An Approach to Unveiling Personality Trait Classification from Bangla Speeches of Political Leaders

by

Swachcha Podder
21301196
Ajfar Alam
21301127
Swoperjita Tanha Akond Rodela
21301117
Ishama Fatima Ismail
21301118

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
June 2025

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:

Swachcha Podder	Ajfar Alam
21301196	21301127
Swoperjita Tanha Akond Rodela	Ishama Fatima Ismail
21301117	21301118

Approval

The thesis titled "An Approach to Unveiling Personality Trait Classification from Bangla Speeches of Political Leaders" submitted by

1. Swachcha Podder (21301196)
2. Ajfar Alam (21301127)
3. Swoperjita Tanha Akond Rodela (21301117)
4. Ishama Fatima Ismail (21301118)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 24, 2025.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Dibyo Fabian Dofadar

Lecturer
Department of Computer Science Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

The power of speech lies in its ability to communicate with people. Political speeches can influence people's minds through their persuasive language. The general public needs to know about the political personnel and their actions. They have huge responsibilities surrounding different sectors such as the education sector, healthcare sector, etc., which affect people's daily lives, and their future might depend on it too. Speech is a way for political leaders to express themselves, and through their speeches, we will be trying to decipher their personality traits. To comprehend the personality traits through the text samples, we use the NEO-FFI personality traits (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism) to determine the personality of a particular political person. In our research, we mainly focus on Bangla speeches, and the data is collected from various authentic platforms in the audio format, which is then converted to a transcript. Politicians effectively utilizing their speech can influence the opinions of the general public. So, political speeches are a great way to understand the leaders' personalities by the general masses to be more conscious and vote wisely. A comparative analysis is shown between different statistical and transformer based classifiers that are used to show the performance after training the model on our primary dataset.

Keywords: Bangla Speech, Political Speech, OCEAN, Bangladesh Political Leaders, NEO-FFI, Deep Learning, Transformers, Multi-Head Attention, Attention-based Mechanism, BanglaBERT, XLM-RoBERTa

Acknowledgement

First and foremost, we would like to thank the Almighty for helping us grant the patience and strength to complete our thesis.

We would like to sincerely thank my supervisor, Dr. Md. Golam Rabiul Alam Sir for his insightful suggestions,constant support and guidance throughout the thesis journey.

We would also likely to forward my gratitude towards our co-supervisor, Mr. Dibyo Fabian Dofadar Sir, who provided constant encouragement and support.

Special thanks to Md. Sajeebul Islam Sk, our research assistant, who has helped us with his active involvement and made a significant contribution to our thesis.

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

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

DL Deep Learning

F_1 Balanced F1 Score

NEO-FFI NEO Five-Factor Inventory

NLP Natural Language Processing

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machine

TF-IDF Term Frequency-Inverse Document Frequency

Chapter 1

Introduction

The dataset that has been made is composed of 557 Bengali speeches and interviews, predominantly Bengali, that are from several Bangladeshi political figures. The gathered speeches are then converted to 2367 transcriptions. The intention behind the construction of this dataset is to study the traits of the speeches of political figures and to analyze the speeches by applying the NEO-FFI traits. After collecting speeches, the speeches are divided into time intervals, where each section is annotated to identify the speech's traits, such as openness, agreeableness, extraversion, conscientiousness, and neuroticism. The time intervals are random, and the intervals are being decided on the speaker's change in topic, tone, or a visible change in emotions. With the data being essentially sourced from YouTube and various media outlets, the dataset establishes the foundation to examine the verbal remarks of political leaders and associate them with models, which allows a better understanding of their personality traits through public discourse.

1.1 Research Problems

In political and public affairs,the style and narrative of speeches of the political figures help us to give a slight perception into the nature and personality of the person. It also gives us an insight into how an authority figure like them can impact a huge number of people to making decisions. As it stands, the political discourse in Bangladesh is very complicated, which is why there is scarcity in data around Bangladeshi political leaders. All the prior existing datasets mostly focus on English datasets based on politicians in the West. Hence, due to the lack of existing available datasets, it was decided by us to create one centered around mostly Bangladeshi Politicians.

It can be said that politics in Bangladesh is much more multifaceted than politics in the West, due to our historical and cultural background, as well as the subtle rhetoric. Due to this, it is possible that personality traits are assessed differently among Bangladeshi politicians compared to Western ones. So it is entirely possible that the already existing models for personality trait recognition that are built for only Western datasets and is in the English language may not be able to completely grasp the personality for a dataset that is in the Bengali language. Furthermore, there is also an insufficiency of Bengali datasets, which centers around Bangladeshi political leaders.

Our research problem revolves around our construction of the Bengali dataset of political leaders from their speeches and interviews, where we are applying the Big Five model traits from NEO-FFI and how their personality is represented in the

dataset. It takes into account how the lack of a dataset is an issue and how we can create one to detect the traits while analyzing and annotating the data in the most meticulous way possible.

1.2 Research Objective

While exploring the topic, it has been gathered that despite Bangla being the 7th most spoken language, it still lacks resources. Also, as we know, the political speech of political leaders in Bangladesh has always played a pivotal role and they have been deeply interconnected with Bangladesh's political history. From the pre-independence era, speeches have been used as a strategic tool to influence the mass public and drive collective action. One of the most notable examples is Bangabandhu's historic speech on 7th March 1971, which inspired collective resistance from the and served as a catalyst of people moving towards liberation. Correspondingly, modern-day speeches from political leaders influence the general public's opinion and steer the course of political events.
For our research, the main objective is to create a comprehensive and well structured Bengali dataset that consists of transcriptions from speeches of Political leaders, specifically from Bangladesh. By applying the Big Five traits and as well as NLP techniques, the study aims to predict traits from the speeches of the leaders and analyze political discourse. Our study specifically aims to evaluate whether the dataset we have created for Bangladeshi Political leaders' speech using the NEO-FFI traits is actually effective for personality traits analysis. More precisely, it explores "How suitable is our Bangla textual dataset for Bangladeshi Political Leaders for actually predicting the NEO-FFI personality traits by training models?" Here, in our analysis, we have tried to approach this by using a statistical model-SVM as a baseline, and two transformer based models, BanglaBERT and XLM-RoBERTa, as well, to observe how our dataset has performed for each of these models.

The key contributions of this research are as follows-

1. **Bangla Political Speech Dataset Curation:**
We have collected speeches of different political leaders from different political parties and annotated them using the NEO-FFI traits. After annotation, using a Google Speech Recognition API, we have converted them into transcriptions.
2. **Implementation of Models:**
After curating our dataset, using traditional models such as SVM as well as transformer-based models such as BanglaBERT and XLM-RoBERTa, we have applied a comparative analysis and assessed personality trait classification.
3.**Dataset Applicability Assessment:**
For our dataset, we have tried to validate its applicability by achieving meaningful scores on transformer-based models such as BanglaBERT and XLM-RoBERTa for personality trait classification.

Chapter 2

Literature Review

For our literature review, we have read through some papers that are related to the analysis of speeches. We found 14 papers that were published from around 2015 to 2024 that were either related to political speeches or Bengali speeches or papers that have tried to explore context through different models, and through these papers, we tried to learn how to progress our research.

2.1 Big Five Traits of OCEAN

As our objective is to create our own primary dataset and then detect personality traits of the speeches of the political leaders, we have tried to read through the papers that incorporate NEO-FFI personality traits. The authors here have focused upon assigning NEO-FFI Big Five personality traits on Bangla speech[2]. Additionally, the authors decided to create their own Bangla dataset where 90% of their content has been sourced from well-known Bengali newspapers and 10% of it has been from Bengali novels. As per the author's knowledge, there has been no previous research done to detect personality traits from Bengali speech, and so this paper contributes to narrowing this gap by creating a primary Bangla dataset. The dataset consists of 1,750 speech-to-text samples and 1,000 recorded spoken samples, all of which were annotated with the Big Five traits. To identify the personality traits, two ensemble models were built, including DistilRo (based on Distilbert and Roberta) for the transcribed texts and BiG(based on Bi-LSTM and GRU) for audio data to ensure robustness and accuracy. In addition to that, feature-level fusion techniques such as MFCC, LPC, Wiener Filtering, and Morlet Wavelet transformation were used for raw data. The authors used feature sets of MFCCS, MELP, MEWLP AND MoMF. However, even with all the techniques, this paper included a few limitations. The speakers for the acted speeches were non-professional, which made it difficult to include diversity in the dataset. Also, as the speakers were anonymous, the human annotators also solely relied on the audio to annotate, which could result in inconsistency. Also, the texts were short, and the models were confused among the neuroticism and the openness trait, making it difficult to annotate. Even with all these constraints, the authors were able to achieve a 90% with MEWLP which was the best overall, and then an 89% F1-score with DistilRo and 81% F1-score with MFCCs, 84% with MoMF, 88% with MELP. It could detect Extraversion strongly, however, it had difficulty in detecting Openness.

As we move on to the next paper, it can be seen that a primary Benchmark Bangla dataset is created by collecting 3000 informal texts from various social media platforms which are then annotated by humans using the Big Five personality[10]. As we are planning on doing human annotation as well, this paper helps us understand how to approach annotation. This paper introduces the first benchmark dataset for automatic personality detection using the Big Five personality traits. The data preprocessing includes tokenization, stemming, special characters, and stopword removal. For feature extraction, the authors have used Bag-of-Words with TF-IDF for statistical models and word embedding for Deep learning models. For classification, both types of models are used; statistical models such as SVM, Naive Bayes, Random Forest, and SGD are used for baseline comparison, and deep learning models such as MLP,FastText,C-LSTM are used for contextual order. Here, amongst all the statistical methods, SVM had the best F1-score out there. However, amongst all the models, C-LSTM, which is a deep learning, outperformed every model and had the best overall F1-score. Even though there are several issues, such as no handling of sarcasm in text, code mixing issues, or unreliability of human annotation, this paper still provides a comprehensive study on personality detection and a complete dataset, which helps us to understand how to approach our own research.

In reference to informal texts which are used in [4], we reviewed another paper which assigned Big Five traits over a pre-existing dataset of short social media texts collected from Twitter[4], however, this paper has a multilingual dataset compared to [10] which has a Bangla dataset. The pre-existing dataset consisted of 14,166 English tweets from 152 users, 9,879 Spanish tweets from 110 users, and 3,687 Italian tweets from 300 users, for a total being 27,732 tweets. This paper aims to detect personality traits from short texts using advanced deep learning techniques, which includes language independent and compositional models. It discusses the issues that current deep learning models face with text representation, where a commonly used approach is to use dense vectors to represent words. These methods were useful,however,it has issues, such as it represents unseen words as unknown vectors, which causes them to lose their meaning. This paper introduces a new language-independent hierarchical model, C2W2S4PT (Character to Word to Sentence for Personality Traits), which understands character, word, and sentence hierarchically using Character-level Bi RNN, Word-level Bi RNN, and MLP Classifier, thus solving out-of-vocabulary issues. So, C2W2S4PT provides State-of-the-Art performance across 3 languages and 5 traits, and it surpasses existing feature-heavy baselines despite having limited Italian data and short texts.

As we are dealing with transcriptions, we also thought it would help us to review papers focusing on how manipulating large language models can be used to express context through Big Five traits [15]. The authors used GPT-2 and BERT as their language model to be assessed on, where they asked 50 questions with prefixes such as 'I' or 'I am', and the answers were based on a 5-point Likert scale. The authors here used both primary and secondary datasets. The primary dataset was made of 404 human participants who wrote 75-150 word descriptions of their own personality, and the secondary dataset consisted of Reddit Personality descriptions. To interpret the results, they compared them with the human population distribution of Big Five traits where they used a survey of 1015000 individuals consisting of Reddit personality descriptions or self-reported descriptions from individuals. Eventually, they could conclude that the results with different contexts provided variations in

the scores. When the language models incorporate personality cues, both BERT and GPT-2 show predictable shifts in traits, with strong median correlation of 0.84 and 0.81. Furthermore, when given 404 human texts, the models predicted and scored 0.48 and 0.44, respectively. GPT-2 also maintained the consistency in traits while generating these texts, achieving a correlation of 0.49 between the input personality and output text. Even though this paper focuses on a smaller model like GPT-2 which could be confuse with double negatives and has limited generalizability, it helped us learn how we could utilize language models for personality traits.

We also analysed another survey paper to identify personality traits from an existing secondary dataset, which comprises 86,000 data points from Facebook and 1.2 million English Tweets already annotated with MBTI labels[12]. This paper highlights the disconnection between psychological theory and computational modeling..This paper discusses various models, and it can be seen that models that focused on the Big Five and were trained on Facebook data, accomplished a correlation of 0.56, surpassing human evaluation. Conversely, MBTI models trained on Twitter data showed a weaker accuracy of 72.5% for certain traits only such as Extraversion. However, deep learning models that could combine both audio, text, and video could get accuracy up to 91%. Even though the MBTI model is easier to understand as it categorizes people into different types using a typology system, the binary approach might make it lose valuable information. The Big 5 model provides continuous scales and also an intensity-based score for each dimension so it tends to be more precise but it can also have measurement errors and biases.Another comparison in this paper is shown where it dictates that although both models deal with extraversion and introversion, the MBTI model claims the person's energy is related to the outer world(extraversion) and the inner world(introversion) while the Big 5 model claims the person's energy is related to measurement of energy levels and their social behavior where extraverted personality refers to being energetic and sociable and the introverted personality refers to being shy and reserved. Therefore, the paper concludes that MBTI traits are harder to detect from short texts and raises concerns over the dataset quality as well as ethical issues, such as user privacy concerns related to the dataset. The author discusses that future research may include different frameworks or models with more parameters to classify personality traits.

In this study of personality trait recognition, the paper aims to detect it from literary characters[7]. This paper uses a secondary dataset for training, which consists of 2,400 essays per user, which are annotated using the Big Five traits, and another secondary dataset for testing, which consists of dialogues from characters named in 38 plays of William Shakespeare. As there has been no prior work done on analyzing personality from literary works, especially considering the fact that there is a distinction between the texts used in the training set and the theatrical texts. The model here uses Scikit-learn to implement supervised learning integrated with bag-of-words features with TextPro-generated annotations such as POS tags, chunks, and entities. For each trait, there are various models used, such as SVM, Random Forest, and Multinomial Naive Bayes. However, amongst them, Random Forest achieved the best F1-score ranging around 0.45(Conscientiousness) to 0.66(Openness). Even though the model was able to achieve some plausible results, the lack of trait diversity implies that the Big Five may not be efficient enough to handle the nuances and complexities of the fictional characters. Also, the model was trained on

modern texts which have quite a different writing style compared to Shakespeare's literary texts. Additionally, there are no labels given for the Big Five traits for the characters, and the model outputs only "yes" or "no", oversimplifying the complexity of the characters. Therefore, the author suggested that in future research, more detailed trait models should be used while also integrating contextual data of the character[7].

2.2 Sentiment and political discourse

We also went through papers that researched into context or sentiment analysis in Phase 1, as we were unsure how we would progress our thesis. Thus, we tried to find papers that focused on Political figures to get more ideas since our dataset consisted of South Asian Political figures. In this paper, the authors try to classify the context and emotion of the Political Leaders around the world through their political speeches[16]. The dataset is of 2010 speeches (2000-2021) from the official websites of the governments of China, Russia, the UK, and the USA, labeled with context and emotion. Two ensemble models named EMPOLITICON-CONTEXT and EMPOLITICON-EMOTION are developed by the researchers,both of which utilize soft voting classifier techniques to merge predictions from XGBoost, CatBoost, Nu-Support Vector Classification, and Linear Discriminant Analysis algorithms. For the long documents which have tokens from 500-1600 on average, Longformer is used for generating contextual embeddings so that it could overcome the BERT's 512 token limitation. The speeches are classified into five emotion categories(joy, optimism, neutral, and upset) and five context categories (nationalism, international affairs, development, extremism, and others). Even though the EMOTICON-EMOTION was found challenging to classify emotion and achieved an accuracy of 53.07% only, the EMPOLITICON-CONTEXT accomplished an accuracy of 73.13% for context classification. However, there were several issues to be noted, including the inadequate performance of extremism and nationalism, as well as difficulties in distinguishing between traits such as openness and neuroticism. There is also the unreliability aspect of annotation, as annotation has been done by humans, so there is a bias to be considered. Despite the complexities, the contextual model could attain a respectable accuracy.

In this next paper, the authors assess whether the influence of political parties and their election outcomes can be determined from the sentiment of the tweets[9]. Previously in [4], the authors also created a multilingual dataset from tweets to detect personality traits, and in [9], the dataset created is primarily in English .The author also examines whether the sentiment observed in the tweets is actually correlated to their election results. In this study, sentiment analysis is done, which is an NLP technique to observe the Anambra State gubernatorial election using tools such as TextBlobs Naive Bayes Classifier and SentiWordNet. The authors collected 33,502 tweets on election day. Then, after preprocessing, 7,430 tweets were kept that had the #AnambraDecides2017 from formerly known as Twitter, now known as X. The tweets were analyzed using Polarity Sentiment Analysis(PSA) and Subjectivity Sentiment Analysis(SSA). In PSA, the tweets are examined to see if they are positive,negative or neutral. In SSA, the tweets are examined to see if they are opinionated or fact-based over time, based on segments. Here, the word frequency analysis is also used to recognize the most common words that are linked to each

political figure. The topic modeling algorithm is used to find the most debated political figure as well as their sentiment. In this study, it was found that the positive sentiments which were towards the APGA party helped benefit the winning candidate, Willie Obiano. The tweets about AGPA did have more positive sentiments and were less opinionated emotionally compared to other parties. Due to having higher proportions of neutral classifications (53.44%), TextBlob was chosen for further analysis even though SentiWordNet had a greater proportion of positive tweets (39.25%). So, the study suggests that when the candidate's qualities and traits match with the party's values,it can impact the election results significantly. There are some limitations in this study. The analysis is based on just tweets from Twitter, and this does not represent the extended population. It also does not take into account the offline influences such as media. Also, here the study is done only in one election campaign, so it is possible that the study could not be applicable to other events.

In reference to political discourse, we also examined another relevant paper that discussed a graph based model named GPolS, which is specifically designed for political debates in Parliament [11]. Since political speeches are often filled with formal language and cues, it is difficult for models to interpret them. So, to deal with this problem, GPolS uses a BERT model, which is fine tuned and trained on the ParlVote dataset. It is a primary dataset consisting of 33,461 debate transcripts from around 1997 to 2019, collected from the UK Parliament. The model uses the textual content such as who the speaker is, what their topic is, and what their political party is, and they take this information to form a graph to represent the speaker's behaviour, their party affiliations, and motion-based content. These information are then analysed using Graph Attention Networks to infer whether a given speaker is in favour of the proposal or is the speaker rejecting it. GPols is able to achieve an accuracy of 80%, surpassing other models such as SVMs and MLPs, and also, the best previous method (DeepWalk) by 6.5%. Even though in this paper, GPolS has worked with English data and has a limitation of 512 tokens only, it could still attain a respectable accuracy, which is a significant step in using contextual data to understand political discourse better.

2.3 Approaches to Text-Based Emotion Detection

While exploring, we found a very interesting paper where the author strives to address the challenges in detecting sentiments from memes, especially in the Bengali language, which has low resources[14]. As memes are growing in popularity every day, it is important to understand the sentiment behind the memes, as people tend to create memes often, and it is a way of expression. For this, the author aims to create Memosen, which is a multimodal dataset that consists of 4368 memes, which were divided into positive, negative, and neutral categories. To evaluate Memosan, a handful of experiments were carried out in order to benchmark it. The memes were collected from various social media sites, and then they were divided into categories based on their content. Out of 4368 memes, 2728 are negative, 1349 are positive, and 291 are neutral. Both unimodal as well as multimodal models were used in this study for sentiment analysis. In unimodal models, there are visual models such as Xception, VGG19, VG16, ResNet50, DenseNet50, and in multimodal models, there are LR, MNV, SVM, Bi-LSTM, CBB, B+MurIL, Bangla-Bert, and XLM-R. The visual models evaluate the image component of the meme, while the textual

model evaluates the text component. The multimodal model combines them both to provide a more thorough evaluation of the sentiment. This analysis shows that the multimodal model outperforms the unimodal model, and there is 1.2% improvement in sentiment classification. For the visual model, ResNet50 performed the best, for the textual, it was MurIL, and for the multimodal, it was ResNet50+CNN. There are also some limitations here. The first one is that the dataset has mostly negative memes, making it imbalanced. The memes are also context-dependent, which makes it difficult for the models to analyze the sentiment. Because of code-mixing and code-switching, there may be difficulties in text extraction. Lastly, in memes, regardless of the context there may be some things which appear consistently, for example, a celebrity's face, which might confuse models and reduce their ability in detecting sentiments.

As we delve more into papers analyzing context and sentiments, it can be discovered that using deep contextual representations in real-world tasks tend to perform better compared to previous methods which only look at the words individually [6]. This paper brings forth Embeddings from Language Model(ELMo)-which is a type of new word embedding method for natural language processing that overcomes the challenges of understanding word polysemy (words with variations having different meanings or contexts) and complex words. It is trained on a very large and structured secondary dataset (1 Billion Word Benchmark), and it utilizes all the layers of a bidirectional Language Model (biLM). Elmo was used across six NLP tasks for assessment-question answering(SQuAD), semantic role labeling(SRL), textual entailment(SNLI), Sentiment analysis(SST-5),named entity recognition(NER), and coreference resolution-for each case, accomplished higher state-of-the-art results, while also reducing the relative error between 6 to 20%. Elmo showcased improved F1 scores for specific NLP tasks; for question-answering, the baseline model had an improved test score from 81.1% to 85.8%, sentiment analysis had an improved accuracy by 1%, named entity recognition achieved a better score than previous state-of-the-art, textual entailment with an increased accuracy by 0.7%, semantic role labeling with a 3% increase and coreference resolution with an increased F1 score from 67.2 to 70.4% However, there is also need for fine tuning on some task specific data which is an inconvenience. This paper further inspects the different kinds of contextual information provided by the bidirectional language model (biLM). After conducting two careful evaluations, it was concluded that the lower layers of biLMs perform better at providing syntactic information while the higher layers tend to perform better at providing semantic information, which relates to the findings from MT encoders. Besides, it is also seen that biLM achieves richer representations than CoVe, which is considered to be a previous method for generating word representations. Hence, while observing the sample efficiency, it is seen that including ELMo in a model makes it much more efficient. For instance, the SRL model reaches a peak performance F1 after 486 epochs of training without ELMo, while with ELMo, the model surpasses the baseline peak performance by the 10th epoch.
Another paper was found where sentiment analysis was done, but with Bangla collected texts[5]. The main objective of this paper is to create a method to classify sentiments from various collected Bangla texts, which is challenging as there is less research on the Bangla language compared to other languages, and then to determine whether the collected text is positive, negative, or neutral. The author collected a small dataset manually from various social media websites of around 1500 sentences of Bangla comments, which was then split into two sets: a training set consisting of

1400 sentences and a test set consisting of 100 sentences. Since words are what people use to express their thoughts and emotions, a feature extraction process called Term Frequency-Inverse Document Frequency (TF-IDF) was used. Here, the TF-IDF helps in recognizing and highlighting the important words by the frequency of appearing words. The authors then try to find a pattern so that the sentences can be classified as positive or negative; they observe the position of the words in the sentence or look at the last 4 words together. The author then evaluates their sentiment analysis algorithm by how well it works. 50 positive and 50 negative sentences are taken for the test, and precision, recall, and accuracy are calculated. Despite the small dataset size and absence of neutral sentiment category, the classification could attain a precision of 81.13%, with a true positive of 86% and a true negative of 80%, with the overall accuracy being 83%.

2.4 Political discourse

As we were researching papers that worked regarding context,we found a paper based on Pakistani Politics. Here, it has been analysed in a Pakistani paper how language is used to exert influence in political discussions by observing Imran Khan's 2014 Dharna Speeches using the FairClough's 3D Critical Discourse Analysis(CDA) model [17]. Here, a primary dataset has been used, which consists of lines from Imran Khan's speeches sourced from platforms such as YouTube and Dailymotion. This paper focuses on language patterns such as pronouns and vocabulary and reveals how using them helps political figures express and connect with the audience. For example, usage of "we" and "us" pronouns highlights unity with the common people and also constantly referencing to Islamic topics and emotional topics strengthens his connection to the audience's cultural values, while also portraying him as superior compared to his opponents. Even though the research has a smaller dataset and is qualitative, it explores how influential text is by applying Fairclough's CDA and how it can make an effect in political discourse.
Lastly, the most recent contribution to our paper review is this notable study, which creates a benchmark primary dataset for political sentiment analysis and also is one of the first studies which applies Large Language Models (LLM) to Bengali political sentiment analysis[19]. The primary dataset here consists of 7,058 entries where 4,132 are positive and 2,926 are negative, and each entry has been manually annotated. The entries have been sourced from various Bangladeshi newspapers, where the commentary has been around political debate. The study validates the performance of this dataset using Pre-Trained Language Models (PLM) such as BanglaBert XLM-RoBERTa, Bangla BERT Base, and sahajBERT, and LLM such as GPT-3.5 and Gemini 1.5. Also, both zero shot and few shot learning techniques are used to see how well the models can actually classify the sentiment. Using traditional learning, the PLM are fine tuned while LLMS are provided with labeled data or provided with none. Amidst PLMS, the BanglaBert surpassed every other model and scored the highest of 88.10%. However, the LLMS outperformed every other PML with Gemini 1.5 scoring 96.33% and GPT-3.5 scoring 94%. The research, however, also has several limitations despite its contributions, as the sentiment labels are binary labels, which reduces the depth in sentiment analysis. Furthermore, only 300 samples were used in LLM analysis due to API cost issues. However, this study provides a valuable insight to sentiment analysis and how it can be more advantageous if worked strategically, while also allowing the scope for future work in multimodal analysis.

Chapter 3

Dataset Collection and Analysis

3.1 Data Collection Approach

As our primary goal was to create our own Bengali dataset consisting of speeches of South Asian Political Leaders, we tried to locate papers that had created their own primary dataset to understand the process of curating and assembling a dataset and how we should progress our own research. Furthermore, we also considered papers that delved into contextual analysis as well as political discourse since it was interconnected to our topic. Reviewing these papers helped us become aware of the internal workings of deep learning models as well as statistical models while also bringing insight into the fact that deep learning models might be better for our research since they are able to handle complex data and can provide better performance.
To detect personality traits of political leaders, we have tried to prepare a Bengali dataset that consists of speeches from 40 Political representatives based in South-Asia, specifically Bangladesh. While assessing, we found that even though there were quite a few Bengali dataset that were curated from short texts collected from social media such as in [4] or from novels and newspapers in [2]; however, to the best of our knowledge, there has been no prior work done on personality trait recognition for Political personals based on Bangladesh. Therefore, our aim was to create a resourceful and rich dataset that could be used in Bengali NLP study.

3.2 Dataset Premise

After conducting a comprehensive analysis, we gathered that there was a lack of pre-existing dataset solely focused on personality detection of South-Asian political leaders, more precisely from political personals from Bangladesh. This encouraged us to create a dataset from which traits can be detected using the Big Five Personality Traits. Even though compiling our dataset proved to be challenging and laborious, we ended up on the decision to create our own dataset. There were several reasons behind this decision. The primary one is that there is a severe scarcity of public dataset availability. Also, most datasets that are usually accessible to the public are built around the English language. As a result, we decided to construct our personality trait recognition dataset, focusing on Bangladeshi political leaders who speak Bengali language, hence, requiring the development of our specific dataset. For our research, the primary dataset curated consists of transcriptions of 557 public, Bengali speeches and interviews, substantially focused on Bangladeshi political personals. Each speech and interview that has been collected is taken and

then manually annotated as they are broken down into random time intervals, which are manually determined, so that the traits can be assessed from the intervals. The dataset provides videos of leaders speaking across several settings such as public assemblies, public rallies, and interviews, which ensures a widespread depiction of speech patterns and style. The goal of this dataset is to represent and label spoken speech through the Big 5 NEO-FFI personality traits and use several deep learning and statistical models to evaluate the personality.

3.3 Data Sources

The process for collecting Bangla speeches involved multiple stages of a selection process and methods based on the relevance of analyzing personality traits. At first, we identified the prominent political parties that have dominated the political landscape including Bangladesh Awami league , Bangladesh National Party, Jatiya Party, Jamaat-e-Islami Bangladesh and the new inclusion, Bangladesh National Citizens Party was considered as well. The key figures or members of these parties who have been contributing to the political dynamics were then identified for their speeches such as Sheikh Hasina, Khaleda Zia and many more. After identification, we conducted a thorough search for each leader's videos and then collected their links in a spreadsheet. For our own feasibility, we used a spreadsheet where in each column, we updated each particular leader's name, their affiliated party, their designation, and their country. We have also kept labeled columns with Big Five traits in order to annotate.

Field	**Field Description**
Country	Name of the country
Designation	Designation of the leader or member
Leaders	Name of political leaders
Speech link	Speech collected from Youtube links
Type	Speech/Interview
Language	Bangla/ English
Traits	5 NEO-FFI personality traits
Interruptions	Interruptions occurred in the audio file (Noise, silence)
Duration	Total Time of the speech collected per minute

Table 3.1: Dataset Field

This helped us maintain a record for each leader and have their required information in an organized manner. Our dataset predominantly consists of speeches from around 2010 to around 2025, with the most recent speeches being from the National Citizens Party. However, there is an exception to this as there are two speeches of Bangabandhu, which are from 1971 and 1974. In our paper, we have currently used political leaders based on those whose speeches are mostly available in public media platforms. For the dataset, we also made sure that the leaders that have been chosen are speaking primarily in Bengali in the targeted speeches. Once the leaders were selected, we sourced the speeches from different platforms such as www.parliament.gov.bd, however, our main platform for collection was mainly YouTube, given its accessibility. However, we needed to convert the audios to transcription as we were mainly aiming to work with text.

3.4 Data Annotation

After the thorough data collection process, we proceeded with our data annotation process, which proved to be a very time consuming process. Also, the approach was based on our dataset field. For our annotation process, we used a spreadsheet where YouTube links were collected for each leader. Each column was labeled with the leader's information, such as their name, their affiliated party, and their designation. For each leader, we have also kept designated columns with the Big Five Traits of OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). Another column was reserved for timestamps, where we input the interruptions or disturbances to discard them. For our dataset, we have used human annotation, which is a very labor-intensive process, however, we thought this process would perform most suitably for us. While listening to the speeches or interviews, the annotator is supposed to keep track of the speaker's content and annotate the content to what they feel aligns most closely to the specific trait. While listening to the speeches and the interviews, the annotation is done from interval to interval, and the interval selected is random by the annotator. This is because the length of every video is different, so it is not possible to have a predetermined interval. Hence, the annotator must determine the segments as he detects changes in the conversation. This results in the human annotation being very subjective, which could create inconsistency and overlapping traits. So, to overcome this inconsistency and to increase the reliability, we watched each video together later for the second time to make sure that the labeling was correct. To allocate traits, one must understand the personality traits well and listen to the videos carefully. Moreover, If there was a confusion regarding determining the trait, the annotators discussed and chose the trait with which the majority agreed with.

Openness: If a political leader is expressing his ideas and thoughts in a creative, imaginative as well as curious way, we consider it openness.
Agreeableness: If a political leader is expressing his ideas in a kind, cooperative as well as compassionate way, we consider it agreeableness.
Extraversion: If a political leader is expressing his ideas in an assertive, talkative as well as outgoing way, we consider it extraversion.
Conscientiousness: If a political leader is expressing his ideas in a reliable, disciplined as well as organized way, we consider it conscientiousness.
Neuroticism: If a political leader is expressing his ideas in an unstable, anxious as well as irritable way, we consider it neuroticism.

Even though having multiple annotators might solve the problem of inconsistency,there still might be some political leaders who show multiple personality traits simultaneously. In this case, the annotators have to figure out which is the primary trait and which is the secondary. As the annotation is quite a labor-intensive process, the annotators need to be at attention at all times as it is very much possible to lose focus while watching a long video.

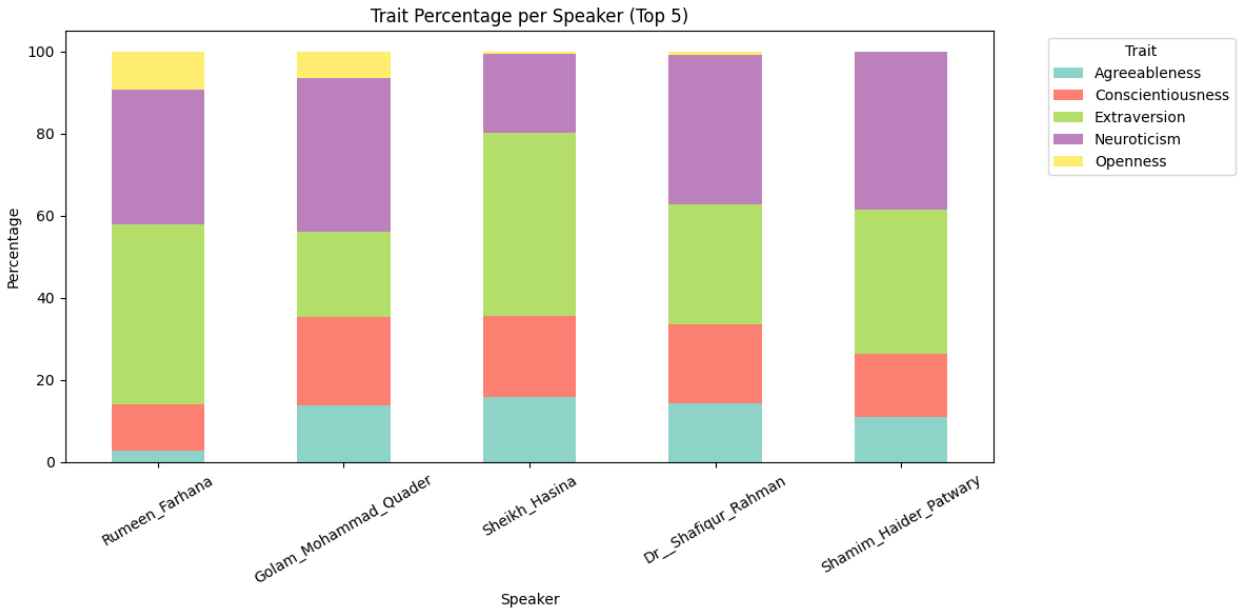

Figure 3.1: Diverse traits among leaders

3.5 Dataset Curation

As we had intended to work with transcriptions, we assembled a transcription system in Python which basically converts the spreadsheet named Annotations.csv that we had organized from youtube links into transcriptions for each leader. This file creates a directory named Each_Leader where there are folders created for each Politician and saved by their name. The audios are downloaded using yt-dlp downloader and saves the audios as wav files (e.g. , clip_1.wav). The audio is then transcribed with the help of Google Web Search API. The system is built for both long and short clips, if clips are longer than 30s, the pydub library was used to split the longer audios to sections and then they are transcribed separately. The divided sections are then concatenated together to form the full transcription which are then saved as .txt files alongside the same folder as the audio files. If there were any errors encountered while processing the data, such as API unavailable, the video was transcribed to ensure data recovery. After all the videos were processed, using the shutil library, they were compressed into a zip file and made available to download.
Then, we made separate csv for each leader where we kept the manually annotated Big Five traits for each leader so that it has the labeled traits and timestamps. Now, the python system we have built finds and retrieves the audios from the folder that correspond to the timestamp in the csv file and transcribe them from audio into Bengali text using Google's Speech Recognition API. It loads the wav files for each leader and aligns with the corresponding csv annotations, where it iterates through each annotated trait and time interval. The time interval is handled using the parse_time function which can handle time formats such as 1:50 to 1 minute 50 seconds. To prevent exceeding Google API's limit, the transcripts are formed into 60s segments and concatenated. For each successfully transcribed segment, the system formats "Leadername_Clipnumber_Trait_SegmentIndex.txt.

Field	Field Description
Text file	Khaleda_Zia_clip_5_Neuroticism_1txt
Text	এরা হলো লুটেরা এরা খুনি এরা হলো দেশের স্বাধীনতার বিরোধী শক্তি
Label	Neuroticism
Word Count	11

Table 3.2: Transcription Dataset Field

3.6 Dataset Quality Control

To ensure the quality control for our dataset, we annotated the dataset twice. At first, every annotator picked a leader and annotated their videos. Then during the second annotation phase, the four of the annotators watched each video together to limit annotation inconsistencies. After the transcription, the annotators checked all the transcriptions manually to ensure they were correctly transcribed and there were no errors.

Text	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Final Labeling
গ্রেনেড হামলারও বিচার আজকে হয়েছে	Neuroticism	Neuroticism	Neuroticism	Agreeableness	Neuroticism
মানুষের ভালোবাসা থাকলে আস্থা থাকলে যেকোনো মানুষ নির্বাচন জয়ী হতে পারে	Extraversion	Extraversion	Agreeableness	Extraversion	Extraversion
উপহার দেবার জন্য আমি মাননীয় প্রধানমন্ত্রী কে কৃতজ্ঞতা জানাই	Agreeableness	Agreeableness	Agreeableness	Agreeableness	Agreeableness
চাঁদের দিকে তাকিয়ে আমাদের লাভ নেই আমাদের গন্তব্যস্থলে আমাদের পৌছাতে হবে	Openness	Conscientiousness	Openness	Openness	Openness
একটা আইন মনে হয় এখন করা দরকার	Conscientiousness	Conscientiousness	Conscientiousness	Conscientiousness	Conscientiousness

Table 3.3: Sample trait annotations by annotators and Final Label

3.7 Limitations

Although the dataset has been carefully curated to provide a structured access to detect traits through speeches of the political leaders, it is not without its notable limitations. The imbalance of data across the leaders has to be taken into account as it can be seen that there are 85 and 63 videos for particular leaders whereas there are only one video each for some leaders. This inequality introduces bias in both mode training and testing analysis as the results will reflect the speech patterns of the overrepresented leaders. Additionally, despite being thorough with annotation, it is possible to have inconsistencies in them as due to personal interpretation. Also, different leaders' speeches may introduce a bias if they are talking about different subject matters. There is also the issue where people having different dialects may lead to lexical and phonetic mismatch. Other constraints include background noise and unclear sound, due to which there were potential inaccuracies in the transcrip-

tion. It is important to be mindful of these factors when one is interpreting the results derived from the dataset.

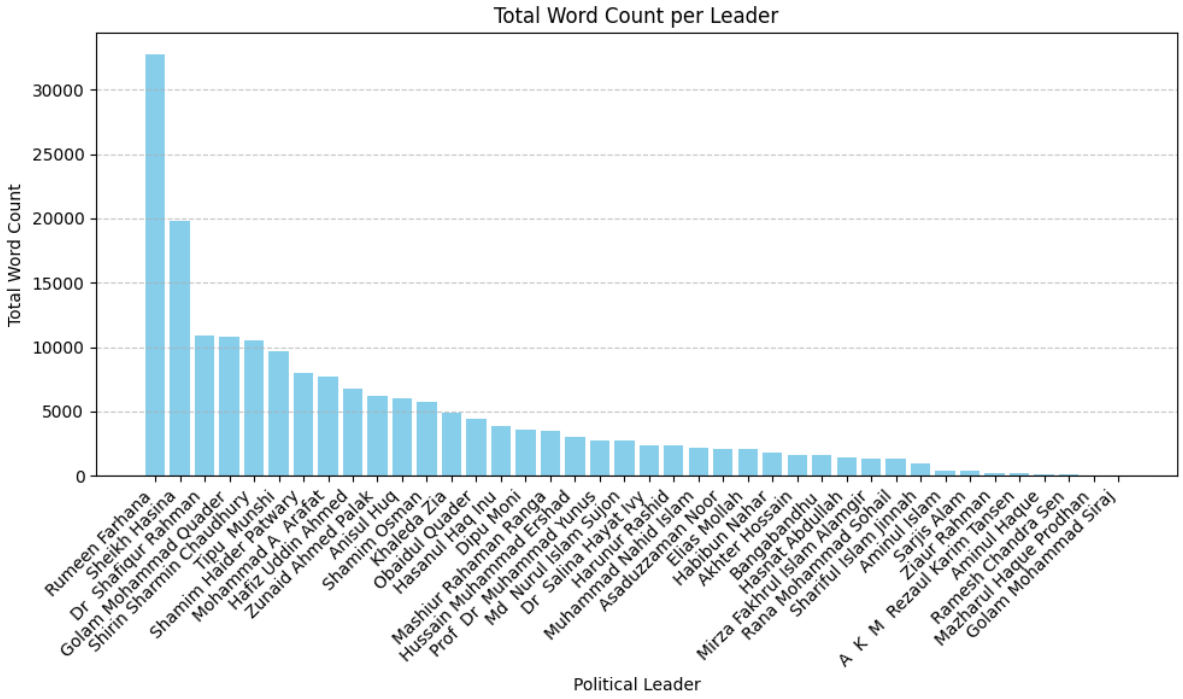

Figure 3.2: Raw Text distribution per Leader

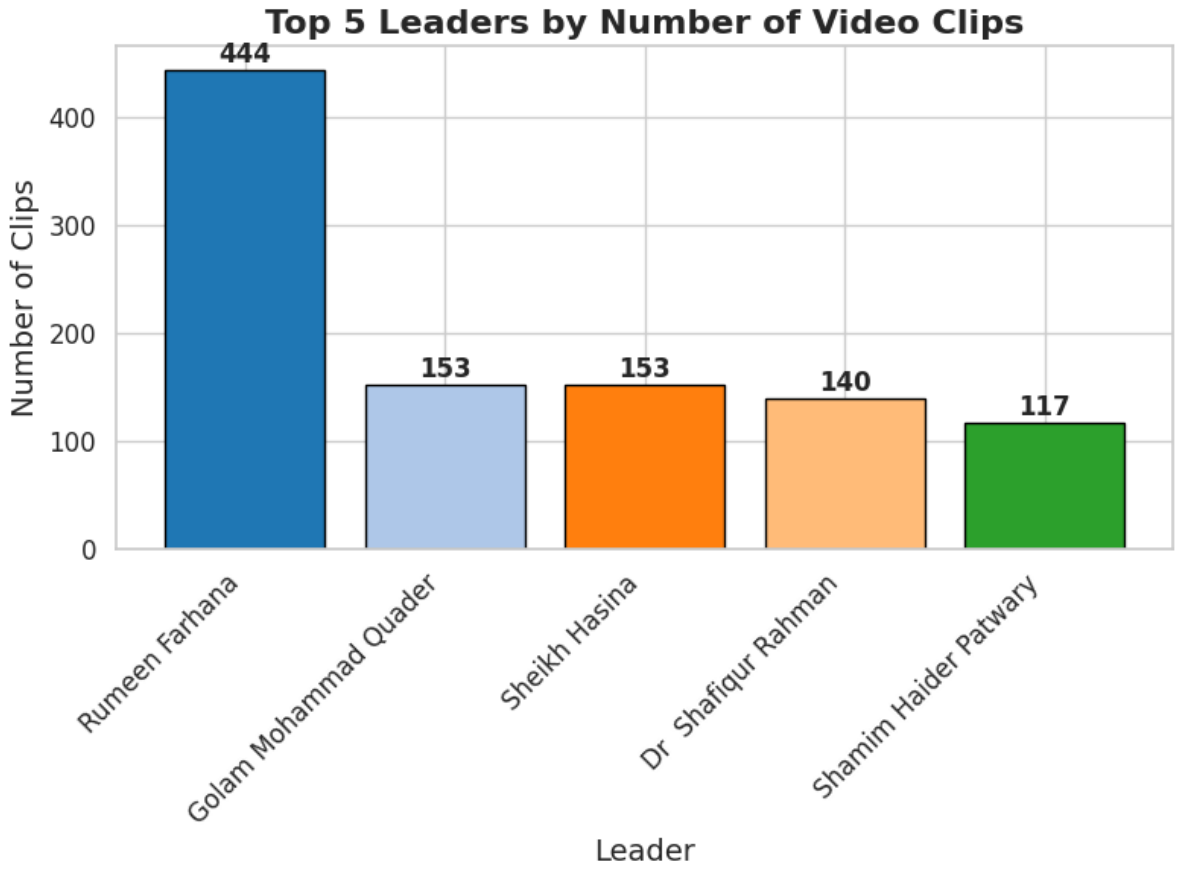

Figure 3.3: Top 5 leaders with most number of videos distribution

Chapter 4

Methodology

The process of data collection began with us collecting YouTube speeches for the Political Leaders, which we then manually annotated multiple times to ensure reliability. After that, we used Google Speech Recognition, where we converted our annotated speeches to transcribed datasets. The transcripts were then manually annotated for reliability, and then label encoding was done. Then, we used the labeled dataset to do data augmentation using SMOTE, which essentially bal- ances the traits across the dataset. After that, cross validation was applied to the dataset for robustness and better generalization. Our baseline system used tra- ditional machine learning, which is an SVM classifier. After establishing baseline metrics, we used advanced transformer-based approaches such as BanglaBert and XLM-RoBERTa. Subsequently, input texts were tokenized using BanglaBert's to- kenizer with sequences padded to 512 tokens, and fine-tuned the models. After all that, the model was trained to classify traits. The research systematically compared all the models' performance through comprehensive metrics (accuracy, F1, recall, and support) and confusion matrices.

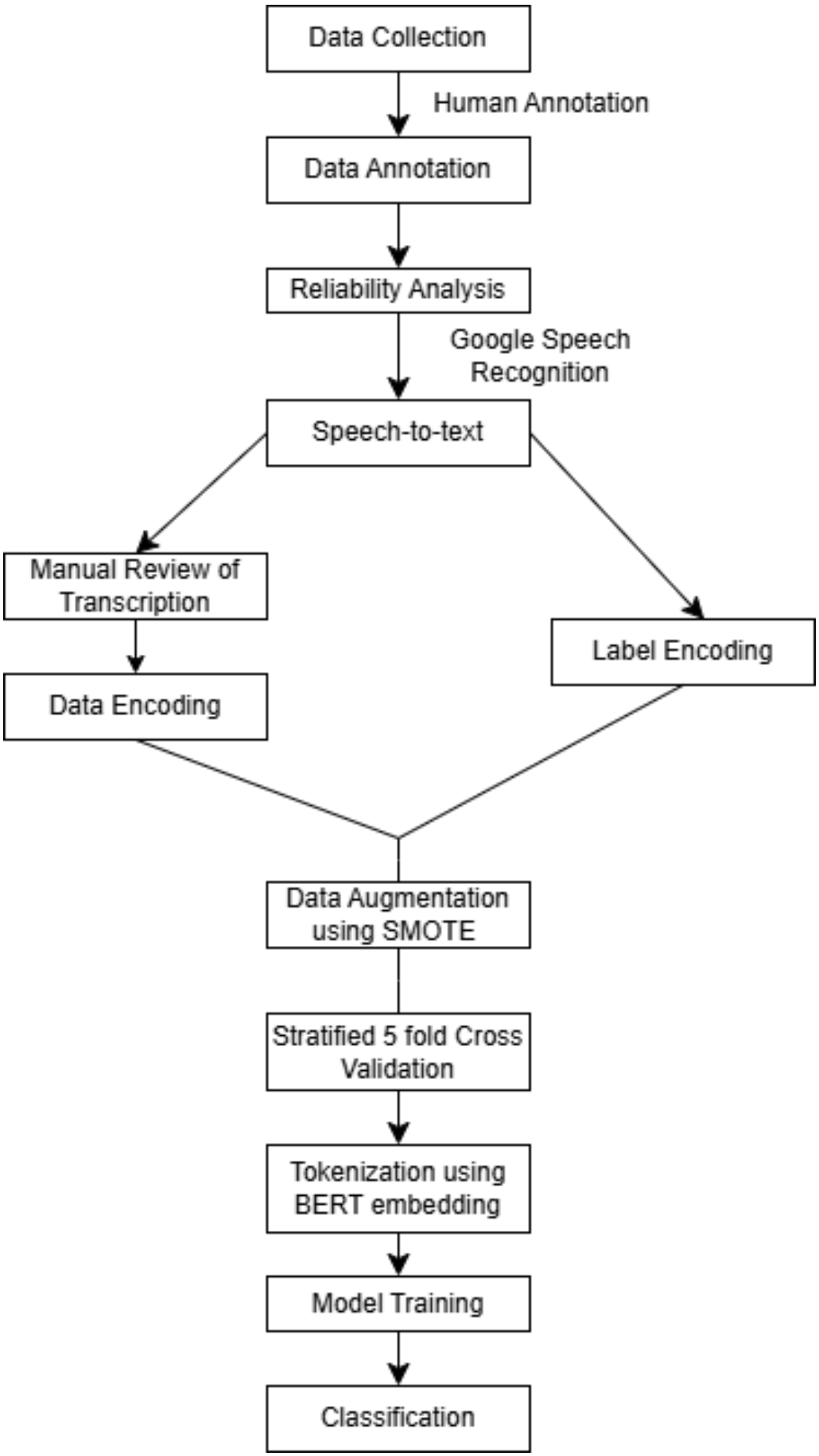

Figure 4.1: Top-level overview of the proposed system

4.1 Preprocessing

Preprocessing is a significant step in organizing raw data for machine learning models. To make sure our transcriptions are as accurate as possible, we implemented cleaning steps very specifically, such as removing duplicate words and extra whitespace using regular expressions to maintain text formatting and uniformity across all chunks. Since we extract raw text from the speech, there were some audio segments that were not clear and were unable to capture certain transcriptions. Instead of skipping the video, the unclear segments were marked [unintelligible]. We manually iterated through the transcripts and reviewed those segments; if identified, then they were corrected with the appropriate words, otherwise removed. Due to time-based segmentation, some sentences had an abrupt interruption or were incomplete. Such cases were manually handled by either completing the sentence through revisiting the actual audio or discarding them if they had no semantic meaning or clarity. Through custom Python scripts and necessary human assessment, we conducted our preprocessing of the dataset.

Before preprocessing	After preprocessing
আমি তার পুত্র আমি আমি কোনদিন মিথ্যা মিথ্যা কথা বলি না	আমি তার পুত্র আমি কোনদিন মিথ্যা কথা বলি না
বাংলাদেশ স্বাধীন হয়েছে মুক্তিযুদ্ধ দিয়ে প্রয়োজনে আবার আরেকটা	বাংলাদেশ স্বাধীন হয়েছে মুক্তিযুদ্ধ দিয়ে প্রয়োজনে আবার আরেকটা মুক্তিযুদ্ধ করব
আমরা বাংলাদেশ কিনে দিয়ে সামনে এগিয়ে যাব মনে রাখতে হবে যেন স্বাধীনতা বিরোধী শক্তি সাম্প্রদায়িক [unintelligible]	আমরা বাংলাদেশ কিনে দিয়ে সামনে এগিয়ে যাব মনে রাখতে হবে যেন স্বাধীনতা বিরোধী শক্তি সাম্প্রদায়িক শক্তি বসে থাকবে

Table 4.1: Sample preprocessed text

4.2 Data Augmentation

In our transcripted dataset, there was considerable class imbalance between the personality traits assigned to the samples, which would introduce a learning bias due to the presence of dominating classes such as Extraversion(658)and Neuroticism(495) here. To prevent this bias, SMOTE (Synthetic Minority Over-Sampling Technique) was used to improve the class imbalance and make all the traits equal. SMOTE resolves this issue by generating new samples synthetically for the minority classes. The SMOTE algorithm generates synthetic samples using the formula:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_k - \mathbf{x}_i)$$

Where:

- $\mathbf{x}_i$: Minority main class instance
- $\mathbf{x}_k$: Randomly picked k-nearest neighbor of $\mathbf{x}_i$ from the minority class
- λ: Random number in between 0 and 1 ($\lambda \sim \text{Uniform}(0, 1)$)
- $\mathbf{x}_{\text{new}}$: Generated synthetic samples in minority class

4.2.1 Text Vectorization and Class Imbalancing

At first, using TF-IDF (Term Frequency- Inverse Document Frequency), the textual data is converted into a numerical vector dataset, and then SMOTE was applied to the vectorized trained dataset, creating synthesized new samples for the minority classes by interpolating between existing samples in the TF-IDF feature space. However, the newly created vectors do not correspond to any actual text data, therefore, a nearest neighbour matching approach was used to retrieve the closest textual counterparts from the original training data. The retrieved text samples and their labels were assembled to form a balanced training set, while still maintaining the integrity of the text samples and creating a balanced dataset where each trait has 658 samples.

The TF-IDF weight for the variable t in the data d from a collection of data D is calculated as follows:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

where:

$$\text{tf}(t, d) = \text{term frequency of } t \text{ in } d$$
$$\text{idf}(t, D) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

with:

- $|D|$ = total number of data
- $|\{d \in D : t \in d\}|$ = number of data has the term t

SMOTE is a preprocessing technique used because we had an imbalanced dataset. It works by identifying minority class samples in the corpus and uses K-NN(K-nearest neighbours). In our case, we chose SMOTE as it performs better in comparison to ADASYN, cause the imbalance is moderate up to 1:5, ADASYN works better for more imbalanced classes like (1:100).

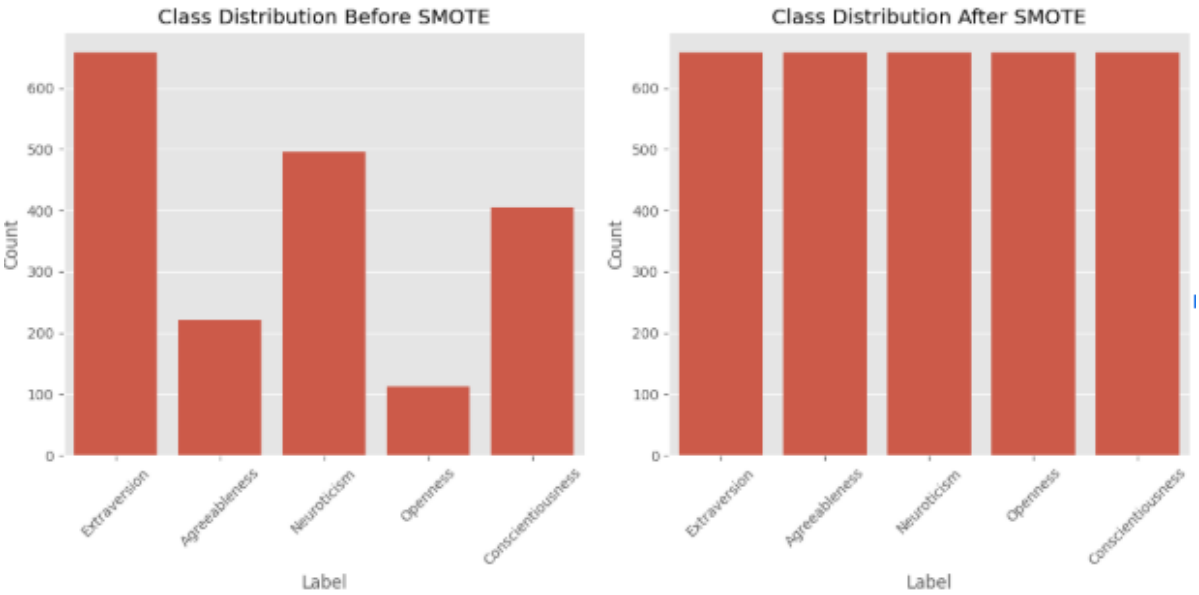

Figure 4.2: Balanced Dataset using SMOTE

4.3 Stratified Cross-Validation

To ensure efficient assessment of model proficiency, a 5-fold stratified cross-validation approach was applied. This method partitions the dataset into five equally sized folds, with each fold maintaining the original class distribution of personality traits. Considering the fact that there was class imbalance in the dataset, where certain traits such as Extraversion or Neuroticism exhibit more prevalence compared to

their counterparts, it was essential to apply stratification to obstruct biased learning and assessment. For each iteration of the cross-validation process, one fold, which is equivalent to 20% of the dataset, was used as the test dataset, and the other four folds, which were equivalent to 80% were used for training and validation. However, within this 80%, an additional 80/20 split was done to the training portion. As a result, each fold consists of 64% of the training data, 20% test data, and 16% validation data. The validation data was used to observe the model's proficiency during training each training epoch, allowing early stopping if there were no further improvements. This helped prevent overfitting. Also, this cross-validation process ensured that every sentence is tested only once in the test dataset while it was added to the training during the four iterations. This approach ensured complete data utilization of the dataset and ensured a comprehensive assessment of the model's performance, even for the underrepresented traits.

4.4 Feature Extraction

Ahead of model training, all the textual training, test, and validation datasets underwent preprocessing through the official tokenizer that has been provided by the model csebuetnlp/banglabert. Exclusively trained for Bangla, this tokenizer uses the WordPiece algorithm, which breaks down words into subunits to productively handle sparse or unseen tokens. Also, each text was tokenized into two components: input_ids, which are numerical representations of tokens and attention_mask, which are binary indicators where 1 indicates actual tokens and 0 indicates paddings. Considering the fact that all Bert models have a standardization of 512 tokens for maximum input length, all text sequences were either padded or truncated to meet this requirement. For longer sequences, they were truncated to fit within the limit, while shorter sentences were padded to maintain uniformity and dimension compatibility. After tokenization, the split training, test, and validation datasets were converted into Hugging Face Dataset objects using the Hugging Face Library for efficient handling. To ensure that these datasets were compatible with Pytorch, the datasets were formatted to keep only certain fields such as input_ids, attention_mask, and labels, which are essential for supervised classification using the AutoModelForSequenceClassification model. As a result of this structured format, the datasets could be compatible with the Hugging Face training pipeline, resulting in more consistent training, evaluation, and batching across folds.

Chapter 5

Description of the model

5.1 Statistical Model

We have also used a statistical model such as Support Vector Classifier (SVM), which is a supervised learning algorithm used in classification and regression tasks. SVM is chosen by us due to its ability to handle high-dimensional feature spaces, such as TF-IDF transformed text. Also, SVM is known to be beneficial for working with sparse, linear or non-linear datasets. As SVM does margin maximization, where it focuses on finding the hyperplane that maximizes the margin between classes, it helps improve unseen data. It also uses kernel functions such as Radial Basis Function (RBF), which it handles the non-linear relationships in data, which is important for trait detection as it can help detect contextual variations [1]. SVMs also handle the high dimensionality of TF-IDF vectors, which makes them suitable for such NLP tasks. For SVM, we have applied data augmentation and stratified cross-validation. We have also balanced the data using SMOTE to handle the data imbalance.

5.2 Transformer

In 2017, Transformer was highlighted as the groundbreaking deep learning approach introduced by Google researchers. It has surpassed all previous approaches, such as RNN, CNN, LSTM, and GRU, since they have the limitation of parallelism and processing data sequentially. For our study, we have utilized two transformer-based models, such as BanglaBERT and XLM-RoBERTa, which are built on the transformer encoder architecture proposed by Vaswani et al. (2017). Transformer contains multiple layers of self-attention and feed-forward networks, which allows it to model semantic relationships without being reliant on recurrence and convolution [8]. It relies on the multi-head self-attention mechanism to understand the words and effectively capture the intricate contextual nuances between them. The multi-head attention mechanism also allows for capturing complex generalization tasks and ensures stability in the input sequence with its encoder-decoder structure. Also, the attention mechanism computes an attention mask which uses it in the encoder-decoder structure to find which words are interconnected [8].
BanglaBERT, which is a pre-trained model, is specifically trained in the Bangla corpus, which makes it highly efficient in detecting Bangla language, as during training, it is able to understand the words and the grammar. In contrast, XLM-Roberta, which is also a pre-trained model, has been trained on a multilingual dataset of over 100 languages, providing it with the ability to generalize across texts,

even in multilingual contexts, and excel in low-resource languages such as Bangla. Here, both of these models are built upon a transformer-based architecture and use the self attention mechanism to understand and learn contextual and semantic representations, which eventually helps us to classify traits according to particular Big Five model traits.

5.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers)[18], a pre-trained transformer-based deep learning model, was developed by Google AI in 2018. This bidirectional model has established benchmarks in NLP tasks such as entity recognition and sentiment analysis. It is due to its bidirectional encoding ability, which allows it to understand the context of a token by looking through preceding and following words while encoding. It is also passed through two training processes: pre-training and fine-tuning. In pre-training, the BERT is trained through a large amount of unlabeled data, where it learns in two ways- Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, random words are masked, and the model tries to guess them from neighbouring words, and in NSP, the model is trained to understand whether one sentence follows another or not. After the pre-training phase, the BERT is fine-tuned using labeled datasets. During fine-tuning, the model integrates task-specific input (text classification, Q/A , NER) layers and output layers. The BERT-based model consists of 12 transformer encoder layers. For each token, it produces 768-sized vectors. Each layer consists of multi-head self-attention, which computes contextual dependencies of a sentence, and a feed-forward neural network that apply non-linear transformation from the self-attention output.

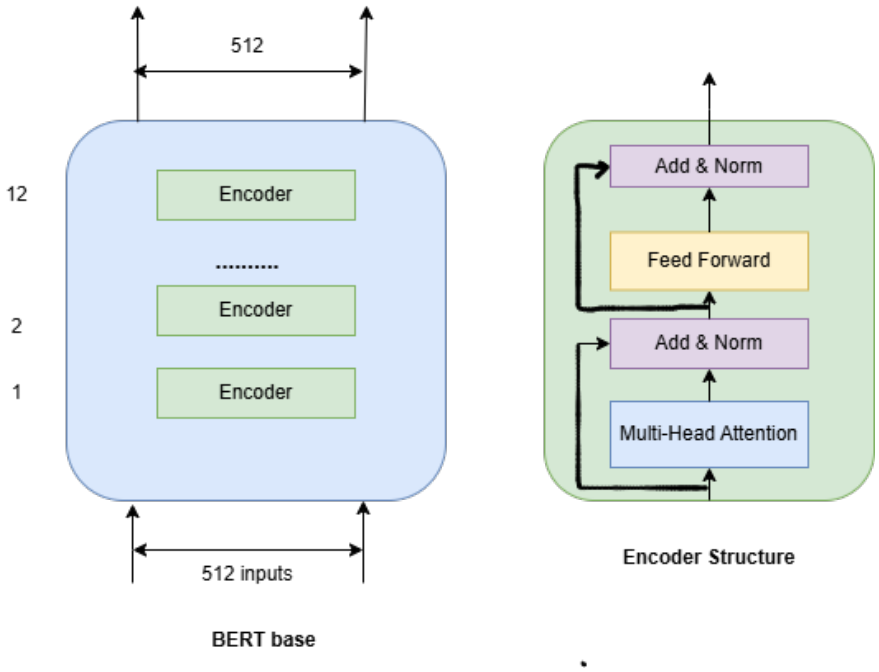

Figure 5.1: BERT model architecture

5.2.2 BanglaBERT

For our research, which is the prediction of traits from the speeches, we have opted for the BERT models. One of them is BanglaBert, which is a Pre-Trained Transformer-based model that is specifically designed as well as fine tuned to handle all the contextual nuances and fine-grained details of the Bengali language by using the BERT architecture, which BanglaBERT has adopted [19]. In this study, the deployment of BanglaBert, that has been used by us was released through the CSE BUET NLP group, who developed this model. Utilizing the BERT framework, BanglaBert employs bidirectional encoding which processes the intricate nuances of Bangla language by analyzing the context that appears before and after each word. This bidirectional encoding allows it to build meaningful, context-aware embeddings that enhance and improve the accuracy in downstream tasks such as classifying traits from transcribed speeches. BanglaBert is based on ELECTRA architecture and due to the unavailability of the Bangla data, the developers of BanglaBERT have gathered diverse data from blogs, news, and ebooks from around 110 Bangla websites as well as Bangla Wikipedia.

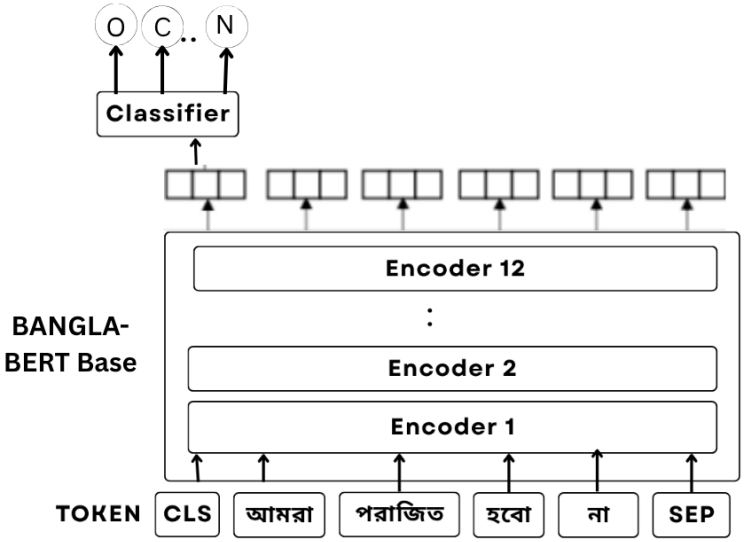

Figure 5.2: BanglaBERT model architecture

5.2.3 RoBERTa

XLM-Roberta [13] is a pre-trained language independent multilingual transformer model that has processed over 100 languages which allows it to understand diverse linguistic patterns. This model is trained on a huge corpus and it employs a supervised learning mechanism where 15% of the tokens within each dataset is usually masked and the model is trained to predict these hidden tokens using the contextual information from the unmasked tokens. This allows XLM-Roberta to understand contextual and semantic nuances without labeled supervision. Also, as this model is language independent, it is efficient for inferring Bengali language which in itself is a low-resource language despite the linguistic diversity being very high.

5.3 Fine Tuning

The fine-tuning procedure involved adapting the pre-trained Bert model for personality classification through supervised training on our labelled dataset. The fine-tuning process was implemented using Hugging Face's Trainer API which provided an optimized framework for efficient training of transformer architecture.
For optimization, AdamW was employed to minimize the categorical cross-entropy loss[3] . The loss function is defined as such:

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}) \tag{5.1}$$

Where:

- N: The number of classes
- C: Number of samples (batch size)
- y_{ij}: Ground truth (1 if class j is correct for sample i)
- $\hat{y}_{ij}$: Predicted output for class j of sample i

5.3.1 Fine-Tuning of BanglaBERT

For BanglaBert, we added a task-specific classification layer which consists of a dense neural network with 768 input dimensions and 5 output units corresponding to our personality traits on top of the existing transformer framework while keeping all transformer layers trainable. The training was conducted for a maximum of 25 epochs although early stopping was applied based on validation performance. To prevent overfitting and ensure computational efficiency, an early stopping mechanism was integrated into the fine-tuning process. Initially testing with patience=2, it emphasized that the training was set to terminate if the validation F1 score showed no improvement over two successive epochs but we observed early termination before the model could fully converge. Hence, the patience was increased to 3 epochs.This configuration allowed for consistent and controlled training across all folds in the cross-validation pipeline.
During hyperparameter tuning, we tested weight decay values with 0.01 and 0.1 to evaluate the regularization strength. After evaluating, we retained the default L2 regularizer with a value of 0.01 as this prevented overfitting by constraining weight magnitudes while maintaining model capacity. To ensure training stability and prevent parameter update explosions, we applied gradient clipping with a maximum norm of 1.0. By constraining the norm of the gradients, this technique ensured that the parameter updates remained within a stable range which is especially crucial when fine-tuning large pre-trained models like BanglaBert.

5.3.2 Fine-Tuning of XLM-RoBERTa

The XLM-RoBERTa was fine-tuned to classify Bangladeshi political leaders' speech texts into Big Five personality traits. We initialized the pretrained xlm-roberta-base model and added a task specific classification task for our five traits. To improve the model's understanding of Bengali, each input text was prefixed with the <bn> language tag, enabling more effective multilingual contextualization. We maintained the identical optimization parameters from our BanglaBert implementation, including weight decay of 0.01, gradient clipping with a maximum norm of 1.0, maximum epochs of 25 and early stopping patience of 3. However, the learning rate was kept 2e-5 as XLM-RoBERTa's multilingual nature requires gentler fine-tuning.

During implementation of both the models, evaluation was conducted at the end of each epoch and using F1-score as the primary metric to select the best-performing model. The load_best_model_at_end=True ensured that the model used for final evaluation was restored from the epoch that achieved the highest f1 score. Besides, both the models prioritized the highest validation F1 score by setting greater_is_better=True which was the model checkpoint with the highest validation F1-score. To manage storage efficiently, we limited the number of saved model checkpoints by setting save_total_limit=2. This configuration ensured that only the two best model parameters were retained for each fold while older checkpoints were automatically discarded. Furthermore, to optimize computational efficiency during fine-tuning, we enabled mixed-precision training (FP16) through the fp16=True flag in Hugging Face's TrainingArguments and this resulted in faster training and lower memory use.

Hyperparameter	**Value**
Learning Rate	3e-5
Training Batch Size	16
Eval Batch size	32
Epochs	25
Weight Decay	0.01
max_grad_norm	1.0

Table 5.1: Hyperparameter of BanglaBERT

Hyperparameter	**Value**
Learning Rate	2e-5
Training Batch Size	8
Eval Batch size	16
Epochs	25
Weight Decay	0.01
max_grad_norm	1.0

Table 5.2: Hyperparameter of XLM-RoBERTa

Chapter 6

Result analysis

In this study, we begin our analysis with a Support Vector Machine(SVM) classifier for personality trait detection from Bangladeshi political speech data to establish preliminary performance metrics, followed by transformer-based approaches: BanglaBert (csebuetnlp/banglabert) and XLM-RoBERTa base. By using the SVM model as a traditional machine learning baseline, it allows us to evaluate how much transformer based models have improved in performance and enables us to analyze how much semantic and contextual modeling adds to trait prediction. Moreover, it also helps evaluate whether large pre-trained models can actually analyze traits from political discourse. Equations 6.1,6.2, 6.3,6.4 shows the metrics and how they give quantitative scores to assess each class.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6.3}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.4}$$

Where:

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

6.0.1 Baseline Performance

The SVM model established an average accuracy of 0.4584 and an F1 score of 0.4259 across five test folds. This demonstrates that while the model showed some ability to distinguish between traits, especially traits like Extraversion and Neuroticism, it struggles with more nuanced traits such as Openness. The F1 score is lower than accuracy, which reveals poor performance on minority classes such as Openness, and hence confirming SVM struggles with imbalanced traits despite using SMOTE. However, the training time of SVM shows fast training which confirms SVM'S computational efficiency for mid-size text data.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVM	0.4584	0.4545	0.4584	0.4259

Table 6.1: SVM Classification Report

6.0.2 Transformer-based model Performance

After establishing our baseline with the SVM model, we evaluated on a more context-aware deep learning approach using the csebuetnlp/banglabert. The results demonstrate a significant performance improvement across key metrics.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Extraversion	0.8089	0.8433	0.8257	823
Agreeableness	0.8201	0.8201	0.8201	278
Neuroticism	0.8914	0.8900	0.8907	618
Openness	0.8673	0.6901	0.7686	142
Conscientiousness	0.8124	0.8043	0.8083	506

Table 6.2: Classification Report of BanglaBERT

Accuracy			**0.8357**	**2367**
Macro avg	0.8400	0.8096	0.8227	2367
Weighted avg	0.8360	0.8352	0.8349	2367

Table 6.3: Accuracy of BanglaBERT

Compared to the baseline SVM model, which achieved a weighted F1 score of 0.4259, the csebuetnlp/banglabert model effectively doubled the performance. This confirms the benefit of using transformer-based models that capture the semantic meaning in texts. We observed in Table 6.2 that the highest F1 score of 0.8907 for Neuroticism,indicating that this trait was most consistently identified, followed by Extraversion, which had an F1 score of 0.8257. On the other hand, the model showed a comparatively lower recall of 0.6901 for Openness which indicates that it missed identifying a notable number of true cases leading to under-representation of this class in the predictions.

To further improve performance on the SVM baseline, we evaluated another transformer-based multilingual model on the XLM-RoBERTa base. The model achieved an overall accuracy of 0.7858, with a macro F1-score of 0.7792 and weighted F1 score of 0.7860. Agreeableness has the lowest precision of 0.6826, suggesting that nearly 32% of its predicted instances are incorrect. These results indicate a substantial improvement over the baseline SVM approach. However, the csebuetnlp/banglabert model still outperformed XLM-RoBERTa across all metrics. This indicates that csebuetnlp/banglabert, being a monolingual Bangla model, showed better generalization on the trait classification task. In contrast, XLM-RoBERTa, a multilingual model, lagged slightly due to its more generalized language representation. Hence, for our research, monolingual pretrained models are clearly advantageous.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Extraversion	0.8016	0.7412	0.7702	823
Agreeableness	0.6826	0.8201	0.7451	278
Neuroticism	0.8398	0.8398	0.8398	618
Openness	0.8739	0.6831	0.7668	142
Conscientiousness	0.7477	0.8024	0.7741	506

Table 6.4: Classification Report of XLM-RoBERTa

Accuracy			**0.7858**	**2367**
Macro avg	0.7891	0.7773	0.7792	2367
Weighted avg	0.7904	0.7858	0.7860	2367

Table 6.5: Accuracy of XLM-RoBERTa

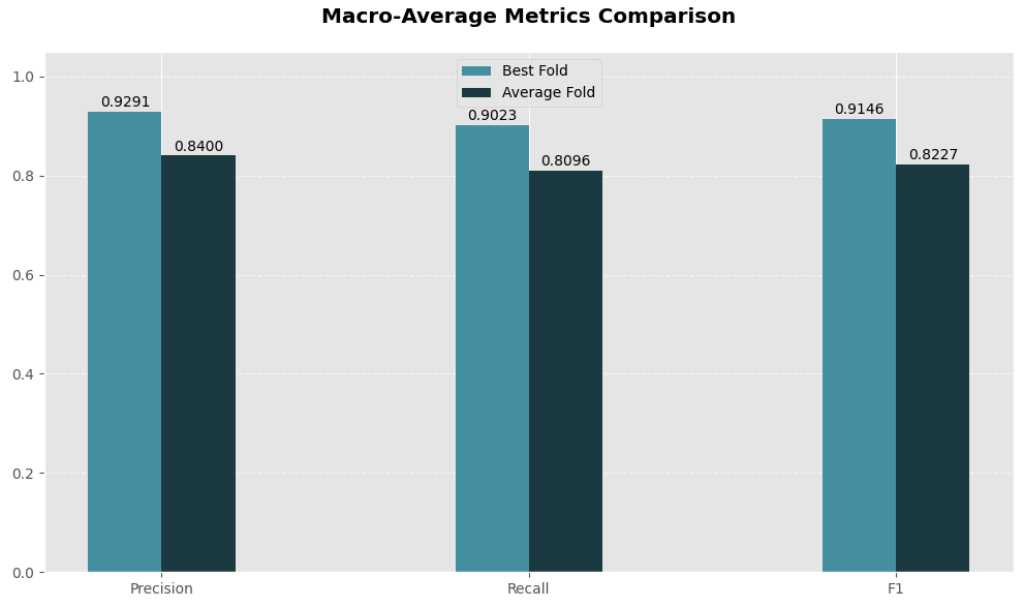

Figure 6.1: Macro average metrics comparison of BanglaBERT

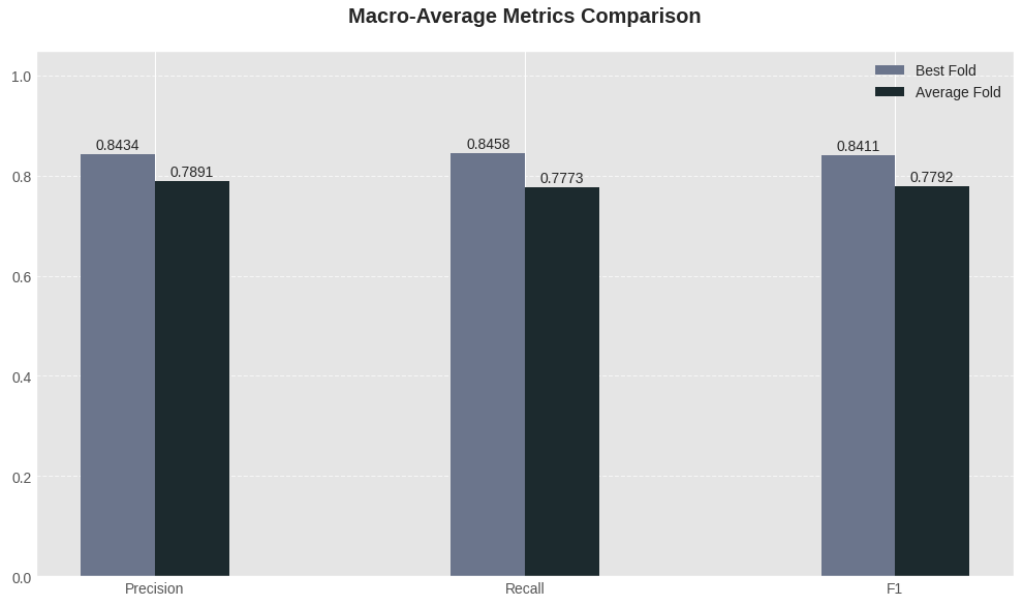

Figure 6.2: Macro average metrics comparison of XLM-RoBERTa

The above graphs provide a comparison of the macro average metrics (Precision, Recall, F1) for BanglaBERT model shown in Figure 6.1 and XLM-RoBERTa shown in 6.2. They are both evaluated on the aggregated folds and the best fold where for each metric, left bar represents the Aggregated fold and the right bar represents the Best Fold. Notably, for BanglaBERT, among all metrics , Precision had the best score for the Aggregated and Best Fold whereas for XLM-RoBERTa, Recall had the highest score for Best fold and achieves high score in Precision for Aggregated folds.

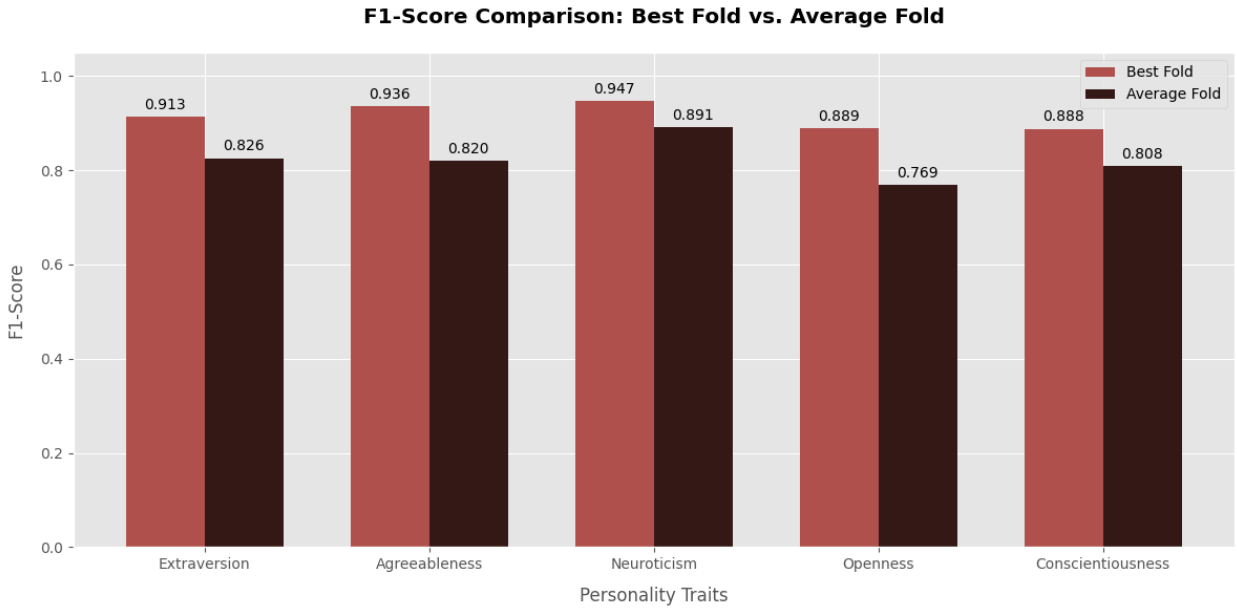

Figure 6.3: F1 score comparison by Traits

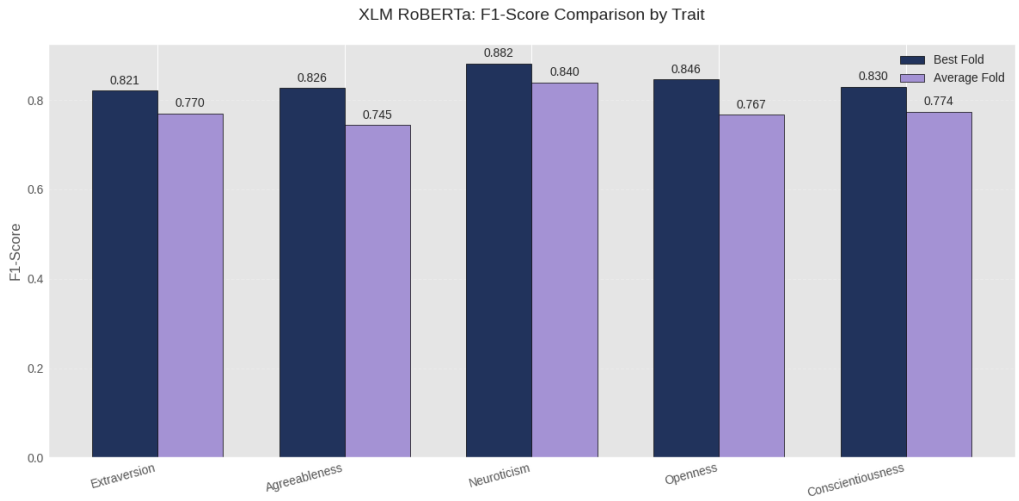

Figure 6.4: F1 score comparison by Traits of XLM-RoBERTa

Figure 6.3 and 6.4 illustrates Big Five Personality Traits comparing F1 scores between average across all folds and best fold. The model demonstrates strong performance for Neuroticism trait in both XLM-RoBERTa and BanglaBERT showing the robustness and stability of detecting the dominant trait of the classifier.

6.0.3 Confusion Matrix

As seen from all the confusion matrices generated, it is seen that despite using SMOTE to balance our data numerically, the models consistently underperformed in detecting Openness compared to other traits like extraversion. This addresses a critical limitation that, while SMOTE addressed numerical imbalance by generating synthetic samples, it failed to resolve the underlying semantic imbalance. Possible reasons for this occurring might be due to using SMOTE with TF-IDF, which duplicates new sentences without adding truly distinct Openness markers. Additionally, political speeches naturally emphasize leadership (Extraversion) but rarely include clear Openness phrases. Future work should prioritize collecting richer Openness sentences beyond synthetic samples.

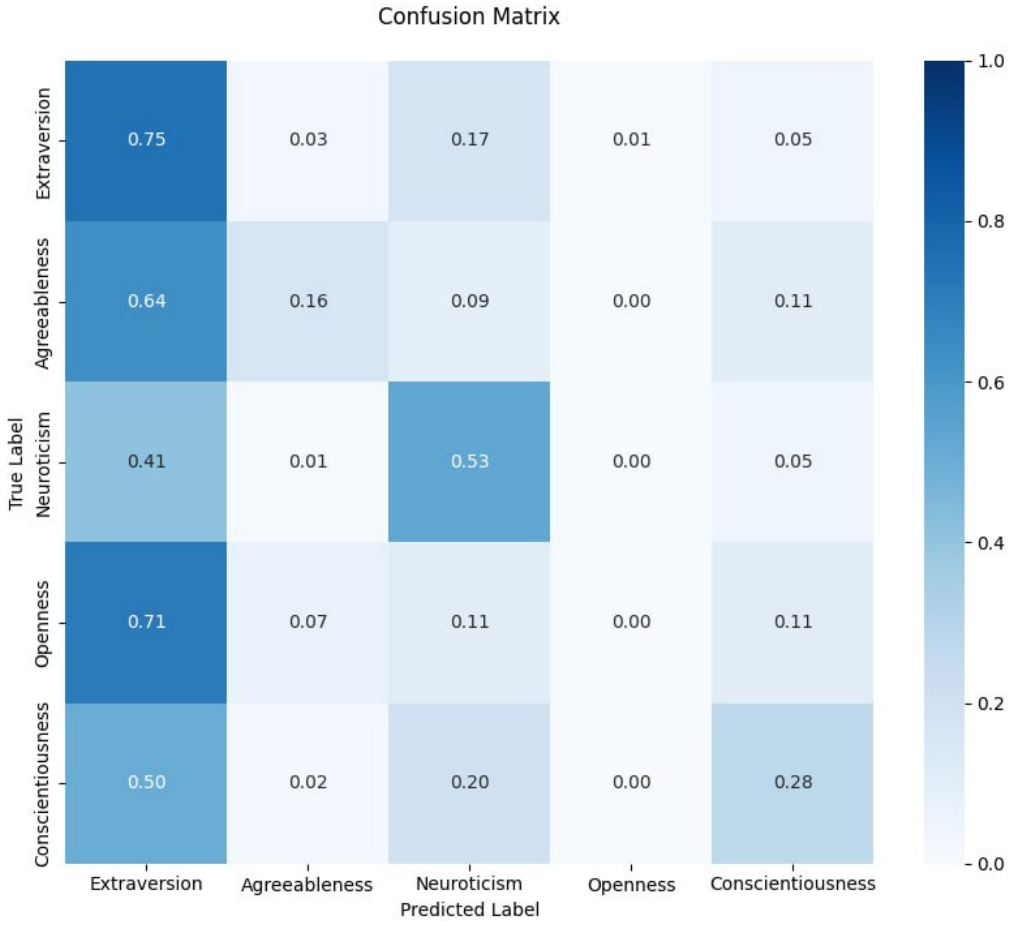

Figure 6.5: Confusion Matrix for SVM

Figure 6.5 displays the confusion matrix for the baseline SVM model. The confusion matrix shows that the SVM model performed relatively well with Extraversion (75%) and Neuroticism (53%) but struggled with Agreeableness, Openness and Conscientiousness. This suggests that the SVM model was biased toward predicting Extraversion and it lacks the contextual understanding required for reliable personality trait classification.

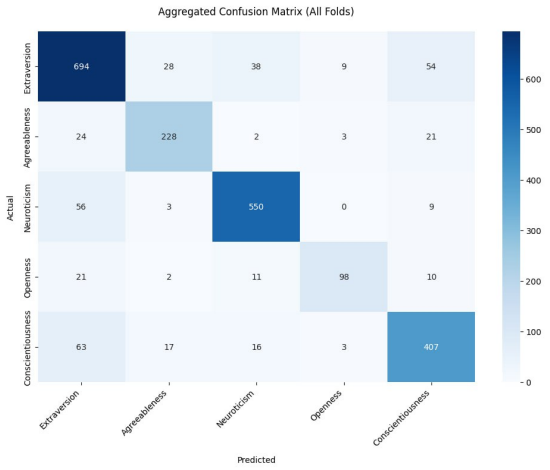

Figure 6.6: Confusion Matrix for Aggregated BanglaBERT

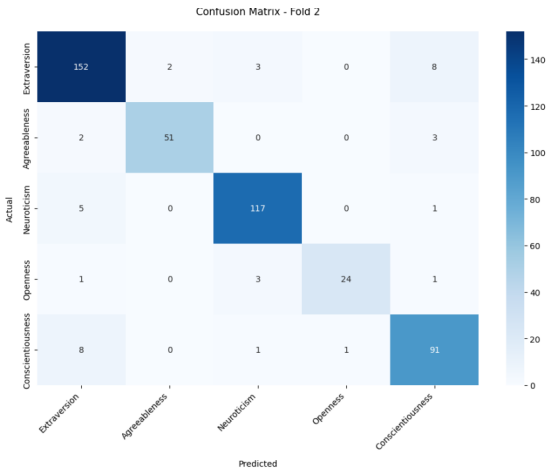

Figure 6.7: Confusion Matrix for BanglaBERT Best Fold

Confusion matrix of BanglaBERT Aggregated folds:
Figure 6.6 shows the aggregated confusion matrix across all cross-validation folds for the BanglaBert model. The model performed particularly well in predicting traits like Extraversion, Neuroticism and Conscientiousness. Predictions for Agreeableness and Openness also showed noticeable improvements over the SVM model. While BanglaBert outperformed SVM, the aggregated confusion matrix reveals a few recurring misclassifications.

Confusion matrix of BanglaBERT Best fold:
Figure 6.7 displays the confusion matrix for the best fold where BanglaBert achieved its best performance. Misclassifications were limited and mostly occurred between traits like Extraversion and Conscientiousness.

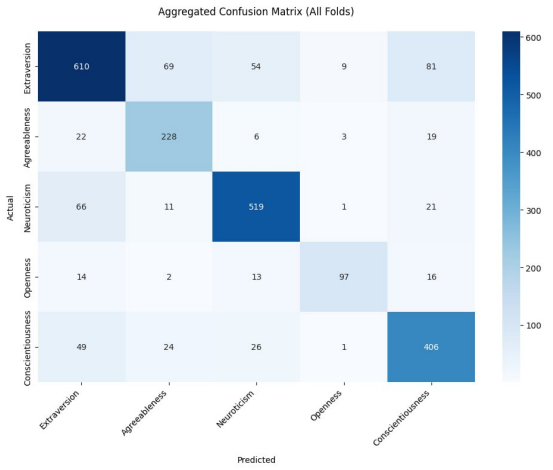

Figure 6.8: Confusion Matrix for Aggregated XLM-RoBERTa

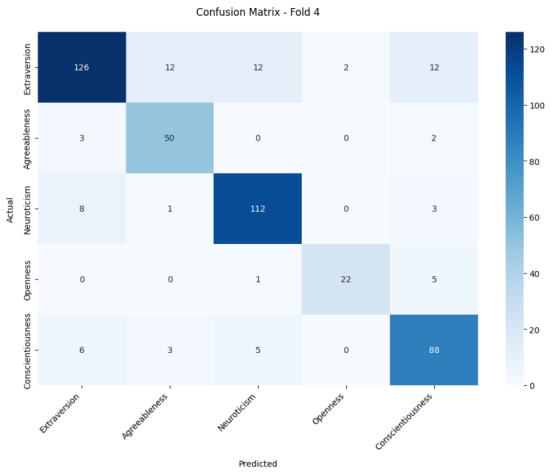

Figure 6.9: Confusion Matrix for XLM-RoBERTa Best Fold

Confusion matrix of XLM-RoBERTa Aggregated folds:
Figure 6.8 demonstrates the aggregated confusion matrix across all cross-validation folds for the XLM-RoBERTa model. The model did a decent job in predicting traits like Extraversion and Neuroticism. Agreeableness and Openness were harder to predict thus leading to incorrect predictions.

Confusion matrix of XLM-RoBERTa Best fold:
Figure 6.9 shows the confusion matrix from the best fold which represents the peak performance of XLM-RoBERTa. The model showed strong performance on Extraversion, Neuroticism and Conscientiousness. Agreeableness was moderately accurate while Openness was the most difficult to classify due to its low number of correct predictions.

The model struggles with label confusions which can be caused by several reasons like in our case one of the reasons could be overlapping traits, Insufficient training data, data drift etc. Though the aggregated fold confusion matrix shows higher misclassification for both the models(BanglaBERT and XLM-RoBERTa) in comparison to the best fold of the confusion matrix, we at the end took the aggregated fold approach as our final result, keeping two considerations in mind:

1. The aggregated approach is a comprehensive way to represent model performance across all the data trait, as it reduces bias.
2. Generalized robustness is given more priority over choosing any single fold's performance.

The more misclassification in the aggregated fold represents overfitting in the best fold performance, making the aggregated fold more reliable for estimating true trait performance.

Paper	**Model**	**Classification**	**F1 Score (%)**	**Accuracy (%)**	**Dataset**
Personality Traits Detection in Bangla (IC-CIT 2020) [10]	Naive bayes Logistic Regression SVM CNN LSTM BiLSTM	Text	64.8 67.4 69.6 71.4 70.9 73.3	66.3 68.5 70.8 72.2 71.7 74.5	Primary (Bangla)
Unveiling Personality Traits through Bangla Speech [2]	DistilRo (DistilBERT+ RoBERTa) BiG (BiLSTM +GRU)	Speech-to-Text	89	NA	Primary (Bangla)
EMPOLITICON: Context and Emotion Classification of Political Speeches [16]	Empoliticon-Context (Ensemble) Empoliticon-Emotion (Ensemble)	Context Emotion	NA	73.13 53.07	Primary (English)
Proposed Model	SVM BanglaBERT XLM-RoBERTa	Text (Speech Transcripts of Politics)	42.59 82.27 77.96	45.84 83.57 78.58	Primary (Bangla)

Table 6.6: Comparative Analysis with Existing Studies

In this section, we aim to compare our analysis with the existing studies on personality traits. In our study, where we have curated a textual dataset consisting of Bangla speeches from Political leaders, prominently from Bangladesh, we have used the traditional model SVM which has obtained an accuracy of 42.59%, however, the transformer-based models used such as BanglaBERT and XLM-RoBERTa achieved 89.07% and 78.58% respectively. In a similar study, where both traditional models and deep learning models have been used, the deep learning model Bi-LSTM outperformed by achieving an accuracy of 74.5% and F1-score of 73.3%[10]. On the other hand, [2] used two ensemble models such as DistilRo (DistilBERT + RoBERTa) and BiG (Bi-LSTM + GRU) for trait classification on Bangla speech and attained substantially higher F1-scores of 89% and 90% respectively, showing potential for speech-based approaches. Alongside this, we have [16] which explored political discourse and introduced EMPOLITICON-context and EMOPOLITICON-emotion, two ensemble models which classify context and emotion from long political transcripts. Here, [10] and [2] primarily concentrate on text-based data, similar to us, whereas [16] demonstrates the advantages of integrating multimodality and speech-based features on trait classification.

Chapter 7

Conclusion

While our study has presented valuable observations into trait detection using the Big Five Traits using the statistical model and the transformer-based deep learning models, there is still considerable scope for future research and refinement. Since our research has been done on a textual dataset, there is further possibility of research for multimodal analysis surrounding the tone, pitch, facial expressions, and gestures. Also, integrating multimodal analysis would greatly improve the accuracy of trait classification. Moreover, our dataset is only focused on Bangla speeches. In the future, the analysis could be extended by making the dataset multilingual, which would add more variations to the classification of trait analysis. Additional ensemble or hybrid models could also be used to improve the performance. Our paper assessed how to detect traits using the Big Five personality traits, such as Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, from the speeches of the political leaders of Bangladesh while creating our own primary dataset. It explored how to apply NLP techniques while using traditional machine learning models such as Support Vector Machines (SVM) and transformer-based deep learning models such as BanglaBert and XLM-RoBERTa. The results from our analysis indicate that using textual data, BERT models perform moderately well due to their ability to understand contextual data, whereas SVM performs competitively when integrated with TF-IDF. SVM achieved an F1-score of 0.4259, demonstrating that this model struggles with class imbalance and nuanced traits. For our dataset, the transformer models significantly outperformed SVM, with BanglaBert reaching an F1-score of 0.8907 and XLM-R reaching 0.8398, both in neuroticism, respectively. While there are certain limitations, this research offers a data-centric approach to analyzing traits of the political leaders through traits, while also opening up pathways for further analysis and exploration. While our study has presented valuable observations into trait detection using the Big Five Traits using the statistical model and the transformer based deep learning models, there is still considerable scope for future research and refinement. Since our research has been done on a textual dataset, there is further possibility of research for multimodal analysis surrounding the tone, pitch, facial expressions, and gestures. Also, integrating multimodal analysis would greatly improve the accuracy of trait classification. Moreover, our dataset is only focused on Bangla speeches. In the future, the analysis could be extended by making the dataset multilingual, which would add more variations to the classification of trait analysis. Additional ensemble or hybrid models could also be used to improve the performance.

Bibliography

[1] V. Apostolidis-Afentoulis, *Svm classification with linear and rbf kernels*, Jul. 2015. DOI: 10.13140/RG.2.1.3351.4083.

[2] D. Shell, "Unveiling personality traits through bangla speech using morlet wavelet transformation and bigs," *JOURNAL OF LATEX CLASS FILES.*, 2015.

[3] A. Hassan, M. R. Amin, A. K. A. Azad, and N. Mohammed, "Sentiment analysis on bangla and romanized bangla text using deep recurrent models," in *2016 International Workshop on Computational Intelligence (IWCI)*, 2016, pp. 51–56. DOI: 10.1109/IWCI.2016.7860338.

[4] F. Liu, J. Perez, and S. Nowson, "A language-independent and compositional model for personality trait recognition from short texts," Oct. 2016.

[5] M. Mahmudun, M. Tanzir, and S. Ismail, "Detecting sentiment from bangla text using machine learning technique and feature analysis," *International Journal of Computer Applications*, vol. 153, pp. 28–34, Nov. 2016. DOI: 10.5120/ijca2016912230.

[6] M. Peters, M. Neumann, M. Iyyer, *et al.*, "Deep contextualized word representations," Feb. 2018.

[7] D. Pizzolli and C. Strapparava, "Personality traits recognition in literary texts," in *Proceedings of the Second Workshop on Storytelling*, F. Ferraro, K. Huang Ting-Hao, S. M. Lukin, and M. Mitchell, Eds., Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 107–111. DOI: 10.18653/v1/W19-3411. [Online]. Available: https://aclanthology.org/W19-3411.

[8] A. Gillioz, J. Casas, E. Mugellini, and O. Abou Khaled, "Overview of the transformer-based models for nlp tasks," Sep. 2020, pp. 179–183. DOI: 10.15439/2020F20.

[9] I. Onyenwe, N. Samuel, N. Mbeledogu, and O. Grace, "The impact of political party/candidate on the election results from a sentiment analysis perspective using anambradecides2017 tweets," *Social Network Analysis and Mining*, vol. 10, p. 17, Dec. 2020. DOI: 10.1007/s13278-020-00667-2.

[10] U. Rudra, A. N. Chy, and H. Seddiqui, "Personality traits detection in bangla: A benchmark dataset with comparative performance analysis of state-of-the-art methods," Dec. 2020, pp. 1–6. DOI: 10.1109/ICCIT51783.2020.9392722.

[11] R. Sawhney, A. Wadhwa, S. Agarwal, and R. Shah, "Gpols: A contextual graph-based language model for analyzing parliamentary debates and political cohesion," Jan. 2020, pp. 4847–4859. DOI: 10.18653/v1/2020.coling-main.426.

[12] S. Stajner and S. Yenikent, "A survey of automatic personality detection from texts," Jan. 2020, pp. 6284–6295. DOI: 10.18653/v1/2020.coling-main.553.

[13] A. Jana and A. Bhowmick, "Sentiment analysis for bengali using transformer based models," Dec. 2021.

[14] E. Hossain, O. Sharif, and M. M. Hoque, "MemoSen: A multimodal dataset for sentiment analysis of memes," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, *et al.*, Eds., Marseille, France: European Language Resources Association, Jun. 2022, pp. 1542–1554. [Online]. Available: https://aclanthology.org/2022.lrec-1.165.

[15] G. Caron and S. Srivastava, "Manipulating the perceived personality traits of language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2370–2386. DOI: 10.18653/v1/2023.findings-emnlp.156. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.156.

[16] A. A. Efat, A. Atiq, A. Abeed, A. Momin, and M. G. R. Alam, "Empoliticon: Nlp and ml based approach for context and emotion classification of political speeches from transcripts," *IEEE Access*, vol. PP, pp. 1–1, Jan. 2023. DOI: 10.1109/ACCESS.2023.3282162.

[17] A. Liaquat, S. Zaman, and S. Saeed, "Analysis of political leader's speeches in socio-political perspective: A study of critical discourse analysis," vol. 9, pp. 590–597, Oct. 2023. DOI: 10.5281/zenodo.8387697.

[18] N. Farhan, S. Sarker Joy, T. Binte Mannan, and F. Sadeque, "Gazetteer-Enhanced Bangla Named Entity Recognition with BanglaBERT Semantic Embeddings K-Means-Infused CRF Model," *arXiv e-prints*, arXiv:2401.17206, arXiv:2401.17206, Jan. 2024. DOI: 10.48550/arXiv.2401.17206. arXiv: 2401.17206 [cs.CL].

[19] F. Faria, M. B. Moin, R. Mumu, M. Abir, A. Nawar, and S. Alam, *Motamot: A dataset for revealing the supremacy of large language models over transformer models in bengali political sentiment analysis*, Jul. 2024. DOI: 10.48550/arXiv.2407.19528.