly, the two models' predictions are provided as inputs to the input layer. Then the concatenation layer concatenates the two inputs into a single vector. This process is made possible by using the Functional API in Keras, which gives flexibility to define connections between layers. After that the merged inputs are passed into a hidden layer (contains 35 neurons in our case), which is necessary for the model to learn complex relationships between the combined predictions. The neurons in the hidden layers apply ReLU activation function to extract complex patterns. Lastly, the output of the hidden layer is passed to the output layer. The output layer consists of one neuron (because this is a binary classification task) and uses the Sigmoid activation function to produce the final prediction of whether a young voter will vote or not.

3.7 LIME (Local Interpretable Model-agnostic Explanations)

LIME is an explainable AI technique which is used to explain the predictions made by black-box machine learning models [9]. It uses a simple interpretable model to explain the model's prediction locally.
Firstly, we have to choose the instance from the dataset for which we want to understand the prediction. Then, LIME creates multiple perturbed samples (slightly different versions) of the selected instance by changing the feature values. The black-box model uses these perturbed instances to obtain predictions. After that, higher weights are applied to the perturbed samples which are closer to the original instance. This step ensures that the explanation focuses on the local behavior of the model. A simple interpretable model is then trained using the perturbed samples and

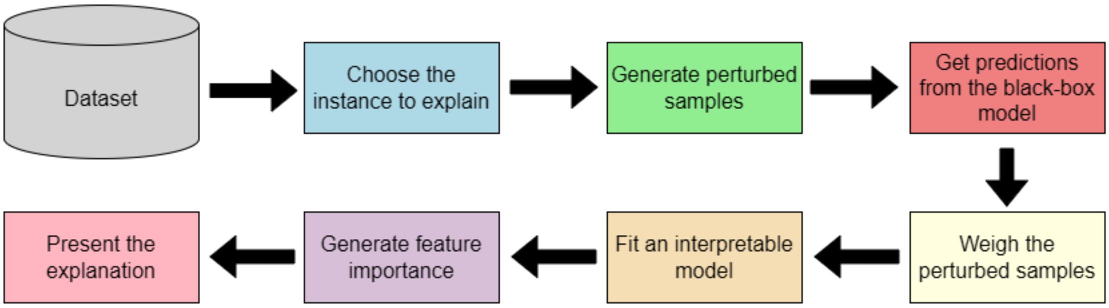

Figure 3.8: Workflow of LIME

their corresponding predictions. Using the interpretable model, the most important features and their contributions are identified. The output of LIME is then visualized with the key features contributing to the prediction.

3.8 SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) is a framework which is used to interpret and explain the predictions of black-box models. It uses the approach of game theory where the final outcome is the result of the contribution of all the players [10]. For example, it assigns a "SHAP value" to each of the features by considering all the feature combinations (subsets of the features). SHAP can not only provide local explanations for a single instance, but also provide global explanations for an overall model.

Firstly, the data, model, and the predictions are inputted. Using these, the baseline prediction is calculated. Baseline prediction, the average prediction of the model across all the instances of the dataset, is necessary because it acts as a reference with which all the features' contributions can be compared. Then subsets of features are created by changing some features' values. After that, for each subset, marginal contribution is computed by comparing the difference between predictions obtained with and without a specific feature in a subset. Finally, the contributions from each feature are combined together to calculate the overall importance in terms of SHAP values.

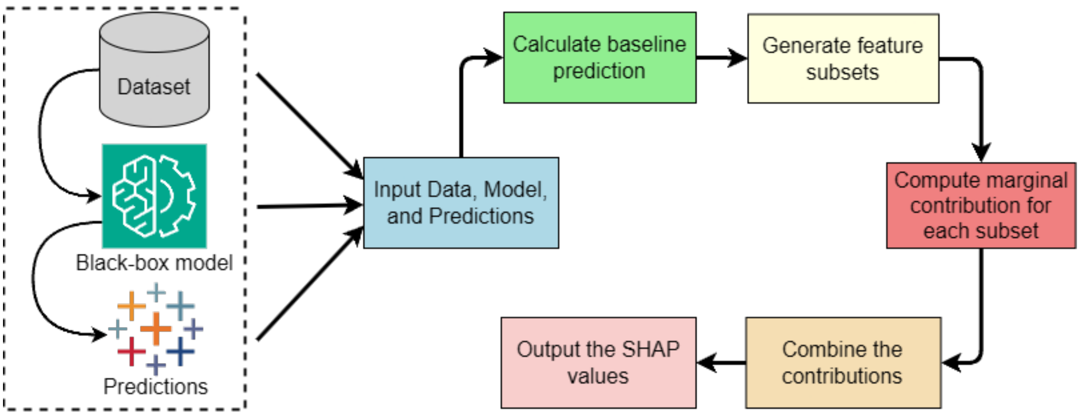

Figure 3.9: Workflow of SHAP

3.9 PCA (Principal Component Analysis)

Principal Component Analysis (PCA) is a dimensionality reduction technique in machine learning. PCA transforms higher dimensional data into lower dimensional data by preserving the variance as much as possible. Principal components are identified, which are linear combinations of the original features, and these components capture the maximum variance of the dataset.

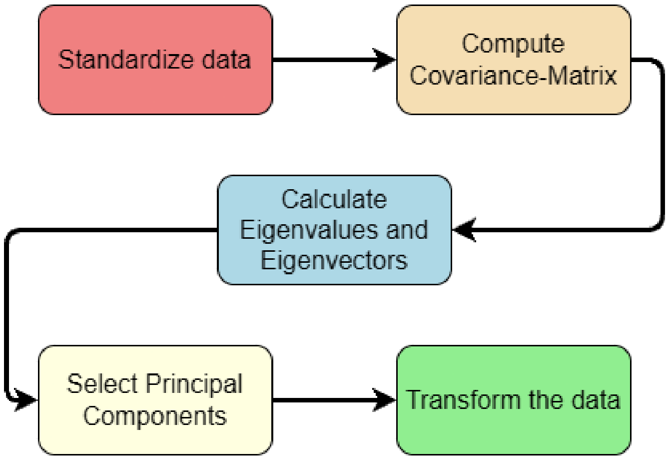

Figure 3.10: Workflow of PCA

The PCA process has several steps. At first, the data is standardized to ensure all the features' contributions are equal. Next, the relationship between the features are measured by calculating the covariance matrix of the standardized data. After that, the eigenvalues (the amount of variance each component captures) and the eigenvectors (the direction of the component) of the covariance matrix is computed to identify the principal component. Lastly, the data is transformed using the selected principal components into a new space with reduced dimension.

Chapter 4

Methodology

4.1 Workflow

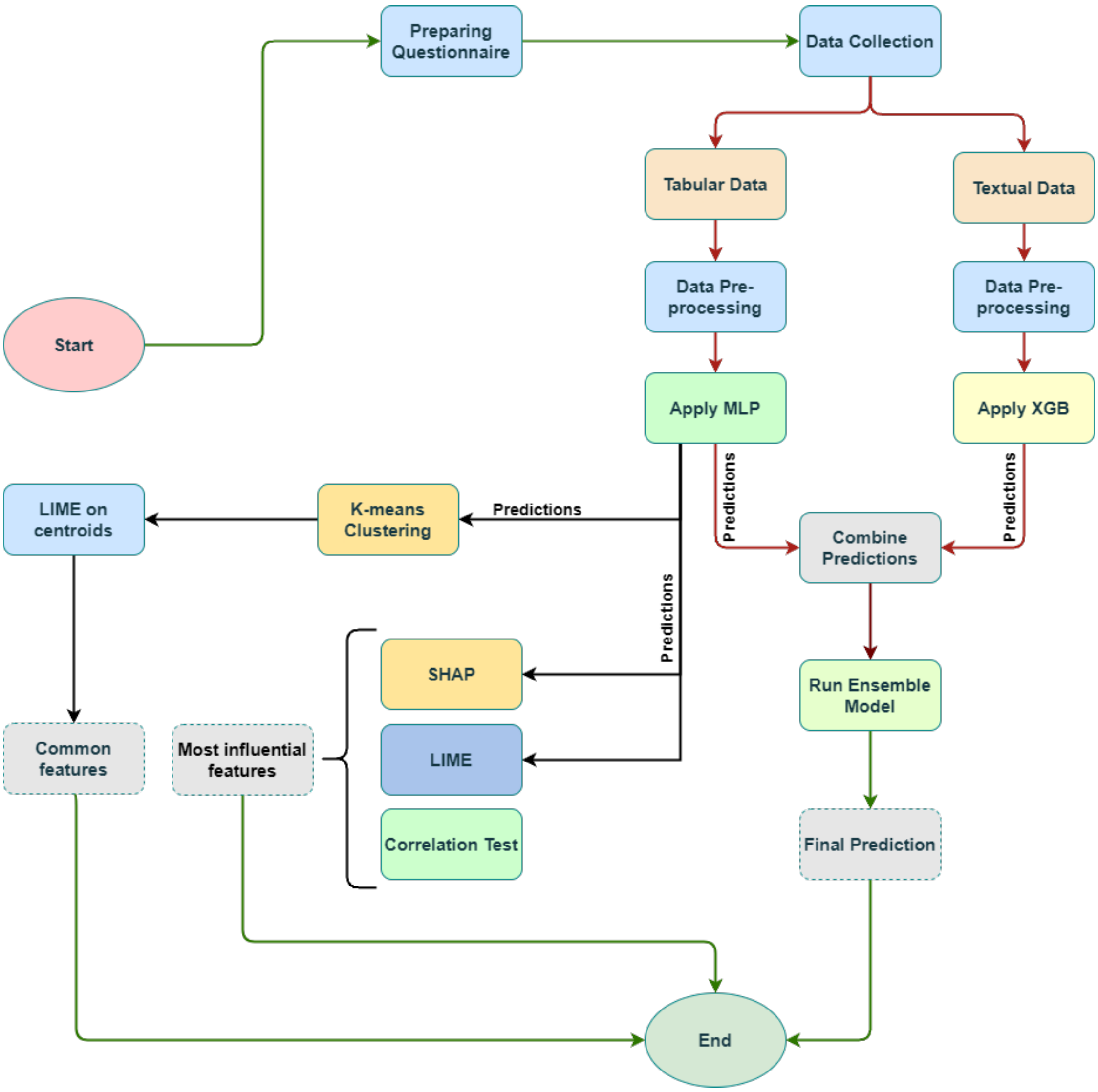

Figure 4.1: Workflow

4.2 Data Collection

Having a dataset is the most important thing in research, because it lets us backup our ideas with real evidence, test hypotheses, and draw meaningful conclusions. However, due to an inadequacy of previous work on this particular topic, we are using primary data which we have collected from only the students of BRAC University using Google Forms. In our Google form, we have prepared 36 questions in total, where 34 were MCQs and the remaining two were short questions. For this research, we have successfully collected more than 1000 data from the students.

4.3 Description of the data

The questions revolved around different factors like personal demographic, educational qualification of the guardians, financial stability, previous voting experience and socio-economic perspectives. Firstly, educational qualifications can determine whether he/she is aware of the voting system or not. Usually more educated individuals are more likely to participate in voting. However, the opposite is also present. So, this question will help us determine if an individual's educational background has any influence on his or her voting behavior. Participants were also asked about their eligibility to vote in the most recent election. If the person was eligible, it can either mean the person was at the age of participating in a voting procedure legally or can mean was simply eligible but was disinterested to vote. If he or she was not eligible, it can mean the person simply was not of the proper age to vote. Also, if the person abstained from voting in the most recent election, we aim to figure out if it was their own will or were they simply not eligible.
Learning about one's occupation will help us determine whether the person will vote or not. This is because laborers are at higher risk of financial losses due to centers remaining closed during the day of voting. Monthly salaried workers or students may choose to vote or not vote. One other factor such as living with family can mean the person is not financially independent yet, the person can also tell if his/her family has previous voting experience or not; if otherwise, the occupation of the person can be known. If a person lives in their own house, it may mean the person is financially stable. However, staying in a rented house may mean that the person may not be financially stable even if he is a property owner. This may mean the person may or may not vote, depending on his or her will. Also, a relation can be established between the influence of parents and an individual's voting behavior. This will reflect if living with family or independently is the cause behind one's motivation to vote or not vote. Moreover, peer pressure from conservative/male dominated families can act as a hindrance for females to vote.
Aside from the factors above, the area the person chooses to vote in, whether it is their hometown or current area of living also exert their influence on voting. If the person wants to go to their hometown, it can mean the person has a higher dedication to participate. The questionnaire asked if the person is willing to vote in the immediate next election, with which we want to analyze if he or she is eligible and wants a fair share of experience as a new voter.

Furthermore, parents' educational backgrounds can determine whether the father is aware of the voting system, the procedures and its need to shape the nation which can naturally influence the person to vote. However, there may not be a positive correlation between a father's education and its influence on a child's voting behavior. A mother being highly educated in a family can also convey that the family is academically qualified because in countries like Bangladesh, higher education of females is still considered a taboo in many regions. If parents are further enthusiastic about voting, then there can be possibilities that their children are also enthusiastic about voting and vice versa. This question will help establish a connection between a guardian's educational background and their enthusiasm. Family members' voting can concurrently cause peer pressure and also interest and knowledge about voting. It can also convey that if they did not vote or talk about politics much in households, the person may naturally not be interested in voting or have knowledge of it. Someone close, such as a relative who is involved in politics, might impose positive or negative thoughts on an individual's mind about voting.

About political awareness, a person well affiliated with ongoing affairs of the system or politics themselves might have higher chances of participating in the voting process. The reverse can also apply as knowledge of corruption/dictatorship in the world of politics might grow a sense of cynicism in the person's mind and they might pull away from attending it. Adverse experience in the last election can make the person indifferent towards the upcoming election. Besides, if the person did not participate in the last one, they may tend to be unenthusiastic about the next one. Previous voting experience means the person knows about its system and was eligible to vote if yes. If no, we will ask if it is because he or she was not eligible or was it his or her own will. It can also lead to reluctance to vote if the experience was bad.

Apart from the above-mentioned, factors such as how the participants spend their leisure time, subscription towards online news channels, awareness of the political landscape nationally and globally, and engagement on social media were taken to scrutinize their political standpoint and speak of their awareness about the country's parliamentary and judicial systems. More use of social media might substantially increase the chance of getting and hearing news related to elections which in turn can influence a person to vote or to not vote. With more people switching to digital devices nowadays, reading newspapers may mean the person still has traditional approaches of being informed about the current world affairs, regularly reading newspapers can reflect one's interest in staying informed about a country's system. Besides newspapers, if a person stays updated about news through social media platforms, it can reflect a person being more enthusiastic about political affairs and is interested to learn more. On the flip side, if a person chooses no, it may mean the person considers social media not a suitable platform for news updates as they might be forged or not convey the current information. A relationship can be established between this answer and the amount of time this person spends on social media. Additionally, watching television news and staying up to date with online news can multiply their awareness about the country's political and judicial systems. Television news channels tend to broadcast news that are verified and thus substantially influence the lives of its daily viewers (citizens of the country). Hence if a person chooses "yes", it can mirror the person being more politically aware or knowledgeable about the ongoing affairs of the country. This can determine one's interest to vote or simply keeping aloof from it. As mentioned before, considering online news as a source of news shows the person has a good amount of information about the political system. If not, it may mean the person does not trust online news as a valid source or simply relies on other sources like newspapers or television as sometimes they may display false information.

To analyze the results more accurately, the questionnaire also asked if the person is willing to go and vote on the day of election if it is a holiday. If the person chooses yes, it may mean he or she went to vote simply because they have a day off, so it was merely for the sake of an experience or to see how voting works. Or, they are genuinely interested to vote. If the answer is no, then it shows the person is not interested in voting. They were then asked if they have a very hectic schedule or are busy with their personal affairs, would they participate in the election. If the person chooses yes, it can reflect that he or she shares a good amount of interest about participating in the election amid workload. If not, it can be either because the person is simply not interested or could not participate due to time constraints. In addition, the questionnaire asked about their participation amid political turmoil. Pre-election violence can lead to voter apathy about the immediate next election. They might also refuse to vote at all as they consider it a risk to their lives. On the contrary, the voter might still go and vote if he or she is used to seeing it in their area.

Despite violence on the election day, if the person goes and votes, it reflects on the importance they place on participation and the value of exercising their right to vote, even in challenging circumstances. Fear of clashes can make a person indifferent towards the voting system. However, if the answer is yes, it would show that the person is very much interested in voting. If the answer is no, it would show that the person is not interested in voting. Both the answers might be influenced by some factors like peer pressure, previous experience, economic or financial situation etc.

Voter indifference is a major reason behind one's lack of participation. One might think their voting won't make a difference if a certain party has strong reign or superiority over others in a particular area, so it is kind of pre-decided who will win. Also, there can be other factors like peer pressure which can make someone think that their vote will or will not make a difference.

The last two questions were opinion based and using one of these we will perform sentiment analysis. The survey asked if there is any additional factor that would affect their voting behavior. This might reflect as to what can be the other evident reasons for the person being willing or unwilling to vote. Lastly, they were asked to share their views on what they think about the immediate next election and the answer to this particular question will be used for sentiment analysis. Based on the comments, we can sense if the person is looking forward to voting in the next election or is simply indifferent.

4.4 Data pre-processing

After successfully collecting data from the participants, we began the pre-processing steps to prepare the data for analysis. But before pre-processing, we split the dataset into two parts: tabular dataset and textual dataset. Initially, the dataset had 1003 entries. However, after careful examination, the size of the datasets reduced to 869. This reduction was due to the removal of the rows containing irrelevant responses (e.g. "N/A", "No comment", random emojis and alphabets) or null values in response to the question "*In a few words, comment on what you think about the next election.*".

Additionally, we have done a renaming process for the feature labels. The original feature names were long (the questions in the questionnaire) and was impacting readability during the processes. To address this issue, we shortened the names to more concise labels. For example, the feature originally labeled "*What is your age?*" was renamed simply as "*Age*", and similarly all other labels were renamed. This renaming step provided a clearer format for analysis.

In order to handle the categorical data effectively, we applied several encoding techniques. For binary categorical features with options "Yes" and "No", we used a binary encoding scheme, mapping "Yes" to "1" and "No" to "0". This transformation has helped us transform the data to numerical values which are required as input for machine learning models.

For categorical features with more than two values, we used One-Hot Encoding. This approach created additional binary columns for each category, ensuring that each instance was represented by only 0s and 1s. Specifically, for the feature "Age", we first categorized the ages into three distinct age groups (**Age1**: *19 to 21*, **Age2**: *22 to 23*, and **Age3**: *greater than 23*) to provide a clear grouping based on age ranges. After that , we applied One-Hot Encoding to represent these age groups as binary features. This transformation helped in correctly preserving the differences between age groups.

We know that the tabular data-frame consists of 54 features in total along with the class label, and the textual data-frame contains the texts and the class label. However, as our research is multi-modal, we still have to follow some pre-processing steps to prepare the textual data for sentiment analysis. During cleaning, we noticed that there were answers written in Banglish. Hence, we translated the answers to English.

To perform sentiment analysis, it is required to convert the text data into suitable numerical format. We have used Python's Natural Language Toolkit (NLTK) to download English stop words and applied some techniques to clean the text data. Each sentence was converted to lowercase and using regular expressions we have removed the non-numeric characters. Then each word of the sentences was converted to its root form using the Porter Stemmer and the common stop words were also removed. The resulting text data was stored in a *corpus*.

To convert the words of the corpus into numerical format, we have applied TF-IDF Vectorizer which creates a matrix (X) by considering both the frequency and significance of words throughout the corpus. The class labels were extracted from the numerical data-frame into another variable (y).

After these transformations, we carefully reviewed the datasets for any redundant or unnecessary columns which might add noise to the analysis. We removed these columns to ensure only the necessary relevant features were present. Lastly, to ensure that the data was evenly distributed, we shuffled the datasets.

4.5 Addressing Class Imbalance

Upon checking the distribution of positive and negative classes in the class label ("*Are you willing to vote in the immediate next election?*"), we observed an imbalance. Specifically, the ratio of 0 to 1 was around 0.398:0.601 which indicated that the dataset was skewed towards the positive class. This imbalance can make the models biased towards the majority class, resulting in a lower accuracy for the minority class predictions. To address this issue, we applied SMOTE (*Synthetic Minority Oversampling Technique*) to the training data.

Before oversampling	
0 (No)	0.398159
1 (Yes)	0.601841

Table 4.1: Ratio of 0 to 1 in the class label before oversampling

After oversampling	
0 (No)	0.5
1 (Yes)	0.5

Table 4.2: Ratio of 0 to 1 in the class label after applying SMOTE

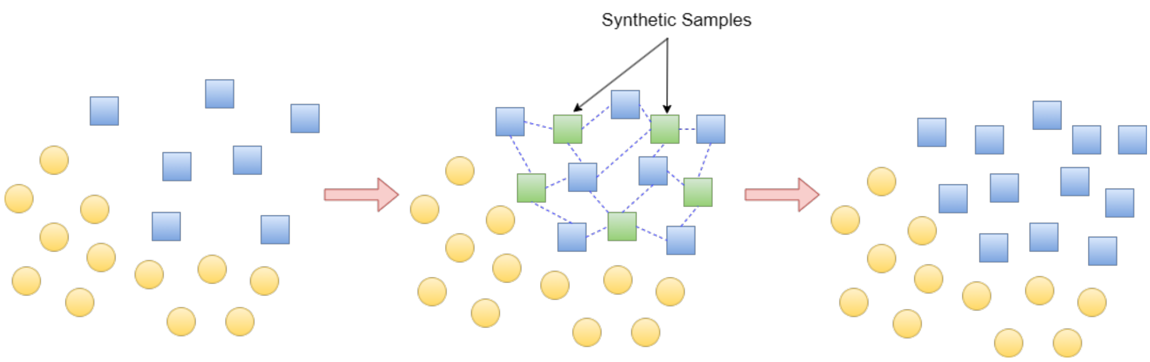

Figure 4.2: Synthetic Minority Oversampling Technique

SMOTE is an oversampling technique designed to balance class distribution by generating synthetic samples for the minority class. Instead of duplicating, SMOTE creates new artificial data by interpolating between the samples of the minority class. At first it identifies pairs of close data points in the minority class and then generates synthetic samples along the line segment connecting them. This ensures that the new data points increase the number of minority samples and preserves the original data distribution. As shown in the figure, SMOTE creates new synthetic samples (green squares) between the existing samples (blue squares). This technique was applied after we split the data into training and testing sets, ensuring only the training data was oversampled.

4.6 Sentiment analysis

Sentiment analysis is a technique in NLP which is used to identify sentiments or opinions in texts. The sentiments can be positive, negative, or neutral. Earlier methods such as Naive Bayes have been replaced by new approaches like BERT which can understand language contexts better and result in higher accuracy. In our research, we will be performing sentiment analysis on the comments made by the participants about the next election. At first we will compare the accuracies between several machine learning models (such as Random Forest, XGBoost, SVM, BERT-base) and then select the best model for the sentiment analysis task. The chosen model will be used to get predictions for each instance of the test-set.

Chapter 5

Model Implementation

5.1 Model evaluation

5.1.1 Multilayer Perceptron

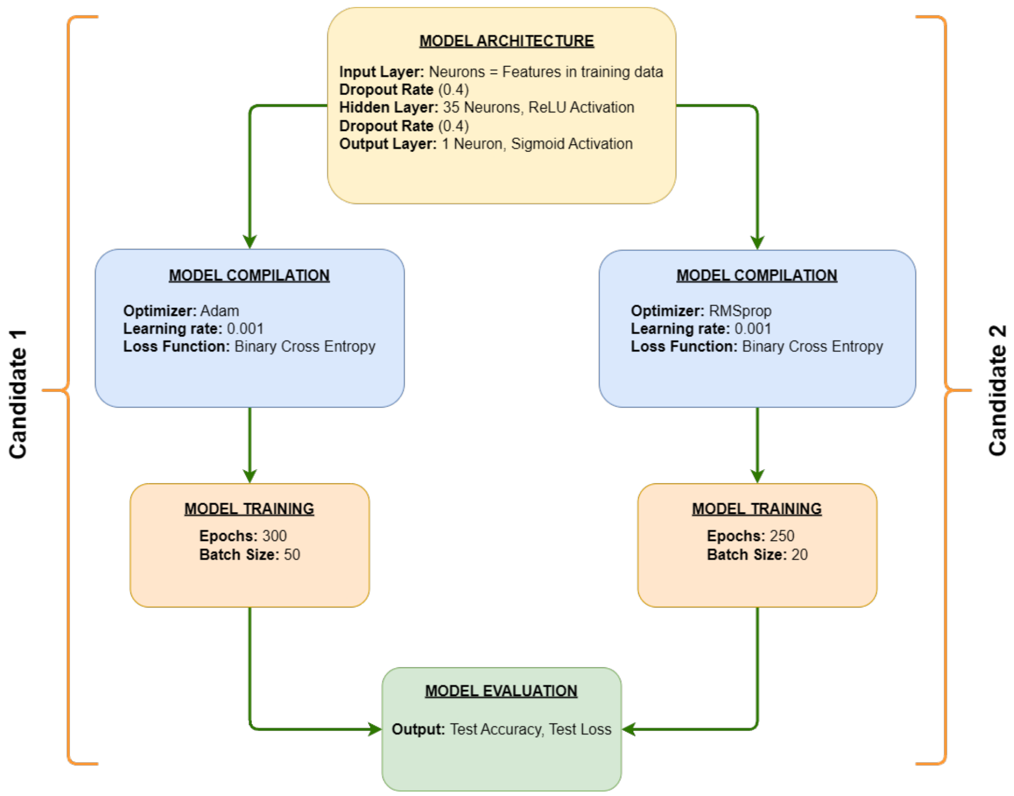

Figure 5.1: Workflow of the candidates of MLP

In our research, we used GridSearchCV several times to find out the optimized hyperparameters. From the tuned hyperparameters, we figured out two candidate models whose performance and test accuracy were the best among all the other candidates. Both candidates have 54 neurons in the input layer matching the size of the features in the dataset, 35 neurons in the hidden layer activated by ReLU, 1 output layer activated by Sigmoid activation function, and binary cross-entropy as the loss function with a learning rate of 0.001. To reduce overfitting, two dropout layers are being used with a dropout rate of 0.4 in both the candidates. However, the key differences between candidate 1 and candidate 2 are in the type of Optimizer, the value of epoch and the batch-size which are Adam, 300, 50 for candidate 1 and RMSprop, 250, 20 for candidate 2 respectively.

5.1.2 XGBoost for Sentiment Analysis

To determine the most effective model for sentiment analysis, we evaluated a set of models: BERT-base, Random Forest (RF), XGBoost (XGB), and Support Vector Machine (SVM) with both linear and polynomial kernels. All the models were trained on the matrix (X) which was obtained after applying TF-IDF on the corpus. However, for the BERT-base model, TF-IDF was not used, as BERT processes raw text data directly. BERT creates tokens and utilizes the embedding process on its own which allows it to capture semantic relationships.

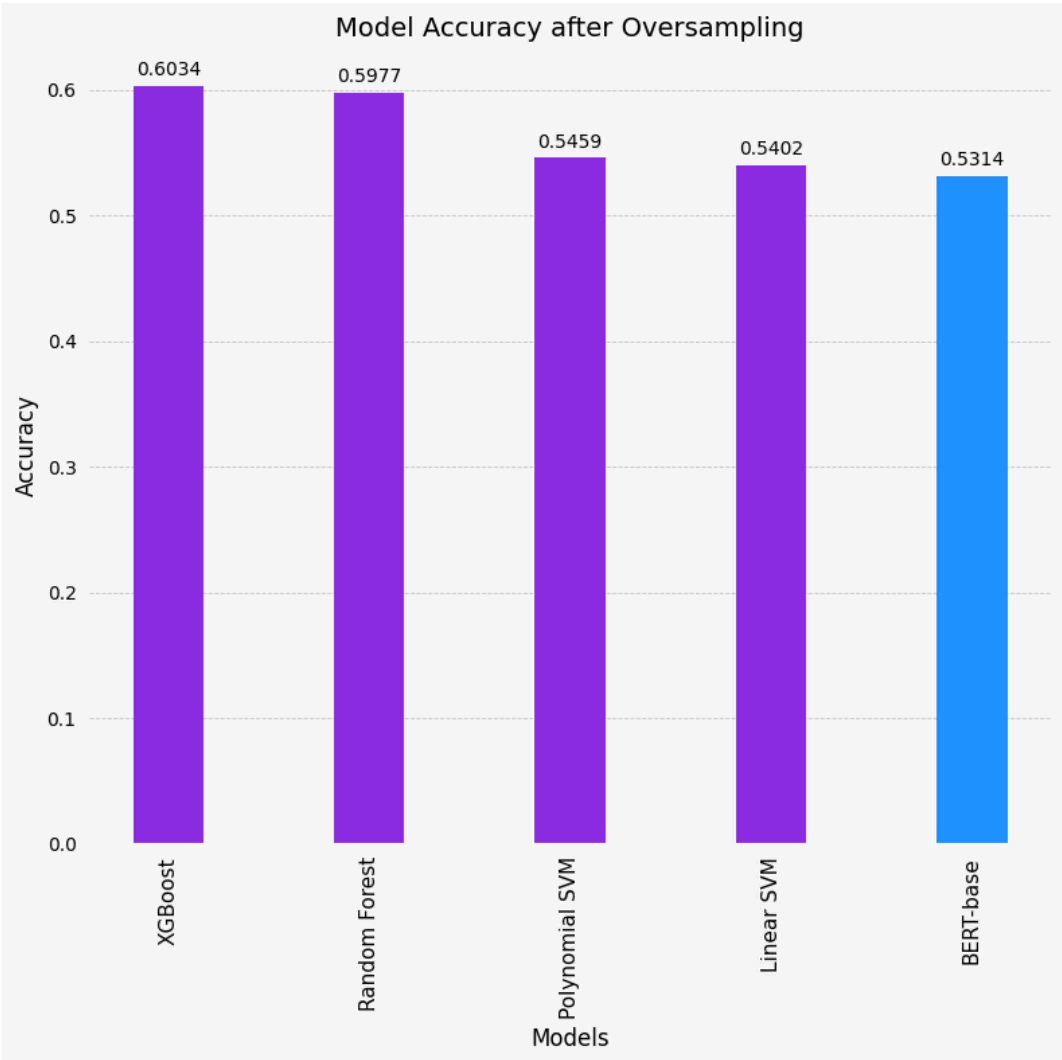

Figure 5.2: Comparison of the accuracies of the NLP models

The bar chart 5.2 shows the accuracy of the NLP models applied after oversampling the dataset using SMOTE. Among the models tested, XGBoost achieved the highest accuracy at 60.34%, indicating its strong performance in handling complex patterns in the data. Random Forest closely followed with an accuracy of 59.77%. Moreover, for the SVM models, Polynomial SVM slightly outperformed Linear SVM with accuracies of 54.60% and 54.02%, respectively. Finally, the BERT-base model achieved an accuracy of 53.14%, which is lower than the other NLP models. This result may be due to the fact that BERT does not use the TF-IDF transformation applied to all the other models. Another reason for BERT's lower performance could be the small size of our dataset as these transformer models are trained on large amounts of data. However, transformer models also face challenges due to the limitation of extracting the exact sentiment from complex expressions, and adapting to new types of data.

Chapter 6

Result Analysis

6.1 Results from Multilayer Perceptron

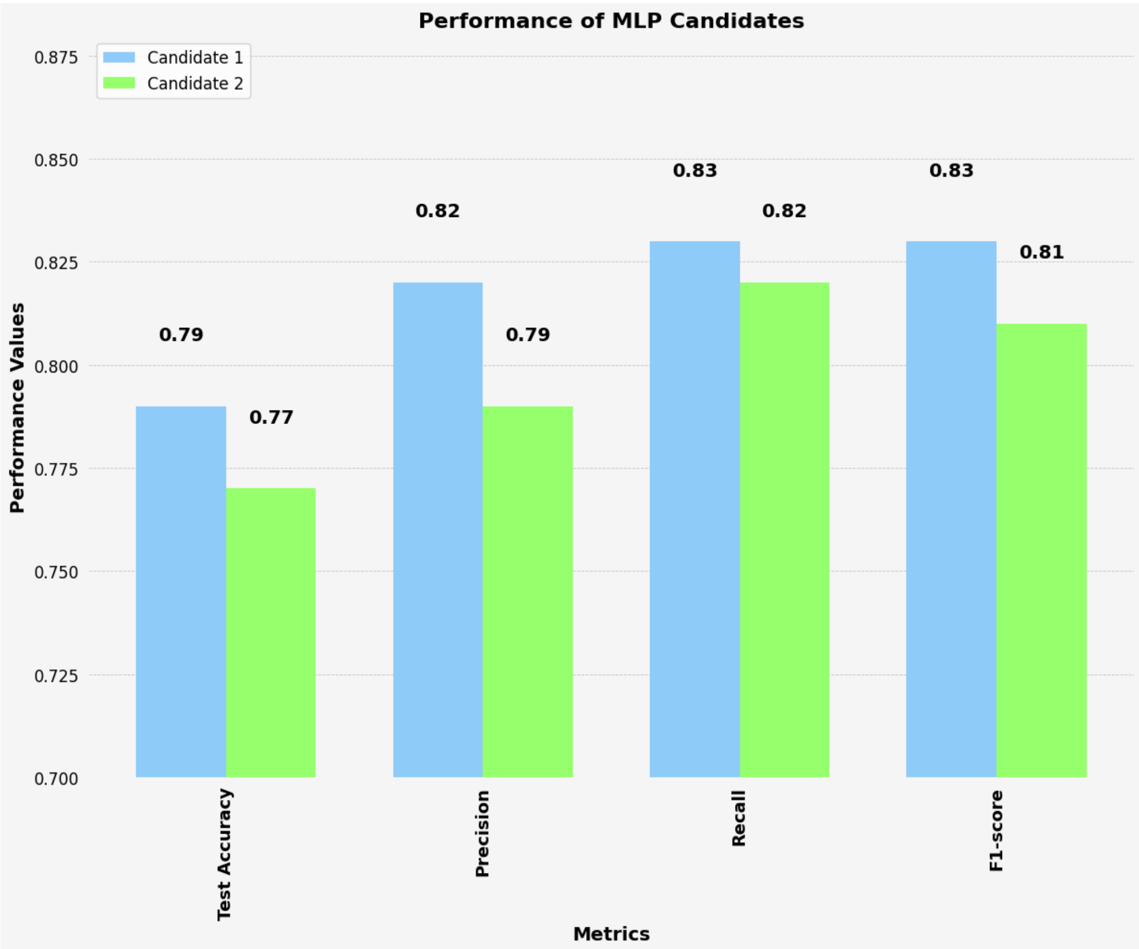

Figure 6.1: Performance of the MLP candidates

Our aim is to reduce the test loss and increase the test accuracy of the model. Upon testing on the tabular dataset, we observed that both the candidates outperformed. The test accuracy (the ratio of correctly classified samples) of candidate 1 is slightly higher than that of candidate 2 whereas candidate 2 has higher precision. Candidate 1 achieved an accuracy of 0.79 which is greater than Candidate 2's accuracy (0.77).

In terms of precision (the number of true positive predictions among all the positive predictions), candidate 1 outperformed which indicates that candidate 1 is better at minimizing false positives. Similarly, candidate 1 showed a higher value for recall (the ratio of true positives to all the actual positives), indicating it can identify positive instances more effectively. Lastly, the F1-score of candidate 1 (0.83) is greater than candidate 2's 0.81.

According to the results, candidate 1's performance is better as it outperformed in the metrics which are essential for accurate and balanced predictions. Therefore, we decided to use candidate 1 as the MLP model for the classification using tabular data.

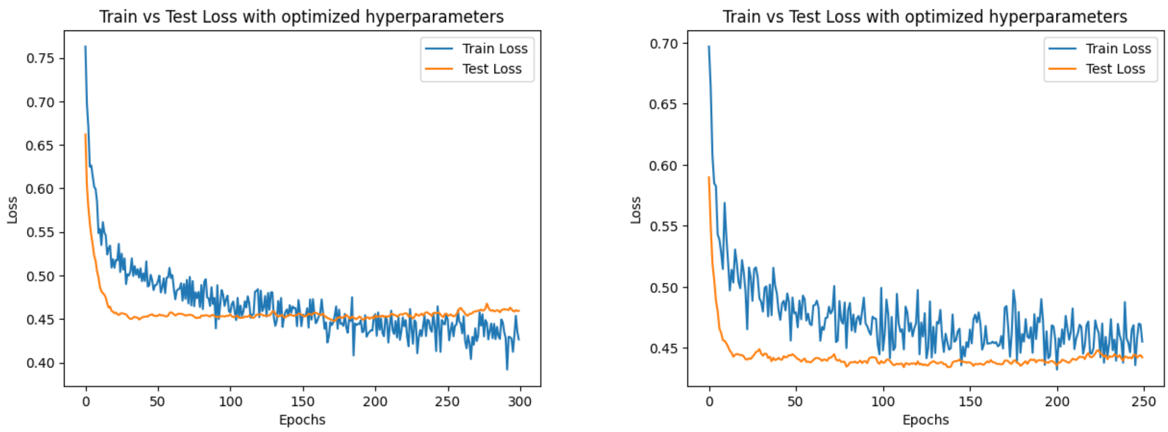

Figure 6.2: Train vs Test Loss Graphs for Candidate 1 and Candidate 2 respectively

6.2 Analysing feature importance

6.2.1 Correlation

The main objective of our research is to find out the important features affecting young voter enthusiasm. Therefore we have used correlation testing to see how each feature is related to the true labels of the initial dataset. Tables 6.1 and 6.2 show the correlations.

The feature "Day off" has the highest correlation with prediction (0.57), which means that if a person has a day off, they will most probably go to vote, showing their eagerness to participate in the election. Comparatively, the feature "Television" has a weak positive correlation (0.14), implying that watching news updates from television does not fully convince the voter to participate in voting. From positive correlations, the feature "Owning a business" has a weak correlation (0.05), meaning that a young voter's enthusiasm to vote hardly depends on their occupation such as owning a personal business.

The feature "Supporting democratic notion_Not supportive" has the highest negative correlation (-0.17) which means that if a young voter is not supportive about the democratic notions, they will most likely refrain from participating. Another feature "Time on socials_I don't use social media" has a weak negative correlation(-

0.08), implying that if no time is spent on social media, a young voter's enthusiasm to vote is low. The weakest negative correlation is found in the feature "Father education_HSC / A levels" (-0.01) which means that if a person's father's highest educational qualification is HSC or A levels, their enthusiasm to vote is the lowest.

Feature	**Correlation**
Day off	0.57
Vote makes difference	0.40
Parent enthusiasm	0.38
Busy schedule	0.28
Lack of transport/far	0.27
Family's recent vote	0.27
Pre-election violence	0.26
Mother vote representation_Encourages you to vote	0.25
Father vote representation_Encourages you to vote	0.25
Voted in last election	0.25
Election day violence	0.19
Voting satisfaction_Moderately satisfied	0.17
Newspaper	0.14
Television	0.14
Effect of having political member_Their involvement moderately influences me to vote	0.13
Effect of having political member_Their involvement strongly influences me to vote	0.09
Online news	0.09
Social media	0.07
Supporting democratic notion_Moderately supportive	0.06
Own house	0.06
Info about parliamentary system_Strongly informed	0.05
Occupation_Own a business	0.05
Financial stability	0.04
Occupation_No	0.03
Father education_SSC / O levels	0.03
Gender	0.02
Eligible in last election	0.01
Mother education_Primary School	0.01
Time on socials_>3 hours	0.01
Mother education_Masters / PhD	0.01
Member in politics	0.00
Info about judicial system_Strongly informed	0.00

Table 6.1: Positive Correlation of Features

Feature	**Correlation**
Supporting democratic notion_Not supportive	-0.17
Effect of having political member_Their involvement has made me disinterested in voting	-0.17
Voting satisfaction_Never voted before	-0.15
Father vote representation_Neutral	-0.14
Info about parliamentary system_Not at all informed	-0.14
Info about judicial system_Not at all informed	-0.11
Mother vote representation_Neutral	-0.10
Political awareness_Not at all aware	-0.10
Time on socials_I don't use social media	-0.08
Age_Age3	-0.08
Voting satisfaction_Highly unsatisfied	-0.06
Occupation_Monthly salaried worker	-0.06
Age_Age2	-0.04
Political awareness_Strongly aware	-0.04
Mother education_SSC / O levels	-0.03
Reside with family	-0.02
Mother education_HSC / A levels	-0.02
Current city	-0.02
Time on socials_<1 hour	-0.01
Father education_Masters / PhD	-0.01
Father education_Primary School	-0.01
Father education_HSC / A levels	-0.01

Table 6.2: Negative Correlation of Features

6.2.2 SHAP

In order to see the global explanations behind the MLP model's predictions, we have used SHAP and obtained the following results as shown in the table 6.3.

Table 6.3: Feature Importance Based on SHAP Values

Feature	**mean(|SHAP value|)**
Day off	0.09
Vote makes difference	0.07
Parent enthusiasm	0.03
Pre-election violence	0.03
Lack of transport/far	0.02
Mother vote representation_Encourages you to vote	0.02
Busy schedule	0.02
Family's recent vote	0.02
Voted in last election	0.02
Info about parliamentary system_Not at all informed	0.02
Voting satisfaction_Moderately satisfied	0.02
Father vote representation_Encourages you to vote	0.01
Voting satisfaction_Highly unsatisfied	0.01
Age_Age2	0.01
Member in politics	0.01
Election day violence	0.01
Occupation_Monthly salaried worker	0.01
Age_Age3	0.01
Father education_SSC / O levels	0.01
Effect of having political member_Their involvement moderately influences me to vote	0.01
Supporting democratic notion_Not supportive	0.01
Info about judicial system_Not at all informed	0.01
Occupation_Own a business	0.01
Online news	0.01
Time on socials_<1 hour	0.01
Supporting democratic notion_Moderately supportive	0.01
Time on socials_>3 hours	0.01
Mother education_Masters / PhD	0.01
Effect of having political member_Their involvement has made me disinterested in voting	0.01
Mother education_HSC / A levels	0.01
Other 24 features	0.09

According to SHAP, the top two most influential features are: *Day off* and *Vote makes difference*. In other words, when an individual gets a day off (has free time) on the day of election or when an individual thinks that his vote would make a difference, these two scenarios contribute mostly to the decision of whether the decision will be *Will Vote* or *Will Not Vote*. Moreover, *Parent enthusiasm* and *Pre-election violence* have the same level of contribution to the final prediction. From these results, we cannot directly infer in which direction (positive or negative prediction) a particular feature is putting more importance, but these SHAP values help us to identify the overall global importance of each feature on the MLP predictions.

6.2.3 LIME

In order to see the local explanation of the predictions of the MLP model, we have used LIME. To find out the feature importance on the individual voting behavior, we have separated the MLP predictions into two datasets: a dataset containing the correctly classified *Will Vote* predictions and another dataset containing the correctly classified *Will Not Vote* predictions. In both the prediction datasets, we applied LIME and observed the explanations for each feature. To calculate the contribution, we summed each feature's LIME explanations for all the instances of the datasets. For the *Will Vote* dataset, if the summation equaled a positive number, we labeled it as WILL VOTE and vice versa. Similarly, for the *Will Not Vote* dataset, if the summation equaled a positive number, we labeled it as WILL NOT VOTE and vice versa. The results are shown in the tables 6.4, 6.5, 6.6, and 6.7.

Table 6.4: Positively influencing features from the LIME explanations of "WILL VOTE" dataset

Feature	**Influence**	**Contribution**
Lack of transport/far	Will Vote	2.827017
Voted in last election	Will Vote	2.060544
Father education_Masters / PhD	Will Vote	1.858267
Mother vote representation_Neutral	Will Vote	1.844057
Political awareness_Not at all aware	Will Vote	1.546433
Occupation_No	Will Vote	1.420636
Info about judicial system_Not at all informed	Will Vote	1.270031
Television	Will Vote	0.771476
Online news	Will Vote	0.753023
Info about parliamentary system_Strongly informed	Will Vote	0.672338
Father vote representation_Neutral	Will Vote	0.604528
Political awareness_Strongly aware	Will Vote	0.5771
Info about parliamentary system_Not at all informed	Will Vote	0.574516
Mother education_Masters / PhD	Will Vote	0.530048
Supporting democratic notion_Not supportive	Will Vote	0.507168
Age_Age3	Will Vote	0.440447
Newspaper	Will Vote	0.391798
Mother education_SSC / O levels	Will Vote	0.366949
Mother education_Primary School	Will Vote	0.196791
Info about judicial system_Strongly informed	Will Vote	0.102734
Parent enthusiasm	Will Vote	0.0478
Father education_SSC / O levels	Will Vote	0.030398
Occupation_Own a business	Will Vote	0.006532

From the explanations of the *Will Vote* dataset (tables 6.4 and 6.5), we see that the feature with most contribution is if a person will go and vote if there is a lack of transport on the day of election or if the voting station is far from home. The second most contributing factor influencing towards will vote is if a person has voted in the last election. Apart from these, we can see that mother's neutral voting representation also acts as a positively influencing factor according to the LIME explanation. Other factors like considering newspaper and television as the source of news also contributes towards voting. On the other hand, having family members involved in politics reduces the likelihood of voting, influencing individuals towards not voting. Similarly, if there is pre-election violence or if someone is highly unsatisfied with past voting experience, then that person is likely to abstain from voting.

From the explanations of the *Will Not Vote* dataset (tables 6.6 and 6.7), we see that a person is unlikely to vote in the city where they currently live. Also, those who are

Table 6.5: Negatively influencing features from the LIME explanations of "WILL VOTE" dataset

Feature	**Influence**	**Contribution**
Gender	Will Not Vote	-15.866127
Time on socials_>3 hours	Will Not Vote	-6.632886
Busy schedule	Will Not Vote	-2.716221
Member in politics	Will Not Vote	-2.045502
Reside with family	Will Not Vote	-1.930497
Effect of having political member_Their involvement moderately influences me to vote	Will Not Vote	-1.871561
Mother vote representation_Encourages you to vote	Will Not Vote	-1.674407
Current city	Will Not Vote	-1.1751
Election day violence	Will Not Vote	-1.04384
Pre-election violence	Will Not Vote	-0.930412
Vote makes difference	Will Not Vote	-0.892402
Father vote representation_Encourages you to vote	Will Not Vote	-0.727468
Effect of having political member_Their involvement has made me disinterested in voting	Will Not Vote	-0.686775
Voting satisfaction_Never voted before	Will Not Vote	-0.548906
Occupation_Monthly salaried worker	Will Not Vote	-0.425605
Financial stability	Will Not Vote	-0.397282
Mother education_HSC / A levels	Will Not Vote	-0.344232
Social media	Will Not Vote	-0.314452
Family's recent vote	Will Not Vote	-0.310061
Voting satisfaction_Moderately satisfied	Will Not Vote	-0.276839
Effect of having political member_Their involvement strongly influences me to vote	Will Not Vote	-0.247279
Father education_Primary School	Will Not Vote	-0.240518
Age_Age2	Will Not Vote	-0.229706
Voting satisfaction_Highly unsatisfied	Will Not Vote	-0.16382
Day off	Will Not Vote	-0.139181
Father education_HSC / A levels	Will Not Vote	-0.12018
Supporting democratic notion_Moderately supportive	Will Not Vote	-0.055919
Eligible in last election	Will Not Vote	-0.053941
Time on socials_<1 hour	Will Not Vote	-0.030793
Time on socials_I don't use social media	Will Not Vote	-0.018715
Own house	Will Not Vote	-0.003104

not at all informed about the judicial system or the parliamentary system, according to the LIME explanations, will abstain from voting. Moreover, if someone close is involved in politics, an individual does not show interest to go and vote. People are also unlikely to vote if they consider online news and social media as their source of news about the country. On the other hand, family's recent vote acts as a factor which influences towards voting. Similarly, when parents are enthusiastic about voting, it is likely that children will be motivated to vote in the election.

Table 6.6: Negatively influencing features from the LIME explanations of "WILL NOT VOTE" dataset

Feature	Influence	Contribution
Family's recent vote	Will Vote	-1.502441
Own house	Will Vote	-1.251249
Age_Age2	Will Vote	-1.064212
Supporting democratic notion_Not supportive	Will Vote	-0.979986
Day off	Will Vote	-0.876931
Mother vote representation_Encourages you to vote	Will Vote	-0.838581
Busy schedule	Will Vote	-0.768331
Parent enthusiasm	Will Vote	-0.601999
Mother education_SSC / O levels	Will Vote	-0.459413
Pre-election violence	Will Vote	-0.437147
Time on socials_>3 hours	Will Vote	-0.423077
Financial stability	Will Vote	-0.407623
Time on socials_<1 hour	Will Vote	-0.377222
Father education_Primary School	Will Vote	-0.365091
Mother education_Primary School	Will Vote	-0.258782
Effect of having political member_Their involvement strongly influences me to vote	Will Vote	-0.185337
Effect of having political member_Their involvement moderately influences me to vote	Will Vote	-0.17036
Info about judicial system_Strongly informed	Will Vote	-0.169122
Voting satisfaction_Highly unsatisfied	Will Vote	-0.158357
Info about parliamentary system_Strongly informed	Will Vote	-0.157313
Occupation_Monthly salaried worker	Will Vote	-0.111456
Age_Age3	Will Vote	-0.102205
Voting satisfaction_Moderately satisfied	Will Vote	-0.067941
Father education_SSC / O levels	Will Vote	-0.051303
Voted in last election	Will Vote	-0.047942
Political awareness_Strongly aware	Will Vote	-0.045885
Occupation_No	Will Vote	-0.040704
Lack of transport/far	Will Vote	-0.020407

Table 6.7: Positively influencing features from the LIME explanations of "WILL NOT VOTE" dataset

Feature	Influence	Contribution
Gender	Will Not Vote	8.91297
Social media	Will Not Vote	3.542327
Info about judicial system_Not at all informed	Will Not Vote	2.740582
Current city	Will Not Vote	2.497865
Reside with family	Will Not Vote	1.298174
Mother vote representation_Neutral	Will Not Vote	1.149554
Mother education_HSC / A levels	Will Not Vote	0.856696
Member in politics	Will Not Vote	0.76161
Political awareness_Not at all aware	Will Not Vote	0.75095
Television	Will Not Vote	0.749422
Mother education_Masters / PhD	Will Not Vote	0.57832
Online news	Will Not Vote	0.550143
Effect of having political member_Their involvement has made me disinterested in voting	Will Not Vote	0.440887
Eligible in last election	Will Not Vote	0.406611
Info about parliamentary system_Not at all informed	Will Not Vote	0.369499
Voting satisfaction_Never voted before	Will Not Vote	0.331919
Father education_HSC / A levels	Will Not Vote	0.299852
Father vote representation_Neutral	Will Not Vote	0.175484
Occupation_Own a business	Will Not Vote	0.146243
Supporting democratic notion_Moderately supportive	Will Not Vote	0.108779
Father education_Masters / PhD	Will Not Vote	0.104103
Newspaper	Will Not Vote	0.0527
Time on socials_I don't use social media	Will Not Vote	0.034305
Vote makes difference	Will Not Vote	0.023557
Election day violence	Will Not Vote	0.013663
Father vote representation_Encourages you to vote	Will Not Vote	0.011168

To know more precisely which factors are influencing towards 'Will vote' or 'Will not vote', we need to take a look at the data points more closely. On each dataset, we applied the elbow method to find out the optimal number of clusters (shown in figures 6.5 and 6.6). As our dataset has many features, we used PCA (Principal Component Analysis) to identify the top two components for visualizing K-means clustering. To achieve this, we plotted one principal component on the x-axis and the other on the y-axis. The figures 6.3 and 6.4 illustrate the clusters.

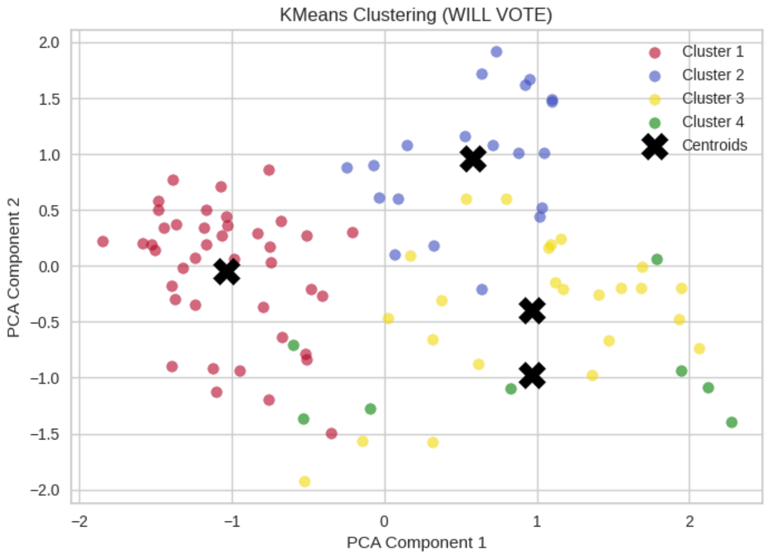

Figure 6.3: Clusters of 'Will Vote' predictions

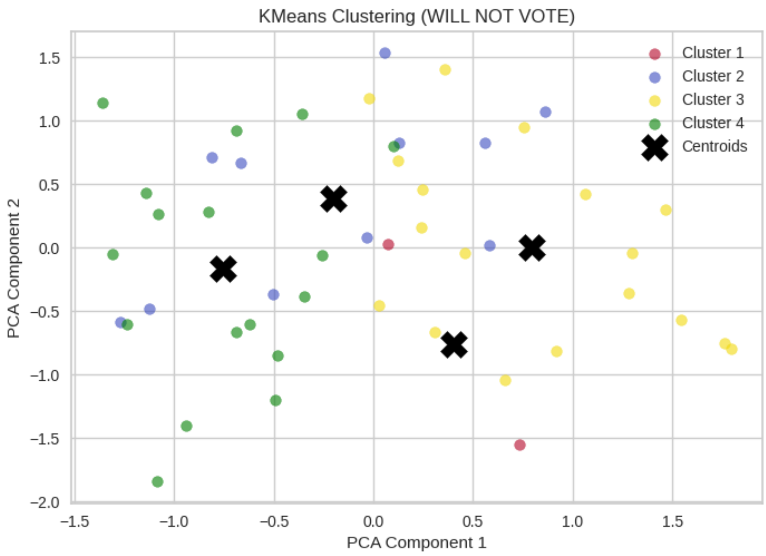

Figure 6.4: Clusters of 'Will Not Vote' predictions

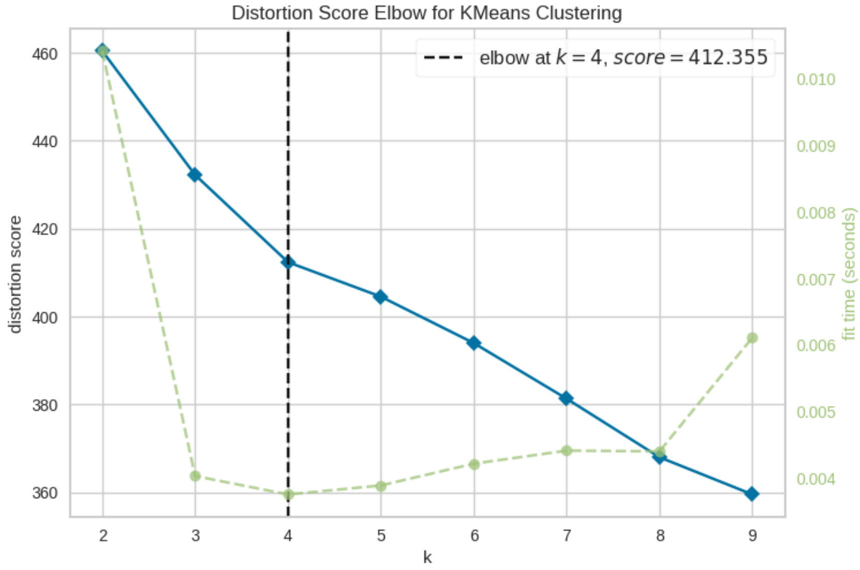

Figure 6.5: Optimal number of clusters for *Will Vote* dataset

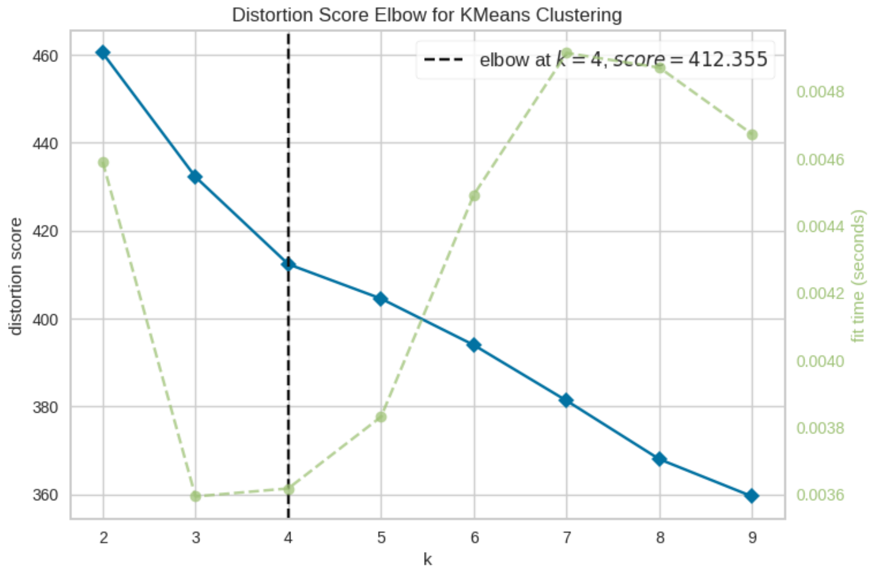

Figure 6.6: Optimal number of clusters for *Will Not Vote* dataset

From the clusters of each dataset, we applied LIME on the centroids, because the centroid would provide us with an idea of the entire cluster.

Table 6.8: Positively influencing features from the LIME explanations of "WILL VOTE" clusters

Feature	**Influence**	**Contribution**
Lack of transport/far	Will Vote	0.119428
Voted in last election	Will Vote	0.101538
Mother vote representation_Neutral	Will Vote	0.090511
Father education_Masters / PhD	Will Vote	0.083415
Info about judicial system_Not at all informed	Will Vote	0.078098
Occupation_No	Will Vote	0.076564
Political awareness_Not at all aware	Will Vote	0.069048
Television	Will Vote	0.045737
Info about parliamentary system_Strongly informed	Will Vote	0.034895
Mother education_Masters / PhD	Will Vote	0.027909
Online news	Will Vote	0.027334
Age_Age3	Will Vote	0.025515
Info about parliamentary system_Not at all informed	Will Vote	0.020305
Political awareness_Strongly aware	Will Vote	0.017902
Newspaper	Will Vote	0.017266
Mother education_Primary School	Will Vote	0.015671
Voting satisfaction_Highly unsatisfied	Will Vote	0.012432
Own house	Will Vote	0.01188
Father vote representation_Neutral	Will Vote	0.007565
Info about judicial system_Strongly informed	Will Vote	0.006237
Parent enthusiasm	Will Vote	0.004656
Mother education_SSC / O levels	Will Vote	0.003256
Supporting democratic notion_Not supportive	Will Vote	0.003166
Eligible in last election	Will Vote	0.002031

From the explanations of the *Will Vote* clusters (tables 6.8 and 6.9), we can see that the factor "Lack of transport/far" has the most positive contribution towards voting. The second most influential factor is if a person voted in the last election. If the person has voted in the previous election then the person is likely to go and vote. Other factors such as parents' enthusiasm also acts as a positively influencing factor towards voting. On the other hand, factors like having a political member in family, and having a busy schedule are more likely to make a person abstain from voting.

From the explanations of the *Will Not Vote* clusters (tables 6.10 and 6.11), we can see that family's recent vote influences a person to vote. Secondly, if a person is an owner of a house, he is more likely to go and vote. Other factors like having a day off (has free time) on the day of election or if there is a busy schedule, then people are more likely to participate in the election. On the other hand, if someone considers social media or television as the source of news, then the person is not likely to participate in election. Also, having a political member in the family demotivates people from voting.

Table 6.9: Negatively influencing features from the LIME explanations of "WILL VOTE" clusters

Feature	Influence	Contribution
Gender	Will Not Vote	-0.701947
Time on socials_>3 hours	Will Not Vote	-0.30862
Busy schedule	Will Not Vote	-0.119809
Effect of having political member_Their involvement moderately influences me to vote	Will Not Vote	-0.092679
Reside with family	Will Not Vote	-0.083858
Mother vote representation_Encourages you to vote	Will Not Vote	-0.077764
Member in politics	Will Not Vote	-0.068115
Election day violence	Will Not Vote	-0.053819
Current city	Will Not Vote	-0.047374
Effect of having political member_Their involvement has made me disinterested in voting	Will Not Vote	-0.047095
Pre-election violence	Will Not Vote	-0.043788
Voting satisfaction_Never voted before	Will Not Vote	-0.042346
Vote makes difference	Will Not Vote	-0.038988
Father vote representation_Encourages you to vote	Will Not Vote	-0.035427
Family's recent vote	Will Not Vote	-0.027973
Occupation_Monthly salaried worker	Will Not Vote	-0.023963
Mother education_HSC / A levels	Will Not Vote	-0.022921
Age_Age2	Will Not Vote	-0.022026
Father education_Primary School	Will Not Vote	-0.019397
Financial stability	Will Not Vote	-0.016234
Social media	Will Not Vote	-0.014984
Effect of having political member_Their involvement strongly influences me to vote	Will Not Vote	-0.014676
Time on socials_I don't use social media	Will Not Vote	-0.014126
Occupation_Own a business	Will Not Vote	-0.010489
Day off	Will Not Vote	-0.007487
Father education_SSC / O levels	Will Not Vote	-0.00632
Voting satisfaction_Moderately satisfied	Will Not Vote	-0.004522
Supporting democratic notion_Moderately supportive	Will Not Vote	-0.004311
Father education_HSC / A levels	Will Not Vote	-0.003605
Time on socials_<1 hour	Will Not Vote	-0.003157

Table 6.10: Positively influencing features from the LIME explanations of "WILL NOT VOTE" clusters

Feature	Influence	Contribution
Gender	Will Not Vote	0.733316
Social media	Will Not Vote	0.298869
Info about judicial system_Not at all informed	Will Not Vote	0.243364
Current city	Will Not Vote	0.211808
Mother vote representation_Neutral	Will Not Vote	0.100197
Reside with family	Will Not Vote	0.088355
Member in politics	Will Not Vote	0.078417
Mother education_Masters / PhD	Will Not Vote	0.077602
Television	Will Not Vote	0.071861
Online news	Will Not Vote	0.070029
Political awareness_Not at all aware	Will Not Vote	0.059614
Mother education_HSC / A levels	Will Not Vote	0.054056
Voting satisfaction_Never voted before	Will Not Vote	0.033454
Effect of having political member_Their involvement has made me disinterested in voting	Will Not Vote	0.030684
Father vote representation_Neutral	Will Not Vote	0.026906
Father education_HSC / A levels	Will Not Vote	0.023657
Supporting democratic notion_Moderately supportive	Will Not Vote	0.023352
Info about parliamentary system_Not at all informed	Will Not Vote	0.021618
Eligible in last election	Will Not Vote	0.021433
Occupation_Own a business	Will Not Vote	0.020112
Father vote representation_Encourages you to vote	Will Not Vote	0.013881
Voting satisfaction_Moderately satisfied	Will Not Vote	0.00773
Father education_Masters / PhD	Will Not Vote	0.004369
Father education_SSC / O levels	Will Not Vote	0.003175
Age_Age3	Will Not Vote	0.002254
Newspaper	Will Not Vote	0.00156
Occupation_Monthly salaried worker	Will Not Vote	0.000336
Election day violence	Will Not Vote	0.00011

Table 6.11: Negatively influencing features from the LIME explanations of "WILL NOT VOTE" cluster

Feature	Influence	Contribution
Family's recent vote	Will Vote	-0.136161
Age_Age2	Will Vote	-0.101523
Own house	Will Vote	-0.092687
Mother vote representation_Encourages you to vote	Will Vote	-0.075892
Supporting democratic notion_Not supportive	Will Vote	-0.074645
Busy schedule	Will Vote	-0.074288
Day off	Will Vote	-0.071171
Pre-election violence	Will Vote	-0.047896
Father education_Primary School	Will Vote	-0.047632
Financial stability	Will Vote	-0.047328
Parent enthusiasm	Will Vote	-0.032555
Time on socials_>3 hours	Will Vote	-0.031756
Mother education_Primary School	Will Vote	-0.026229
Voting satisfaction_Highly unsatisfied	Will Vote	-0.023182
Mother education_SSC / O levels	Will Vote	-0.022193
Time on socials_<1 hour	Will Vote	-0.02153
Info about parliamentary system_Strongly informed	Will Vote	-0.015832
Time on socials_I don't use social media	Will Vote	-0.012564
Vote makes difference	Will Vote	-0.011639
Info about judicial system_Strongly informed	Will Vote	-0.01043
Effect of having political member_Their involvement strongly influences me to vote	Will Vote	-0.009892
Lack of transport/far	Will Vote	-0.009272
Effect of having political member_Their involvement moderately influences me to vote	Will Vote	-0.008485
Political awareness_Strongly aware	Will Vote	-0.0057
Voted in last election	Will Vote	-0.00202
Occupation_No	Will Vote	-0.000254

6.2.4 The Influential Factors

According to the global contribution of features provided by SHAP, we can see that the most influential feature is *day off*. Similarly, the feature *day off* shows the strongest positive correlation with the class label. Also, in the LIME explanations, the feature day off shows higher positive influence in the *Will Vote* dataset and in the *Will Not Vote* cluster. Likewise, the feature *parents' enthusiasm* has a positive contribution according to SHAP value, correlation, and LIME value for the *Will Vote* dataset and Will Not Vote cluster. The feature *lack of transport/far* has a SHAP value of 0.02, correlation of 0.27, and also shows positive influence in all the *Will Vote* and *Will Not Vote* clusters and datasets. The feature *vote makes difference* has a greater SHAP and correlation value, but the lime value shows that there is a negative influence of this feature according to the values of the *Will Vote* dataset, and *Will Vote* cluster.

6.3 Results from Sentiment Analysis

After identifying XGBoost as the top-performing model, we decided to further improve its accuracy through hyperparameter tuning using GridSearchCV (GSCV). GSCV searches the optimized values of each hyperparameter from a given set of values that yields the highest model accuracy. The optimal hyperparameters obtained were:

Hyperparameter	**Optimal Value**
colsample_bytree	0.8
learning_rate	0.2
max_depth	7
n_estimators	57
subsample	1.0

Table 6.12: Optimized Hyperparameters for XGBoost from GridSearchCV

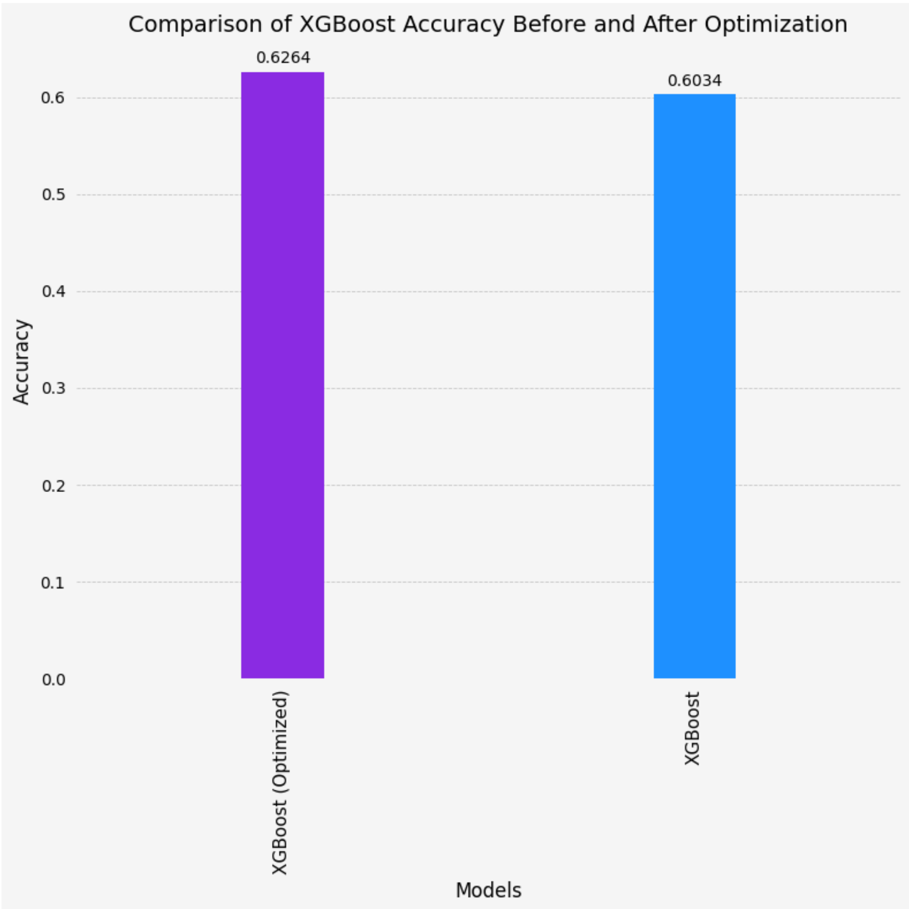

Figure 6.7: Performance of XGBoost with and without GSCV

Using these optimized hyperparameters, XGBoost achieved an accuracy of 0.6264 which was better than the previous score of 0.6034.

6.4 Results from the ensemble model

The ensemble model took inputs from the two base models: the XGBoost model and the MLP model, and concatenated them using a concatenate layer which resulted in a two-dimensional input. There was a dense layer with one neuron, using sigmoid function, to process the concatenated outputs and generate the final prediction. With this initial structure, we have achieved an accuracy of 0.7759.
To increase the accuracy, we have applied GridSearchCV. After hyperparameter tuning, the structure of the model was enhanced. The optimized model had an additional dense layer (hidden layer) with 35 neurons right after the concatenated layer which is using the ReLu activation function. The final output layer remained a single layer with one neuron, using the sigmoid function, responsible for final prediction. This hidden layer improved the models accuracy to 0.8218.

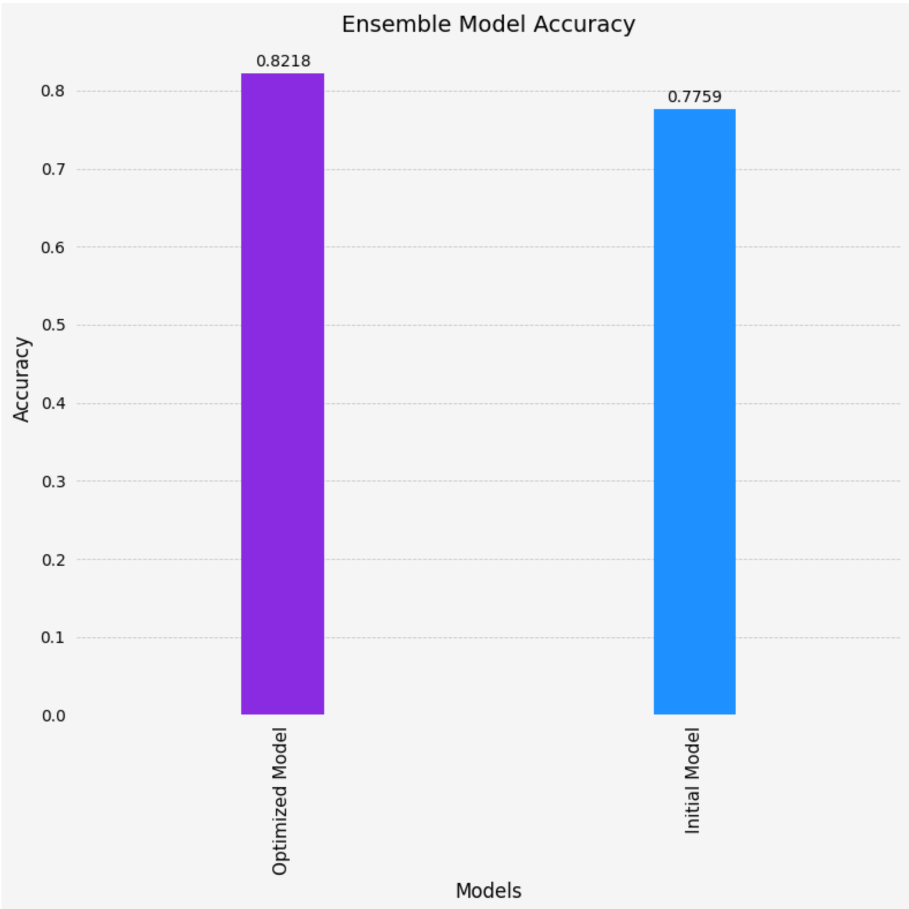

Figure 6.8: Comparison of the ensemble model accuracies

Layer (type)	Output Shape	Param #	Connected to
xgb_input (InputLayer)	(None, 1)	0	-
mlp_input (InputLayer)	(None, 1)	0	-
concatenate (Concatenate)	(None, 2)	0	xgb_input[0][0], mlp_input[0][0]
final_output (Dense)	(None, 1)	3	concatenate[0][0]

Figure 6.9: Initial structure of the ensemble model

Layer (type)	Output Shape	Param #	Connected to
xgb_input (InputLayer)	(None, 1)	0	-
mlp_input (InputLayer)	(None, 1)	0	-
concatenate_1 (Concatenate)	(None, 2)	0	xgb_input[0][0], mlp_input[0][0]
hidden_layer (Dense)	(None, 35)	105	concatenate_1[0][0]
final_output (Dense)	(None, 1)	36	hidden_layer[0][0]

Figure 6.10: Structure of the ensemble model after optimization

6.5 Comparison of MLP with the Ensemble Model

Metric	MLP	Ensemble Model
Accuracy	0.79	0.82
Precision	0.82	0.81
Recall	0.83	0.86
F1-Score	0.83	0.84

Table 6.13: Performance comparison of MLP and Ensemble Model

From the results, we can say that our proposed ensemble model works better than MLP. The accuracy of the ensemble model is higher compared to the MLP model. Though MLP model is slightly better at predicting true positives, the ensemble model is better at identifying all the actual positive cases.

Chapter 7

Final Remarks

7.1 Conclusion

Voting is an important part of a democratic nation, and understanding which factors motivate young voter to vote is also very crucial. This research, through the intersection of machine learning and artificial intelligence, tried to find out the most influential non-political factors that influence young voters' enthusiasm. We used the MLP model on the tabular data and achieved an accuracy of 0.79, whereas the accuracy obtained from sentiment analysis using XGB is 0.62. After using these predictions as the combined input in the ensemble model, we obtained a final accuracy of 0.8218 which shows that our multi-modal research has the capability to improve the models' results. Apart from accuracy-scores, SHAP and LIME were used on the MLP predictions to find out the feature importance. Not only that, but the correlation between each feature and the class label was also calculated. After comparing the SHAP values with the values of correlation, and by seeing their influence from LIME explanations, we identified some of the most influential factors which have positive or negative influence on a young voter's voting behavior.

7.2 Future Work

While conducting this research, we identified some of the most influential factors that motivate or discourage young voters from voting. Our aim is to expand and continue this work by collecting more primary data. We plan to gather a larger dataset not only from the students of BRAC university but also from other colleges and universities as well. This will enrich the dataset and allow us to capture more information on the voting nature. Furthermore, we will be implementing Causal Discovery to understand the causal relationship in the data and make predictions using Causal Inference. Moreover, we plan to use more complex models for performing sentiment analysis. Besides, the limitation we mainly faced was that even after providing careful instructions, many answers given for the opinion based question were random (e.g. N/A, No comments, emojis, dots, "Banglish" sentences etc). Therefore, to overcome this, participants should be more careful about responding correctly to the survey-questionnaire.

Bibliography

[1] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002, ISBN: 9780262194754.

[2] A. Blais, "What affects voter turnout?" *Annual Review of Political Science*, vol. 9, no. 1, pp. 111–125, 2006. DOI: 10.1146/annurev.polisci.9.070204.105121.

[3] D. E. Glasford, "Predicting voting behavior of young adults: The importance of information, motivation, and behavioral skills," *Journal of Applied Social Psychology*, vol. 38, no. 11, pp. 2648–2672, 2008. DOI: 10.1111/j.1559-1816.2008.00408.x.

[4] H. Coffé and M. Voorpostel, "Young people, parents and radical right voting. the case of the swiss people's party," *Electoral Studies*, vol. 29, no. 3, pp. 435–443, 2010. DOI: 10.1016/j.electstud.2010.03.015.

[5] M. Wagner, D. Johann, and S. Kritzinger, "Voting at 16: Turnout and the quality of vote choice," *Electoral Studies*, vol. 31, no. 2, pp. 372–383, 2012. DOI: 10.1016/j.electstud.2012.01.007. [Online]. Available: https://doi.org/10.1016/j.electstud.2012.01.007.

[6] A. Fowler, "Electoral and policy consequences of voter turnout: Evidence from compulsory voting in australia," *Quarterly Journal of Political Science*, vol. 8, no. 2, pp. 159–182, 2013.

[7] C. Sandvoss, "Toward an understanding of political enthusiasm as media fandom: Blogging, fan productivity and affect in american politics," *Participations: journal of audience and reception studies*, vol. 10, no. 1, pp. 252–296, 2013. DOI: 99512911102346.

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, New York, NY, USA: ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[11] M. Zolghadr, S. a. A. Niaki, and S. T. A. Niaki, "Modeling and forecasting us presidential election using learning algorithms," *Journal of Industrial Engineering International*, vol. 14, no. 3, pp. 491–500, 2017. DOI: 10.1007/s40092-017-0238-2. [Online]. Available: https://doi.org/10.1007/s40092-017-0238-2.

[12] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means clustering method and elbow method for identification of the best customer profile cluster," in *IOP Conference Series: Materials Science and Engineering*, vol. 336, 2018, p. 012 017. DOI: 10.1088/1757-899x/336/1/012017. [Online]. Available: https://doi.org/10.1088/1757-899x/336/1/012017.

[13] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019. DOI: 10.1016/j.procs.2019.05.008. [Online]. Available: https://doi.org/10.1016/j.procs.2019.05.008.

[14] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, Minneapolis, Minnesota, vol. 1, 2019, p. 2.

[15] S. S. Kim, R. M. Alvarez, and C. M. Ramirez, "Who voted in 2016? using fuzzy forests to understand voter turnout," *Social Science Quarterly*, vol. 101, no. 2, pp. 978–988, 2020. DOI: 10.1111/ssqu.12777.

[16] L. Moses and J. M. Box-Steffensmeier, "Considerations for machine learning use in political research with application to voter turnout," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:232306017.

[17] R. Chandra and R. Saini, "Biden vs trump: Modeling us general elections using bert language model," *IEEE Access*, vol. 9, pp. 128 494–128 505, 2021. DOI: 10.1109/access.2021.3111035.

[18] K. Hua, J. Jamil, and I. Shaharanee, "A decision tree classifier for predicting voter turnout in malaysian general election," *Central Asia and The Caucasus*, vol. 22, no. 5, pp. 816–830, 2021.

[19] D. Mardaoui and D. Garreau, "An analysis of lime for text data," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 130, 2021, pp. 3493–3501. [Online]. Available: https://proceedings.mlr.press/v130/mardaoui21a.html.

[20] C. Zhu, "A study of key factors influencing youth voter turnout," 2021.

[21] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing bert against traditional machine learning models in text classification," *Journal of Computational and Cognitive Engineering*, 2023. DOI: 10.47852/bonviewjcce3202838.

[22] M. P. R. Archive, *Prediction for the 2020 united states presidential election using machine learning algorithm: Lasso regression*, https://mpra.ub.uni-muenchen.de/103889/, n.d.

Questions of the survey:

1. What is your age?
2. What is your gender?
3. Along with being a university student, do you have any other occupation?
4. Do you live with your family?
5. Does your family live in their own house?
6. What is your father's educational background?
7. What is your mother's educational background?
8. How much time do you spend daily on social media?
9. In your opinion, is your family financially stable?
10. Are your parents enthusiastic about voting?
11. Have your family members voted in the most recent election?
12. How does your father represent voting to you?
13. How does your mother represent voting to you?
14. Are any of your family members or someone close to you involved in politics?
15. How does their involvement affect your voting behavior?
16. How politically aware do you consider yourself?
17. Have you ever voted before? If yes, how satisfied were you with the voting experience?
18. How informed do you consider yourself about the parliamentary system of our country?
19. How informed do you consider yourself about the judicial system of our country?
20. How supportive are you about the notion of "Democracy"?
21. Would you consider the 'newspaper' as a source of your news related to the country/political system?
22. Would you consider 'social media' as a source of your news related to the country/political system?
23. Would you consider 'television' as a source of your news related to the country/political system?
24. Would you consider 'online news' as a source of your news related to the country/political system?

25. If the voting center is far away and you lack adequate transport, would you still be interested to vote?

26. If you get a day off on the day of election, would you participate in the election?

27. If you have a very busy schedule or are busy with personal affairs, would you participate in the election?

28. If there is pre-election violence, would you still go and vote?

29. If there is violence on the day of election, would you still go and vote?

30. Do you plan to vote where you are currently living or in your hometown?

31. Were you eligible to vote in the most recent election?

32. Did you vote in the most recent election?

33. Are you willing to vote in the immediate next election?

34. Do you think your vote will make a difference?

35. Except for the above-mentioned factors, is there anything else that would affect your voting behavior?

36. In a few words, comment on what you think about the next election.

Feature names after encoding:
Age_Age2, Age_Age3, Busy schedule, Current city, Day off, Effect of having political member_Their involvement has made me disinterested in voting, Effect of having political member_Their involvement moderately influences me to vote, Effect of having political member_Their involvement strongly influences me to vote, Election day violence, Eligible in last election, Family's recent vote, Father education_HSC / A levels, Father education_Masters / PhD, Father education_Primary School, Father education_SSC / O levels, Father vote representation_Encourages you to vote, Father vote representation_Neutral, Financial stability, Gender, Info about judicial system_Not at all informed, Info about judicial system_Strongly informed, Info about parliamentary system_Not at all informed, Info about parliamentary system_Strongly informed, Lack of transport/far, Member in politics, Mother education_HSC / A levels, Mother education_Masters / PhD, Mother education_Primary School, Mother education_SSC / O levels, Mother vote representation_Encourages you to vote, Mother vote representation_Neutral, Newspaper, Occupation_Monthly salaried worker, Occupation_No, Occupation_Own a business, Online news, Own house, Parent enthusiasm, Political awareness_Not at all aware, Political awareness_Strongly aware, Pre-election violence, Reside with family, Social media, Supporting democratic notion_Moderately supportive, Supporting democratic notion_Not supportive, Television, Time on socials_<1 hour, Time on socials_>3 hours, Time on socials_I don't use social media, Vote makes difference, Voted in last election, Voting satisfaction_Highly unsatisfied, Voting satisfaction_Moderately satisfied, Voting satisfaction_Never voted before