

An Explainable Transformer Based Architecture for Detecting Sentiment from Bengali Comments

by

**MD. Rokibul Hasan (171014033)
MD. Armanur Rakib (192014032)
Shajedul Islam (192014033)
Tion Das (201014044)**

*Capstone Project Report (CSE499) submitted in partial fulfillment of the
requirements for the degree of*

Bachelor of Science in Computer Science and Engineering

Under the supervision of

Sovon Chakraborty



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF LIBERAL ARTS BANGLADESH**

FALL 2024

© *MD. Rokibul Hasan, MD. Armanur Rakib, Shajedul Islam, Tion Das*
All rights reserved

DECLARATION

Project Title	An Explainable Transformer Based Architecture for Detecting Sentiment from Bengali Comments
Authors	MD. Rokibul Hasan, MD. Armanur Rakib, Shajedul Islam, Tion Das
Student IDs	171014033, 192014032, 192014033, 201014044
Supervisor	Sovon Chakraborty

We declare that this capstone project report entitled *An Explainable Transformer Based Architecture for Detecting Sentiment from Bengali Comments* is the result of our own work except as cited in the references. The capstone project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

MD. Rokibul Hasan
171014033

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

MD. Armanur Rakib
192014032

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Shajedul Islam
192014033

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Tion Das
201014044

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Date: December 26, 2024



Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Mohammadpur, Dhaka - 1207

CERTIFICATE

This is to certify that the capstone project report entitled **An Explainable Transformer Based Architecture for Detecting Sentiment from Bengali Comments**, submitted by **MD. Rokibul Hasan** (171014033), **MD. Armanur Rakib** (192014032), **Shajedul Islam** (192014033), **Tion Das** (201014044) is a undergraduate students of the **Department of Computer Science and Engineering** has been examined. Upon recommendation by the examination committee, we hereby accord our approval of it as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of *Bachelor of Science in Computer Science and Engineering*.

Sovon Chakraborty, Lecturer

Dept. of CSE, University of Liberal Arts Bangladesh

Dr. Nafees Mansoor, Associate Professor

Dept. of CSE, University of Liberal Arts Bangladesh

Dr. Mohammad Ashikur Rahman, Assistant Professor

Dept. of CSE, University of Liberal Arts Bangladesh

Suravi Akter, Lecturer

Dept. of CSE, University of Liberal Arts Bangladesh

Place: Dhaka

Date: December 26, 2024

ACKNOWLEDGEMENTS

We are most grateful to our research supervisor *Sovon Chakraborty*, for providing us the opportunity to perform research under him. His incredible leadership, wise feedback, and steadfast support have been a major driver of the work we do. His commitment and encouragement taught us to aim high and tackle our research with concise clarity and confidence. It has been a pleasure to learn from him and be mentored by him throughout my time.

We owe our special thanks to the honorable faculty members of the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh. They have helped shape our lives through their teaching, guidance, and encouragement.

To our parents, thank you for your unconditional love and support without which, this would not be possible. A big thank you to our friends and family as well for their perpetual support and encouragement on this journey.

Most importantly, we would like to thank all of the people who have helped us learn and succeed in this pursuit.

MD. Rokibul Hasan, MD. Armanur Rakib, Shajedul Islam, Tion Das

University of Liberal Arts Bangladesh

Date: December 26, 2024

Dedicated to
My Family

– MD. Rokibul Hasan

Dedicated with affection to
My Sister

– MD. Armanur Rakib

Dedicated
To my beloved parents

– Shajedul Islam

Dedicated to
my loving Father

– Tion Das

ABSTRACT

In this work, we try to design a reliable system for hate speech detection and biases in Bengal text data by employing advanced machine learning and natural language processing (NLP) methodologies. Considering the objective to reveal political, religious and gender bias as well as general hate speech the research responds to the questions regarding the linguistic complexities of the Bengali language and the scarcity of its sources. To obtain a dataset, the author gathered information from various sources and was careful to label all videos correctly. Prespecified models like GRU, LSTM were used and BanglaBERT was used to evaluate contextual embedding and bidirectional processing which outperformed other models. Techniques like LIME to improve the understanding and trust of the users on predicted outcomes by the proposed system were incorporated. The findings of this research have key applications for establishing safer digital environments for the Bengali speaking population and provide main steps to future work on detecting hate speech in low resource language.

Keywords: Hate Speech and biases, Religious Bias, Gender Bias, Bengali Text Analysis, Natural Language Processing, Machine Learning

Contents

1	Introduction	1
1.1	Problem statement	2
1.2	Aims, objectives and Motivation	3
1.2.1	Aims	3
1.2.2	Objective	4
1.2.3	Motivation	5
1.3	Project Specification	5
2	Background Study and Analysis	7
2.1	Overview of Bengali Hate Speech Detection	7
2.2	Datasets in Bengali Hate Speech Detection	8
2.3	Models and Techniques	9
2.3.1	Traditional Machine Learning Models	9
2.3.2	Deep Learning Approaches	9
2.3.3	Feature Extraction Techniques	10
2.3.4	Multilingual and Transfer Learning Models	10
2.4	Analysis of Datasets	10
2.5	Performance of Models	11
3	Methodology	15
3.0.1	Precision, Recall, and F1-Score	15
3.0.2	Precision	16
3.0.3	Recall	16
3.0.4	F1-Score	16
3.1	Dataset description	17
3.1.1	Characteristics of the dataset	17

3.1.2	Dataset Fields	17
3.1.3	Selection Criteria for Data	18
3.1.4	Bias Categorization	18
3.1.5	Applicability	19
3.2	Data collection Procedure	19
3.2.1	Identification of the Source	19
3.2.2	Annotation Process	20
3.2.3	Preprocessing	20
3.2.4	Ethical Considerations	20
3.2.5	Challenges	21
3.3	Data quality	21
3.3.1	Language-Specific Challenges	21
3.3.2	Labeling and Categorization	22
3.3.3	Data Imbalance	23
3.3.4	Data Preprocessing Quality	24
3.3.5	Data Quality Assurance	24
3.4	Final data outlook	25
3.4.1	Data Structure and Columns	25
3.4.2	Dataset Size	26
3.4.3	Preprocessing Steps Applied	27
3.4.4	Data After Preprocessing	27
3.4.5	Handling Challenges Specific to Bengali Language	28
3.5	Utilized model	29
3.5.1	Random Forest Classifier (RF)	29
3.5.2	Logistic Regression (LR)	30
3.5.3	Naive Bayes (NB)	32
3.5.4	Long Short-Term Memory (LSTM)	33
3.5.5	Gated Recurrent Unit (GRU)	34
3.5.6	BERT (Bidirectional Transformers for mask language models)	37
3.5.7	DistilBERT	38
3.5.8	The Explainable AI-enabled BanglaBERT	40

4	Design and Development	43
4.1	Data Collection and Preprocessing	43
4.2	BanglaBERT Model	44
4.2.1	Pretraining and Fine-Tuning	44
4.2.2	Advantages of BanglaBERT	44
4.3	Explainability with LIME	44
4.4	Evaluation and Metrics	45
4.5	Deployment	45
5	Result analysis	46
5.1	BanglaBERT with Explainable AI	46
5.2	GRU Model	47
5.3	LSTM Model	47
5.4	Explainable AI in Action	48
6	Conclusions	49
6.1	Social Issues	49
6.2	Legal Issues	50
6.3	Ethical Impact	50
6.4	Environmental Impact	51
6.5	Brief Summary	51
6.6	Future works	52
	References	53

List of Figures

3.1	Methodology	15
3.2	Dataset	26
3.3	Post-processed data example	28
3.4	Random Forest Architecture (Source: https://tinyurl.com/mr4baddu)	30
3.5	Logistic Regression Model (Source: https://tinyurl.com/bde9vrpa)	31
3.6	Naive Bayes Model (Source: https://tinyurl.com/p6y6auyv)	33
3.7	LSTM Architecture (Source: https://tinyurl.com/y7ndjt5u)	35
3.8	GRU Architecture (Source: https://tinyurl.com/mxrvbj6a)	36
3.9	BERT Architecture (Source: https://tinyurl.com/2sc6pemw)	38
3.10	DistilBERT Architecture (Source: https://tinyurl.com/3ut3vy47) . .	39
5.1	BanglaBERT with Explainable AI classification report	46
5.2	GRU Model Calculate precision, recall, and f1-score	47
5.3	Evaluate the LSTM Model	48
5.4	BanglaBERT with Explainable AI	48

List of Tables

2.1	Litarature Review	14
3.1	Dataset Size	26
3.2	Comparison of Models	42

Chapter 1

Introduction

Modern society is untraceable without the Internet and social networks have become a natural and essential tool for communication, where people can state personal feelings or share opinions anytime. These systems foster extensive communication between numerous groups of the population, which make them promising areas for various types of positive emotions and worthwhile experiences, as well as negative phenomena, which include, for example, hate speech and language abuse [Emon et al. \(2019\)](#), [A. K. Das et al. \(2021\)](#). Sentiment analysis helps in improving customers satisfaction, content filtering, and decision making processes for both businesses and government.

Bengali is a language used by 300 million people online and is considered one of the most used languages on the internet. That said, there is a long way to go for sentiment analysis research in Bengali as it is far behind the commonly used language, English. Added to the richness of the Bengali script, and morphological and dialectal variation, aspects that are known to complicate the language itself presents a host of problems to models of NLP [M. Das et al. \(2022\)](#). Moreover, code-mixed language, especially the Bangladeshi variant known as Banglish (Bengali in the Latin script), is widely used and makes it difficult to detect sentiments because of the informal and unstandardised written form used [Rahman et al. \(2024\)](#).

With the most recent improvement in deep learning, especially transformer models like BERT, sentiment analysis forms the new digital forefront where machine learning models can learn contextual and semantic information in the text [Keya et al. \(2023\)](#). These models have shown great accuracy in different NLP tasks, notably hat-speech and abusive language detection in Bengali [Emon et al. \(2019\)](#), [A. K. Das et al. \(2021\)](#), [M. Das et al. \(2022\)](#), [Keya et al. \(2023\)](#). Nonetheless, owing to

their ‘black box’ nature, such models are less effective in situations where model interpretability and explainability are critical. To overcome this challenge, a novel area called Explainable AI or XAI has been developed to offer understanding of model prediction results to encourage trust and enhance the decision-making processes [Keya et al. \(2023\)](#).

Although great progress has been made in the identification of hate speech and abusive language in Bengali text, there is still a large loophole in the whole context of the sentiment analysis, especially in the biased nature of the models. The majority of the existing research focuses on the classification accuracy at the cost of understanding the decision-making processes which limits the application of the sentiment analysis systems in critical and high risk contexts [Emon et al. \(2019\)](#), [A. K. Das et al. \(2021\)](#), [M. Das et al. \(2022\)](#), [Rahman et al. \(2024\)](#), [Keya et al. \(2023\)](#). Furthermore, there is a lack of availability of qualified, annotated Bengali data for the same purpose which adds further complexity to build more reliable and generalizable models.

According to these challenges, this research introduces an Explainable Transformer-Based Architecture to classify sentiment from Bengali comments. Extending the prior research in the hate speech and abusive language datasets, our study aligns the transformer models with the methods of decision explanation to accomplish sentiment analysis with high accuracy and to provide real-time interpretability of the outputs. We believe that by doing so, we can contribute to the improvement of the reliability and integrity of opinion mining systems, from which practical applications are expected, for which understanding of why certain results come up is critical.

1.1 Problem statement

In the past few years, hate speech, and other biased presentations in digital arrays have become an emerging social issue that affects people and various groups. Despite the amount of work done for detecting hate speech in highly resourced languages such as English, low resource languages like Bengali are still fairly unexplored. The lack of moderation from Bengali-speaking users exposes millions of the population to such negative content thus enhancing discrimination and social enmity.

Specifically, complexity is incurred with the language and culture of Bengali since the language has a rich and diverse nature. As the problem of bias involves political, religious or gender biases, such risks can only be defined with reasonable cause and a certain level of insight. This is compounded by a dearth of annotated datasets, language specific NLP tools, and effective and efficient detection models.

This particular work aims at tackling these problems through the proposed reliable Bengali hate speech detection and classification using machine learning. Therefore, this paper aims to achieve accurate identification while avoiding cultural bias by using the high-quality annotated dataset and advanced techniques of BanglaBERT. The use of explanation models provides increased reliability and clarity in predicting situations, thus advancing the application of the system in practical use cases.

1.2 Aims, objectives and Motivation

1.2.1 Aims

The broad goal of this work is therefore to build a good and explainable Bengali hate speech detection system. Specifically, it seeks to:

- **Identify and Categorize Hate Speech:** Develop an effective model for identifying hate speech in Bengali text and classify it as political, religious discrimination or gender or general.
- **Leverage Advanced NLP Models:** Use better understanding or more advanced models such as BanglaBERT which are particularly designed for Bengali language solving the language barriers of the mentioned language.
- **Enhance Explainability:** Include Explainability AI tools for LIME, in order to increase understanding of model-related decisions in the system and improve end-users' confidence.
- **Develop Practical Applications:** Design an easy to use platform and application programming interface for implementing the developed system in

actual-use contexts such as enforcing moderation of social media platforms, facilitating research work, etc.

- **Promote Safer Digital Spaces:** To support reducing the amount of hatred speech on the Internet, in enfranchising the Bengali speaking population, and lessening continued creation of toxic content.

1.2.2 Objective

- **Data Collection:** Gather and correct a thorough collection of Bengali text data about hate speech and related biases from news sources, internet forums, and social media platforms.
- **Annotation:** Hand-annotate the dataset to guarantee accurate classification into gender prejudice, religious bias, political bias, or generic hate speech without any particular bias.
- **Text Preprocessing:** Use NLP methods like stop-word removal, tokenization, stemming, and advanced feature representation with Word2Vec or Transformer-based embeddings.
- **Model Development:** Create and optimize a variety of machine learning models, such as deep learning architectures (like LSTM and BERT) and conventional classifiers (like SVM and logistic regression) that are suited for Bengali text.
- **Hyperparameter tuning:** To avoid overfitting, optimize model performance using approaches such as cross-validation and grid search.
- **Evaluation:** Use metrics such as accuracy, precision, recall, and F1-score to assess models' effectiveness in hate speech identification.
- **Demonstration System:** Create a user-friendly interface or API to show the system's functionality in real-world circumstances.
- **Future Scope:** Extend the system to include code-mixed text and other low-resource languages, as well as update datasets and models to reflect changing hate speech trends.

1.2.3 Motivation

The fast growth of web-based services has opened a great opportunity for communication and interconnection in Bengali-speaking areas. But it also reminds of hate speech and biased content that in recent years have found fertile ground for development, deepen social splits, and encourage discrimination. Political, religious, and gender biases are especially typical, and many of them reveal essential social injustice, which requires efficient intervention measures.

However, Bengali is spoken by millions of people all over the world, but it is paid very little attention in NLP. Existing tools, mainly designed in the English language, do not cover up the characteristic features of Bengali text. This is not only replicating culture exclusion but is also hampering efforts of better online climates.

This project is motivated by the weakness of the current hate speech detection methods that only apply general solutions in the English language. Using progressive methods of machine learning and Explainable AI this study will assist Bengali communities to navigate a safer digital world.

1.3 Project Specification

The project specification outlines the key components and requirements necessary to achieve the stated aims:

Functional Requirements

- **Hate Speech Detection:** The system needs to be capable of correctly classifying hate speech within Bengali text and categorizing into pre-defined biases, such as political, religious, gender and other.
- **Explainability:** LIME is a technique of Explainable AI that the system has to support in order to offer clear and measurable justification of its decisions.

Non-Functional Requirements

- **Performance:**

- Achieve those characteristics of high accuracy and precision as well as having good rates of Recall and F1-Score.
- Reduce delay to cater real time oriented services.
- **Scalability:**
 - Of particular importance is the ability of the system to process large amounts of data, comparable to those of social networks.
- **Security and Privacy:**
 - The rights are, respectively, the protection of user data and ensuring such data's anonymity.
 - Contact law regulations such as GDPR.

Technical Specifications

- **Data:**
 - **Dataset:** Gathered from several sources reviewed and tagged in terms of bias categories and sentiment polarity.
 - **Size:** More than ten thousand of annotated samples for sufficient training.
- **Models:**
 - Our model BanglaBERT for hate speech detection.
 - Explaining the integration of EXPLAINABLE AI with the built LIME model.
- **Technologies:**
 - **Programming Languages:** Python (primary).
 - **Frameworks:** TensorFlow, PyTorch, Hugging Face Transformers.
 - **Deployment:** Docker for containerization process, Flask/Django for building API.
- **Infrastructure:**
 - Remote host servers for training of the system and hosting the same.
 - Introducing Graphics Processing Unit acceleration for both model training and model inference.

Chapter 2

Background Study and Analysis

This chapter presents a wide review of related works on Bengali hate speech detection, highlighting advancements in datasets and models, identifying key challenges, and potential advances in detection accuracy, fairness, and scalability that might avail in the years ahead, as we seek to make progress on serious online toxicity challenges.

2.1 Overview of Bengali Hate Speech Detection

Detection of hate speeches in Bengali has become a trending topic in recent times due to the growing incidence of online toxicity. Hate speech data sources are mostly social media platforms including Facebook, Twitter, and YouTube. Hate speech detection is an important field of Natural Language Processing (NLP) that aims to detect any kind of harmful, offensive, or discriminatory language on social media. The use of automated hate speech detection systems is becoming vital given the extreme scale of content propagation on social media such as Twitter and Facebook [Davidson et al. \(2017\)](#). Despite this, the scarcity of resources, such as annotated datasets, and advanced language models has limited advancements. Although considerable progress has been made towards such detection for languages such as English, hate speech detection particularly designed for Bengali is an unexplored space that needs more attention and this type of detection requires specific solutions.

Few studies have attempted to tackle Bengali hate speech detection and sentiment analysis, and there has been disappointing advancement in both the creation

of datasets and the formulation of the models [Haque et al. \(2023\)](#). Bengali language is morphologically rich which makes it challenging to detect hate speech in Bengali [Al Maruf et al. \(2024\)](#). Several studies have highlighted the need for language-specific datasets and customized detection models. Moreover, researchers should design multilingual frameworks, construct context-sensitive representations, and promote fairness in detectors. Overcoming these challenges can help the field progress towards a more accurate and inclusive hate speech detection system that enhances the safety of digital environments for Bengali-speaking populations.

2.2 Datasets in Bengali Hate Speech Detection

This is crucial for machine learning to ensure basing on hate speech detection and sentiment analysis system datasets. The existing datasets for hate speech detection in Bengali are limited in size and scope. For example, the biggest size dataset of 30,000 comments was created by focusing on hate speech categories across domains like politics, religion, and entertainment. It consists of 10,000 hate speech comments and 20,000 non-hate speech comments [Romim et al. \(2021\)](#). This data set is a valuable resource, but has some limitations as well. The biggest issue with human labeling is that an annotator may have different ideas of what hate speech is than the typical labeling tool, which can lead to very different results. The data is also limited in its coverage because the dataset is focused on specific domains, and the hate speech seen across other content types online may not generalize. Other, smaller annotated datasets (5,000 to 15,000 comments) have targeted specific types of hate speech, e.g., speech in religious or political context [Al Maruf et al. \(2024\)](#). To address the unique challenges posed by code-mixed content, researchers have compiled datasets specific to Banglish, which were further categorized into 5,000 comments with sexual, racial, and religious hate speech labels [M. H. Islam et al. \(2024\)](#). In a related paper, we discuss how the nature of the data Banglish makes detection more complex, as models need to recognize hate speech in both Bengali and English [M. H. Islam et al. \(2024\)](#). This has made it even more challenging to develop precise detection systems for Banglish as the text often contains more than regular slang, informal expression, and non-standard words.

2.3 Models and Techniques

Bengali hate speech detection and sentiment analysis has been addressed using various machine learning and deep learning models [Nath et al. \(2024\)](#). Various machine learning (ML) and deep learning (DL) models have been proposed to detect hate speech, especially in the Bengali language [Ishmam & Sharmin \(2019\)](#), including classical ones such as Support Vector Machines (SVM), to advanced Approaches, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and hybrid CNN-BiLSTM architectures.

2.3.1 Traditional Machine Learning Models

In these early studies, traditional machine learning models, such as SVMs, were utilized giving moderate results while suffering from the limitation of manual feature extraction . Early analyses worked with models such as Support Vector Machines (SVM), Random Forest and Logistic Regression. These models did moderately well, with SVM models achieving up to 87.5% accuracy on balanced datasets. However, their performance on detecting nuanced or implicit hate speech was not very promising [Romim et al. \(2021\)](#).

2.3.2 Deep Learning Approaches

Recently, deep learning (DL) architectures, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM), have gained attention [Al Maruf et al. \(2024\)](#) [Jahan & Oussalah \(2023\)](#). For example:

G-BERT: This model was used combining the Berts and Gpus models with great results! A Bengali-specific variant, G-BERT, has achieved competitive performance in hate speech detection tasks by exploiting the contextual ability of transformer-based architectures [M. H. Islam et al. \(2024\)](#) [WATANOBE \(n.d.\)](#). This shows the promise of incorporating deep learning and transfer learning approaches to enhance accuracy and robustness in hate speech detection for Bengali.. It uses learned embeddings from pre-trained vectors that can capture complex linguistic patterns and contextual dependencies [M. H. Islam et al. \(2024\)](#).

CNN-BiLSTM Hybrid Models: Recent works have proposed hybrid models that combine architectures. For instance, the CNN-BiLSTM model introduced in the work by Biswas et al. This approach is effective for text-based detection, which is a sequential data, as CNNs can be used to extract local features from the input whereas, BiLSTMs can catch sequential dependencies in the text [M. H. Islam et al. \(2024\)](#). These models combine the advantages of CNN for local feature extraction and BiLSTM for long range dependencies. They excel in identifying subtle and context-aware hate speech [Rahman et al. \(2024\)](#).

2.3.3 Feature Extraction Techniques

Effective hate speech detection relies on robust feature extraction methods. Common techniques include Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BoW), and custom word embeddings like BengFastText. BengFastText, trained on 250 million Bengali words, has been instrumental in improving the performance of classifiers [Al Maruf et al. \(2024\)](#) [Keya et al. \(2023\)](#).

2.3.4 Multilingual and Transfer Learning Models

Leveraging multilingual pre-trained models such as mBERT and XLM-RoBERTa has shown promise in tackling data scarcity by transferring knowledge from resource-rich languages to Bengali [M. H. Islam et al. \(2024\)](#).

2.4 Analysis of Datasets

The majority of a hate speech detection and sentiment analysis model's effectiveness comes from the data quality and data diversity. impose on the data Using diff types Like Religion, Politics, and entertainment to make the model robust but some type of genres can be overrepresented while some are underrepresented, which can restrict generalizability [Romim et al. \(2021\)](#) [Jobair et al. \(2023\)](#). Banglish datasets, on the other hand, are unique, and come with their own preprocessing, and feature extraction processes [M. H. Islam et al. \(2024\)](#) which give rise to new classes of exciting problems and techniques, towards being able to parse multilingual content. Along with expanding and diversifying datasets, this data assists researchers to represent the intricacies of hate speech in Bengali in a better way.

2.5 Performance of Models

Traditional machine learning approaches have always been performed by hybrid models, like CNN-BiLSTM, in terms of performance. These most successful models learn local and global patterns in text and can reach 90% accuracy in benchmark datasets [Romim et al. \(2021\)](#). Multilingual embeddings have been increasingly effective in improving detection accuracy in low-resource settings through transfer learning methods. Use of custom embeddings such as BengFastText have also improved linguistic feature representation and through that improve task performance [Badri et al. \(2022\)](#). It emphasizes the need for efficient architecture with effective feature extraction methods for hate speech detection.

Literature Review Table

Literature Review:

Paper Name	Year and Publisher	Contribution	Shortcomings
Automated hate speech detection and the problem of offensive language	2017, ResearchGate	Proposed a dataset for automated hate speech detection and distinguished between hate speech and offensive language.	i) Difficulty in detecting context-dependent hate speech;
Unmasking Toxicity: A Comprehensive Analysis of Hate Speech Detection in Banglish	2021, Elsevier	Focused on hate speech detection in Banglish highlighting unique challenges in this hybrid language.	i) Small dataset; ii) Limiting generalizability; iii) Focuses mainly on Banglish;
Measuring the impact of bias in NLP systems	2018, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing	Explored the impact of bias in NLP systems, particularly in models for hate speech detection, emphasizing the need for fairness and transparency.	i) Primarily focused on English; ii) Few practical implementations;

Paper Name	Year and Publisher	Contribution	Shortcomings
Understanding religious bias in social media text	2020, ACM	Investigated religious bias in social media texts, highlighting the role of NLP in identifying and addressing such biases.	i) Focuses only on religious bias; ii) Ignoring other types of bias; Lack of diverse social media sources;
Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation	2020, Journal of Language Technologies	Introduced a large Bengali dataset for hate speech detection.	i) Does not fully address informal language;
Bangla Hate Speech Detection on Social Media Using Attention-Based Recurrent Neural Network	2020, De Gruyter - Journal of Intelligent Systems	Proposed an attention-based RNN model using GRU and LSTM to classify hate speech into seven categories.	i) Moderate accuracy (77%); ii) Lack of comprehensive dataset preprocessing; iii) Insufficient explainability of model decisions;
G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media	2023, IEEE	Developed the G-BERT model, combining BERT and GRU for hate speech detection, achieving 95.56% accuracy.	i) Limited to specific dataset characteristics; ii) Not tested on Bengali informal writings.
Detection of Hate Speech Text in Hindi-English Code-Mixed Data	2020, Elsevier	Proposes an SVM-based model using fastText embeddings for Hindi-English code-mixed hate speech.	i) Focus on binary classification only; ii) Lacks real-world deployment results.
Hate Speech and Offensive Language Detection in Bengali	2022, arXiv	Creates a 10K dataset and evaluates various transformer models.	i) Limited to Twitter data; Focuses primarily on transformer-based models; ii) Limited exploration of baseline techniques;

Paper Name	Year and Publisher	Contribution	Shortcomings
Multimodal Hate Speech Detection from Bengali Memes and Texts	2022, RWTH Aachen University	Built a multimodal hate speech detection system using text and images, achieving an F1 score of 0.83.	i) Complexity of multimodal fusion; ii) Limited generalizability to text-only cases;
Hate Speech Detection in the Bengali Language: A Comprehensive Survey	2024, Journal of Big Data	Provided a survey of Bengali hate speech detection methods, datasets, and challenges.	i) Lacks implementation of novel techniques; ii) No original dataset contribution; iii) Survey mainly descriptive.
Deep Learning Based Cyberbullying Detection in Bangla Language	2024, Annals of Emerging Technologies in Computing (AETiC)	Advances cyberbullying detection in Bengali by employing a bidirectional LSTM model, focusing on performance evaluation and overfitting mitigation.	i) Overfitting remains a concern despite attempted mitigations; ii) Further investigation is necessary to ensure the model's applicability to diverse datasets;
Combining FastText and Glove Word Embedding For Offensive and Hate Speech Text Detection	2022, Elsevier	Significant contributions to NLP and content moderation by addressing the critical issue of hate speech detection.	i) Language dependency; ii) Dataset dependency and limited scope;
Bengali Hate Speech Detection with BERT and Deep Learning Models	2023, ResearchGate	Compiled a dataset of 8,600 user comments and categorized them into distinct topics; evaluated five different models, including LSTM, Bi-LSTM, GRU, CNN, BERT.	i) Potentially overlooks diverse hate speech instances; ii) Broad categorization may miss nuances;
A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing	2023, Published by Elsevier	Comprehensive review of hate speech detection techniques focusing on NLP and deep learning.	i) Limited coverage of low-resource languages; ii) Relies on existing reviews; iii) Future scope lacks implementation examples.

Paper Name	Year and Publisher	Contribution	Shortcomings
Hateful Speech Detection in Public Facebook Pages for the Bengali Language	2019, IEEE	Identified linguistic as well as some quantitative features for six classes in the social context of Bangladesh.	i) Small dataset; ii) Accuracy reported 70.10% for GRU and Random Forest Model is 52.40%;
Sentiment Analysis on Bangla Text using Extended Lexicon Dictionary and Deep Learning Algorithms	2022, Published by Elsevier	Proposed hybrid DL models (LSTM, Bi-LSTM, Capsule Networks) for Bangla sentiment analysis with rule-based methods.	i) Focused only on sentiment analysis; ii) Moderate accuracy (84%);
Identifying Hate Speech of Bangla Language Text using Natural Language Processing	2024, Brac University Thesis	Developed a hate speech detection system for Bangla using machine learning and deep learning models (mBERT, CNN-LSTM).	i) Limited to social media datasets; ii) Multiclass accuracy remains low (64%); iii) Scope for dataset expansion;
A Deep Learning Approach to Detect Abusive Bengali Text	2019, Published by IEEE	Introduced stemming rules for Bengali and evaluated multiple algorithms for detecting abusive Bengali texts.	i) Limited dataset; ii) No comparison of transformer-based models;
A Deep Learning Approach to Detect Abusive Bengali Text	2019, Published by IEEE	Introduced stemming rules for Bengali and evaluated multiple algorithms for detecting abusive Bengali texts.	i) Limited dataset; ii) No comparison of transformer-based models;

Table 2.1: Literature Review

Chapter 3

Methodology

The presented part of The methodology section provides the information about detecting the hate speech in Bengali comments collected during the research. They are an ideal pipeline since it selects the data, preprocesses it, trains the models and evaluates them. The following outlines the escalating procedure below:

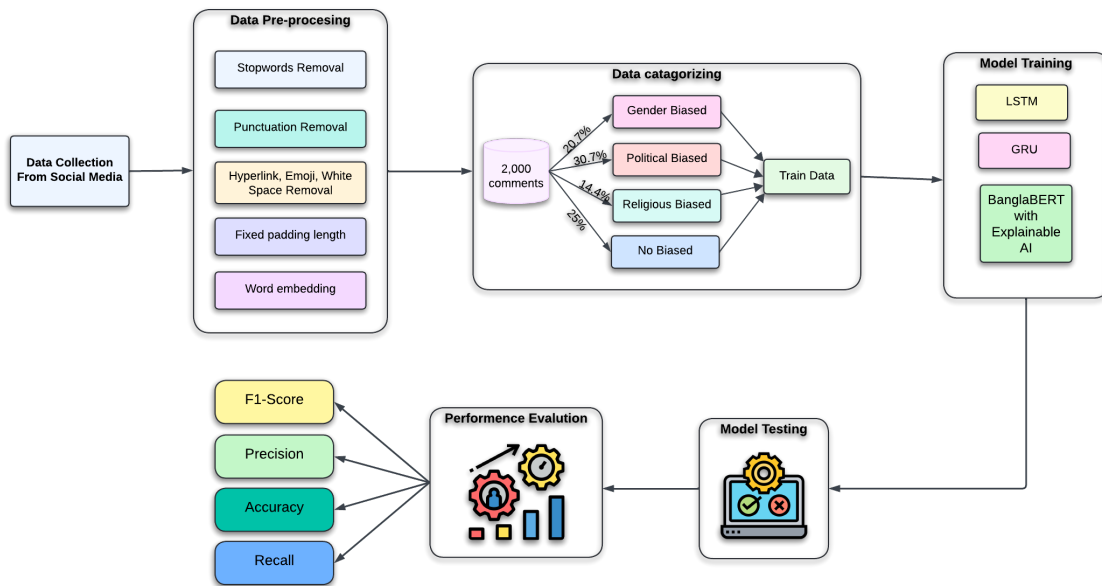


Figure 3.1: Methodology

3.0.1 Precision, Recall, and F1-Score

After training as well as making predictions using the above models, then the evaluation metrics will describe the performance of the models. In classification

problems, especially in hate speech detection, the most used measures are Precision, Recall, and F1-Score.

3.0.2 Precision

Its measure is the number of correctly identified positive cases divided by the total of all positive predictions within the test data set, that is, true positive over total positive. It undertakes the assessment of the positive predictions' reliability.

$$Precision = \frac{Truepositives}{Truepositives + Falsepositives} \quad (3.1)$$

- **High Precision** means that whenever it is labeled as hate speech then it will be right.
- **Low Recall** means the model can find hate speech where in fact there is no hate speech present.

3.0.3 Recall

Recall is defined as the total number of correctly predicted positive Observations (true positives) to the total actual positive Observations (true positives + false negatives). It shows how good the model is in detecting what is considered as real hate speech.

$$Precision = \frac{Truepositives}{Truepositives + Falsenegatives} \quad (3.2)$$

- **High Recall** means the model is able to distinguish the most of the hate speech cases in the set.
- **Low Recall** implies that the model is likely to miss out on a large quantity of true hate speech.

3.0.4 F1-Score

F1-Score is the average of Precision and Recall using the Formula $Precision * Recall / (Precision + Recall)$. It is a measure of intrinsic significance as it offers an average of Precision and Recall and has an added advantage when the classes are skewed in negative samples such as hate speech.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.3)$$

- **High Recall** means the model is able to distinguish the most of the hate speech cases in the set.
- **Low Recall** implies that the model is likely to miss out on a large quantity of true hate speech.

Hate speech of low F1-Score depicts a correct evaluation of Precision and Recall that means the model works effectively in classification of hate speech without going wrong too many times.

3.1 Dataset description

Our dataset focuses on sentiment analysis and biased comments in Bangla internet discourse identifying the frequency of biased language on social media with the goal of classifying the bias that are presented there. We gathered more than 2,000 comments from YouTube, Facebook, Instagram, Twitter, and TikTok, among other platforms. All the comments were written fully in Bangla. But some comments may have emojis or special symbols that match the actual internet communication style. A wider representation of online communication styles and potential biases is ensured by the variety of sources of online communication.

3.1.1 Characteristics of the dataset

- **Language:** The dataset presents all possible ways of language usage, such as the substantially formal and informal language and also emojis, which is more efficient for further analysis using high-level NLP approaches.
- **Balance:** To avoid the model bias towards the specific categories and sentiment polarities, equal numbers of samples from each of the bias categories and both polarity sentiments have been included in the dataset.

3.1.2 Dataset Fields

- **Text:** Only comments written fully in Bengali.
- **Bias Category:** Four types of bias that were defined.

3.1.3 Selection Criteria for Data

- **Language:** The tweets selected were only the ones that were completely in Bangla script, and thus the comments section. Comments in mixed languages were not considered. Owing standards to make sure the dataset was relevant and of high quality:
- **Bias Visibility:** We limited ourselves to cases where bias is quite overt, and from which it becomes easy to directly deduce a path through which the model and others can learn and act accordingly.
- **Emojis and Special Characters:** Since emojis and special characters frequently use for emotional or contextual meaning for that they were added in comments.
- **Sentiment and Bias Categorization:** To guarantee correct processing every comment was reviewed and annotated to ensure accuracy.

3.1.4 Bias Categorization

We identified four different categories of bias:

- **Political Bias:** This bias category includes all the comments containing people's prejudice or animosity towards some political Parties, ideology, politicians or certain political events. Examples include comments that support a given political party condemning the other or those which follow the trend of tweeting unconfirmed facts concerning political issues.
- **Religious Bias:** This category contains such comments as thinking something is unfair, disrespect of a religion an individual holds. It can be Offensive about certain religions disseminating stereotypes or even inciting people to hate certain religious groups.
- **Gender Bias:** This category refers to commentaries portraying gender stereotypes, prejudice or discrimination. This can include making offensive comments about women, using women's bodies as a tool or promoting the unhealthy roles of male and female.

- **No Bias:** In this bias type all comments which don't fit in the other type of bias are placed in this category. Importantly, this transmits hate speech which although negative does not relate fully to gender, religious or political prejudice. This category also includes comments that are neutral and show no bias.

A more complete understanding of online discourse is possible when these two classification techniques are combined. For example a comment could be both Religious Bias and Polarity 2 (Negative) indicating a negative comment expressing religious biases.

3.1.5 Applicability

- Improved generalization for hate speech detection tasks in Bengali using the proposed machine learning models.
- Survey on Paper Research on bias detection and its relation with sentiments.
- Engineering of transparency of fully explainable AI systems to perform ethical content moderation. training of machine learning models for hate speech detection in Bengali.
- Research on bias detection with sentiment analysis

3.2 Data collection Procedure

Data collection was conducted following systematic and ethical procedure for ensuring reliability and relevance of obtained dataset. The procedure is described in detail below:

3.2.1 Identification of the Source

Widely used social medias and video streaming site for their diverse user on the Bengali-speaking community in order to collect comments

- **Facebook:** Primarily in open profile make comments on political religion and social challenges.

- **Instagram:** For such data might be available in the comments section.
- **TikTok:** The users' responses to other posted video content.
- **YouTube:** Extracted from videos concerning politics, religion and gender showing hate to others.

3.2.2 Annotation Process

Bias Categorization:

- Annotators were briefed on how to recognize such comments, and then sort them based on one of the four biases.
- Inter-annotator validation was done to maintain compliance in the number of annotations.

Sentiment Polarity Labeling: It based the sentiment from user purpose from the text. Cases of such inconsistencies were shown that meant that the disagreements were resolved by coming to a point with other reviewers.

3.2.3 Preprocessing

- **Text Clearing:** Removal of non-Bengali characters, irrelevant symbols or special characters and extra spaces.
- **Normalization:** Standardization of Bengali script to handle the variations on collected data.
- **Tokenization:** Divide text into tokens and reduce words to their base forms, addressing the nature of Bengali language.

3.2.4 Ethical Considerations

1. Only publicly available comments were collected to respect user privacy and policy.
2. No personal information was included.

3. The dataset complies with ethical standards for hate speech and sentiment analysis research.
4. Validation and Integration
5. Comments were cross-validated by multiple annotators to ensure accuracy and for the best uses of the collected data.

3.2.5 Challenges

- **Linguistic Complexity:** The dataset captures variation such as sarcasm, different kinds of bias, and expressions.
- **Contextual Diversity:** Comments have a variety of topics ensuring relevance across the dataset.
- **Data Imbalance:** Balanced representation of four different categories to prevent bias in different machine learning models.

3.3 Data quality

Availability of good quality datasets is a big determinant of the success of any machine learning project particularly for tasks such as hate speech detection. Comment data collected from Facebook, TikTok, and some YouTube channels in the Bengali language pose some limitations and advantages within this research. QA suggested below is crucial in order to achieve a high quality dataset which can be used to support modeling processes: Below are the factors influencing the quality of the Bengali-language dataset:

3.3.1 Language-Specific Challenges

The basic language of the dataset is Bengali, which has its special issues that need to be discussed in the process of data quality assessment. These challenges include:

- **Morphological Complexity:** This is true because Bengali is a morphological language with elaborate forms of words which arise because of an affix and conjugations. of the verb may vary for example The verb for eating

“খাওয়া”

has the form like

“খেয়েছি”

meaning I ate or

“খাবে”

meaning he/she will eat. It is crucial to guarantee that these various forms are processed adequately during preprocessing, in order to have high levels of data quality.

- **Code-Switching:** One such feature closely observed in the comments section of Bengali sites and pages is code switching, where users partly as well as sometimes fully switch between Bengali and English. This mixing can make the comments sometimes conflicting, which often makes it hard for the models to determine the sentiment behind the text or the category of the text. That is why, much attention was paid to the receipt and further work with these cases during preprocessing.
- **Slang and Informal Language:** Social media comments are informal, contain language which is not typical to other types of text containing slang, colloquialisms, and abbreviations. For instance ‘GOVT’ instead of ‘government’ or ‘BD’ instead of ‘Bangladesh’. Consequently, these abbreviations and slang are normalized to guarantee that the data is comprehensible to modeling.

3.3.2 Labeling and Categorization

Hate speech labeling aids in accomplishing the feature extraction task of the supervision assessment in supervised learning models. In this dataset, comments are labeled according to four categories of hate speech:

- Political Bias
- Religious Bias

- Gender Bias
- No Bias

The terms used to label these sets affect the performance of the models in direct proportion. Care was taken to ensure:

- **Consistency in Labeling:** The labeling was done using standard definition for each type of bias with a view of making the comments uniformly categorized. This task was also accomplished with the help of a team of annotators from Bangladesh, who know Bengali and the context of hate speech.
- **Inter-Annotator Agreement:** In order to increase the quality of the labels, all the annotations were done by multiple annotators and all the differences between them were discussed to ensure high inter-annotator agreement. It means that the identities of the leptons that are used to train the models of FG and BG are more reliable.

3.3.3 Data Imbalance

One key problem associated with datasets used in the hate speech classification problem is the fact that there are many more non-emotional comments than actual hate comments. Since not all comments on social media are hate speech, there is a need to balance the dataset to prevent model bias on non hate speech class. The following steps were implemented to maintain data balance:

- **Oversampling of Underrepresented Categories:** The dataset is oversampled to get enough data points for the minority classes (similar to gender bias) to train our model with.
- **Undersampling of Dominant Categories:** In cases where some of the more general categories had an infestation of examples, techniques of undersampling were used in order to avoid the skewing of the model towards the more common categories.

3.3.4 Data Preprocessing Quality

Given the Bengali text, the following preprocessing steps were crucial in ensuring data quality:

- **Tokenization:** Bengali text was mostly segmented into words or subwords by using share-csj as well as word segmentation models. Particular attention was paid to Bengali characters, punctuations and space to keep the consistency.
- **Stop Word Removal:** Technical and Bengali specific stop words having less feature discriminating capability including

“এবং”, “অথবা”, “করা”

were complete removed out from the text. All these words are rather used often and are not directly associated with the changes in meaning as far as hate speech identification is concerned.

- **Text Normalization:** As the comments were unstructured, informal language with the use of emoticons, and the comments are in both Bengali and English, it was crucial to standardize the data. Terms such as ‘govt’ short for ‘government’ were converted to full form in order to eliminate confusion.
- **Handling of Misspellings and Grammar Variations:** The Bengali language in particular and more especially in social media is full of a lot of spelling errors and grammatical irregularities. Spell correcting tools were employed in an effort to preprocess the text, to remove any spelling distortions while retaining the semantic content of the messages.

3.3.5 Data Quality Assurance

To ensure the integrity and accuracy of the dataset, the following quality assurance steps were employed:

- **Manual Inspection:** To verify the accuracy of the labels, tokenization and other preprocessing steps a subset of the comments was visually examined. Any differences detected were spotted and corrected by the annotators.

- **Automatic Data Quality Checks:** Maintenance checks included the necessity of having none of the data entries to be repeated or else having some of the characters like the HTML tags or even corrupted texts. For the model training, only clean, well processed data was used.

3.4 Final data outlook

The last data situation summarizes the general spatial arrangement and organization of data, as well as its overall features, after data cleansing and quality control measures. It will also make the data ready for model training and evaluation to be conducted on it.

3.4.1 Data Structure and Columns

The dataset will consist of the following key columns:

1. **Comment:** The raw text of the comment, cleaned and preprocessed.
2. **Category:** The target label representing the type of hate speech or bias. Categories include:
 - Political Bias
 - Religious Bias
 - Gender Bias
 - No Bias
3. **Polarity:** The target label representing the type of hate speech or bias. Categories include:
 - 0 = Neutral
 - 1 = Positive
 - 2 = Negative
4. **Polarity1:** The target label representing the type of hate speech or bias. Categories include:
 - 0 = Neutral

- 1 = Positive
- 2 = Negative

Comment	Category	Polarity	Polarity1
কোন লাভ নেই, আমাদের দেশে সব জায়গায় রাজনীতি।	Political Bias	1	1
বিমান স্টাফদের ব্যবহার ভালো না	No Bias	2	2
যারা নারীর স্বাধীনতা খুঁজে, তাদের প্রতিটি পরিবারে এমন ঘটনা হওয়া দরকার।	Gender Bias	2	2
দনীতি বাজ সরকার দনীতিবাজ কর্মকর্তা কর্মচারীদের গাফিলতির কারণে বাংলাদেশ বহির্বিশ্বে সম্মান নষ্ট হচ্ছে	Political Bias	2	2
মাথায় সুনতি লেবাস, কিন্তু কথা চাড়ালের মত	Religious Bias	2	2

Figure 3.2: Dataset

3.4.2 Dataset Size

The final dataset should ideally have Five thousands of comments to ensure that the model is trained on a diverse and large enough set of data to generalize well. A balanced distribution across categories is essential for training an unbiased model. For example, the dataset should have around 2,000 comments per category (Political, Religious, Gender, and No Bias) to avoid imbalance.

Dataset size Table:

Category	Count
Political Bias	607
Religious Bias	288
Gender Bias	415
No Bias	520

Table 3.1: Dataset Size

This distribution is balanced so that the model chooses to learn to check each type

of hate speech and non-bias content without having to overfit with any particular class.

3.4.3 Preprocessing Steps Applied

Given that the dataset is in Bengali, the following preprocessing steps will be crucial:

- **Tokenization:** Converting the text of Bengali either into words or tokens. Extra precaution will be taken about handling words and characters which are particular to Bengali.
- **Lowercasing:** Again, to minimize variability in the data, all text is preprocessed to be in lowercase to remove any variations in consideration of case.
- **Stop-Word Removal:** Filtering out normal daily Bengali stop words which does not have any relation with sending hate speech.
- **Stemming and Lemmatization:** Converting all Bengali words into base form in order to come to the standardized form of any word.
- **Handling Special Characters and Emojis:** As social media comments contain emojis, hashtags, and URL, then these will be either stripped off or replaced by the necessary text form.
- **Language Segmentation for Code-Switching:** There were uses of comments in Bengali which involve interlanguage switching, the use of Bengali and English in a comment. The preprocessing step will then detect such cases and deal with them in the right manner.
- **Noise Filtering:** Trimming pre and post processing for noisy text by scraping off added punctuations, strange symbols or characters repeatedly used on social media.

3.4.4 Data After Preprocessing

Finally after all these preprocessing steps the dataset will be neat and tidy and is well suited for training many models. The text will be preprocessed in Bengali language; it will be tokenized, normalised and deletion of unwanted noises which

will be helpful in feeding to the Machine learning & Deep learning models. The last formatted dataset for analysis will be 3 with target labels being categories of hate speech and all the comments ready for classification.

Comment	Category	Polarity	Polarity1
সরকারের কোনো পরিকল্পনা নেই।	Political Bias	2	2
ধর্মের প্রতি শ্রদ্ধা থাকা উচিত।	Religious Bias	0	0
নারীরা সবসময় আবেগপ্রবণ।	Gender Bias	2	2
এই উৎসবটি অনেক আনন্দদায়ক।	No Bias	1	1

Figure 3.3: Post-processed data example

3.4.5 Handling Challenges Specific to Bengali Language

There are several specific challenges related to Bengali hate speech detection, such as:

- **Word Variations and Morphology:** Bengali has a kind of morphological structure such that the same word has several variants that differ in their forms, conjugations or declensions. Stemming or lemmatizing allows the model to identify all the forms of the word, when done in this manner.
- **Bengali-English Code-Switching:** As has been explained before, code-switching is typical for Bengali comments that are posted to social media platforms. The preprocessing pipeline will tackle this by appropriately segmenting and normalizing (particularly English and Bengali) text to feed the model comments it can interpret.
- **Shortened Words and Slang:** People in social media have a tendency to employ symbols, abbreviations, kr intertwining letters (for example ‘govt’ instead of ‘government’, ‘bd’ instead of Bangladesh). Enhanced attention will be paid to formalize these forms.

3.5 Utilized model

Detection of hate speech especially in a language such as Bengali calls for the use of various models of machine learning capable of handling textual data. Preemptive and Classical models alongside Deep learning models are used for this research. Below, we provide a detailed explanation of each model and how they can be utilized to classify hate speech into four distinct categories: Political bias, religious bias, gender bias and no bias were the other options.

3.5.1 Random Forest Classifier (RF)

Random Forest is a technique, which combines many decision trees during the training phase and makes a decision with the help of the majority of these trees. It operates in two main phases:

- **Training Phase:**

- Decision trees that are numerous, are trained using subsets of the data and of features at hand (bagging).
- All trees make predictions independently, and then the one with the highest frequency of class for classification problems is chosen.
- It is quite easy to model: decision trees are graphical structures in which each vertex looks like a decision “Is the word ‘terrorist’ located in the comment?” They are constructed by recursively splitting the data on feature space that has less error in classification.

- **Prediction Phase:**

- Each tree in the forest receives a new input (a comment and makes a prediction based on that input.
- The last result of the process is the class which has the most votes in all trees in the decision tree.

- **Key Characteristics:**

- Ensemble Learning: Random Forest decision tree consists of multiple weak classifiers whereby a combination of these makes a strong classifier.

- Handling Overfitting: Where other decision trees can be overfitting and can give inaccurate results, Random Forest uses an average of multiple trees to overcome this challenge and give the best results.

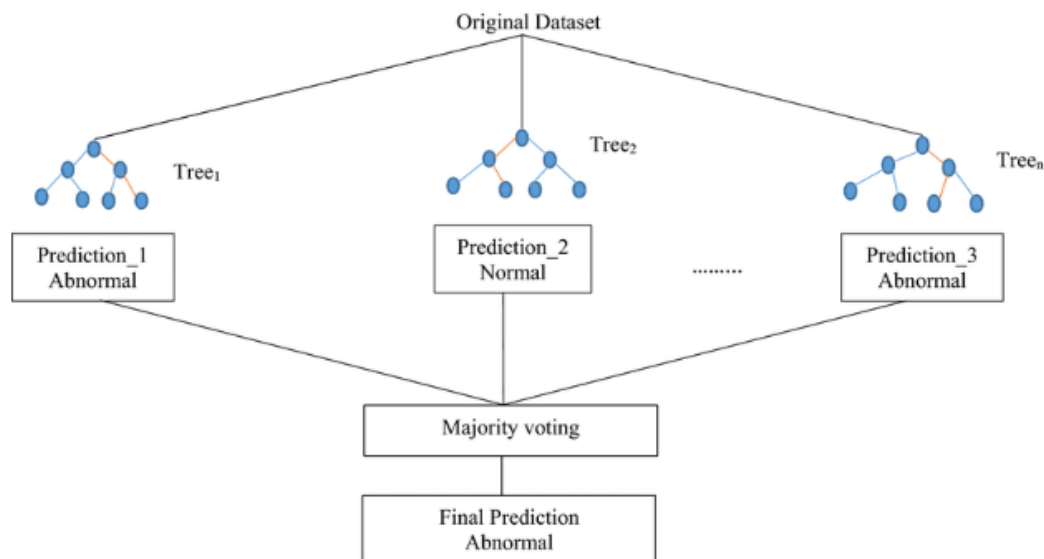


Figure 3.4: Random Forest Architecture (Source: <https://tinyurl.com/mr4baddu>)

3.5.2 Logistic Regression (LR)

Logistic Regression is a type of a linear model used when the target variable is categorical and binary or multiclass. It operates on the premise of estimating the likelihood that an input will belong to a particular category into which new text may be classified. The model is grounded at the logistic function (sigmoid), which makes a real number and turns it into a probability of a class label in the 0,1 range [Biswas & Rabbi \(2024\)](#).

- **Training Phase:**

- This is done by searching the best weights of the provided input features so that the weighted sum of the features can be mapped to a probability value between 0 and 1 on which the model is trained. This is done through passing the sum of weight multiplied by features through a sigmoid function.

- During the training phase the parameters (weights) are tuned using optimization algorithms such as: Gradient Descent or to minimize the logistic loss (cross entropy) to get the most probable given data.

- **Prediction Phase:**

- As for comment-ja new input, and features are mapped, for instance, with the help of TF-IDF or Word2Vec; then the weighted sum of features is again passed through the sigmoid function and gives the probability.
- They forgot that if the probability obtained is greater than 0.5 (for binary classification), the model assigns the input to a specific class (political bias, religious bias and so on).

Key Characteristics:

- **Probabilistic Output:** Logistic Regression predicts the class by outputting the probability of every class and the model itself can estimate how confident it is about the prediction.
- **Linear Decision Boundary:** The best use of logistic regression is with datasets in which the separation between classes is reasonably linear.

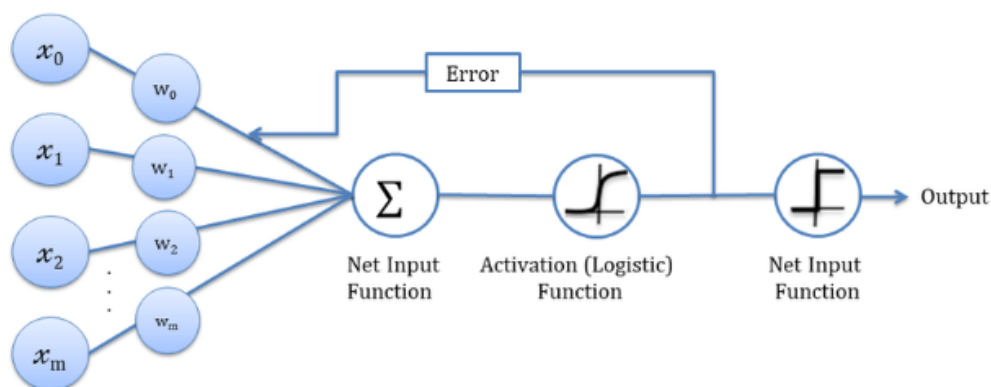


Figure 3.5: Logistic Regression Model (Source: <https://tinyurl.com/bde9vrpa>)

3.5.3 Naive Bayes (NB)

In the current case, Naive Bayes is a classification technique because it computes the class of input features based on its likelihood. The “naive” part comes with a basic assumption that the features, (words in a comment), are independent of each other given the class.

- **Training Phase:**

- For each class (e.g., Political Bias, Religious Bias), the model calculates the prior probability $P(\text{Class})$, and the likelihood $P(\text{Feature}; \text{Class})$ for each feature (word)
- Bayes’ Theorem is used to compute the posterior probability $P(\text{Class}; \text{Feature}_1, \text{Feature}_2, \dots)$ for each.
- The intent is to compute this posterior probability to its maximum for each class.

- **Prediction Phase:**

- As for a new input (comment), the model calculates the probabilities for each class with reference to the frequencies of the used words then assigns the comment to the class with the probability highest to it.

Key Characteristics:

- **Independence Assumption:** Uses the naive assumption that features (words) are independent, which uniformly is not the case, but for practical purposes is fine.
- **Fast and Efficient:** Naive Bayes algorithm is especially fast and is suitable for use with big data, especially in text classification.

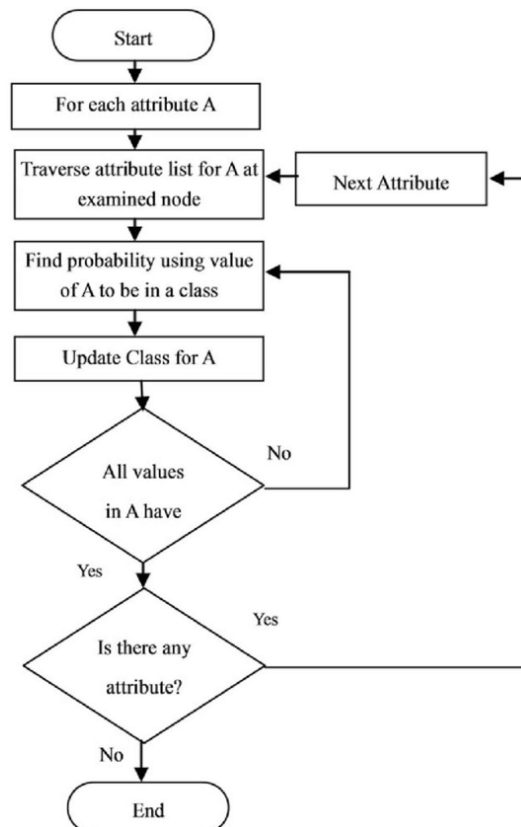


Figure 3.6: Naive Bayes Model (Source: <https://tinyurl.com/p6y6auyv>)

3.5.4 Long Short-Term Memory (LSTM)

LSTM is a kind of Recurrent Neural Network (RNN) that is tailored to cope with the problems, long-term dependencies which normal RNN's have. It incorporates specific gating which allows it to have a memory of information as it moves across sequence.

- **Training Phase:**

- LSTM networks take input in form of text data in a sequence of words, and it processes data word by word and the current word, which is the input at a single instance; the subsequent words are stored as cell states. An important part is the gates controlling how much is passed from the previous time step, represented by the input, forget, and output gates.
- In training LSTM is trained to be able to modulate the weights of these

gates in order to generate a best guess of the class label (bias category) for the provided input.

- **Prediction Phase:**

- For a new input which is a comment, LSTM breaks down the words and operations on the sequence of words and revises memory every operation to capture context.
- The final output looks at which class the model learned sequentially with the highest probability at that order.

Key Characteristics:

- **Capturing Long-Term Dependencies:** LSTMs are tweaked dependently to capture long range practices in text which make them very vital in analyzing context hate speech.
- **Memory Cells:** They allow its component, the memory cell, to retain essential information through sequences of long texts.

3.5.5 Gated Recurrent Unit (GRU)

It is an application of LSTM but is less complex than LSTM, taking shorter time to make its computations. Like LSTM it uses the gates to manage the information, but the structure is less complex than LSTM [Trang \(2023\)](#) [Zulqarnain et al. \(2024\)](#).

- **Training Phase:**

- Like LSTM, GRU processes text sequentially but uses two gates: the update gate and the reset gate. When one or the other of these gates is open it can allow leakage currents in the circuit. It is composed of the update gate which allows a decision on how much of the past information will be transmitted to the future and the reset gate which also decides on how much of the past memory should not be transferred to the future.
- GRU makes the selection of which part of the input sequence to pay attention to and what part is irrelevant [Trang \(2023\)](#).

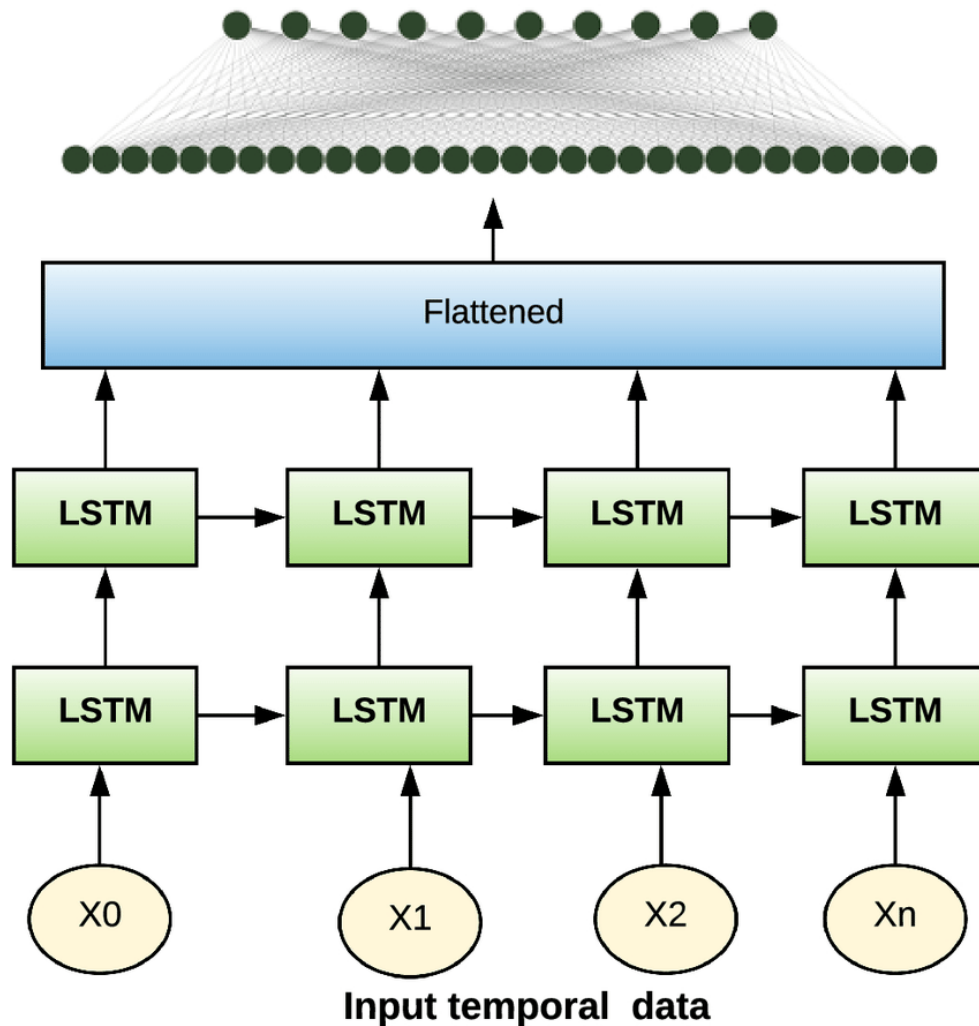


Figure 3.7: LSTM Architecture (Source: <https://tinyurl.com/y7ndjt5u>)

- **Prediction Phase:**

- About a new input (comment), GRU feeds the words one by one, using the gates of the current step to update memory and the final class is produced after the end of the sequence [Trang \(2023\)](#).

Key Characteristics:

- **Computational Efficiency:** LSTMs take longer time to train than GRUs hence GRUs are best suited for real time applications [Trang \(2023\)](#).
- **Simpler Structure:** They also know that GRU requires fewer gates and param-

eters than LSTM, yet are still quite effective for modeling sequential data Xu et al. (2024).

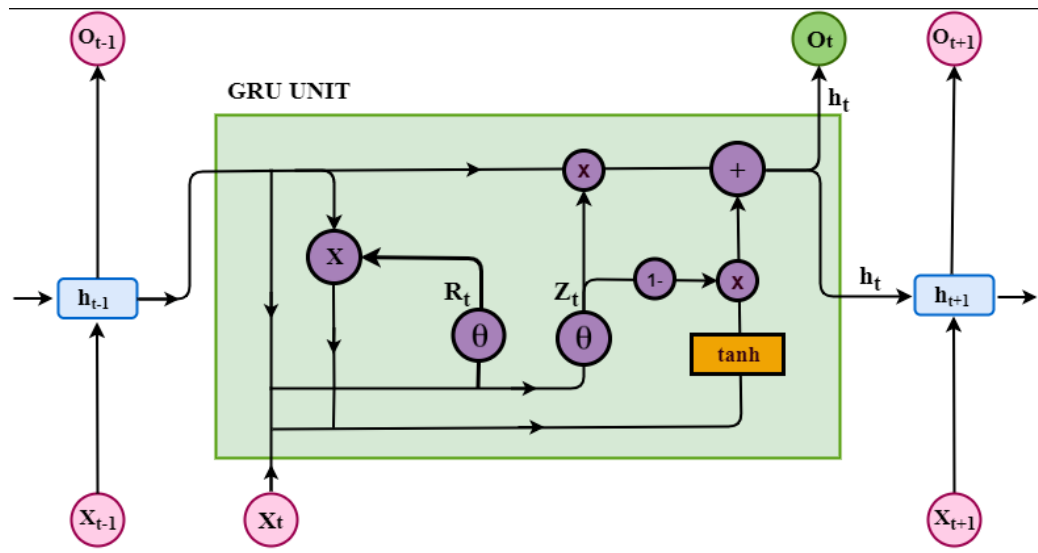


Figure 3.8: GRU Architecture (Source: <https://tinyurl.com/mxrvbj6a>)

3.5.6 BERT (Bidirectional Transformers for mask language models)

BERT is based on the transformer model where rather than going sequentially from left to right as in 'vanilla' seq2seq models, or in left to right as in RNNs, all words in a sentence can be processed at the same time through self-attention mechanisms. The most significant feature distinguishing BERT from previous models is that it retains background information about text as it processes it both from left to right (and vice versa).

- **Pretraining Phase:**

- is pre-trained on large corpora in an unsupervised manner using two tasks: two pre-training tasks namely: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, something is masked, and the model learns to predict that something." In NSP, the model thus tends to learn some correlation between the two subsequent sentences.

- **Fine-Tuning Phase:**

- Therefore, while pretraining BERTs can be fine tuned on any particular task such as hate speech. For this, BERT has learned the correlation between the contextual embeddings of the words and the class labels (Political Bias, Religious Bias, Gender Bias, No Bias).

Key Characteristics:

- **Bidirectional Context:** BERT uses the information from the left neighboring word and also from the right neighboring word to give a meaning to context.
- **Powerful Transfer Learning:** BERT can be trained on specific tasks thus making it very versatile in many NLP tasks.

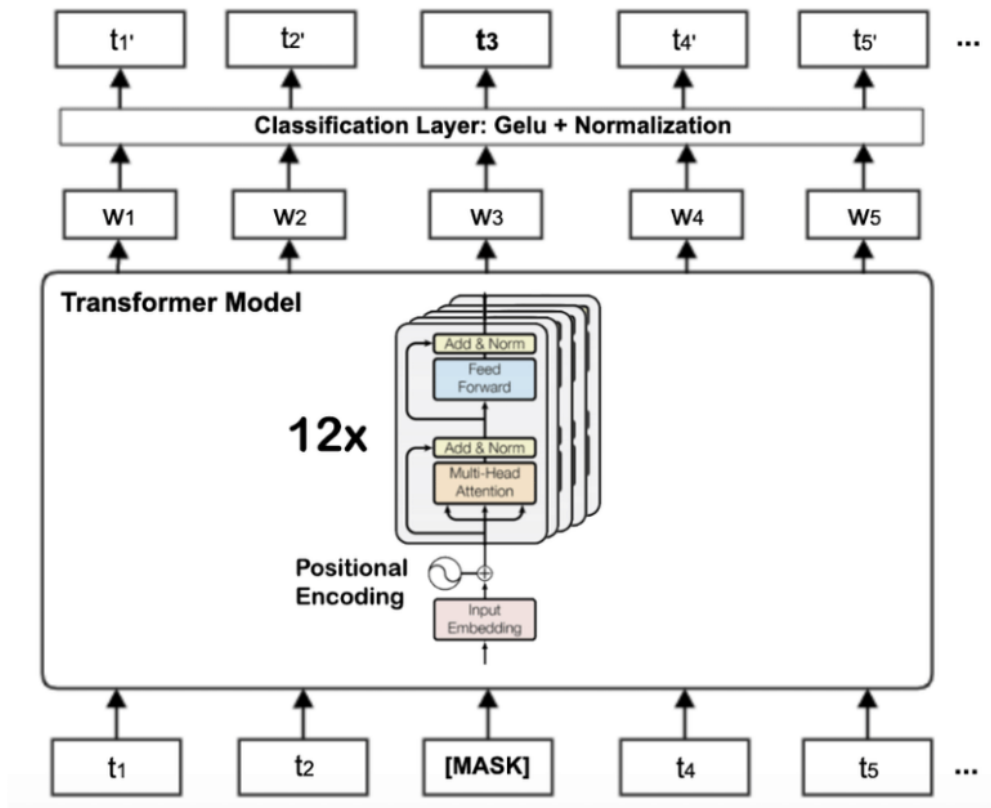


Figure 3.9: BERT Architecture (Source: <https://tinyurl.com/2sc6pemw>)

3.5.7 DistilBERT

DistilBERT is similar to BERT in that it is a compressed version of the model through knowledge distillation keeping much of the functionality of BERT [Raihan et al. \(2023\)](#).

- **Training Phase:**

- DistilBERT is trained with the help of BERT as the teacher model. Furthermore, it reproduces the behavior of BERT while sharing the knowledge from the large model to the small model.

- **Fine-Tuning Phase:**

- Like other variations of BERT, DistilBERT is fine-tuned into specific tasks but uses lesser computational resources and can be deployed more effectively.

Key Characteristics:

- **Smaller and Faster:** DistilBERT is faster but accurate as the regular BERT with a few trade-offs to achieve the gain in speed.
- **Efficiency for Real-time Applications:** The advantage of DistilBERT particularly is its size and speed it makes predictions much faster making it ideal for use where there needs to be almost immediate determinations such as hate speech in the comments section of social media posts.

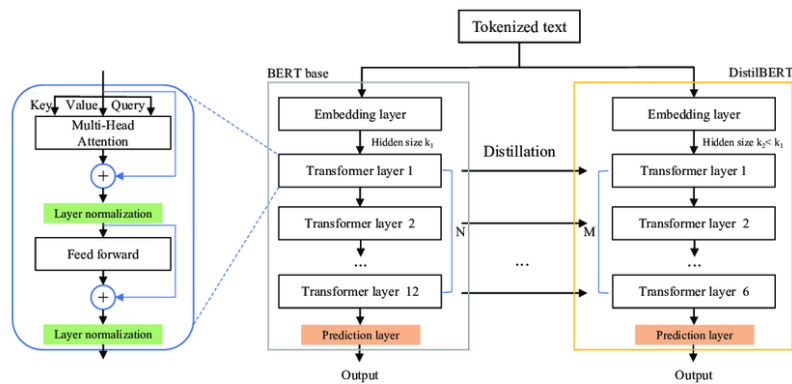


Figure 3.10: DistilBERT Architecture (Source: <https://tinyurl.com/3ut3vy47>)

3.5.8 The Explainable AI-enabled BanglaBERT

BanglaBERT is developed from the BERT model using a corpus called Bangla Folks. This corpus includes a vast number of Bengali documents and helps BanglaBERT to grasp the language peculiarities and the context in which it is used [Anik et al. \(2023\)](#).

- **Pretraining Phase:**

- is BanglaBERT is the BERT model fine-tuned on a large amount of Bengali text corpus using two types of objective, MLM and NSP. These tasks help the model to capture bidirectional contextual information encodings of the Bengali language, which is a low resource language [Imbwaga et al. \(2024\)](#); ?.

- **Fine-Tuning Phase:**

- is In use cases such as hate speech detection, we fine-tune BanglaBERT using labeled Bengali text datasets. It is capable of detecting political, religious, and gender biasing by making use of semantically analyzing the word and phrases in Bengali.

- **Explainability:**

- In order to provide interpretability, the LIME (Local Interpretable Model-agnostic Explanations) was incorporated [Banshal et al. \(2023\)](#).
- LIME visualizations make known the importance of the particular words or phrases inputted for the classification.

Key Characteristics:

- **Bengali Language-Specific:** Compared to BERT, BanglaBERT is especially tailored to enhance the understanding of Bengali text, to deal with specific issues.
- **Bidirectional Context:** Similar to other transformers models like BERT , BanglaBERT takes into consideration left and right contextual data thus it is best suited for understanding language constructs such as sarcasm and bias common in hate speech datasets.

Comparison of Models

Comparison of Models:

Model	Strengths	Weaknesses	Best Use Case
Random Forest	Good results when handling high-dimensional data and does not overfit with ensemble learning.	Less interpretable, slow with very large datasets.	General purpose classification, balanced data set.
Logistic Regression	A Simple, fast, and interpretable.	Less susceptible to pronounced nonlinearity, and may fail to capture extremely complicated patterns in decision boundaries.	Quick baseline model, binary or multi-class classification.
Naive Bayes	Easy to implement, time efficient, compatible with textual data, suitable for datasets of large size.	Fail to account for associations between the features, incorrect assumption.	Quick and accurate categorization for textual issues.
LSTM	A Very useful when working with sequences in a long-term dependency setting.	Needs more training data, training time is also long.	A Language model, Sequential texts, Contextual dependent problems. Talaat (2023)
GRU	It is also faster and far more efficient than LSTM.	LSTM can be less effective than this one in case of highly complex.	Real-time applications requiring sequential context.

Model	Strengths	Weaknesses	Best Use Case
BERT	High performance, both left and right context, trained on many big texts.	Highly computational and demand a lot of resources.	High-end NLP activities, problem scenarios involving context.
DistilBERT	Faster training version of BERT, retains most of the BERT strengths.	Slightly reduced performance compared to BERT.	Fast inference in resource constrained environmentsNg et al. (2023).
BanglaBERT	Picked specifically for Bengali, does a better job understanding linguistic nuances.	It is still very dependent on large Bengali corpora for training.	Bengali language NLP tasks, hate speech detection.

Table 3.2: Comparison of Models

Chapter 4

Design and Development

This system incorporates an Improved BanglaBERT, a transformer-based model used for supervised learning of Bangla text with Explainable AI features in its design. Below are the key stages:

4.1 Data Collection and Preprocessing

- **Data Sources:** The required legal comments and posts were extracted from the widely used social media sites like Facebook, YouTube and Twitter from Bengali content only.
- **Annotation:** Native Bengali speakers are used to annotate the dataset into the four categories as Political Bias, Religious Bias, Gender Bias, and No Bias. The entries were also given a polarity as general (positive, neutral or negative).
- **Preprocessing:**
 - Banned words as well as symbols like #, @, URLs and any text which is not in Bengali.
 - Tokenization and stemming by utilising the Bengali NLP libraries.
 - Code switching one of them and or normalization of slangs the other one.

4.2 BanglaBERT Model

4.2.1 Pretraining and Fine-Tuning

- BanglaBERT, developed from the BERT model architecture, was further pre-trained on enormous volumes of extensive Bengali corpus to understand the linguistic and contextual semantics [Chowdhury et al. \(2024\)](#).
- Specifically, fine-tuning done on the annotated set helped in enhancing the model for hate speech detection and bias categorisation.

4.2.2 Advantages of BanglaBERT

- Focuses on using bidirectional context to recognize and identify other forms of hatred not necessarily in clear speech [Lasri et al. \(2023\)](#).
- Since the model is designed for Bengali only it's more effective for this task rather than using multilingual models for this job.

4.3 Explainability with LIME

Integration of Explainable AI:

- Interpretability was addressed by adding LIME (Local Interpretable Model-agnostic Explanations) to the suggested system.
- To every prediction, LIME broke down visual explanations and provided the words or phrases that contributed most to the model's decision.

Benefits:

- Increased confidence and openness of the system.
- It lets stakeholders perceive and verify decisions made when building a model, especially regarding misclassification.

4.4 Evaluation and Metrics

Performance Metrics:

- The efficiency of the proposed model was measured using Precision, Recall, F1-Score, and Accuracy.
- Across all the assessments, BanglaBERT surpassed baseline models in terms of the performance measures.

Error Analysis:

- Misclassified instances were further explained using LIME in order to come up with corrective action.

4.5 Deployment

API Development: To connect the developed BanglaBERT-based system with applications and platforms, an API has been designed.

User Interface: The following goals were to be met when designing the web based dashboard:

- Offer an opened text to the users while allowing them to submit their text for analysis.
- Submit elaborate results that are accompanied by bias categorization and explainability reports.

This design makes the solution very accurate, understandable, and real-world deployable for detecting hate speech and biases in Bengali, pivotal for researchers, moderators, and citizens.

Chapter 5

Result analysis

5.1 BanglaBERT with Explainable AI

- Precision: 95
- Recall: 95
- F1-Score: 95
- Accuracy: 95

	precision	recall	f1-score	support
0	0.96	0.94	0.95	497
1	0.94	0.96	0.95	506
accuracy			0.95	1003
macro avg	0.95	0.95	0.95	1003
weighted avg	0.95	0.95	0.95	1003

Figure 5.1: BanglaBERT with Explainable AI classification report

Thus, BanglaBERT performed satisfactorily to distinguish hate speech, and with similar accuracy ratios in all categories. The use of LIME to ascertain an explainable

AI showed which words or phrases were most important in the classification process and thus gave the necessary transparency and reliability. You will discuss the results of your experiments/critically analyze the performance of your developed system. You are also compare the results/performance with those of the existings.

5.2 GRU Model

- Precision: 74.74
- Recall: 84.69
- F1-Score: 79.40

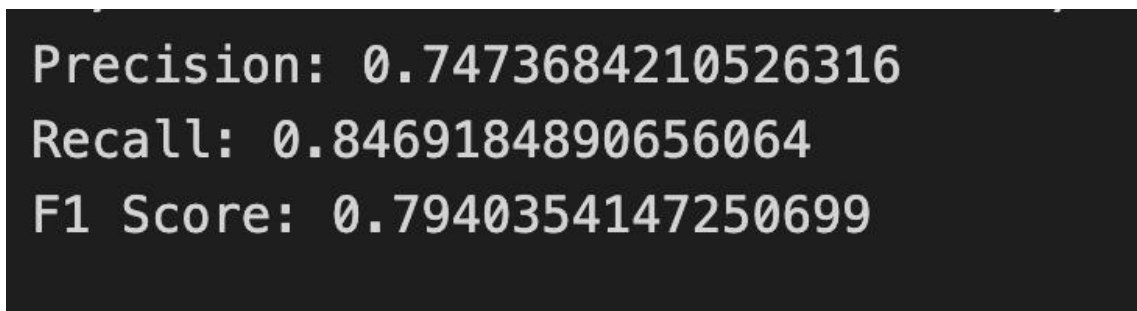


Figure 5.2: GRU Model Calculate precision, recall, and f1-score

The experiment demonstrates that GRU mostly reached high results, predetermining its effectiveness in the detection of evident hate speech. It again performed well regarding sequential patterns in less lengthy texts but was slightly off from the BanglaBERT in comprehending biases.

5.3 LSTM Model

- Accuracy: 84.69
- Loss: 79.40

LSTM found long-term dependency well, obtaining good accuracy and loss. It was slightly better than GRU in terms of identifying implicit prejudice, such as slang and veil-ed hatred [Haque et al. \(2023\)](#).

Test Accuracy: 0.7797619104385376, Test Loss: 0.4648233652114868

Figure 5.3: Evaluate the LSTM Model

5.4 Explainable AI in Action

Unlike the general approach exhibited by this kind of model, BanglaBERT paid particular attention to explainability. LIME was used to:

- Find out those words or the phrases that have the greatest impact on a classification.
- This is done effectively and each of the models should display heat maps showing the weight assignment on core token predictions.
- Augment the model's interpretability further so that the stakeholders have confidence in the mechanisms of the system.

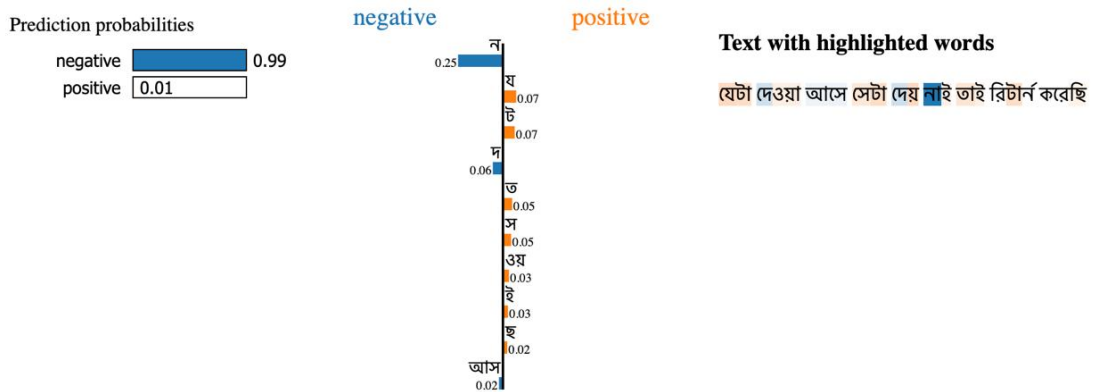


Figure 5.4: BanglaBERT with Explainable AI

For example, in a comment labeled with a bias called “Religious Bias,” LIME explained terms that are linked to religious extremism so that the moderators could control the problem adequately.

Chapter 6

Conclusions

This research was able to use three state-of-the-art machine learning algorithms, BanglaBERT, GRU, and LSTM, to detect hate speech in Bengali. Among the models, BanglaBERT had the highest accuracy, and precision and recall values, due to its contextual embedding ability and use of bidirectional language processing. Accountability tools such as LIME further improved BanglaBERT's decision-making necessity by providing a deep explanation of its prediction results for content moderation [Khanam et al. \(2024\)](#).

LSTM and GRU model with a slightly lower accuracy was also a reasonable solution for sequential text patterns as well as for detection of explicit hate speech. Due to its ability to manage implicit and subtle bias, LSTM is best used as an improvement strategy [M. S. Islam & Alam \(2023\)](#).

In this study, significant issues such as the dearth of datasets and language-dependent preprocessing are mitigated, advancing the formation of more secure online spaces for Bengalis.

6.1 Social Issues

- **Impact on Society:** The following research project is to help decrease the spread of hate speech on social media related platforms and make the digital environment a safer and friendlier place for the Bengali speaking people. Regarding uses of language that is toxic, when done, it enhances social inclusion hence enhancing community cohesion [Sultana et al. \(2023\)](#).

- **Cultural Sensitivity:** It is important to show respect for the culture and language of Bengali with its rich linguistic diversities. It is crucial to avoid misclassifications on the existence of any collections that would worsen the social divide in the society.
- **Accessibility:** All tools and interfaces designed within the course of this project should be easy to use and understandable for any user – technically proficient and laypeople.

6.2 Legal Issues

- **Data Privacy:** The project is designed so that all used data conforms with the privacy regulations, privacy of user data is preserved, and the principles of the General Data Protection Regulation (GDPR) are followed.
- **Copyright and Ownership:** External, as well as internal, datasets are used in accordance with the guidelines set by the university and do not violate any intellectual property rights.
- **Regulatory Compliance:** The project only allows posts that do not violate countries legal requirements particularly the ICT Act of Bangladesh and International laws against hate speech to eliminate possible legal confrontations.

6.3 Ethical Impact

- **Bias and Fairness:** a lot of effort is spent to avoid the reinforcement of bias into the data sets and the models. Inclusion of large and medium data sets as well as small data sets guarantees non-discrimination in the treatment.
- **Transparency:** The integration of Explainability results in using techniques such as the Local Interpretable Model-agnostic Explanations (LIME) can give an understanding of how the decision of the system is made.
- **Accountability:** There exist ways of handling wrong classifications within the system so that whoever submits content to the platform can dispute wrong flagging.

- **Freedom of Speech:** The rules are implemented in order to control dissemination of hate speech while preserving freedoms of each person and not infringing their rights.

6.4 Environmental Impact

- **Energy Consumption:** Training of large scale models such as BERT, and BanglaBERT, takes a lot of computational power and, in turn, a lot of energy and that energy results in carbon emission.
- **Mitigation:** Use of energy efficient hardware and fine tuning of the algorithms is possible to minimize the effects on the environment.
- **Alternatives:** There are lighter models like DistilBERT which are considered for use where possible so as to consume limited resources.
- **Sustainability:** To a certain extent, it is necessary to provide long-term management and scalability so that the advantages of the system will continue to be sustainable without a high amount of cost to the environment.

6.5 Brief Summary

This particular project is therefore about the detection of hate speech in Bengali text including considerations of social, legal and ethical issues as well as the impact on the environment. Digitally, socially, it helps reduce forms of generic divisional language and at the same time promotes reconciliations of plural cultural differences. From a legal perspective, it is legal in all aspects regarding data privacy as well as regulatory in terms of the ethical use of data. Professional issues encompass avoiding prejudice, reporting objectively and ensuring free speech. In this respect, the project bears relatively low carbon costs in the training of predictive models and, as a project, is environmentally sustainable. The recommendations serve as an emphasis on the applicability of interdisciplinary integration and energy saving measures, audience participation, and frequent updates of the system considering its responsible use in countering hate speech.

6.6 Future works

To enhance the scope and impact of this study, future work will focus on:

1. Expanding Dataset Scope:

- Extend the datasets to include more documents which contain multilingual and code-mixed text (Bengali and English).
- The bias categories should be broken down into fine categories like socio-economic bias or age bias.

2. Real-World Deployment:

- Create the APIs and plugins to disseminate the hate speech detection system into the most visited social site such as Facebook and Youtube.
- Develop a real time hate speech monitoring dashboard.

3. Improved Model Architectures:

- gestures = CNN-RNN hybrid structures for feature- rich text representation 4
- Try other transformer models such as XLM-R or the multilingual BERT model where there is a need for cross linguistic use.

4. Ethical and Bias Mitigation:

- Overcome discrimination by incorporating bias reduction measures so that the models perpetrate the negative stereotype.
- Work with other linguist and social scientific experts in order to enhance the cultural competency or language model used in the model development.

5. Explainability and Trust:

- Improve visual explanation by adding more tools to LIME to make them interactively understandable.
- For the different explainability methods, they can use SHAP (Shapley Additive Explanations).

References

- Al Maruf, A., Abidin, A. J., Haque, M. M., Jiyad, Z. M., Golder, A., Alubady, R., & Aung, Z. (2024). Hate speech detection in the bengali language: a comprehensive survey. *Journal of Big Data*, 11(1), 97.
- Anik, S. H., Zamal, J. I., & Shikder, M. W. (2023). *Sentiment analysis on bengali news comments using transformer based deep learning model* (Unpublished doctoral dissertation). Department of Computer Science and Engineering (CSE), Islamic University of
- Badri, N., Koubi, F., & Chaibi, A. H. (2022). Combining fasttext and glove word embedding for offensive and hate speech text detection. *Procedia Computer Science*, 207, 769–778.
- Banshal, S. K., Uddin, A., & Piryani, R. (2023). Analyzing the performance of bert for the sentiment classification task in bengali text. In *International conference on advanced network technologies and intelligent computing* (pp. 273–285).
- Biswas, M. R., & Rabbi, M. F. (2024). Annotated bangla natural language processing (bnlp) by using python or machine learning.
- Chowdhury, M. S., Sultan, T., Jahan, N., Ali, M. A., Hasan, K. T., & Ahmed, N. (2024). A multilingual intelligent system for objectionable text recognition utilizing an explainable ai-supported deep learning model on bengali, bengali transliteration, and english embedded text.
- Das, A. K., Al Asif, A., Paul, A., & Hossain, M. N. (2021). Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1), 578–591.
- Das, M., Banerjee, S., Saha, P., & Mukherjee, A. (2022). Hate speech and offensive language detection in bengali. *arXiv preprint arXiv:2210.03479*.
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international aaai conference on web and social media* (Vol. 11, pp. 512–515).

- Emon, E. A., Rahman, S., Banarjee, J., Das, A. K., & Mitra, T. (2019). A deep learning approach to detect abusive bengali text. In *2019 7th international conference on smart computing & communications (icscc)* (pp. 1–5).
- Haque, R., Islam, N., Tasneem, M., & Das, A. K. (2023). Multi-class sentiment classification on bengali social media comments using machine learning. *International journal of cognitive computing in engineering*, 4, 21–35.
- Imbwaga, J. L., Chittaragi, N. B., & Koolagudi, S. G. (2024). Explainable hate speech detection using lime. *International Journal of Speech Technology*, 27(3), 793–815.
- Ishmam, A. M., & Sharmin, S. (2019). Hateful speech detection in public facebook pages for the bengali language. In *2019 18th ieee international conference on machine learning and applications (icmla)* (pp. 555–560).
- Islam, M. H., Farzana, K., Khalil, I., Ara, S., Shazid, M. R. A., & Mehedi, M. H. K. (2024). Unmasking toxicity: A comprehensive analysis of hate speech detection in banglish. In *2024 6th international conference on electrical engineering and information & communication technology (iceeict)* (pp. 963–968).
- Islam, M. S., & Alam, K. M. (2023). Sentiment analysis on bangla food reviews using machine learning and explainable nlp. In *2023 26th international conference on computer and information technology (iccit)* (pp. 1–6).
- Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232.
- Jobair, M., Das, D., Islam, N. B., & Dhar, M. (2023). Bengali hate speech detection with bert and deep learning models. In *International conference on big data, iot and machine learning* (pp. 845–859).
- Keya, A. J., Kabir, M. M., Shammey, N. J., Mridha, M. F., Islam, M. R., & Watanobe, Y. (2023). G-bert: an efficient method for identifying hate speech in bengali texts on social media. *IEEE Access*.
- Khanam, F., Chakraborty, A., Habib, M. A., & Iqbal, M. S. (2024). Bangla sentiment analysis on highly imbalanced data using hybrid cnn-lstm & bangla bert. In *2024 3rd international conference on advancement in electrical and electronic engineering (icaeee)* (pp. 1–5).
- Lasri, I., Riadsolh, A., & Elbelkacemi, M. (2023). Self-attention-based bi-lstm model for sentiment analysis on tweets about distance learning in higher education. *Int. J. Emerg. Technol. Learn.*, 18(12), 119–141.
- Nath, S. S., Karim, R., & Miraz, M. H. (2024). Deep learning based cyberbullying detection in bangla language. *arXiv preprint arXiv:2401.06787*.

- Ng, S. Y., Lim, K. M., Lee, C. P., & Lim, J. Y. (2023). Sentiment analysis using distilbert. In *2023 ieee 11th conference on systems, process & control (icspc)* (pp. 84–89).
- Rahman, M., Jui, R. S., Sakib, C. N., Ridoy, F. A., & Ananya, T. T. (2024). *Identifying hate speech of bangla language text using natural language processing* (Unpublished doctoral dissertation). Brac University.
- Raihan, M. N., Goswami, D., & Mahmud, A. (2023). Mixed-distil-bert: Code-mixed language modeling for bangla, english, and hindi. *arXiv preprint arXiv:2309.10272*.
- Romim, N., Ahmed, M., Talukder, H., & Saiful Islam, M. (2021). Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of international joint conference on advances in computational intelligence: Ijccaci 2020* (pp. 457–468).
- Sultana, S., Redoy, M. O. F., Al Nahian, J., Masum, A. K. M., & Abujar, S. (2023). Detection of abusive bengali comments for mixed social media data using machine learning.
- Talaat, A. S. (2023). Sentiment analysis classification system using hybrid bert models. *Journal of Big Data*, 10(1), 110.
- Trang, N. T. M. (2023). A hybrid model of roberta and bidirectional gru for enhanced sentiment analysis. *International Journal of Open Information Technologies*, 11(11), 19–24.
- WATANOBE, Y. (n.d.). G-bert: An efficient method for identifying hate speech in bengali texts on social media.
- Xu, W., Chen, J., Ding, Z., & Wang, J. (2024). Text sentiment analysis and classification based on bidirectional gated recurrent units (grus) model. *arXiv preprint arXiv:2404.17123*.
- Zulqarnain, M., Ghazali, R., Aamir, M., & Hassim, Y. M. M. (2024). An efficient two-state gru based on feature attention mechanism for sentiment analysis. *Multimedia Tools and Applications*, 83(1), 3085–3110.