Computational Analysis and Detection of Bengali Communal Violent Speech

By

Abdullah Khondoker
20301065
Enam Ahmed Taufik
20301398
Md. Iftekhar Islam Tashik
20301078
S M Ishtiak Mahmud
20301071

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

Department of Computer Science and Engineering
Brac University
June 2024

Declaration

It is hereby declared that

1. The thesis submitted is 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.

5. This thesis contains bad language from the datasets used; readers' discretion is advised.

Student's Full Name & Signature:

Abdullah Khondoker

Abdullah Khondoker
20301065

ENAM AHMED TAUFIK

Enam Ahmed Taufik
20301398

Md Iftekhar Islam Tashik

Md. Iftekhar Islam Tashik
20301078

Ishtiak Mahmud

S M Ishtiak Mahmud
20301071

Approval

The thesis titled "Computational Analysis and Detection of Bengali Communal Violent Speech" submitted by

1. Abdullah Khondoker (20301065)
2. Enam Ahmed Taufik (20301398)
3. Md. Iftekhar Islam Tashik (20301078)
4. S M Ishtiak Mahmud (20301071)

Of Spring, 2024 has been accepted as satisfactory in fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 24, 2024.

Examining Committee:
Supervisor:
(Member)

Farig Yousuf Sadeque

Dr. Farig Yousuf Sadeque
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

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

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson & Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Communal violence is intensified by the widespread use of cyber hate, leading to aggression and increased conflicts among different religious, ethnic, and social groups, creating a barrier to social harmony. This research focuses on evaluating Bengali textual data sourced from Twitter and Reddit comments. The primary objective of this study is to enhance the accuracy of detecting communal violence-inciting speech. To achieve this, we employed and fine-tuned large language models, specifically the pre-trained BanglaBERT, aiming for a significant improvement over existing detection methodologies. Improving the detection of communal violent speech will help content moderation systems to effectively moderate and remove content linked to communal violence, thereby fostering communal peace in the Bengali-speaking regions.

Keywords: Communal violence; Bangla Language; Religious groups; Ethnic groups; Social groups; Social media posts; Computational analysis; Violent speech; Machine learning(ML); Natural Language Processing(NLP)

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

AD	Augmented Data
BB	BanglaBERT Base
BBL	BanglaBERT Large
CR	Classification Report
EC	Ethno Communal
FTM	Fine-Tuned Model
$mBERT$	Multilingual BERT
ND	Noncommunal
ND	Nondenominational Communal
Pr	Paraphrased Data
RC	Religio Communal

Chapter 1

Introduction

Communal violence is when different ethnic or communal groups fight each other, driven by the group loyalty and hatred, and target victims by their group identity [1]. A nation's social harmony and stability can be severely damaged by communal violence, especially in a country like Bangladesh with a diverse and dense population. The rapid expansion of digital media and communication platforms has played a pivotal role in intensifying this threat, as more than 44.7 million people in the country actively engage in social media [36]. This surge in online activity has made it easier for hateful and inflammatory textual content, known as cyber hate, to spread. These harmful contents have already incited numerous incidents of physical violence against various ethnic, religious, and social communities, causing widespread social unrest. Therefore, proactively identifying cyber hate is crucial to prevent its continued contribution to communal violence. While significant efforts have been made to identify cyber hate of many kinds, we observed a notable research gap specifically focused on communal violence. To bridge this gap, we will concentrate primarily on the critical task of identifying community-targeted cyber hate. In pursuit of this objective, we will utilize a comprehensive dataset [43] comprising various forms of textual content, including comments and posts extracted from social media platforms. The professionals have categorized the data and each category of the dataset is carefully designed in terms of covering the wide range of communal violent speech and particular topics that can be attributed to this context. This categorization is useful for building our NLP model and discussion that follows. After that, we will turn to the creating a state-of-the-art machine learning techniques and Natural Language Processing (NLP) advancements. These models will be thoroughly trained on the vast dataset to detect and generate communal violent speech, which has subtle linguistic characteristics and contextual indications. Our ultimate aim is twofold: firstly, to determine whether a certain text can be categorized as communal violent speech and secondly, to identify the exact category within the communal violent speech spectrum. We hope that by reaching these goals we will be able to offer a useful mechanism for fighting communal violence spread through social channels. Our research aims to achieve not only to fill the existing research gap but also to add to the effort of proactively preventing occurrence of communal violence by increasing knowledge on the different linguistic indicators that causes it.

1.1 Research Problem

In this research, we aimed to address the issue of detecting violent speech in the context of communities by using comments obtained from social media platforms and using powerful pre-trained large language models capable of classifying speech with high accuracy and predicting its specific thematic classification.

1.2 Research Objective

- Some linguistic methods will be used to conduct data analysis on a huge and heterogeneous text corpus.
- State-of-the-art natural language processing models will be applied to extract relevant information and insights from the text data.
- We will attempt to detect comments inciting communal violence more accurately and identify the specific category of violent speech.
- The performance and effectiveness of the proposed methods and models will be evaluated using appropriate metrics and benchmarks.
- We aim to share the research findings and contributions by publishing them in a prestigious journal or presenting them at a reputed conference.

1.3 Research Structure

The remainder of this paper is structured as follows: In Chapter 2, we provide a comprehensive literature review and our motivation. Chapter 3 defines the classes and labels in the dataset with examples. Chapter 4 presents an overview and analysis of the dataset, detailing rigorous preprocessing steps to ensure data cleanliness and various augmentation techniques, including SMOTE, zero-shot, few-shot, paraphrasing, and manual augmentation. In Chapter 5, we introduce the models employed, specifically BanglaBERT, Multilingual BERT (mBERT), and an ensemble model incorporating a Multi-Layer Perceptron (MLP) classifier. Chapter 6 offers a detailed analysis of the results, focusing on four-class and sixteen-class metrics and exploring different ensembling methods like mean value, voting system, and MLP classifier. Chapter 7 delves into error analysis, addressing dataset limitations such as class imbalance and data annotation anomalies, as well as the limitations of pre-trained models. Finally, Chapter 8 concludes the paper and suggests directions for future work, emphasizing strategies for reducing class imbalances and improving the model based on error analysis, identifying areas for future research in Bengali communal violent speech detection.

Chapter 2

Literature Review

In our pursuit to investigate communal violence speech detection, we dedicated our efforts to thoroughly examining research papers within the domains of hate speech and communal violence. Our selection process was intensive, particularly emphasizing papers containing high citation counts and recent publication dates. We have organized these selected papers chronologically based on their publication date, enabling us to build a cohesive and current knowledge foundation for our research efforts.

Basave et al. [3] introduced the Violence Detection Model (VDM), a novel probabilistic modeling framework designed to tackle the identification of violent content and the extraction of violence-related topics from social media, particularly Twitter. The primary objective was to enhance classification and topic coherence in the context of rapidly changing social media events. To conduct their experiments, the researchers collected data from three datasets: TW, DB, and DCH. TW contained over 1 million Twitter tweets, categorized as violent and non-violent, collected over two months. DB was derived from DBpedia and included articles related to violence and non-violent documents from TW. DCH was created by transforming violent DBpedia documents into tweet-sized chunks. Their training dataset included 10,581 tweets, and the test dataset comprised 759 tweets for violence detection and topic analysis. Preprocessing steps involved cleaning the text data by removing various elements like punctuation, numbers, non-alphabet characters, stop words, user mentions, links, and hashtags, and applying Lovins stemming to reduce vocabulary size. Additionally, low-frequency words (less than 5 occurrences) were eliminated to address data sparsity. The core methodology of VDM was a weakly supervised approach that associated violence labels with documents, topics with violence labels, and words with both violence labels and topics. A switch variable was used to determine whether words were generated from a background distribution or a category-specific topic distribution. Model parameters were estimated using Collapsed Gibbs Sampling for latent violent categories and topics, given the observed data. In their comprehensive evaluation of violence classification models, including VDM, ME-GE, ME-PR, and JST, they primarily used the F1 score as the performance metric. Notably, VDM outperformed all other models, demonstrating clear superiority. In the IG phase, VDM consistently performed well with F1 scores ranging from 0.7566 to 0.8309, with the chunking strategy (DCH) achieving the highest F1 score. In the RWE strategy, F1 scores ranged from 0.7969 to 0.8575, with TW achieving the highest F1 score of 0.8575. Furthermore, the authors conducted a comparative analysis of violence classification accuracy using VDM, PLDA, and JST with varying numbers of topics. All three models initially performed similarly with one topic. PLDA's performance declined as the number of topics increased, while VDM remained stable and JST excelled with one topic but dropped with more. VDM also generated representative violent and non-violent topics related to specific events and performed well in topic coherence using Pointwise Mutual Information (PMI). However, the study lacked online learning strategies for VDM to dynamically adjust its parameters, which may limit its real-time use in violence detection from social streaming data.

Agarwal and Sureka [4] in their paper, tried to figure out how to use computers to find mean and extreme tweets, and they looked at how these tweets talked and what they looked like. The authors used a big set of real tweets to test if their computer system, which used math and patterns to learn, could tell them which tweets were bad and also they tried to compare models performance. Authors implemented two independent one-class classifiers (KNN and Lib- SVM) to classify tweets as hate-promoting or unknown. The authors first proposed method focused on automatically classifying tweets as either promoting hate or being of an unknown nature. It operated based on the one-class k-nearest neighbors (KNN) algorithm and required inputs such as a pre-processed training dataset, a testing dataset, the number of nearest neighbors, and a threshold measure for identifying outliers. Each tweet in the testing dataset was represented as a feature vector, with each feature denoted as fi, and m representing the total number of discriminatory features. The algorithm calculated the Euclidean distance between a testing instance and all instances in the training dataset. A distance matrix was created for each instance in the testing dataset, allowing for the determination of the nearest neighbor in the training data. To manage the large testing dataset, K was set to 100. The algorithm computed an average distance among all K nearest neighbors. If the ratio of distances was lower than the threshold, the instance was labeled as promoting hate; otherwise, it was marked as unknown. The threshold value was computed using the harmonic mean of training dataset distances. In another approached algorithm, One-class Support Vector Machine (SVM) was a supervised learning method used to estimate the distribution of a given training dataset. LibSVM provided a wrapper class for implementing one-class SVM classifiers. In this approach, all SVM formulations were framed as quadratic minimization problems. In this algorithm, they comprised essential modules of LibSVM and they input both the training dataset and testing dataset into the algorithm. They trained a one-class SVM and KNN on 10, 486 positive class tweets and observed an F-Score of 0.83 and 0.60 respectively. Authors identified that unlike KNN classifier, presence of internet slangs and question mark played an important role in LibSVM classifier.

Thomas Davidson et al. [6] addressed the challenge of distinguishing hate speech from offensive language on social media, which was crucial for effectively detecting hate speech. They employed a lexicon of hateful words, a multi-class classifier, and an analysis of the classification results. They gathered a hate speech lexicon compiled by Hatebase.org and collected a sample of tweets from 33,458 Twitter users, resulting in a corpus of 85.4 million tweets. They then randomly chose 25k tweets and hired CrowdFlower workers to label them as hate speech, offensive, or neither. They used the majority vote of the workers to label the tweets, resulting in 24,802 labeled tweets. Also, they applied various methods to categorize tweets into either hate speech, offensive but not hate speech, or neither. They used logistic regression with L1 regularization to reduce the dimensionality of the data and then tested different models such as naive Bayes, decision trees, random forests, and linear SVMs. They selected logistic regression with L2 regularization as the best model based on its performance and interpretability. The authors preprocessed each tweet by lowercasing, stemming, creating n-gram features, adding POS tag features, and computing readability and sentiment scores. They also included features for hashtags, mentions, retweets, URL, and the length of each tweet. The authors trained one classifier for each class and predicted each tweet's class label based on the assigned probability. The authors shared the outcome of their hate speech detection model; generally, the outcome accuracy was high of 0.90, though having a precision of 0.44 and recall of 0.61 for the hate class. They proved that the model misinterpreted hate speech as offensive language in the majority of cases when tweets included several slurs or were aimed at other users. They also reported that their model was able to identify the common manifestation of hate speech like racism and homophobic manifestation rather than the less common ones like sexism and misogyny. The authors also noted some limitations of the proposed hate speech detection method including the overlap between hate speech and offensive language, the ambiguity of the hate speech lexicon, the subjectivity and variability of hate speech labeling, context and application diversity of hate speech, and insufficient and unreliable non-linguistic features. They suggested some future works for hate speech detection on social media, such as identifying and correcting social biases in the detection algorithms, distinguishing between different uses and contexts of hate speech, capturing the nuances and subtleties of hate speech beyond explicit keywords or offensive language, and understanding the characteristics and motivations of hate speakers.

The Paper by Kumara [12] addressed the transformative impact of social media on information dissemination while highlighting the concerning rise of harmful and hurtful content propagation. The study presented a meticulously crafted approach, featuring a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model for detecting hate and offensive content across three languages: English, Hindi, and German. The advent of social media had revolutionized how information was shared, but it had also brought to the forefront a detrimental issue—the widespread dissemination of mean-spirited and offensive messages. In response to this challenge, Kumara et al. proposed a BERT-based solution. The authors employed a rich vocabulary to convert text that includes emoticons and emojis to equivalent text in order to preprocess the data for the BERT model. The authors used "distilbert-base-uncased" for English and "distilbert-base-multilingual-cased" for Hindi and German. Further, they used a maximum word length of 30 tokens for each text to standardize the input size of the model for different texts. The important step of tuning was explained wherein the pre-trained models were tailored to the specific task of identifying the hate and offensive speech. The authors used a batch size of 32 and a learning rate of 2×10^{-5} to train the model for 20 epochs to achieve a balance between model capacity and computational power. The results of the research were summarized in the evaluation of the fine-tuned BERT model for several sub-tasks and languages using precision, recall, and F1-score metrics. Impressively, the model consistently produced favourable results, with F1-scores ranging from 0.1623 to 0.5300. While the study showcased notable accomplishments in multilingual hate speech detection, it also acknowledged its limitations. These constraints included the possibility of dataset narrowing, the complex problem of emoticon and emoji translation, and the necessity for further in-depth analyses of negative effects and obstacles encountered during fine-tuning. More comprehensive testing might have helped the selection of model training parameters, and future research possibilities required more in-depth talks about potential difficulties.

This research [13] proposed a machine learning-based event categorization and analysis method to identify violence-related events in Bengali newspaper articles, a large corpus of unstructured text. This study's dataset was compiled from a range of online Bengali newspapers. This thorough effort yielded a diverse dataset of news articles on many topics and events. A complete text preparation procedure was used to modify and prepare news items for analysis after data collection. This involved removing irrelevant letters and sentences. The news stories were categorized as either violent or neutral. Based on the nature of the incident, violent events were further divided into many categories. The train set of 600 news articles, each categorized, and the test set of 300 news articles were generated using this thorough dataset generation method. The researchers employed a systematic methodology to analyze news articles with the aim of identifying instances of violence. They began by gathering news data and then focused on extracting n-grams from stories containing specific triggering phrases. These n-grams varied in length from 2 to 5 words and followed the same word order as the source articles. Researchers described the dataset using these extracted n-grams and a binary vector technique to signal n-grams in news items. By considering the term frequency of n-gram features, they reduced the feature set dimension to improve dataset quality. Subsequently, six different machine learning algorithms, including K-Nearest Neighbor, AdaBoost Classifier, Decision Tree, Random Forest, Support Vector Machines with a linear kernel, and Logistic Regression, were applied to the training dataset, which was a multi-class classification problem with violence categories. To assess model performance, a separate test dataset of 300 articles from Bengali newspapers was created. Among the evaluated machine learning models for classifying violence-related events in news articles, Logistic Regression achieved the highest accuracy at 83.4%. Random Forest and Decision Tree models also performed strongly, with accuracies of 81.7% and 78.5%, respectively. AdaBoost Classifier yielded a respectable accuracy of 76.4%, while Support Vector Machines with a linear kernel achieved 78.2%. The K-Nearest Neighbor model had a lower accuracy of 48.9%. To validate the accuracy of Logistic Regression, a test was conducted, involving a random selection of 50 news articles from each class type. These articles were measured against their predicted labels. The accuracy rate was 97.4%, with 341 of 350 articles classified correctly according to their predicted labels. Defining and categorizing event types may have been difficult. Rare events were hard to predict due to limited historical data.

The authors [18] introduced a novel method called DeepHateExplainer, which utilized a collection of BERT variants to identify hate speech in the Bengali language.

Additionally, the approach offered transparent explanations for the generated predictions. The Bengali Hate Speech Dataset (BHSD) was expanded by including an additional 5,000 labelled examples. The dataset classified the observations into five categories of hate: political, personal, geopolitical, religious, and gender-abusive. The dataset's text was collected from various online sources, such as Facebook, YouTube comments, and newspapers. The dataset included specific types of texts containing common slurs and terms directed towards a specific person or entity or generalized towards a group. They proposed DeepHateExplainer, which leveraged a neural ensemble method of transformer-based neural architectures, such as Bangla BERT-base, multilingual BERT-cased/uncased, and XLM-RoBERTa, to classify Bengali texts into different categories of hate speech, such as political, personal, geopolitical, and religious hates. The texts were first preprocessed comprehensively to remove noise and irrelevant information. Then, sensitivity analysis and layer-wise relevance propagation (LRP) were applied to identify the important terms in the texts that contributed to hate speech detection and provide human-interpretable explanations. Comprehensiveness and sufficiency scores were used to measure the quality of explanation which referred to how true the explanations were to the model's prediction. They also evaluated DeepHateExplainer against other ML and neural networks-based baselines such as linear and tree-based models, CNN, Bi-LSTM, Conv-LSTM using word embedding. They argued that DeepHateExplainer would be able to achieve F1 scores of 78%, 91%, 89%, and 84% for these categories than ML and DNN baselines. They also evaluated the quality of the explanations by calculating Other completeness and sufficient other metrics that show how close the explanations are to the model's predictions. DeepHateExplainer They demonstrated that it was superior to several baseline methods in terms of F1 rates for various categories of hate speech, including machine learning and neural network models. The authors faced the issue of limited publicly available hate speech annotated data for the task of Bengali hate speech classification. They had only a limited amount of data and may possibly not have included a broad and challenging set of hate speech in Bengali. They also did not go into further details as to how their method can solve other general problems such as its applicability to other domains and languages, its robustness to adversarial examples, and others, including its societal impact. The given aspects could be strengthened in future by gathering more information, carrying out more experiments and also find the dangers and benefits of the approach.

This study [19] aimed to fill the gap in the literature for detecting biased and aggressive contents on the platforms such as YouTube, Twitter, Facebook, and Telegram. The ComMA project was meant to create a system that filtered out such content that is with gender and religious orientation. This paper formulated a multi-label classification problem to annotate the comments as aggressive, gender biased, and communally charged and hence is a useful resource for the researchers working in the area of hate speech and text classification. The corresponding dataset for this study is called ComMA v 0. 2, was developed by Kumar et al. (2021b) for the multilingual assessment of aggression, gender bias, and use of communal language on social media platforms. The sentences were then classified using the spacylangdetect toolkit to aid in standardizing the language; this generated tags for Hindi, Bangla, and English. The sentences tagged as Hindi and Bangla were further subjected to transliteration to get a uniform multilingual corpus in English. Text data was preprocessed by using the Count Vectorization method to transform data into frequency-based vectors in numerical form. These vectors were then fed into state-of-the-art models, including XGBoost, LightGBM, and the traditional Naive Bayes, forming an ensemble voting classifier to determine the final label. The study assessed performance across tasks, revealing distinct results. Aggression analysis scored lowest with F1 scores: Multilingual (0.361), Bangla (0.442), and Hindi (0.402). Conversely, Gender Bias demonstrated higher effectiveness with F1 scores: Multilingual (0.632), Bangla (0.669), and Hindi (0.702). Communal Bias achieved notable success: Multilingual (0.777), Bangla (0.866), and Hindi (0.642). The Boosted Voting Ensemble method, employing XGBoost, LightGBM, and Naive Bayes, showed resilience, especially for unknown inputs. However, it faced challenges in classifying aggression due to contextual overlaps. IndicBERT performed strongly in Gender Bias (Multilingual: 0.558, Hindi: 0.796, Bangla: 0.732) and Communal Bias (Bangla: 0.876, Hindi: 0.639, Multilingual: 0.783). Aggression analysis, however, remained the lowest-performing task for Indic Bert (Bangla: 0.341, Hindi: 0.439, Multilingual: 0.357). This study compared two approaches, Boosted Voting Ensemble and IndicBERT, for analyzing aggression, communal bias, and gender bias in multilingual social media data. While IndicBERT excelled in individual tasks, the ensemble approach demonstrated higher adaptability and robustness across all categories. Aggression analysis posed challenges due to contextual overlaps. Future work may have explored advanced embeddings like GloVe and BERT for performance enhancement, emphasizing the complexities in handling such contextual overlaps, as well as considered ensembling techniques in deep learning settings.

The authors [20] initiated the ICON-2021 task, aiming to identify aggression, gender bias, and communal bias in multilingual social media comments. This paper presented task details, dataset sources, and participant engagement. The fact that most systems only predicted a third of three-class classifications was notable. The dataset received significant preprocessing for quality and homogeneity after being cautiously collected from several social media networks. Although detailed information on embeddings and procedures was not expressly supplied, the authors' thorough strategy implied a well-thought-out choice. Modern embedding techniques for natural language processing were probably used in this. They used comprehensive feature engineering and model implementation to detect rage, gender prejudice, and communal bias in multilingual comments. For a detailed understanding of the dataset's preprocessing and feature engineering procedure, Kumar et al.'s work from 2021 was a priceless resource. Different strategies were used by the participating teams to create their systems for the joint assignment. The participating teams' methods incorporated a wide variety of models and strategies. It was common to use BERT-based models that were adjusted for specific language variances and code-mixed data. Other teams examined Random Forest, Logistic Regression, SVM, and Decision Tree classifiers. XLM-RoBERTa, MURIL, and BanglaBERT models were also used, along with bespoke attention and mean-pooling. Some teams used ensemble methods, Logistic Regression with word and character n-grams, and multilingual sentence encoder vectors with pre- and post-aggregation strategies to improve performance. These diverse strategies demonstrated the wide array of approaches taken by different teams to address the challenges presented in the competition. The evaluation metrics including instance F1 and Micro F1 provided a weighted average, especially useful for class-imbalanced cases. All the teams gathered F1-scores on Meitei, Bangla, Hindi & Multilingual Dataset. The highest weighted instance F1-scores were 0.322 for Meitei, 0.292 for Bangla, 0.398 for Hindi, and 0.371 for the multilingual category. The best model used joint training for all subtasks and languages, demonstrating the efficacy of multitasking and multilingual learning in low-resource settings. Well-tuned linear classifier ensembles outperformed Transformers-based systems. The comprehensive error analysis showed that model errors were mostly caused by overfitting to specific linguistic features and limited generalisation to out-of-domain data. To improve model performance, more diverse training data and careful data point selection were needed. Another important factor was contextual knowledge, which required explicit contextual information in the dataset and model.

The paper of Das et al. [17] proposed using an encoder-decoder machine learning model, a well-known tool in the field of NLP, for the categorization of Bengali user comments on Facebook pages. The authors categorised 7,425 Bengali comments into Hate Speech, Aggressive Comments, Ethnical Attacks, Religious, Religious Hatred, Political, and Suicidal Comments. They tokenized and removed bad characters and punctuation during dataset preprocessing. Then they removed all stopwords from their dataset. Bangla text was preprocessed using emoticon and emoji analysis. The author compared CNN, Bi-directional LSTM, GRU, and RNN with Attention Mechanism for abusive text classification in Bangla, including binary and multi-class categories. This paper used a neural network to simulate human brain function, focusing on the Additive Attention Mechanism. This experiment used a simplified Attention Mechanism to emphasise selective activities in a context. Textual characteristics were understood using a bidirectional Recurrent Neural Network (RNN). This choice was made because a sentence's meaning often depended on previous and subsequent text. Thus, the bidirectional RNN merged information from both directions to create a comprehensive feature vector, improving text semantics. Additionally, the neural network model assigned word weights using an attention mechanism. This weight assignment was crucial because it emphasised important keywords and devalued less important ones. Long source sentences were the main focus of the attention mechanism in neural machine translation. Instead of using the encoder's final hidden state, it connected the context vector to the entire source input. The authors described the encoder layer, a bidirectional RNN (LSTM) with forward and backward hidden states concatenated to form the encoder state. This unique design combined preceding and following words into one word. An alignment model scored input-output pairs based on compatibility in this model. The weight coefficients (α) determined the contribution of each source's hidden state to the output. These alignment scores were calculated using a feed-forward network with one hidden layer trained with other model components. A score represented the scoring function, where va and wa were alignment model training weight matrices. Finally, the model generated new words in the output sequence using alignment scores and context vectors to predict speech category. This comprehensive method enhanced the model's text reading and translation abilities, making it useful in natural language processing. The decoder used LSTM and GRU models and an attention mechanism to improve performance. EarlyStopping and an 80% training 20% testing data split yielded 74% accuracy for the LSTM and GRU models and 77

Ebipatei et al. [22] research paper focused on spotting subjective bias in text. The objective of the authors was to develop a practical model for the detection of biases in text with particular attention to biases that could be identified in phrases from Wikipedia. The authors used a number of strategies. They used a WNC dataset of 360,000 utterances split half-and-half between biased and neutral terms. The three groups of biases – framing, epistemological, and demographic – were identified as 57%. 7%, 25%, and 11.7%, respectively. The suggested model included the use of both BERT and BiLSTM neural network architectures which are two of the most popular neural network designs. BERT was employed to pre-process the text data and it effectively captured the surrounding details. The authors further made the model more superior by incorporating attention processes that helped to pay attention to important textual elements while categorizing. The authors employed two important measures: accuracy and an F1 score to evaluate their method. Their proposed model based on BERT and BiLSTM with attenuation was impressive with an accuracy of 0.89 and F1 score of 0.90. Also, this combined model was superior to both individual BERT and BiLSTM with Glove embeddings as well as BERT combined with BiLSTM without attenuation. This demonstrated the ability of the model to identify biases with high efficiency. Notably, these results also outperformed those of other popular models, which is a significant development for state-of-the-art bias detection systems. The paper also outlined future research prospects in this area. It recommended improving the model by adding the support of other languages instead of being limited to English only. Moreover, the authors suggested that the current bias detection should be developed to a document level, to identify larger contexts of biases.

The authors [21] undertook a crucial examination of the profound changes observed in the religious and social behaviours of two distinct communities in the aftermath of the 2012 Ramu violence incident. This study endeavored to delve into the psychological and sociological aspects of the minority community's mindset following the violence, while also shedding light on the evolving patterns of social interactions between minority and majority communities in the post-riot scenario. The 2012 Ramu violence stood as a significant event in the context of Bangladesh, marked by a series of orchestrated attacks on Buddhist monasteries, shrines, and residences of Buddhist residents in the Ramu Upazila. The violence erupted abruptly in the late hours of September 29, 2012, with local mobs embarking on a destructive rampage. Surprisingly, a total of 12 Buddhist temples and monasteries, as well as 66 residential houses, were specifically targeted and subjected to acts of vandalism. The commencement of this series of devastating events was triggered by a visually stimulating depiction, illustrating the act of disrespect towards the Quran, which had been disseminated on a fabricated Facebook profile erroneously associated with a male adherent of Buddhism. As the manifestation of violence transpired, its consequences extended beyond the Ramu Upazila, impacting the southern districts of Bangladesh as well. In this context, it is noteworthy that Buddhist monasteries, Sikh Gurudwaras, and Hindu temples were subjected to deliberate acts of aggression, thereby intensifying the existing religious animosity and societal disharmony. Soruar et al. research sought to provide invaluable insights into the aftermath of the Ramu violence, particularly focusing on the response of the minority community and the evolving dynamics between minority and majority communities in the wake of this tragic incident. By examining the changes in religious and social behaviours, the study contributed to their understanding of the profound impact such communal violence can have on the fabric of society.

This paper [32] centered on detecting Bengali hate speech on public Facebook pages, tackling challenges such as limited datasets and the intricate linguistic landscape. This paper delved into the effectiveness of machine learning models and emphasized the linguistic diversity inherent to the Bengali language. In this study, they veraciously prepared their dataset sourced from Facebook, employing the FacePager program for data collection. Since informal language was common online, rigorous manual examination excluded non-Bengali comments. To keep the language consistent, they included impure text and even dialect. Refined further entailed the removal of emojis, punctuations, numerical values, and special characters. This careful selection resulted in a final set of 10,133 Bengali comments that they made publicly available in their GitHub repository. To distinguish hate speech, they referred to Facebook community standards and group those comments that directly harm individuals or group of people according to characteristics such as race, ethnicity, nationality, religion, or sexual orientation. Several categories were made due to the dynamic nature of hate speech that was most of the time mixed with other aspects of speech like sarcasm. Pre-processing further involved character and punctuations removal, tokenization, stemming, and stopword's removal. Also, they used pre-trained Bengali word embedding from fastText for Linguistic representation. Using numerous machine learning and deep learning models, they carefully tuned the hyperparameters of models, which include Support Vector Machine, Random Forest, Convolutional Neural Network and more. The results presented an extensive assessment of machine learning and deep learning architectures. Among machine learning techniques, SVM with RBF kernel demonstrated exceptional performance with an F1 score of 0.87, achieving an accuracy of 87.22%. This was closely followed by SVM with linear kernel, Random Forest (RF), and K-Nearest Neighbour (KNN) with commendable F1 scores of 0.83. SVMs excelled at class separation, helping balance datasets. Decision Tree (DT) performed worst due to its susceptibility to deviation and increased complexity with more data. Multilayer Perceptron (MLP) outperformed other deep learning methods with an F1 score of 0.86 and accuracy of 85.64%. The 1-dimensional CNN performed well with an F1 score of 0.84 and an accuracy of 84.70%. Although computationally intensive, LSTM had an F1 score of 0.77. These findings demonstrated the efficacy of SVM with RBF kernel and MLP in machine and deep learning. The study found that online hate speech presented unique challenges like misspellings, grammatical errors, and sarcasm. Emoji sentiment analysis was a future research topic. Study limitations included expanding the dataset, exploring multivariate categorization, and deepening sentiment analysis.

This study [27] introduced a novel approach to multi-label learning, which enabled a more nuanced categorization of hate speech compared to traditional single-label methodologies. It also addressed the various challenges associated with text mining in social media data, including the diversity of platforms, the subjectivity and complexity of language, and the significance of context and bias. The research utilized a unique dataset that included multi-social media data which was semi-automatically labeled by the use of a semi-supervised approach. The study also applied data augmentation and balancing techniques to overcome the limitations of scarce and imbalanced data availabilities. The study conducted multi-label hate speech severity classification on social media data using the BERT model. In this regard, the study deployed semi-supervised approaches to annotate a newly developed dataset that contains information from various social media channels, including YouTube, Reddit, Twitter, and Kaggle. The study also aimed at improving the dataset through the use of data augmentation and balancing techniques such as GPT-2 and the AugLy Facebook library. The research employed a methodical approach encompassing various stages such as experimental design, text mining, data augmentation, pseudo-labeling, algorithm selection, testing, prediction, and evaluation. The model exhibited varying performance across different categories of hate speech, namely toxic, severe toxic, obscene, threat, insult, and identity hate, with scores ranging from 0.734659 to 0.948695. Using BERT and data augmentation, the authors proposed a multi-label learning solution for hate speech detection. They showed that their model could classify hate speech into different severity levels and provide probability scores for each label. They also improved the minority classes by generating synthetic hate comments using GPT-2 and AugLy. They also analyzed the dependencies between hate data and their classes and collected data from multiple platforms to enhance generalization. However, their study had some limitations, such as using synthetic data that may not match real-world hate speech, relying on subjective human-annotated ranking for evaluation, omitting an ablation study to examine the model components, using a mixed dataset that may vary in quality and representativeness, and focusing on English social media only. They suggested future directions to address these limitations, such as conducting an ablation study to identify the model's most important factors or components and extending the model to different languages or cultural contexts.

The study by Mohammad Javed [31] examined transformative shifts in Bangladeshi nationalism, spotlighting the significant role of social media platforms like Facebook and Twitter. These online platforms shaped secularism, atheism, and the Shahbag movement. Employing a mixed-method approach encompassing online observation and offline interviews, the research underscored the intricate interplay between virtual and physical spaces in molding nationalist sentiments within everyday life. This paper examined the impact of social media on Bangladeshi nationalism and religious sentiments, drawing from anthropological participant observation techniques and netnographic research methods. Influenced by Kozinets' work on netnography, the study explored the online self in an interactive manner, utilizing both qualitative and quantitative approaches. The research, sensitive in the context of Bangladesh, avoided direct comments on online platforms but raised issues during observations for offline interviews. It encompassed 14 in-depth interviews conducted in Sylhet and Dhaka, shedding light on the relationship between online and offline activities, religious views, perceptions of nationalism, and emotional attachment to religion. Content and social network analyses were employed to discern patterns in online activism, particularly among religiously motivated pro-nationalists. The study showed that social media, particularly Facebook, was shaping Bangladeshi religious and nationalist sentiments. Bangladesh's social media boom, led by Facebook, showed its power to mobilise social movements and shape religious beliefs.The paper also examined how social media had exacerbated racial tensions, particularly after sensitive events. Online activism promoted anti-atheist and anti-secular views, showing how the online and real worlds interacted. Overall, the study showed how social media shaped Bangladesh's nationalist and religious narratives. It also discussed how secularism and atheism were used interchangeably in Bangladesh. Atheism rejected all faiths, while secularism was a political theory that tolerated many religions. Abdul Hannan, a madrasa instructor and informant, expressed the widespread belief in Bangladesh that secularists and atheists threatened Islam. The teacher claimed secularism might reduce Islam. Social media cyberbullying and physical harm against atheists and secularists reinforced the stigma. The internet also promoted religious nationalism by emphasising the need to expel atheists and secularists.

Pareek et al. [30] addressed the significant problem of hate and aggression on social media, particularly when expressed in code-mixed language. The authors proposed a solution based on deep learning algorithms, specifically Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) models, with support from natural language processing techniques. Authors used dataset in their paper of HASOC. This dataset was a collection of social media posts that had been used in many different intentions, which had mostly used multilingual word and language. Authors organized this paper into three critical phases: pre-processing, feature extraction and embedding layer creation, and model training. Pre-processing involved data cleaning, tokenization, and encoding, while the embedding layer was built using Fasttext word embeddings. Two deep learning classification models, CNN and Bi-LSTM, were utilized, with specific parameter settings such as epochs, batch size, activation functions, dropout, and maximum text length defined for model training. The research introduced two distinct classification tasks: Task 1 distinguished between Hate and Non-Hate posts, while Task 2 further categorized Hate posts into Hate, Offensive, and Abusive categories. This granularity enabled a more nuanced analysis of social media content and varying degrees of negativity. The experimental result illustrated the adequacy of both CNN and Bi-LSTM in recognizing hate in code-mixed dialect, with CNN imperceptibly outflanking Bi-LSTM. CNN achieved F1 scores of 0.70 and 0.71 for Tasks 1 and 2, respectively, while Bi-LSTM attained F1 scores of 0.73 and 0.67 for the same tasks. One notable limitation of this study was the potential dataset bias, as it may not fully represent the complexity and diversity of social media content and user interactions. The dataset's composition and source could introduce sampling bias, limiting the generalizability of the developed models to a broader range of real-world scenarios. Authors suggested for future work to apply this method by working on its dataset and creating more labels and posts.

Khan et al. [28] study addressed the critical challenges of detecting cyber aggression in social media, emphasizing the dire need for more effective filters. Their innovative approach introduced novel emotional features within a streamlined deep neural network (DNN) model, outperforming existing methods. The profound impact of cyber aggression on mental health underscored the urgency for improved detection techniques in this domain. The study primarily focused on leveraging the Cyber-Troll dataset, comprising 20,001 tweets categorized as either cyber-aggressive (7822) or non-cyber-aggressive (12,179). Preprocessing involved removing extraneous information using NLTK, while feature extraction encompassed discrete emotions and

Word2Vec embeddings, significantly enhancing model effectiveness. To tackle overfitting, mutual information-based feature selection identified the top 25 features, which were then combined with eight emotional features. Seven machine learning algorithms (SVM, LR, KNN, DT, NB, GB) and four deep learning models (LSTM, BiLSTM, CNN, DNN) were implemented, with the DNN model, characterized by three dense layers, ReLU activation, and sigmoid output, demonstrating promise in classifying aggressive and non-aggressive tweets. The inclusion of emotional features and Word2Vec embeddings significantly contributed to model performance, holding potential for real-world applications in curbing cyber-aggression and enhancing online discourse. In the context of aggression detection in textual data, especially within social media, this paper advanced existing methodologies. The GB classifier's maximum F1 score was 77%, demonstrating that the traditional machine learning methods could not be relied upon to accurately detect violence. Also, it championed a streamlined yet efficient DNN model, effectively learning and identifying aggressive language even with fewer layers. Integration of eight emotional features and Word embeddings enhanced classification capabilities, achieving an impressive F1 score of 97% (DNN), surpassing other models in accuracy and training time. This signified significant progress in aggression detection, providing a more effective solution compared to previous methods. The research improved previous methods and opened up multi-lingual applications, especially in non-native English-speaking regions. Future work included scaling up to larger datasets for more comprehensive analysis. Online aggression could be better understood by including other languages and social media platforms.

This study [29] employed machine learning and natural language processing to automatically recognise hate speech and objectionable words that insulted or attacked someone based on their religion, ethnicity, or race. To lessen the detrimental effects of online hate speech and build algorithms to adapt to its changes. A Twitter dataset from Kaggle was used to build and evaluate a system for detecting hate speech and offensive language. The dataset had two CSV files with three columns: id, label, and tweet for training and testing. Over 3000 people downloaded, and 12 contributed to the dataset. The paper described how to clean, preprocess, and classify dataset tweets. NLP techniques were used to correct spelling errors, remove stop words, and assign hate speech percentages to the tweets. Word cloud visualization showed the dataset's most frequent and relevant words. Oversampling was performed to balance the number of hate speech and regular tweets in the training dataset. ML classifiers included SVM, RF, multinomial NB, XG Boost, and Logistic regression. Classifier accuracy, recall, F1 Score, and support were measured using accuracy score, classification reports, and confusion matrix. SVM was found to be the best classifier with 98% accuracy, followed by Random Forest and Logistic Regression with 97% accuracy each. The authors identified several challenges in generalizing hate speech detection, such as defining hate speech, creating unbiased datasets, evaluating models, and addressing ethical and social issues. They used Twitter data, which may not reflect other languages or settings. Their system's authenticity and accuracy may be affected by their limited hate speech definition. Their dataset and annotation had biases, which may limit model generalizability and fairness. They ignored ethical and social issues including false positives and bias amplification in their method. The computational resources needed for their system were not reported. They proposed testing their system on additional platforms, developing more diverse and inclusive datasets, addressing ethical issues and biases, exploring new approaches and strategies, and optimising hate speech detection computing resources.

Riley Botelle et al. [25] presented an NLP model to evaluate interpersonal violence in mental healthcare electronic records. They extracted clinical text referring to violence, such as the presence of violence, patient status (perpetrator, witness, victim), and violence type (domestic, physical, sexual). They also considered the high risk of violent victimization for people with mental illness, especially women. They used data from the CRIS database, which contained over 500,000 de-identified EHRs of patients from a specialist mental healthcare provider in London. They searched the CRIS database using 17 keywords and selected 3,771 text extracts from the records of 2,832 patients. They had two clinical medical students annotate the text extracts for reference to violence. They fine-tuned BioBERT, a pretrained transformer model for biomedical text mining, on the annotated dataset. They evaluated the performance of the NLP models using 10-fold cross-validation, measuring precision, recall, and F1 score for each model. They estimated inter-annotator agreement on a subset of the annotated data using agreement and Cohen's kappa. They developed annotation guidelines iteratively based on discussions and queries raised by the annotators. They achieved acceptable scale, efficiency, and accuracy levels in their NLP model, with precision ranging from 89% to 98% and recall ranging from 89% to 97% for different violence labels. They showed high inter-annotator agreement for the annotation labels, ranging from 82-96 (60-85 Cohen's kappa). They filled a gap in research by using BioBERT in mental health applications, which had not been extensively researched before. Their methodology could estimate the occurrence of clinical references to violence in EHRs, providing valuable insights into the prevalence of violence in mental healthcare settings. The authors discussed several limitations of their study, such as the lack of temporality, witnessing, and hidden violence in their NLP models, the annotator disagreements on some labels, and the limited generalizability of their data. They suggested future directions to address these issues, such as integrating the time of violence mentions, modeling the effect of witnessing violence on health, and exploring more categories of violence beyond the selected keywords.

This research [39] addressed the escalating problem of hate speech on online platforms, particularly social media, and aimed to devise more effective methods for its detection and mitigation.The dataset consisted of tweets categorized into three classes: Hate speech (0), offensive language (1), and neither (2), totaling approximately 24,000 entries. Data was divided into training (75%), validation (15%), and testing (10%) sets following specific ratios. Data preprocessing involved several critical steps, including tokenization, lemmatization, and one-hot encoding, aimed at converting the text data into numerical representations. Furthermore, word embedding techniques, specifically Word2vec and Gensim, were applied to capture semantic relationships within the text. The authors compared the performance of three deep learning models for hate speech detection: LSTM (Long Short-Term Memory), Bi-LSTM (Bidirectional Long Short-Term Memory), and a modified Bi-LSTM architecture. LSTM was proficient in dealing with sequences while Bi-LSTM took into account both past and future contexts for better contextuality. The modified Bi-

LSTM had additional layers incorporated to further improve its performance. The research also revealed some other models such as RCNN that effectively merged the powers of CNNs and RNNs to extract spatial and temporal information from the text data. The methodology included training the models using labelled datasets that included instances of hate speech and instances of non-hate speech. Loss functions for the models were then optimized using optimization algorithms. To evaluate the performance of the model, the evaluation metrics like F1-score, recall, accuracy, and precision were measured on a different test dataset. The highest F1-score of 0 was recorded for the RCNN model among the models. 90, which shows that the model is highly accurate at identifying instances of hate speech. The second and third positions were taken by the Modified Bi-LSTM and Bi-LSTM models with F1-scores of 0. 8916 and 0. 8901, respectively. The LSTM also continued to show reasonable performance with an F1-score of 0. 8388. Limitations included the potential inability to detect hate speech presented as sarcasm or irony and an imbalanced dataset affecting performance.

Titli and Paul [41] proposed a multi-modal approach for identifying hate speech in Bengali language, focusing primarily on two consumption types, 'Sadhu' and 'Cholito'. They employed neural ensemble models to analyze both textual and visual inputs, utilizing two distinct datasets that they classified into nine segments. The dataset comprised 30,000 comments, with 10,000 labeled as hate speech, and an additional 3,938 samples in the classified dataset. To extract features for analysis, they employed the Bengali BERT tokenizer and adapted the pre-trained Bengali BERT algorithm with various hyperparameters for training, conducting ten epochs for each model. Specifically, they evaluated two models, BertTwo and BengaliBert, with the latter achieving the highest accuracy of 0.706 using training batch size 16, test batch size 32, learning rate 0.00005, and dropout rate 0.4. The authors identified two key areas for future research: multi-domain data sets and harmful comment categorization. They acknowledged that dataset imbalance posed a substantial challenge, potentially leading to inflated conclusions, and noted that random undersampling, a common approach to address this issue, may result in information loss. To mitigate this limitation, they advocated for exploring more advanced techniques such as oversampling and synthetic data generation. Moreover, they pointed out the importance of improving models' capability in various tasks and their awareness of ambiguous and situational harmful contents. Ethical concerns were identified as a key area in model development as it relates to creating less biased and more fair categorization methods. The authors emphasized the importance of interpretability to ensure users' trust in these models.

This study [37] sought to understand the occurrence of hate speech in social media platform in the Bengali language particularly in Bangladesh. The authors developed a new model called G-BERT which incorporated BERT for text encoding and GRU for hate speech classification. The datasets for the study were collected by data crawling from Bengali online news sites and social media sites such as Facebook and Twitter by employing the BeautifulSoup Python module. Particular time intervals and user participation were applied, together with particular keywords for hate speech and abusive language. A total of 20,000 posts, comments, and memes were analyzed with more than 50% of the contents deemed offensive. Graduate students and Bengali language processing experts processed this data and used it to create the Bengali Offensive Text from the Social Platform (BHSSP) dataset. In order to facilitate fair model assessment and building, the dataset was divided into training, testing, and validation sets. Emoji translation, hashtag translation, and removal of Bengali stop words were performed as pre-processing. The study utilized N-gram features, TF-IDF weighting, and a combined model called G-BERT. The N-gram features from unigrams to trigrams were adopted to extract the necessary language patterns and weighted by TF-IDF values. BERT and GRU were integrated in the G-BERT model with the aid of the power of pre-trained contextualized language model. BERT was applied for contextualized word embeddings and GRU was used to classify these embeddings. In the process of effective detection of hate speech in Bengali writings, the hybrid approach made effective use of both models. The proposed G-BERT model was tested using critical parameters like accuracy, precision, recall, F1-score. Thus, G-BERT was effective in detecting hate speech. G-BERT was able to surpass many other existing models such as Bangla BERT, LSTM-BERT, AdaBoost-BERT, and L-BOOST with an overall accuracy of 95.56% Further, the results of recall and accuracy of G-BERT were also found to be very high at 95.07% and 93.63%, respectively. The greatest F1-score, 92.15% illustrated this great performance and showed that G-BERT has more accuracy and recall abilities compared to the mentioned numbers. The authors highlighted a number of limitations. Due to its training dataset, the G-BERT model's linguistic variety was restricted. Dependence on a small sized dataset might limit generalizability. There was no real-time performance evaluation, and the emphasis on hate speech might obscure other offensive linguistic patterns.

The work by Dimosthenis et al. [33] addressed the challenge of automatically detecting hate speech in online social media. The main objective was to test hate speech recognition algorithms on multiple datasets with different biases and traits. Researchers collected 13 hate speech datasets with various forms and temporal features. It was crucial to use this approach to evaluate hate speech detection algorithms across a wider range of online discourse. The datasets that made up this group included MHS, CMS, HTPO, HateX, AHSD, HSHP, AYR, MMHS, HatE, HASOC, DEAP, and LSC. The researchers put a preprocessing pipeline in place to standardize the datasets and assure consistency. This required concentrating on English-language tweets on Twitter, excluding non-English and non-Twitter tweets using a language identifier and eliminating duplicate entries. Furthermore, URLs and mentions were removed, and datasets with tweet IDs and labels were removed. Then, the datasets were either used directly for binary classification or mapped to binary classification for hate speech. For balance in the multiclass scenario, the subclasses of hate speech were divided into seven categories. The aggregated dataset, which included 83,230 tweets from 13 different source datasets, gave a complete depiction of hate speech, with almost 33% of the tweets labelled as hate speech. BERT-base, RoBERTa-base (general-purpose models), BERTweet, SVM, and TimeLMs-21 (specialized in social media, mainly Twitter) were used by the researchers. The models were rigorously tested in both binary and multiclass hate speech categorization contexts, with the macro-averaged F1 score serving as the performance parameter. For single-class classification, TimeLMs outperformed other models, achieving an impressive average F1 score of 70.7, closely followed by BERTweet at 70.1.

RoBERTa and BERT also delivered competitive results with F1 scores of 69.7 and 67.9, respectively, while the SVM model lagged slightly behind at 67.2. In contrast, for multiclass classification, SVM secured the highest F1 score of 81.9, thus highlighting its effectiveness in handling the complex classification task posed by the amalgamated dataset. TimeLMs remained a strong contender with an F1 score of 71.6, while BERTweet and RoBERTa exhibited similar F1 scores of 70.9 and 69.5, respectively. BERT, while still performing respectably with an F1 score of 66.9, appeared to be slightly less effective than the other models within this specific dataset context. Due to computational constraints, the study used base-sized language models and relied on existing datasets, which may introduce biases. It also focused on Twitter and English, limiting generalizability.

Fethi Fkih et al. [35] introduced a sophisticated approach to address the critical issue of threat detection on Twitter. The DetThr model integrated with the semantic network ThrNet, offered an effective measure to discover threats in the vast terrain of social media content. It is worth mentioning that several critical phases were at play in their technique of studying. They first described their steps which included data pretreatment and cleaning such as tokenization and stopword removal to ensure data quality. This first step was essential in creating a ‘clean’ dataset that was ready for analysis. Threat Target Identification was an important aspect in their process. Fkih et al. later discovered that knowing the purpose of a threat was an essential part of threat detection. They did this by combining two techniques: NER/Phrase matching. This method was very careful which ensured that the target of a threat was accurately identified within tweet messages. ThrNet was an essential component of their plan and had the function of a moving x-ray of an interconnected knowledge base that exposed the hidden links between concepts embedded in Tweets. Statistical and graphical data modeling techniques were used to build ThrNet, which proved to be highly effective in identifying and linking concepts in tweets based on frequency of mention. The WordLink module within the WORDij program iteratively constructed and strength-weighted word pairs by using a training dataset to construct this network. It made sure that it eliminated all low frequency occurrences of phrases as well as pairings to ensure that performance is maximised. The Gephi tool provided the visualization of this semantic network where the complexity of this relationship can be studied and analyzed. The DetThr model saw the advancement of threat detection by integrating ThrNet which significantly improved its capacity. It utilized ThrNet’s knowledge to realize threats from non-threats contained in tweet messages. DetThr worked as the sentiment analysis of a given tweet by locating the target in ThrNet and then checking the threat potential of a given tweet by looking at the associated words to the target. This semantic approach outperformed traditional machine learning-based models in empirical experiments with a test accuracy of 76% and F1 score of 75%. One notable constraint was the exclusion of threat targets with a frequency of less than five from ThrNet, as these targets might not have enough interconnections with threat words. As future work, expanding ThrNet’s scope to encompass additional target terms held potential to enhance the model’s effectiveness in detecting threats on Twitter.

Chapter 3

Definitions

The following definitions of the four main classes and their respective sub-classes are included to provide a clear understanding of their meanings, as detailed in the original paper [43]. Additionally, Table 3.1 is included to present examples for each of the sixteen classes which will illustrate how the annotators categorized and annotated the texts, ensuring a comprehensive and cohesive grasp of the annotation process, sourced on the original paper [43].

- **Religio-communal Violence:**
  Religio-communal violence refers to acts or statements deliberately aimed at inciting violence, typically targeting a specific religious group, often minorities, converts, or non-believers.

- **Ethno-communal Violence:**
  Ethno-communal violence refers to acts of violence directed against individuals or groups based on their ethnic or communal identity.

- **Nondenominational Communal Violence:**
  Nondenominational communal violence encompasses various forms of communal violence beyond those based solely on religious denominations.

- **Non Communal Violence (General Violence):**
  Non-communal violence encompasses any form of violence that does not fall into the categories of religio-communal, ethno-communal, or nondenominational communal violence. This category includes a wide range of violent incidents that occur outside of communal conflicts.

- **Derogation:**
  Derogation refers to various actions or behaviors aimed at belittling or demeaning individuals or groups.

- **Antipathy:**
  Antipathy involves various behaviors and attitudes characterized by a strong negative sentiment towards individuals or groups.

- **Prejudication:**
  Prejudication involves various actions or attitudes that support, justify, deny, or falsely accuse intentional or bona fide misdeeds, mistreatments, and discriminations.

- **Repression:**
  Repression involves various actions or expressions indicating an intent, willingness, or desire to cause harm to others.

Table 3.1: Overview of Violence Types and Expressions across our definitions

Communal Violence Class	**Violence Expression Class**	**Example (Bangla)**	**Example (English)**
Religio communal	Derogation	জেই ধর্মে সাধুরা উলঙ্গ হয়ে ঘোরে সেই ধর্মে মানুষেরা হিজাবের মর্মতা কি করে বুঝবে একটা উলঙ্গ ও নিক্রিস্টো ধর্ম হচ্ছে হিন্ধু ধর্ম রামের শিক্ষা আর কতো ভালো হবে ছি ছি	The religion (Hinduism) where the monks move about naked, what would its followers understand about Hijab?
	Antipathy	তাদের সময় গনিয়ে আসছে দেশ ছাড়ার সবকটাকে মালুরদেশ ভারত পটানো হবে আর কটাদিন সবুরকর	It's almost time! All the Malauns (Slur for Hindu folks) will be de- ported to India. You wait!
	Prejudication	লোকগুলো কে মেরে ফেলতে হবে কেনো!মাদ্রাসার বেশির ভাগ হুজুর সমকামী! এদেরকে তো মেরে ফেলার কথা কেউ বলে না!	Why must we kill these people? All the Hujurs (Islamic teachers) in madrasas are homosexual! Why is nobody asking to kill them?
	Repression	এটা বাংলাদেশ হলে গুলি করে মেরে ফেলত, বাংলাদেশে উগ্রবাদী হিন্দুদের সোজা কড়া উচিত, যাতে বাংলাদেশ ঝামেলা করতে না পরে।	If it was in Bangladesh, they would have been shot dead. It is imperative to straighten up the extremist Hindus in Bangladesh, so they can't cause any troubles/
Ethno communal	Derogation	এই রোহিঙ্গা এখন রাজনৈতিক. অর্থনৈতিক সামাজিক ও সংস্কৃতিক উভয় ক্ষত্রেই গলার হাড্ডি হয়ে দেখা দিয়েছে!! এবার বুঝবে এই রোহিঙ্গা কতো প্রিয় জিনিস!!	These Rohingya are now political, economic, moral, social and cultural burden. Now you'll understand how adored the Rohingyas are!
	Antipathy	রোহিঙ্গাদের দেশে রোহিঙ্গাদের পাঠানো হোক	Deport the Rohigyas back to their own country.
	Prejudication	এমন অনেক রুহিঙ্গা ও স্থানীয় ব্যবসায়ী ওখানে আছে যারা প্রতিনিয়ত বড় বড় ইয়াবা চালান আনে।	There are many Rohingyas and lo- cal traders who regularly bring large consignments of yaba (drugs).
	Repression	এই রোহিঙ্গার স্বীকারোক্তি নিয়ে যত দ্রত সম্ভব নাফ নদীতে চুবিয়ে মারেন।	Get a confession out of this Ro- hingya, then immediately drown him in Naaf river.

Continued on next page

Communal Violence Class	**Violence Expression Class**	**Example (Bangla)**	**Example (English)**
Non denominational communal	Derogation	নোয়াখাইল্লা বাটপারের পাল্লায় পরসিলাম	I fell into the clutches of a Noakhailla (region in Bangladesh) swindler
	Antipathy	সিলেটিভাষীদের দেশ থেকে বের করে দাওয়া হোক	Deport the Sylheti Speakers.
	Prejudication	ব্রাহ্মণবাড়িয়া লিখতে হবে!? ব্রাহ্মণবাড়িয়া শব্দটি হিন্দুদের, আর বি-বাড়িয়া শব্দটি হিন্ধু হিন্ধু লাগেনা হিম?!!	So I have to write BrahmanBaria (region in Bangladesh)? Is it because writing B.Bariya doesn't sound Hindu enough?
	Repression	হেট পায়ের মায়া থাকলে বাউনবাইরা লইয়া কিছু কইয়েন না	If you care about your limbs, don't try to slander BaunBaria (region in Bangladesh)
Non communal	Derogation	যারা হা হা রিয়াক্ট দিচ্ছে এদের জন্মে সমস্যা আছে	Those reacting 'Haha' have dubious paternity record.
	Antipathy	যারা সমাজে সমকামিতা প্রচার ও প্রসার কোর্টে চায়, সরকারইতো উচিত ছিল তাদের আইনের আয়তায় আনার।	Those who want to promote and spread homosexuality in the society, the government should have brought them under the law.
	Prejudication	পুলিশের তো এমন লোকেরই দরকার। কারণ তারই ইয়াবার বড় কাস্টমার।	The police need such people be- cause they are Yaba's (drug) big cus- tomers.
	Repression	ইসরাইল সন্ত্রাসী গোষ্ঠী আল্লাহ আপনি এদের ধ্বংস করে দিন।	Israel is terrorist group God destroy them.

Chapter 4

Dataset

4.1 Analysis of Dataset

The dataset [43] consists of comments sourced from social media, reflecting a diverse range of user-generated content . These textual entries have been categorized and annotated by domain experts to identify and classify instances of communal violence. The total dataset contains 12,791 rows and 5 columns. It is divided into three segments: training, validation, and test sets. The training set comprises 7673 rows and 5 columns, while the validation and test sets consist of 2559 rows and 5 columns individually.

Text	**religio-communal**	**ethno-communal**	**non-denominational-communal**	**noncommunal**
একে নিয়ে বন্দুকযুদ্ধে যাওয়া দরকার ছিল.... কিন্তু রেব তো এখন	0	0	0	4
রোহিঙ্গা জনগোষ্ঠী একদিন কাল হবে দেশের জন্য	0	3	0	0
মালায়নরা বাংলাদেশের ইলিশ মাছ খেয়ে সীমান্তে আমাদের হত্যা করে....	3	0	0	0
যাই হয় যাক কিন্তু এই দুর্নীতিবাজ দেশের জন্য যুদ্ধ করবোনা	0	0	3	1

Table 4.1: Short Overview of Dataset

The columns within the dataset include the primary text, four classes of violence: religio-communal, ethno-communal, nondenominational communal, and noncommunal. Specifically, the types of violence are categorized into four distinct sub-classes: 1 (derogation), 2 (antipathy), 3 (prejudication), and 4 (repression).

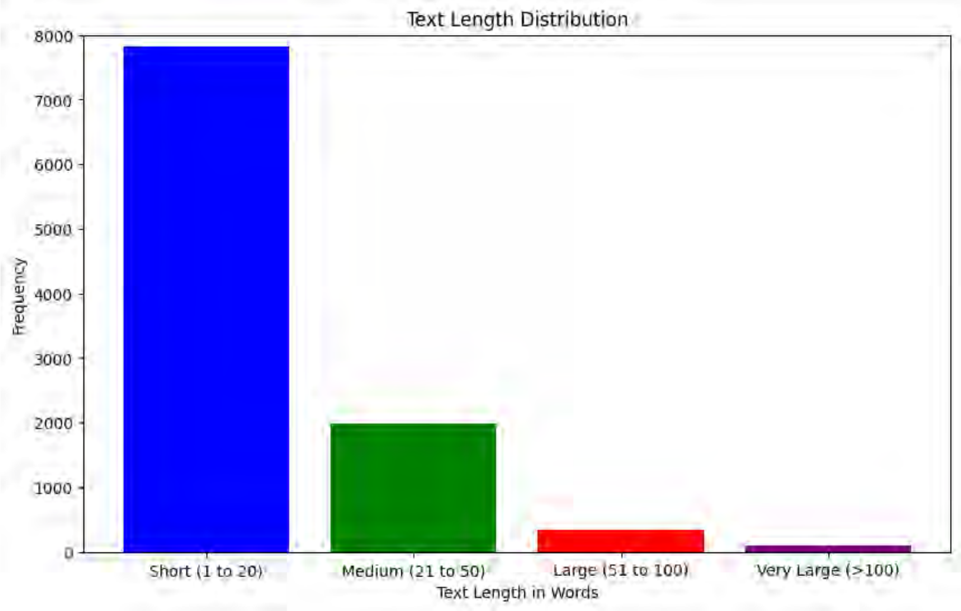


Figure 4.1: Text Length Distribution

As seen from Figure 4.1, the dataset primarily consists of over 7500 short texts that contain 1-20 words. An additional 1978 entries fall within the 21-50 words range. Notably, very few texts surpass 100 words. This distribution indicates a predominant emphasis on concise expressions in the dataset.

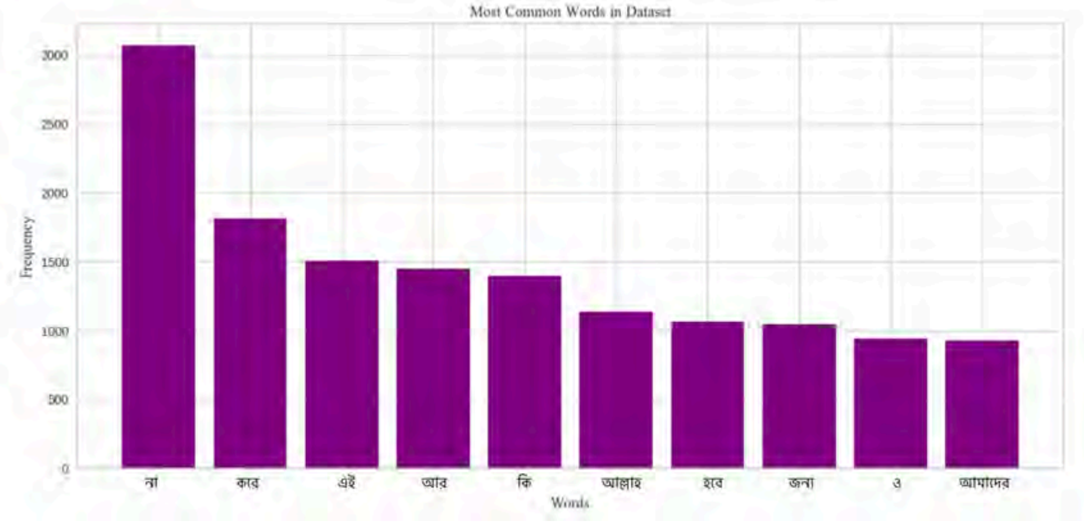


Figure 4.2: Most Frequent Words

From Figure 4.2, the most common word in the dataset is ‘না’ with a frequency above 3000. Other common words display significantly fewer occurrences. For instance, ‘এই’, ‘আর’ and ‘কি’ indicate the roughly equal frequencies distribution around 1500 and 1400. Also, words like ‘আল্লাহ’, ‘হবে’, ‘জন্য’, ‘ও’ and ‘আমাদের’ exhibit frequency distributions varying from approximately 1100 to 950. Some of these common words will be denoted as stop words that will be excluded for further analysis of the dataset.

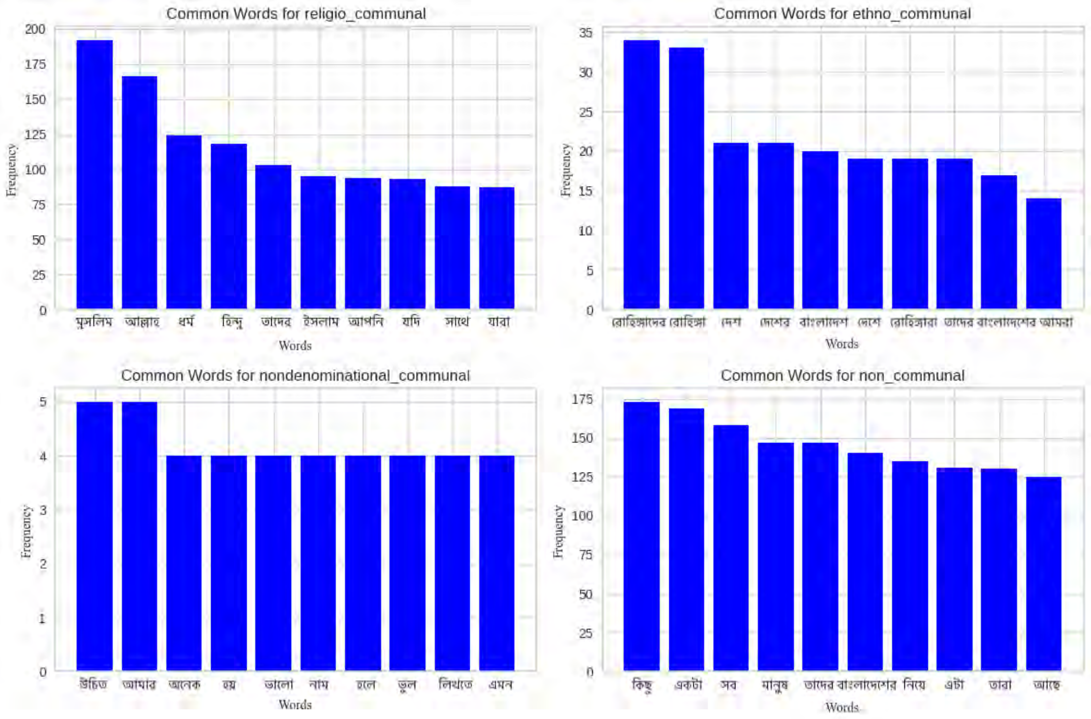


Figure 4.3: Common Words for Four Classes

Delving further into the word frequency analysis, common words for each of the classes of communal speech is shown in Figure 4.3. To make this analysis clearer, frequently occurring but less meaningful words like 'না', 'করে', and 'এই', which appear in all classes, were left out. This approach ensures a more focused and meaningful assessment of the distinct linguistic characteristics inherent to each class, thereby providing a clearer insight into the nuances of communal speech patterns. 'মুসলিম', 'আল্লাহ', 'হিন্দু', 'ইসলাম' are the unique common word for the 'religio-communal' class. These words serve as distinctive identifiers within the religious context. For example, 'হিন্দুরা নিশ্চিহ্ন হয়ে যাবে', 'ঠিক আছে আমি খুব খুশি মুসলিম হাটাও দেশ বাচাও জয় শ্রী রাম'. These lexical markers serve as linguistic cues signifying the ongoing discourse and escalating tensions between the Muslim and Hindu demographics within the community. Meanwhile, 'রোহিঙ্গাদের', 'রোহিঙ্গা', 'রোহিঙ্গারা' are the unique common words for the 'ethno-communal' class. For example, 'রোহিঙ্গা জনগোষ্ঠী একদিন কাল হবে দেশের জন্য', 'নতুন ইতিহাস করবে ওখানে রোহিঙ্গা গিয়ে....বাহ বাহ সুদির বাই', these expressions exemplifies a prevailing trend of 'ethno-communal' violent speech targeted at the Rohingya community. The wordcloud Figure 4.4 presents distinct words crucial to each class, excluding common non-relevant frequent words like stop words [10]. It effectively highlights unique and class-specific linguistic features. Furthermore, the majority of the texts and comments in the dataset are categorized as non-violent, constituting 65% of the total data. The remaining 35% is identified as violent. Within this subset of violent data, four classes of communal violent speech are present. It reveals that 'Non-communal' incidents account for a substantial 67.3%, indicating a significant majority of the dataset which does not exhibit explicit markers of communal violence. 'Religio-communal' violent speech makes up a significant 26.72% of the sample, highlighting the prevalence of religious-based communal occurrences. 'Ethno-communal' incidents are comparatively lower at 4.69%, suggesting a lesser prevalence of ethnic-linked violent speech. 'Nondenominational communal' incidents

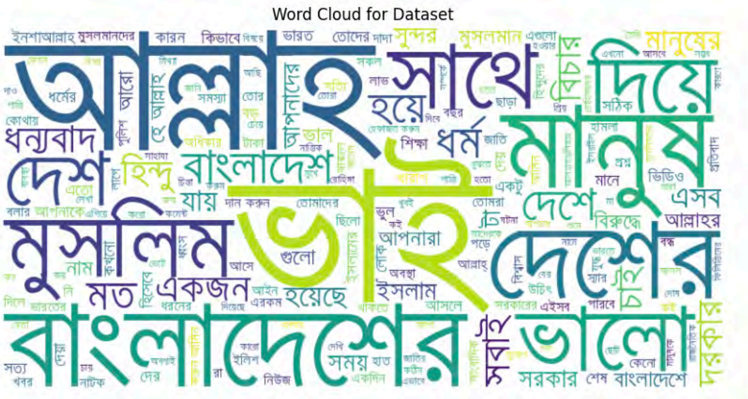


Figure 4.4: Word cloud

are rare, constituting only 1.2%. Also, looking at Fig4.5 each subclass distribution within the classes reveals distinctive patterns. Additionally, in the visual representation generated by the t-SNE graph, a total of 3,695 documents have been effectively categorized into 16 distinct classes. The t-SNE graph figure 4.6 demonstrates a significant level of dispersion, suggesting a diversified and potentially complicated dataset with various debate points and less prominent clustering around specific issues.

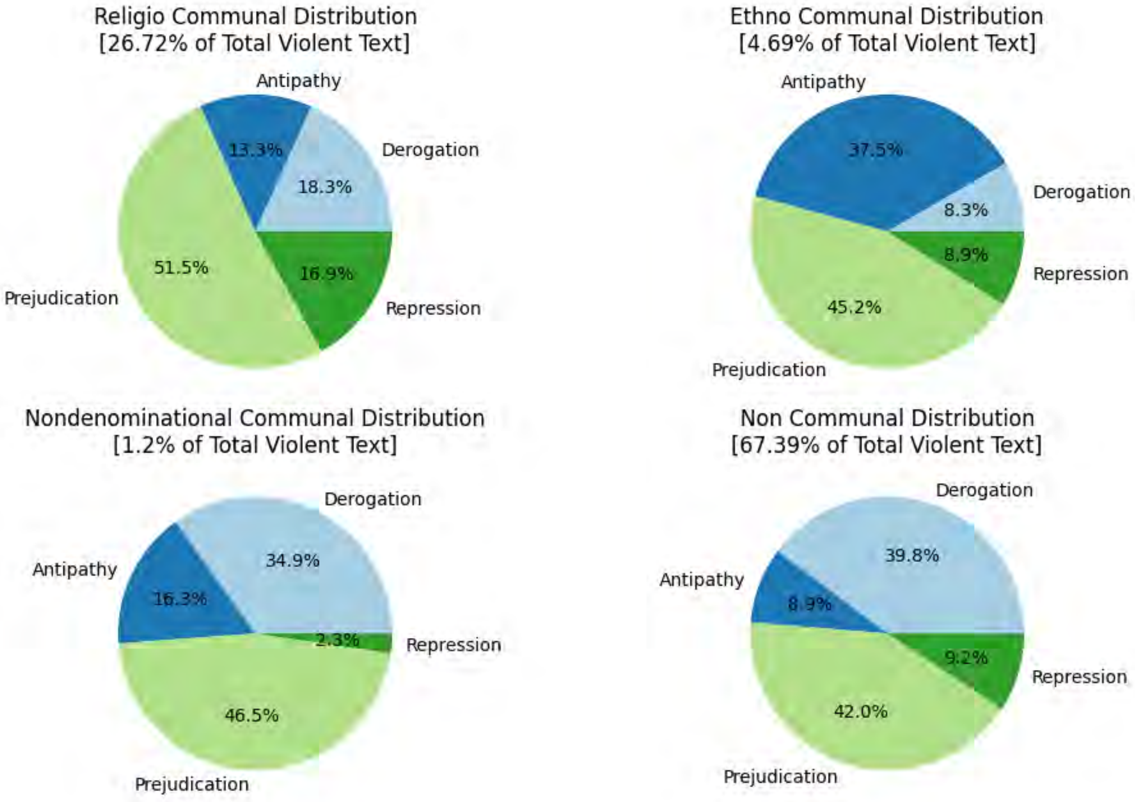


Figure 4.5: Distribution of Four Classes along with Four Sub-Classes

Dense clusters are observed for topics like “Ethno_Prejudication” and “Nondenominational_Repression,” suggesting a high degree of homogeneity in how these subjects are addressed. On the other hand, “Noncommunal_Antipathy” is more diffusely represented, indicating a wider diversity in the discussion of these topics. A notable intermingling of “Nondenominational_Prejudication” and “Noncommunal_Derogation” in the graph’s center suggests these themes frequently intersect within the documents. Outliers in “Religio_Repression” point to exceptional narratives that stand apart from the core discussion.

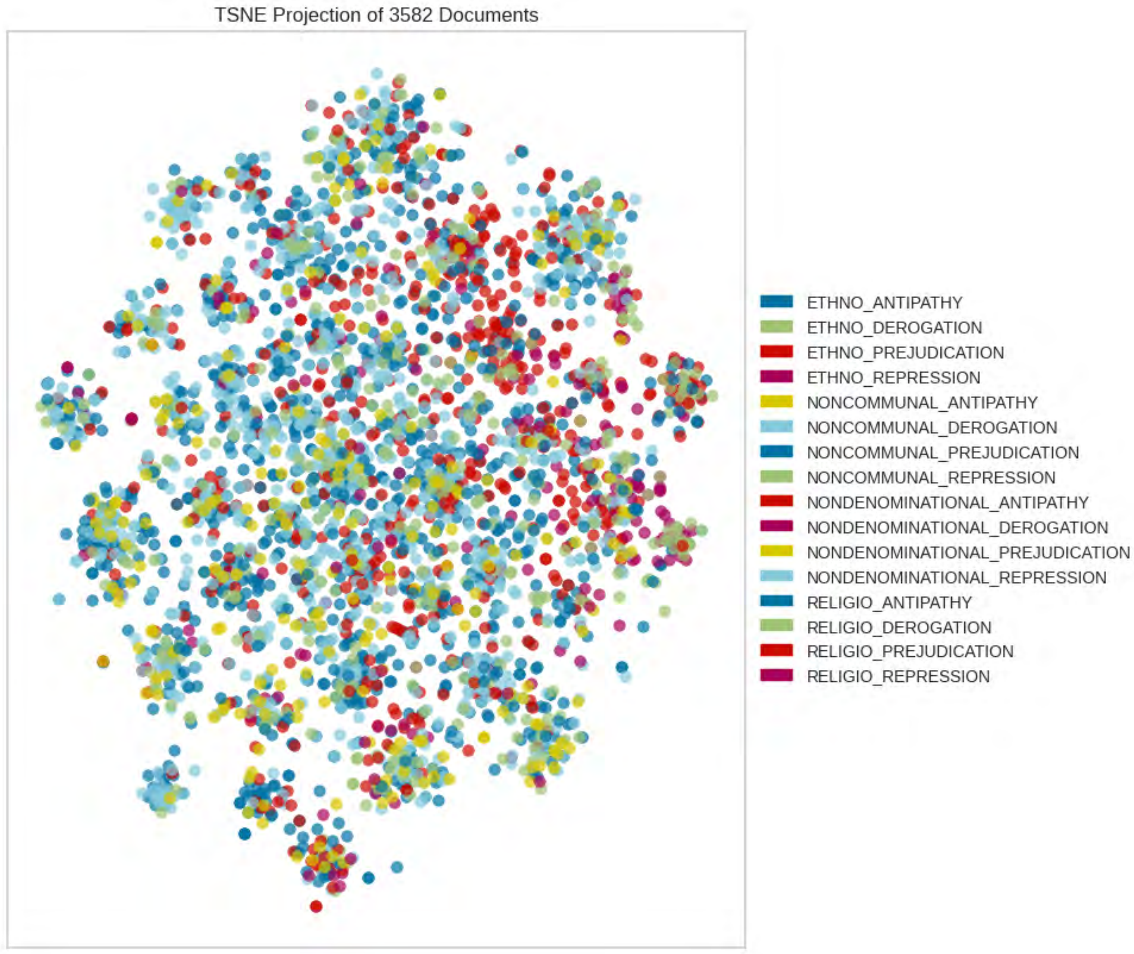


Figure 4.6: Scatter points for Sixteen Classes

4.2 Data Preprocessing

We implemented several preprocessing steps to enhance the quality of Bengali text-comments data. First, we systematically removed stop words [10] to filter out irrelevant terms and improve the overall clarity and relevance of the comments. Secondly, we addressed broken characters within the text, ensuring that all characters are correctly rendered and readable, thereby preserving the integrity of the linguistic content. After that, we cleansed the data of linguistic noise by eliminating punctuation, redundant spaces, and garbage characters, which are typically non-contributory in natural language processing contexts. We also addressed the presence of emojis in the data by replacing each emoji with its corresponding linguistic text meaning. This replacement was executed using ‘Emoji_Dict.p’ from Kaggle [9], which provided a comprehensive emoji-to-text mapping. For example, the ‘🙂’ emoji was substituted with ‘:pleading_face_emoji:’. We embarked on this process understand-

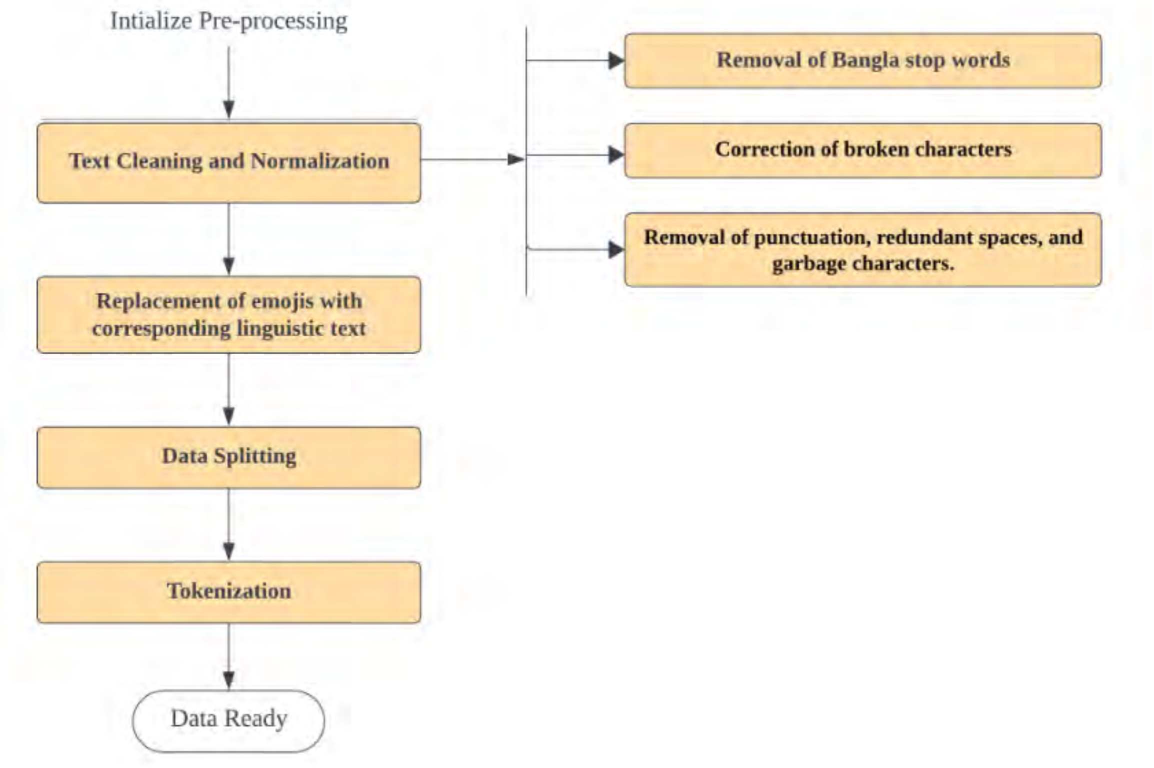

Intialize Pre-processing
Text Cleaning and Normalization
Removal of Bangla stop words
Correction of broken characters
Removal of punctuation, redundant spaces, and garbage characters.
Replacement of emojis with corresponding linguistic text
Data Splitting
Tokenization
Data Ready

Figure 4.7: Data preprocessing workflow

ing that emojis play a huge role in conveying sentiments and expressions. The task of translating emojis to text was to preserve the expressive nuances of the original text in a form convenient for NLP analysis. Then, we divided the training data into two distinct subsets: 85% was used for training and the remaining 15% was used for validation. This division helped to maintain a good balance between the training and the validation procedures. Lastly, we used the 'csebuetnlp/banglabert' [24] tokenizer, where the maximum length was 512 and padding was set to the maximum length. This specific tokenizer was selected because of its design that is optimized for the Bengali language, making it more efficiently in terms of dealing with the unique linguistic features associated with the Bengali language.

4.3 Data Augmentation

To reduce the initial imbalances in our dataset, we conducted a strategic augmentation process aimed at enriching classes with lower number of data. We tried to implement SMOTE, Zero-shot, Few-shot, text generation techniques, and a manual data augmentation approach.

4.3.1 SMOTE Approach

Synthetic Minority Over-sampling Technique(SMOTE), is a statistical technique used in the field of machine learning to address the problem of imbalanced class distributions in classification datasets [16]. However, the application of SMOTE shows unsatisfactory results. When it comes to word embeddings (High-dimensional vectors that contain the information about the meaning and the context of a word) pre-trained models like BanglaBERT or mBERT, using SMOTE may appear several limitations [38]. SMOTE, however, does not take these components into account and is built for numeric data. Additionally, SMOTE operates in the high-dimensional space of embeddings, and for the generated synthetic samples to be valid representations of words, they need to be semantically sensible in the high-dimensional space. In addition, text data is complex and represents relations like syntax and semantics that have nothing to do with simple interpolation methods used in developing SMOTE. This can potentially result in the destabilization of the pre-trained models as the synthetic samples might differ from the patterns observed and learned by the model.

4.3.2 Zero-Shot and Few-shot Approach

We examined two approaches- Zero-Shot and Few-Shot which were used to reduce the amount of data related to the Nondenominational-communal' and Ethno-communal' classes. In responding to the issue of data paucity amongst the 'Non-denominational communal' group, we first considered the Zero-Shot method, which functions upon the premise of transferring information from known to unknown classes [40]. This method is used for performing the model inference on the data classes which have not been encountered before to try to reduce the gap between the existing data classes and the novel data classes using the techniques like semantic embedding or generative adversarial networks. Even after our intervention there was no improvement in the performance of Zero-Shot. Thus, we used the Few-Shot approach, which entails augmenting a few labeled data instances to enable the effective training of models [42]. It chooses sample instances that best represent the entire data set to improve the generalization range of knowledge and improve model performance. In the context of the Few-Shot approach, we used the Z-map method to increase the volume of our dataset. For example, Few-shot is an attempt to achieve model capability to make correct predictions based on few examples and the Z-map technique is a method to map the features of new samples which are not included in the few-shot learning task to the feature space created from a few samples [15]. Nevertheless, the effectiveness of Few-Shot is also not satisfactory despite our attempts. The primary reason behind these shortcomings was found to be the mismatch in semantics between the source and the target data instances generated due to the highly contextual nature of the Bengali language. This discrepancy indicated the difficulties in successfully utilizing data augmentation methods within higher-level languages, including those characterized by significant language complexity, and the need to contemplate the role of linguistic nuances in such endeavors in the future.

4.3.3 Data Paraphasing

We made an effort to paraphrase data for the two minor classes, 'Ethno-communal' and 'Nondenominational communal,' utilizing the 'csebuetnlp/banglat5_bangla paraphrase' [23] model. The primary goal was to augment the data and enhance its variability. However, since the initial quantity of data for these classes was already quite low, the paraphrasing process did not provide substantial benefits. The limited amount of original data resulted in the model having insufficient material to generate varied and meaningful paraphrases. Consequently, the impact of this approach on improving our dataset for these specific classes was minimal. Thus, the scarcity of data remained a significant challenge, limiting the overall effectiveness of our paraphrasing efforts.

4.3.4 Manual Augmentation

We conducted a thorough collecting process by gathering comments from YouTube and Twitter that had the potential for violence. This entailed methodically combing through multiple posts and their associated comments on both platforms, identifying and selecting those that contained evidence of communal violence. Our goal was to capture a wide range of violent and hate speech, with a special emphasis on content that could spark religious, geopolitical, and other types of communal conflict. For that, in addition to our manually curated dataset, we sourced four notable Bengali hate speech datasets from reputable platforms like GitHub and Kaggle, amassing a total of 300,000 comments. These datasets include: ‘Bengali Hate Speech Detection Dataset by UCI’ [26], ‘Bengali Hate Speech Dataset by Nauros from Kaggle’ [14], ‘rezacsedu/Bengali-Hate-Speech-Dataset’ [11], ‘Multi-Labeled Bengali Toxic Comments’ [34]. These datasets were chosen for their direct relevance to our research on hate speech, ensuring a comprehensive and representative sample that aligns closely with our existing dataset. We sifted through the comments of these datasets by employing our baseline model, a fine-tuned Bangla BERT, to classify the texts as communally violent. This process helped us narrow down the comments. Then, these comments underwent another round of manual annotations. We assessed each comment individually, and if all raters agreed on its classification, it was included in our dataset through a blind voting system. To maintain accuracy and avoid fatigue, we shuffled through classes and limited the annotation of data to 50 comments per sitting. This approach enabled us to augment our dataset with an additional 1,794 entries, distributed as follows: 300 in the ‘Religio-communal’ class, 508 in the ‘Ethno-communal’ class, and 1,073 in the ‘Nondenominational communal’ class.

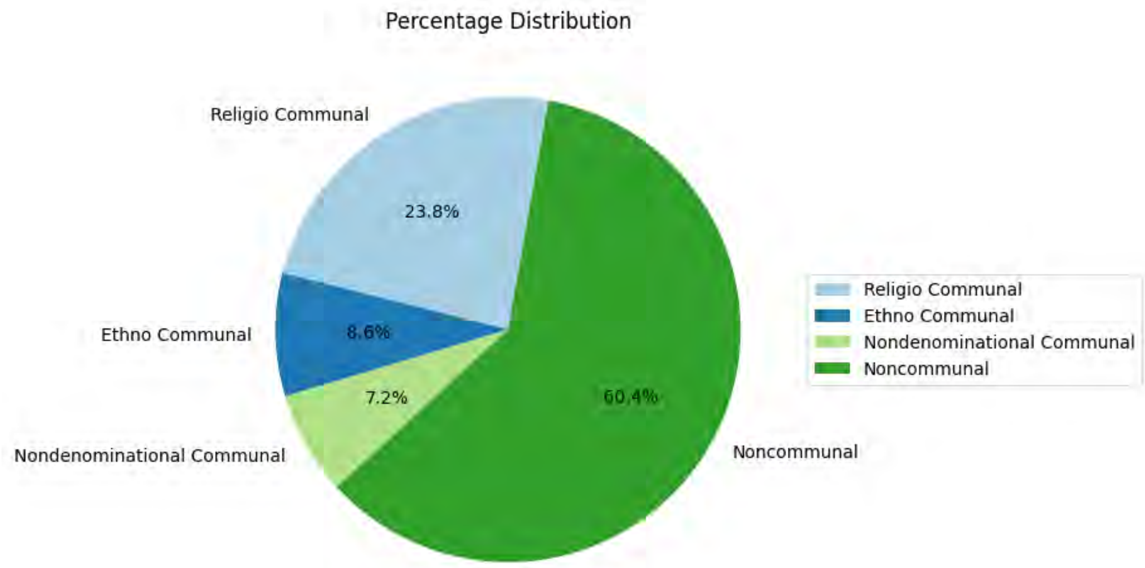


Figure 4.8: Class distribution after manual data augmentation

After this augmentation, the class distribution became more balanced (see Figure 4.8). Although 'Non-communal' and 'Religio-communal' classes remained dominant, there were notable increases in 'Ethno-communal' and 'Nondenominational communal' classes. Specifically, 'Ethno-communal' data rose from 4.69% to 8.6%, and 'Nondenominational communal' data increased from 1.2% to 7.2%, thus reducing the class imbalance.

Chapter 5

Model

5.1 Transformer

Transformers are a new style of sequence processing that overcomes the limitations of classic Sequential Recurrent Neural Networks. They replace recurrent connections with parallel execution and attention methods, allowing the model to focus on many parts of the input sequence at the same time. This captures complex relationships with context using multi-headed attention layers. The encoder generates numerous attention vectors per word, resulting in a weighted average that effectively captures the nuance communicated to the decoder and creates the path for following layers. During training, these attention vectors cover up the next word, allowing the model to predict and compare it to the actual result, improving its ability to recognise hidden trends in sequential data. The impact of BERT on sequence processing has established a new standard in natural language understanding.

5.2 Bidirectional Encoder Representations from Transformers-BERT

BERT [Figure 5.1] shares its foundation with the transformer model, employing a multilayer bidirectional transformer encoder, which has each output element connected to each input element and dynamically calculates weightings between them. This encoder consists of N layers, each having two sub-layers. The first sub-layer consists of a position wise fully connected feed forward network, whereas the second one includes a multi head-self-attention mechanism. Regarding the architecture, it utilizes self-attention on the encoder side and attention on the decoder side. BERT_base specifically features 12 layers in its encoder stack. Additionally, BERT_base has larger feedforward networks with 768 hidden units and more attention heads (12 in the case of BERT_base). These specifications are higher than the transformer architecture first proposed in the original paper that proposed 512 hidden units and 8 attention heads. In addition, BERT_base uses 110 million parameters, reflecting the model's power and ability to generate a subtle understanding of language compared to BERT_large, which increases these features substantially. The encoder stack of BERT_large is 24 layers deep as compared to BERT_base which is half that number. It also has larger feed forward network with 1024 hidden units and increased number of attention heads to 16. This improves configuration

leads to BERT_large to have 340 million parameters which significantly enhance its ability to provide complex language representation and understanding. Two stages comprise the BERT model's architecture: pre training and fine-tuning. In the pre-training stage a huge corpus of unannotated data is used to training the model such that the model is able to learn context representations. BERT's Masked Language Model (MLM) is a clever way for the model to learn about language. Suppose a game where some words are not visible and BERT has to guess them. A special layer is added for predicting the hidden words that can assist in tasks such as text classification in BERT. It then translates its guesses into a space of language, determining the plausibility of each word. BERT gets better by adjusting its guesses during training based on how much off it was from the genuine hidden words. Even though, this makes BERT sluggish in learning but the benefit pays off because the model learns to get good at understanding the context around words and BERT becomes commendable for tasks such as interpreting a particular category in a sentence. After pre-training, it fine-tunes its knowledge for specific tasks. When fine-tuning the probabilities or parameters of BERT, it becomes adaptable and uses its contextual wisdom to stand out in a variety of NLP tasks. BERT is a very powerful and flexible language model as pre-training provides a wide understanding and fine-tuning suits this knowledge for particular applications.
While BERT revolutionized worldwide language understanding, localized differences are critical. Introducing BanglaBERT, a linguistic revolution for Bengali. It bridges the gap between worldwide NLP advances and local language quirks, making it relevant for Bengali audiences and contributing to the reduction of global-local language inequalities [7].

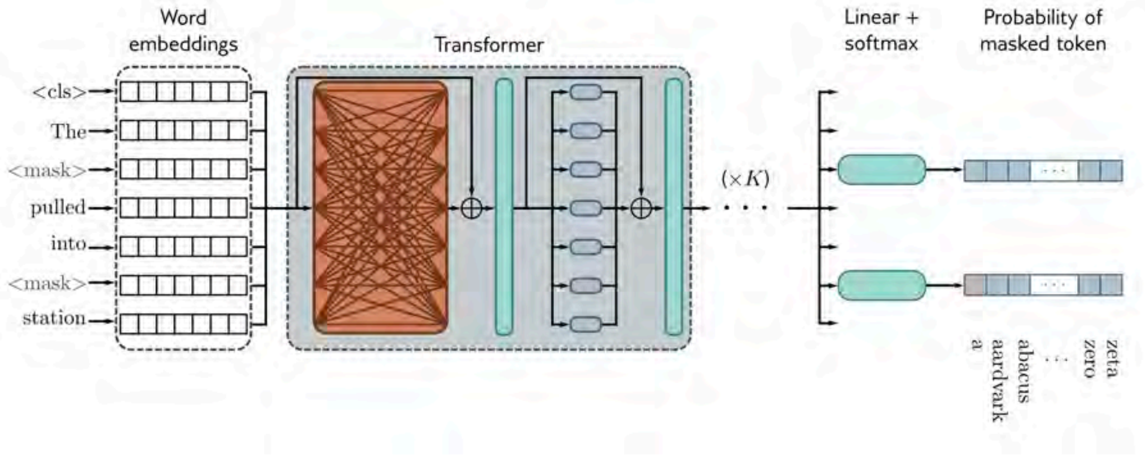
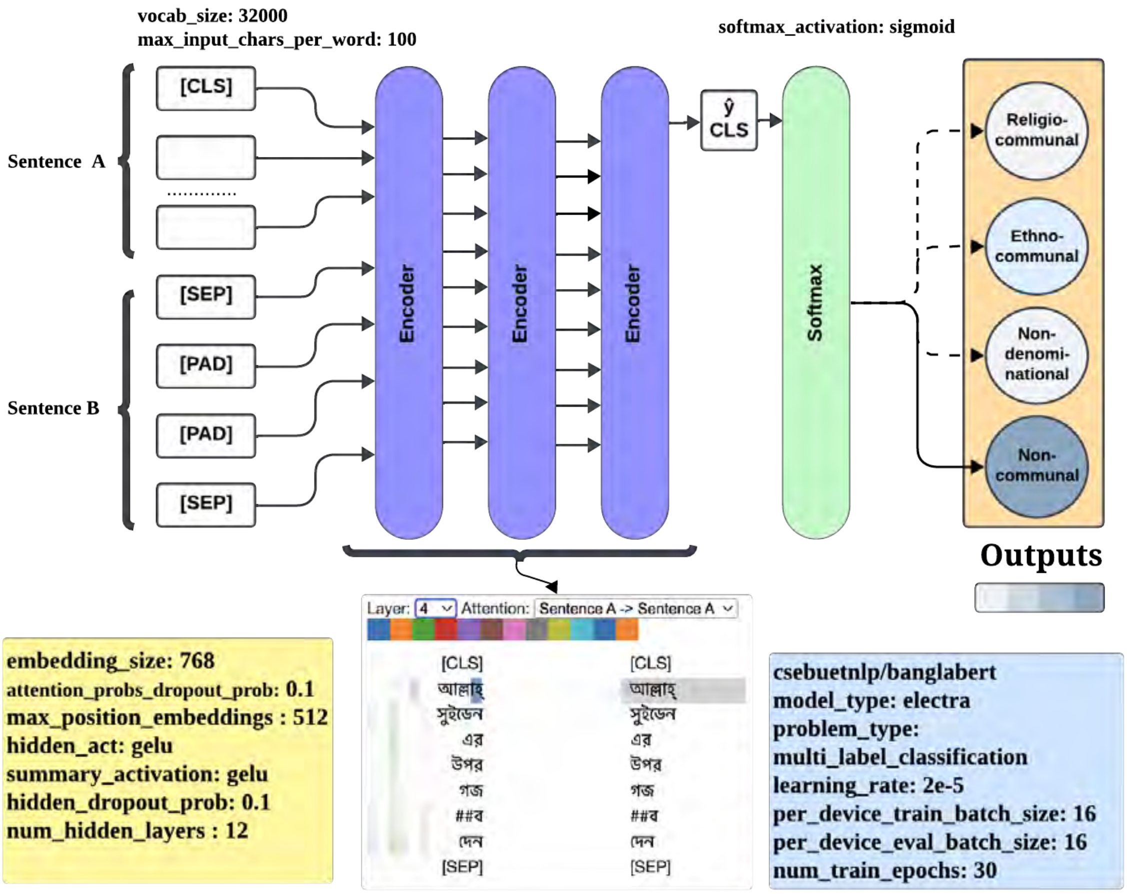


Figure 5.1: Overall structure of BERT model

5.2.1 Bangla-BERT

BanglaBERT [24], a specialized pre-trained language model tailored for the Bengali language, addresses the challenges posed by resource constraints in the field of NLP. Since Bangla is a low-resource language, a large 27.5 GB dataset was specifically gathered from 110 websites and preprocessed for pre-training. BanglaBERT makes use of a generator and discriminator model that are jointly trained together with the use of Leveraging ELECTRA pre-training with the Replaced Token Detection (RTD) setting. During pretraining, Some of the ones present in the input text are substituted with irrelevant or "masked" ones from the generator model. It aims to create dummies within which some fraction of the tokens are artificially substituted. While discriminator must then figure out must then anticipate whether or not each token is from the original sequence. Instead of considering only a 15% sequence slice, the RTD approach back-propagates the loss signals from all tokens in a sequence, providing the model with more signals to learn from. Under supervised fine-tuning, BanglaBERT gives excellent results, beating other models in tasks comprising sentiment classification on BLUB (Bangla Language Understanding Benchmark) instance with an obtained score of 72.89. Given these strengths in mind, BanglaBERT is selected as our model for detecting communal violence speech in the dataset.



vocab_size: 32000
max_input_chars_per_word: 100

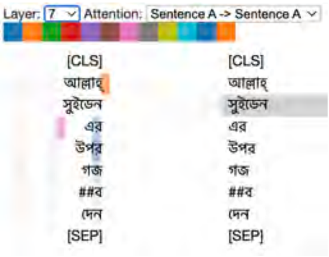
softmax_activation: sigmoid

embedding_size: 768
attention_probs_dropout_prob: 0.1
max_position_embeddings : 512
hidden_act: gelu
summary_activation: gelu
hidden_dropout_prob: 0.1
num_hidden_layers : 12

csebuetnlp/banglabert
model_type: electra
problem_type:
multi_label_classification
learning_rate: 2e-5
per_device_train_batch_size: 16
per_device_eval_batch_size: 16
num_train_epochs: 30

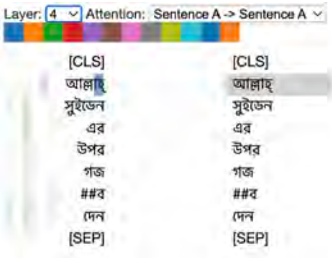
Figure 5.2: Graphical Representation of BanglaBERT Model

We are using the dataset that was initially used by [43] paper to create a baseline model by using BanglaBERT from csebuetnlp and we have polished it after fine-tuning it. Before using BanglaBERT in our model, we carefully followed a multi-step procedure to integrate it smoothly and improve sentence classification performance. Hence, initially we fine-tuned a pre-trained BanglaBERT base model with 12-layer hidden layers and a vocabulary size of 32,000. Subsequently, to further enhance our model's performance, we transitioned to fine-tuning the BanglaBERT



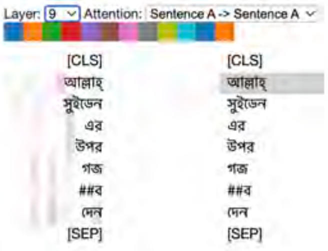
Layer: 7 Attention: Sentence A -> Sentence A

Figure 5.3: Attention to next token



Layer: 4 Attention: Sentence A -> Sentence A

Figure 5.4: Random attentions



Layer: 9 Attention: Sentence A -> Sentence A

Figure 5.5: Attention to previous token

large variant, which offers a more extensive architecture with increased parameters and deeper layers. We carefully partitioned the attorney dataset into 85% training and 15% validation for evaluation. For utilization, we optimized our input layer by tokenizing sentences into subword tokens with a maximum sequence length of 512, according to BanglaBERT's capabilities while using a standard learning rate of 2e-5. 30 epochs were used for training with a batch size of 16. We also delved into possible hyperparameter tunes like dropout rates, optimizer settings, and batch size. We also deployed visualizations to identify significant attention patterns across various layers. In Figure 5.4, Attentions in the case of layer 4 are focused on the same word 'আল্লাহ্' linked with token 'আল্লাহ্' and the same every token from layer 4 pays attention to the following word. On the other hand, in Figure 5.3, 'সুইডেন' is semantically focused on 'এর' and other tokens are all random attention, which implies no significant structure. Moreover, in Figure 5.5, the word 'আল্লাহ্' shows respectively high activation towards the same word 'আল্লাহ্' also the following other tokens, and most of the words shows the information of respectively high activation towards the previous word token. Furthermore, the mentioned approach included transfer learning, optimization of the input layer, and the possibility of hyperparameter tuning, which served a strong base for sentence classification. Also, hidden layers used GELU activation and summaries used GELU activation. Then, the final step provided a vector output of a single vector, whose activation was categorized with sigmoid into a specified category. We also applied an early stopping at 3, stopping the process when no progress was evident in the validation set for three consecutive iterations, thus avoiding overfitting. We have also employed the best-saved model strategy, saving several checkpoints and choosing the one that did the best job on a held-out validation set or evaluation metric.

5.2.2 Multilingual BERT

mBERT, or Multilingual BERT [8], is a significant advancement in natural language processing (NLP), particularly in the areas of multilingual understanding and generation. mBERT, developed as an extension of Google's groundbreaking BERT (Bidirectional Encoder Representations from Transformers) model, stands out for its ability to comprehend and generate text in multiple languages at the same time. Its working principle is built on a detailed structure characterized by several transformer layers with a fixed token length of 768 and a fixed number of self-attention heads. Typically, mBERT entails 12 transformer blocks with every block having 12 self-attention heads. Under this architecture, the self-attention mechanism reduces the computational cost of attention scores for the tokens in the input sequence while modeling the long- and short-range dependency between the input token context. After the self-attention process, feed-forward neural network layers process the gathered data, and thus, output representations for each token are created. The combined use of two fully connected layers in the feed-forward layers with Rectified Linear Unit (ReLU) activation functions enables mBERT to identify nuanced textual patterns in the input text. Most significantly, the implementation of mBERT transcends its technical and theoretical design principles to its provision of a multilingual training corpus. mBERT is thus capable of forming universal representations because of its exposure to multilingual data during training thus making it suitable for any given language. Considering the fact that mBERT is a multilingual model that can leverage similarities and differences in language it is a good choice for Bangla language training. Additionally, mBERT's multilingual nature helps to facilitate cross-lingual transfer learning, which refers to the process by which a model is trained in one language and then transferred to another for subsequent improvements in performance. This kind of cross-lingual transfer learning is of particular significance as far as low-resource languages like Bangla are concerned since annotated datasets for such languages are often comparatively scarce.

5.3 Ensemble Model- The Multi-Layer Perceptron (MLP) Classifier

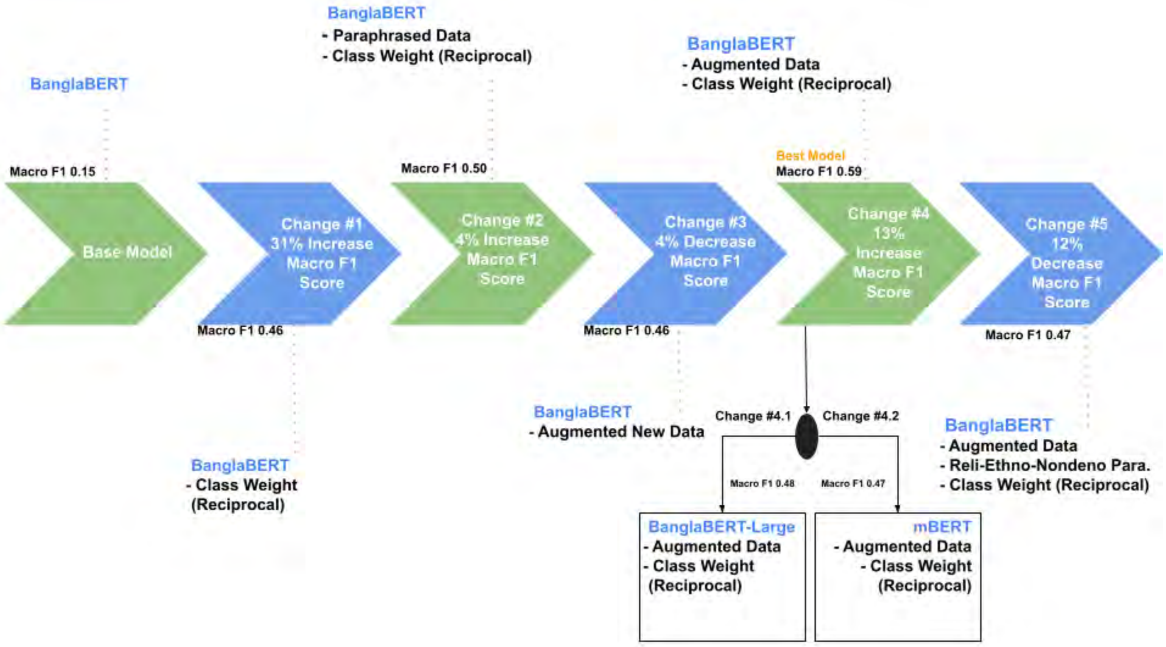
The MLP classifier is an advanced machine learning technique which is adept at capturing patterns and non-linear interactions in data [2]. It has a multi-layer architecture that enables the network to learn hierarchical features representation of the data and therefore does not require manual feature engineering. Unlike simple ensemble methods such as Mean Value Ensemble and Voting System Ensemble that perform prediction by simply averaging results from multiple models, MLP classifiers operate directly on the raw input data by forming hierarchical structures of artificial neurons. One of the most important advantages of using MLP classifiers instead of traditional ensemble methods is their possible application to virtually any type of learning task. MLPs offer parameters that can be tuned, and employ approaches like regularization and dropout to mitigate generalization and overfitting issues. MLP classifiers are of significance in dealing with the complexities of data since they can learn to adjust the weights and biases at every layer and hence enhance the accuracy of predictions. However, MLPs do not scale well and lack the ability to automatically choose parameters based on the dataset or relationship type, which is often necessary when using ensemble methods. Moreover, MLP classifiers can be trained using semi-supervised learning methods for using unlabeled examples, which makes them useful for a broad range of real-world applications.

Chapter 6

Result Analysis

In this research, we evaluated the performance of fine-tuned models across three distinct settings: one with four class, another with sixteen class and finally ensemble methods. For the four-class classification task, we conducted a more exhaustive study, training several models based on the hyperparameters tuned on an optimized set. This optimized set was determined through Bayesian optimization, which is capable of handling hyperparameters such as learning rate and batch size. We also employed class weights and data augmentation techniques. We developed a baseline model for sixteen class, however it could not be further analyzed due to scarcity and inconsistency of data in the subclasses. So, we employed our best fine-tuned models generated from the four-class setting to develop an ensemble model. This approach was used to combine the model with the well-optimized strengths in the hope of achieving overall performance enhancement.

6.1 Four Class Metrics



BanglaBERT
Macro F1 0.15
Base Model
BanglaBERT
- Class Weight (Reciprocal)
Macro F1 0.46
Change #1 31% Increase Macro F1 Score
BanglaBERT
- Paraphrased Data
- Class Weight (Reciprocal)
Macro F1 0.50
Change #2 4% Increase Macro F1 Score
BanglaBERT
- Augmented New Data
Macro F1 0.46
Change #3 4% Decrease Macro F1 Score
BanglaBERT
- Augmented Data
- Class Weight (Reciprocal)
Best Model
Macro F1 0.59
Change #4 13% Increase Macro F1 Score
Change #5 12% Decrease Macro F1 Score
Macro F1 0.47
BanglaBERT
- Augmented Data
- Reli-Ethno-Nondeno Para.
- Class Weight (Reciprocal)
Change #4.1
Change #4.2
Macro F1 0.48
Macro F1 0.47
BanglaBERT-Large
- Augmented Data
- Class Weight (Reciprocal)
mBERT
- Augmented Data
- Class Weight (Reciprocal)

Figure 6.1: Progression Flow for Four Class Metrics

Through Figure 6.1, we see that the dataset had a noticeably low proportion of ‘Ethno-communal’ and ‘Nondenominational’ data at initial stage, resulting in a significant imbalance. This huge skew in data distribution hampered the model’s ability to accurately distinguish texts from these specific classes, so it failed to detect any ‘Ethno-communal’ or ‘Nondenominational’ texts.

Class	Class weight	Learn Rate	Batch size	Epoch	Precision	Recall	F1	Macro F1
Religio-communal	No	2e-5	16	30	0.48	0.50	0.49	0.35
Ethno-communal					0.50	0.35	0.41	
Nondenominational					0.00	0.00	0.00	
Non-communal					0.56	0.48	0.52	

Table 6.1: CR of Bangla BERT (Baseline)

Then, we added reciprocal class weight so that the ‘Ethno-communal’ and ‘Nondenominational’ data which have lower amount of data gets more priority. The class weights were assigned with the following values, Religio-communal (2.78), Ethno-communal (7.49), Nondenominational (9.469), and Non-communal (1.09). After applying this we get a considerable boost in F1 score at Religio-communal and Nondenominational, but a decrease in the other two classes (see Table 6.2). Religio-communal saw an increase from 0.49 to 0.55, Nondenominational l improved from 0 to 0.16. But Ethno communal fell from 0.41 to 0.32, and Non-communal decreased from 0.52 to 0.44. This balanced adjustment ensured a more equitable training process, but while it improved the performance measures of some classes, it negatively affected the others.

Class	Class weight	Learn Rate	Batch size	Epoch	Precision	Recall	F1	Macro F1
Religio-communal	Yes	2e-5	32	30	0.47	0.65	0.55	0.37
Ethno-communal					0.21	0.72	0.32	
Nondenominational					0.11	0.32	0.16	
Non-communal					0.60	0.34	0.44	

Table 6.2: CR of Bangla BERT (Baseline)

To further examine the effect of adding class weights, we introduced paraphrased data using the ‘csebuetnlp/banglat5_banglaparaphrase’ [23] model. However, this addition did not yield favorable results (see Table 6.3). Compared to the previous model, the macro F1 score increased from 0.37 to 0.43. We observed a decline in Religio-communal, where all had a improved F1. Religio-communal saw a slight decreased from 0.55 to 0.53. The F1 score for Non-communal remained same, but Ethno-communal rise from 0.32 to 0.48, and Nondenominational increased from 0.16 to 0.29. Although there was a slight improvement in the macro F1 score, indicating a generally better model performance, there was variation in the performance for each class. The slight reduction in the Religio-communal may may have become overly generalized after integrating paraphrased data, while the remarkable improvement in Ethno-communal and Nondenominational indicate how data augmentation helps in improving the performance of the model. A consistent F1 score for Non-communal shows that the paraphrasing of the data did not hamper its performance for this class. Thus, these changes contributed to the delineation of the model's performance across the categories.

Class	Class weight	Learn Rate	Batch size	Epoch	Precision	Recall	F1	Macro F1
Religio-communal	Yes	2e-5	32	30	0.45	0.66	0.53	0.43
Ethno-communal					0.39	0.60	0.48	
Nondenominational					0.23	0.37	0.29	
Non-communal					0.61	0.35	0.44	

Table 6.3: CR of Bangla BERT (Added Paraphrased data)

After data augmentation, particularly on 'Ethno-communal' and 'Nondenominational communal' texts, the model showed significant improvement in classifying those classes, enhancing the F1 score of Ethno-communal from 0.48 to 0.56 and Nondenominational from 0.29 to 0.61 compared to paraphrased data. (see Table 6.4).

Class	Class weight	Learn rate	Batch size	Epoch	Precision	Recall	F1	Macro F1
Religio-communal	No	2e-5	32	30	0.55	0.39	0.46	0.53
Ethno-communal					0.46	0.72	0.56	
Nondenominational					0.51	0.76	0.61	
Non-communal					0.62	0.38	0.47	

Table 6.4: CR of Bangla BERT (with Augmented Data)

Then, to address the class imbalance that still existed after data augmentation, we assigned class weights to the model based on the inverse frequency of each class. Specifically, weights were determined as follows: Religio-communal (2.78), Ethno-communal (7.49), Nondenominational (9.469), and Non-communal (1.09). This weighting scheme was designed to amplify the influence of underrepresented classes— 'Religio-communal', 'Ethno-communal', and 'Nondenominational communal' —due to their relatively lower prevalence compared to the Non-communal class. After implementing these class weights, we observed a significant improvement in the F1 scores across all classes (see Table 6.5). The F1 score for Religio-communal increased from 0.46 to 0.47, Ethno-communal rose from 0.56 to 0.66, Nondenominational F1 score jumped from 0.61 to 0.69 and Non-communal exhibited a increase from 0.47 to 0.52. This indicates the enhanced model sensitivity and precision for these minority classes. This model achieved the highest Macro F1 score.

Class	Class weight	Learn rate	Batch size	Epoch	Precision	Recall	F1	Macro F1
Religio-communal	Yes	2e-5	32	30	0.51	0.43	0.47	0.60
Ethno-communal					0.61	0.72	0.66	
Nondenominational					0.86	0.58	0.69	
Non-communal					0.54	0.60	0.57	

Table 6.5: CR of Bangla BERT (with Augmented Data)

Subsequent analysis revealed that the model frequently misclassified instances between the ‘Religio-communal’ and ‘Non-communal’ classes, with ‘Religio-communal’ being misidentified as ‘Non-communal’ and vice versa. To mitigate this issue, we undertook the task of paraphrasing the texts from all the classes except Non-communal. We refrained from adding Non-communal paraphrased texts since this class contained a majority of the data. The objective was to enhance the distinctiveness of the Religio-communal data, thereby improving the model’s ability to differentiate it from Non-communal instances. Surprisingly, the results were the opposite of what we expected. The F1 scores decreased for all classes except Ethno-communal (see Table 6.6). Specifically, the F1 score for Religio-communal dropped from 0.52 to 0.46, Nondenominational fell from 0.72 to 0.53, and Non-communal decreased from 0.53 to 0.34. This suggests that paraphrasing did not enhance the distinctiveness as intended and may have introduced more noise into the dataset, confusing the model further and leading to poorer performance overall. This could be due to the fact that the application of paraphrasing techniques for the other classes might have unintentionally generalized the vocabulary within the Non-communal class. This can happen when distinct keywords indicative of the class are replaced with more generic terms that are common across different classes. This loss of unique vocabulary makes it harder for the model to identify instances belonging to the Non-communal class. Additionally, the augmentation of paraphrased data to other classes caused Non-communal class to have lesser class weight, which can be another reason for the drop in the F1 score.

Class	Class weight	Learn rate	Batch size	Epoch	Precision	Recall	F1	Macro F1
Religio-communal	Yes	2e-5	32	30	0.54	0.41	0.46	0.49
Ethno-communal					0.55	0.77	0.64	
Nondenominational					0.38	0.85	0.53	
Non-communal					0.68	0.22	0.34	

Table 6.6: CR of Bangla BERT (with Augmented Data and Paraphrased Data)

While the multi-class paraphrasing strategy did not prove beneficial for ‘Ethno-communal’ and ‘Nondenominational communal’, it emphasizes the need for further exploration and potentially a more nuanced approach to improve ‘Religio-communal’ classification without compromising the performance of other classes. In response to this insight, we decided to pivot our strategy by implementing BanglaBERT Large. This decision was based on our observation that previous iterations, combining augmented data with class weights, resulted in the highest macro F1 scores. Comparing our latest results with a previous approach where only class weights were adjusted

Table 6.5, we observed a decrease in F1 scores for all classes (see Table 6.7). However, there was a noticeable decrease in scores for Religio-communal from 0.52 to 0.49, Nondenominational from 0.72 to 0.61, and Non-communal from 0.52 to 0.35. But the score of Ethno-communal was relatively constant. This suggests persistent challenges in accurately discerning between "Religio-communal" and "Non-Communal" instances, highlighting an ongoing area for improvement.

Class	Class weight	Learn rate	Batch size	Epoch	Precision	Recall	F1	Macro F1
Religio-communal	Yes	2e-5	16	30	0.42	0.60	0.49	0.51
Ethno-communal					0.48	0.77	0.59	
Nondenominational					0.46	0.89	0.61	
Non-communal					0.73	0.23	0.35	

Table 6.7: CR of BanglaBERT Large (with Augmented Data)

In our exploration, we turned to mBERT as it was trained in multiple languages simultaneously, encapsulating knowledge from various linguistic sources. Given our consistent success with augmented data and class weights for optimizing results, we applied these parameters to our implementation of mBERT. When we compared it with our best model Table 6.5, we noticed some changes in the F1 scores (see table 6.8). For 'Religio-communal' and 'Nondenominational communal', the scores dropped by 0.10 for both. Conversely, the F1 score for Non-communal dropped from 0.52 to 0.39, while Ethno-communal experienced a rise, increasing from 0.58 to 0.61. These changes indicate an ongoing conflict between the Religio-communal and Non-communal classes, which results in a lower macro F1 score for this model.

Class	Class weight	Learn rate	Batch size	Epoch	Precision	Recall	F1	Macro F1
Religio-communal	Yes	2e-5	32	30	0.30	0.72	0.42	0.52
Ethno-communal					0.51	0.76	0.61	
Nondenominational					0.70	0.64	0.67	
Non-communal					0.48	0.33	0.39	

Table 6.8: CR of mBERT (with Augmented Data)

This summary Table 6.9 presents the results of various fine-tuned models, including multiple configurations of Bangla Bert, BanglaBERT Large, and mBert. All models were trained with a consistent **learning rate of 2e-5** and over **30 epochs** with **Early stopping** including **patience size of 2**. Additionally, we chose a **threshold of 0.5** because we analyzed the classification report for thresholds ranging from 0.1 to 0.9 and observed that all of the fine-tuned models showed significantly better results at the threshold of 0.5.

Model Name	**Class Weight**	**Batch Size**	**F1**				**Macro F1**
			RC	EC	ND	NC	
Bangla Bert (Baseline)	No	16	0.49	0.41	0.00	0.52	0.35
Bangla Bert (Baseline)	Yes	32	0.55	0.32	0.16	0.44	0.37
Bangla Bert (Pr)	Yes	32	0.53	0.48	0.29	0.44	0.43
Bangla Bert (AD)	No	32	0.46	0.56	0.61	0.47	0.53
Bangla Bert (AD)	Yes	32	0.47	0.66	0.69	0.57	0.60
Bangla Bert (AD + Pr)	Yes	32	0.46	0.64	0.53	0.34	0.49
BanglaBERT Large (AD)	Yes	16	0.49	0.59	0.61	0.35	0.51
mBert (AD)	Yes	32	0.42	0.61	0.67	0.39	0.52

Acronym Meaning: **RC** - Religio Communal, **EC** - Ethno Communal, **ND** - nondenominational Communal, **NC** - Non Communal, **AD** - Augmented Data, **Pr** - Paraphrased Data.

Table 6.9: FTMs Performance Summary Across Four Classes

6.2 Sixteen Class Metrics

The presented Table 6.10 outlines the precision, recall, and macro F1 scores for a baseline model across sixteen class. Notably, classes such as “Religio Repression” and “Non communal Derogation” demonstrate higher precision, recall, and macro F1 scores, suggesting that the model performs relatively well in distinguishing and capturing instances within these classes. Specifically, “Religio Repression” stands out with a precision of 0.52, recall of 0.36, and a macro F1 score of 0.43, indicating robust performance in identifying and correctly classifying instances associated with this class. On the other hand, classes like “Religio Antipathy,” “Ethno Derogation,” “Ethno Antipathy,” and several others exhibit zeros across precision, recall, and macro F1, signifying challenges in correctly classifying instances within these classes.

Class	**Derogation (F1 Score)**	**Antipathy (F1 Score)**	**Prejudication (F1 Score)**	**Repression (F1 Score)**
Religio Communal	0.11	0.00	0.24	0.43
Ethno Communal	0.00	0.00	0.00	0.00
Nondenominational	0.00	0.00	0.00	0.00
Non-Communal	0.36	0.14	0.23	0.32

Table 6.10: Classification Metrics for Sixteen class

For the sixteen class setup, the dataset has a significant class imbalance, much greater than that of the four-class setup. So we applied class weights to address the issue. However, assigning weights to these classes led to poorer overall model performance. The weights, calculated to balance the dataset, ranged from relatively moderate to extremely high values, especially for underrepresented classes with very few instances. For example, classes like 'Ethno_Repression', with only 12 instances, and 'Nondenominational_Repression', with just 1 instance, had weights as high as 181.19 and 2174.25 respectively. This extreme imbalance caused the model to become unstable and less accurate, resulting in an overall poorer performance compared to before adding class weights (see table 6.11). Consequently, while the intention was to mitigate class imbalance, the excessive weights adversely affected the model's performance, highlighting the need for alternative approaches to handle such severe imbalances effectively. In this case, the data is simply not enough to identify the classes accurately.

Class	**Derogation (F1 Score)**	**Antipathy (F1 Score)**	**Prejudication (F1 Score)**	**Repression (F1 Score)**
Religio Communal	0.05	0.00	0.06	0.19
Ethno Communal	0.00	0.10	0.12	0.10
Nondenominational	0.00	0.00	0.00	0.06
Non-Communal	0.02	0.14	0.00	0.27

Table 6.11: Classification Metrics for Sixteen class (With Class Weights)

6.3 Ensemble model

We employed an ensemble approach to enhance the classification performance for communal violence classification tasks using multiple fine-tuned BERT models. By integrating predictions from different models, ensemble techniques can improve the performance and minimize the error of individual models, leading to more robust and accurate results. Specifically, we explored three types of ensemble techniques: the mean value approach, the voting system, and the MLP classifier.

We selected our seven best-performing individual models, focusing on both individual class performance and overall performance. However, the selected models are:

1. BanglaBERT Large (Augmented Data with Class Weights)
2. mBERT (Augmented Data with Class Weights)
3. Bangla BERT (Added Class Weights)
4. Bangla BERT (Paraphrased data with Class Weights)
5. Bangla BERT (Augmented Data and Paraphrased data with Class Weights)
6. Bangla BERT (Augmented Data with Class Weights)
7. Bangla BERT (Augmented Data without Class Weights)

In our research, we implemented an ensemble model which consists of five different fine-tuned models. We used four fine tune models: Bangla BERT (Added Class Weights), Bangla BERT (Paraphrased data with Class Weights), Bangla BERT (Augmented Data and Paraphrased data with Class Weights), Bangla BERT (Augmented Data without Class Weights). For the fifth model in the ensemble, we switched between three different models: BanglaBERT Large (Augmented Data with Class Weights), mBERT (Augmented Data with Class Weights), Bangla BERT (Augmented Data with Class Weights).

6.3.1 Mean Value

The mean value ensemble technique involves averaging the predictions from multiple fine-tuned BERT models to generate a final prediction. This approach helps to smooth out the variances and biases inherent in individual models, leading to more stable and reliable predictions.

FTMs	**Class**	**Precision**	**Recall**	**F1**	**Macro F1**
One FTM from BBL, Four FTMs from BB	Religio-communal	0.54	0.57	0.56	0.61
	Ethno-communal	0.60	0.74	0.66	
	Nondenominational	0.75	0.82	0.78	
	Non-communal	0.72	0.33	0.46	
One FTM from mBERT, Four FTMs from BB	Religio-communal	0.53	0.62	0.57	0.62
	Ethno-communal	0.60	0.74	0.67	
	Nondenominational	0.88	0.71	0.78	
	Non-communal	0.68	0.36	0.47	
Five FTMs from BB	Religio-communal	0.54	0.59	0.56	0.63
	Ethno-communal	0.60	0.74	0.66	
	Nondenominational	0.78	0.75	0.77	
	Non-communal	0.67	0.41	0.51	

Table 6.12: Mean Value Ensemble Models Performance Summary

In the Table 6.12 Mean Value Ensemble Models Performance table, the precision values for the Five Fine-tuned models on BanglaBERT range from 0.54 to 0.78, which are higher than those for BanglaBERT Large (0.54 to 0.72) and comparable to mBERT (0.53 to 0.88). This indicates a slight increase in precision for the Fine-tuned BanglaBERT models. The recall values for these fine-tuned models (0.41 to 0.75) are comparable to the ranges for mBERT (0.36 to 0.74) and BanglaBERT Large (0.33 to 0.82), with significant improvements in the "Ethno-communal" class. The F1 scores of the Fine-tuned BanglaBERT models (0.51 to 0.77) are competitive when compared to mBERT (0.47 to 0.78) and BanglaBERT Large (0.46 to 0.78), with the "Nondenominational" class receiving a high F1 score of 0.78. The ensemble of Five Fine-tuned models on BanglaBERT achieves a macro F1 score of 0.63, which is slightly higher than that of mBERT (0.62) and BanglaBERT Large (0.61), indicating improved overall performance.

6.3.2 Voting system

The voting system in ensemble technique is often considered more reliable than using the mean value because it incorporates the concept of majority rule. In a voting system, predictions are made based on the most common outcome among all the models in the ensemble. Since this approach can be more robust because it reduces the impact of any single model's bias or variance on the final prediction, we used it to achieve more precise results.

FTMs	Class	Precision	Recall	F1	Macro F1
One FTM from BBL, Four FTMs from BB	Religio-communal	0.51	0.58	0.55	0.61
	Ethno-communal	0.60	0.74	0.66	
	Nondenominational	0.72	0.81	0.76	
	Non-communal	0.72	0.36	0.48	
One FTM from mBERT, Four FTMs from BB	Religio-communal	0.50	0.62	0.56	0.62
	Ethno-communal	0.61	0.76	0.67	
	Nondenominational	0.82	0.69	0.75	
	Non-communal	0.65	0.39	0.48	
Five FTMs from BB	Religio-communal	0.52	0.62	0.57	0.63
	Ethno-communal	0.60	0.74	0.67	
	Nondenominational	0.76	0.75	0.77	
	Non-communal	0.64	0.43	0.51	

Table 6.13: Voting System Ensemble Models Performance Summary

In the Voting System Ensemble Models Performance Table 6.13, the precision values for the Five Fine-tuned models on BanglaBERT range from 0.52 to 0.76, which are comparable to mBERT (0.50 to 0.82) and higher than BanglaBERT Large (0.51 to 0.72), indicating a consistent level of precision. The recall values for these fine-tuned models (0.43 to 0.75) are similar to mBERT (0.39 to 0.76) but slightly lower than BanglaBERT Large (0.36 to 0.81), reflecting balanced performance across classes. The F1 scores for the Fine-tuned BanglaBERT models (0.51 to 0.76) are comparable to mBERT (0.48 to 0.75) and BanglaBERT Large (0.48 to 0.76), with the “Non-denominational” class receiving a particularly high score of 0.76. The ensemble of Five Fine-tuned models on BanglaBERT achieves a macro F1 score of 0.63, which is slightly higher than that of mBERT (0.62) and BanglaBERT Large (0.61).

6.3.3 MLP Classifier

The Multilayer Perceptron (MLP) classifier can be more reliable than a voting system or mean value in an ensemble due to its ability to learn complex patterns through its network of neurons and multiple layers. So, we used it to achieve more precise results compared to the mean value and voting system.

FTMs	Class	Precision	Recall	F1	Macro F1
One FTM from BBL, Four FTMs from BB	Religio-communal	0.50	0.42	0.46	0.60
	Ethno-communal	0.68	0.67	0.68	
	Nondenominational	0.78	0.71	0.74	
	Non-communal	0.54	0.48	0.51	
One FTM from mBERT, Four FTMs from BB	Religio-communal	0.51	0.41	0.45	0.60
	Ethno-communal	0.70	0.72	0.71	
	Nondenominational	0.84	0.67	0.74	
	Non-communal	0.55	0.47	0.51	
Five FTMs from BB	Religio-communal	0.53	0.42	0.47	0.61
	Ethno-communal	0.66	0.73	0.69	
	Nondenominational	0.82	0.68	0.74	
	Non-communal	0.56	0.49	0.52	

Table 6.14: MLP Classifier Ensemble Models Performance Summary

In the MLP Classifier Ensemble Models Performance Table 6.14, the precision values for the Five Fine-tuned models on BanglaBERT range from 0.53 to 0.82, showing improvements compared to BanglaBERT Large (0.50 to 0.78) and mBERT (0.51 to

0.84). Recall values for these fine-tuned models range from 0.42 to 0.73, which are comparable to BanglaBERT Large (0.43 to 0.71) and mBERT (0.41 to 0.72). The macro F1 score of this ensemble is 0.61, indicating a balanced performance across all classes and an improvement over previous models. These findings suggest that the Fine-tuned BanglaBERT ensemble enhances overall precision and F1 scores, particularly in maintaining balanced performance across classes.

Ensemble Techniques	**FTMs**	**Class**	**F1 Score**	**Macro F1 Score**
Mean Value	One FTM from BBL, Four FTMs from BB	Religio-communal	0.56	0.61
		Ethno-communal	0.66	
		Nondenominational	0.78	
		Non-communal	0.46	
	One FTM from mBert, Four FTMs from BB	Religio-communal	0.57	0.62
		Ethno-communal	0.67	
		Nondenominational	0.78	
		Non-communal	0.47	
	Five FTMs from BB	Religio-communal	0.56	0.63
		Ethno-communal	0.66	
		Nondenominational	0.77	
		Non-communal	0.51	
Max Voting	One FTM from BBL, Four FTMs from BB	Religio-communal	0.55	0.61
		Ethno-communal	0.66	
		Nondenominational	0.76	
		Non-communal	0.48	
	One FTM from mBert, Four FTMs from BB	Religio-communal	0.56	0.62
		Ethno-communal	0.67	
		Nondenominational	0.75	
		Non-communal	0.48	
	Five FTMs from BB	Religio-communal	0.57	0.63
		Ethno-communal	0.67	
		Nondenominational	0.77	
		Non-communal	0.51	
MLP Classifier	One FTM from BBL, Four FTMs from BB	Religio-communal	0.46	0.60
		Ethno-communal	0.68	
		Nondenominational	0.74	
		Non-communal	0.51	
	One FTM from mBert, Four FTMs from BB	Religio-communal	0.45	0.60
		Ethno-communal	0.71	
		Nondenominational	0.74	
		Non-communal	0.51	
	Five FTMs from BB	Religio-communal	0.47	0.61
		Ethno-communal	0.69	
		Nondenominational	0.74	
		Non-communal	0.52	

Table 6.15: Ensemble Models Performance Summary

In our individual model, we encountered a difficult conflict between “religio communal” and “non communal” data, an anomaly that frequently impacts accuracy in prediction. However, through the process of ensemble learning, we’ve been able to considerably decrease this conflict, though it remains an ongoing challenge in our ongoing research for accurate prediction. Among the ensemble techniques used, the voting ensemble method has proven particularly effective (see Table 6.15), applying the collective learning of five fine-tuned models built on the BanglaBERT. This ensemble approach has produced the best confusion matrix results (see Figure 6.2) so far.

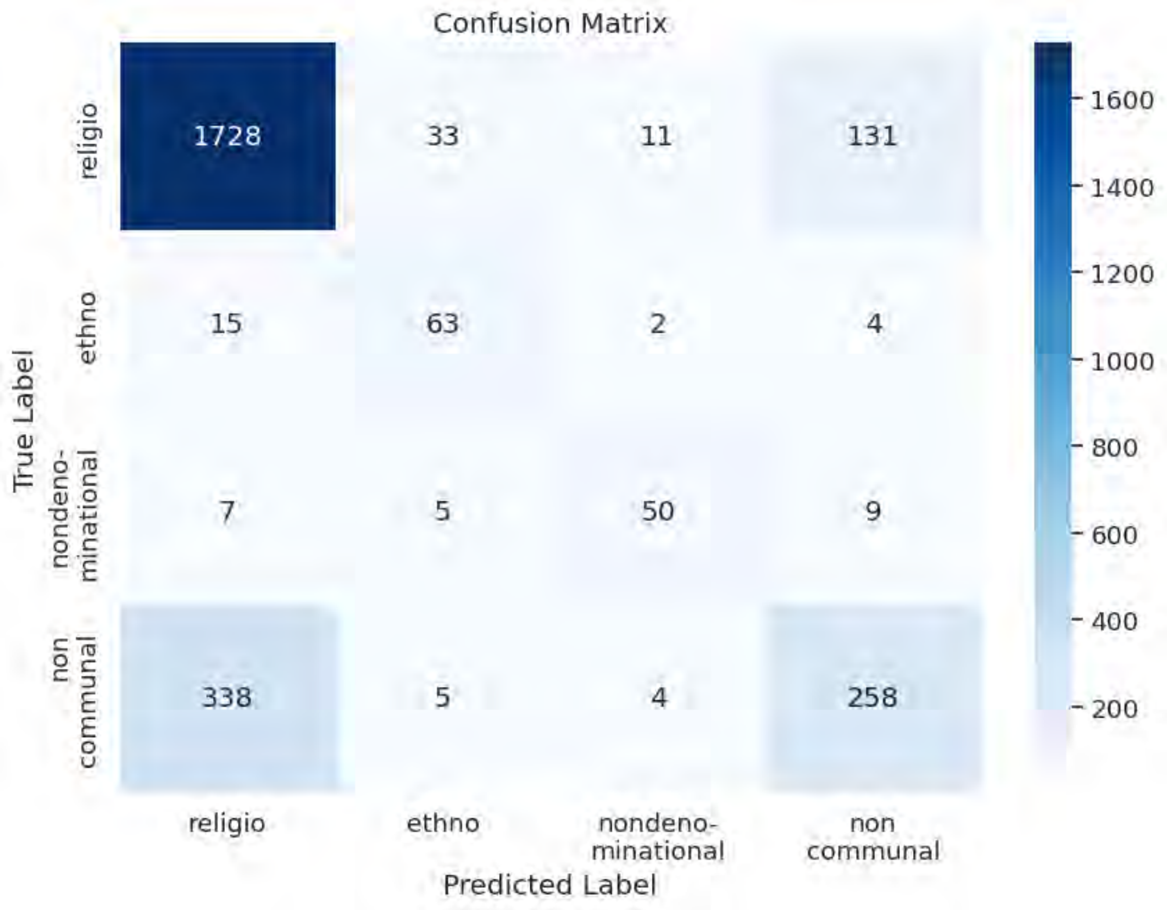


Figure 6.2: Confusion Matrix of Voting Ensemble for Five FTMs From BB

Chapter 7

Error Analysis

7.1 Dataset Limitations and Challenges

7.1.1 Class Imbalance Analysis

Our dataset is divided into four main classes: Religio-communal, Ethno-communal, Nondenominational and Non-communal. Each of these main classes is further divided into four sub-classes as Derogation, Antipathy, Prejudication, and Repression. We found that our model performs better when classifying the four main classes compared to the sixteen classes. Our analysis revealed that this discrepancy is largely due to the insufficient amount of data available for each class in sixteen class setup. One major issue is the uneven distribution of data among the classes. Since religious violence is a more frequent topic on social media, it likely led annotators to categorize 26.72% of data as Religio-communal class. On the contrary, the Ethno-communal and Nondenominational class represent only 4.69% and 1.2% of total text respectively. When further divided into sub-classes, the disparity becomes even more pronounced. The sub-class of Ethno-communal text ranges approximately from 12 to 67 texts whereas Nondenominational sub-classes have only 1-13 texts which is far too low compared to other two classes and very insufficient to train a robust model. (See Figure 7.1)

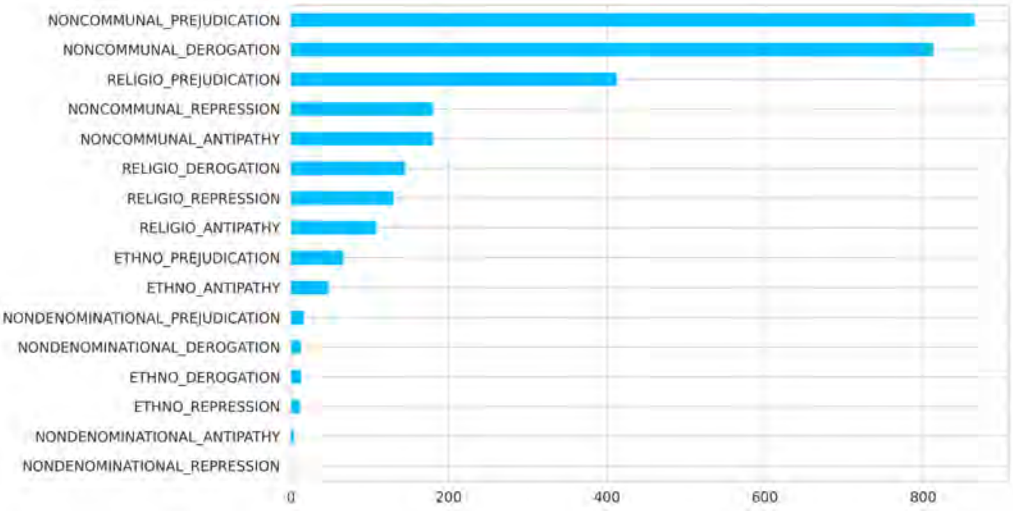


Figure 7.1: Data Quantity of Sixteen Classes

This imbalance creates significant challenges. The model may overfit to the more prevalent classes and underperform on the less represented ones. The lack of sufficient data for the smaller classes makes it difficult for the model to learn meaningful patterns, leading to poor generalization. Additionally, the imbalance can result in biased predictions, where the model disproportionately misclassifies instances from the underrepresented classes.

7.1.2 Data Annotation Anomalies

Another significant challenge we faced was the incorrect annotation of some data points. The annotators of this dataset also mislabeled some of the text, which contributed to our model's difficulty in correctly classifying instances in these particular classes and negatively impacted its overall performance. To investigate this, we collected the misclassified texts by our best performing model and randomly chose 235 texts, then manually reviewed them. According to the definitions provided for each class, some texts were clearly mislabeled, whereas our model was able to correctly classify those data (see Table 7.1). In addition to that, determining the semantic meaning of some texts without proper context proved to be difficult, and some texts did not express any violence but were still categorized in one of the classes. This

Text	Annotation	Model Prediction	Prediction Probability
খাল কেটে রোহিঙ্গা আনা - প্রবাদের নতুন সংস্করণ	No Violence	Ethno	0.9401
বোরখা আর হিজাব ছাড়া নারী মানেই মোল্লাদের চোখে উলংগ নারী।	No Violence	Religio	0.7581
জংগী, ধর্মান্ধ, উগ্রবাদী, ধর্ম ব্যবসায়ীদের উচিত ইসরায়েলের বিরুদ্ধে জিহাদ ঘোষণা করা	Non-communal	Religio	0.6965
রোহিঙ্গারা আমাদের দেশের পরিবেশ খারাপ করবে। আমাদের যুব সমাজ ধ্বংসের কারণ হয়ে দাড়াচ্ছে। এদের জায়গা দেয়া চরম ভুল হয়েছে মনে হচ্ছে। এরা একদিন আমাদের ঘরেই আমাদেরকে ফিলিস্তিনি বানানোর পায়তারা না শুরু করে দেয়! যতদ্রুত সম্ভব এইসব আপদ বিদায় করা জরুরি। ওদের প্রতিটি পদক্ষেপের ওপর নজরদারি বাড়াতে হবে। ওদের নেতার ওয়াশিংটন চলে যাওয়া কি আমাদের ভাবিয়ে তোলে না? ওদের খুন খারাবি আমাদের ভাবায় না? ওদের মাদক ব্যবসায় কি আমাদের বিবেক জাগ্রত করে না?	No Violence	Ethno	0.9254
তুমি নাস্তিক সন্দেহ হয়... বেজন্মা বেজন্মা মনে হয় .. ঠিক জায়গায় তোর জন্ম হয়নি ।	No Violence	Religio	0.7004
আপনি একটা দায়িত্ব নিয়ে, ইসরায়েল কে দমন করুন, মন্ত্রী মহাশয়	No Violence	Non-communal	0.6301
মোল্লা কেনো আল্লাহর গজব থেকে কেউ কখনো পার পেয়েছে কি? যার পরিনাম আপনাদের ভারত এখন বুঝতেছে মন্দির বন্ধ করে মসজিদ খুলে দিচ্ছে কারন ইসলামের বিজয় নিশ্চিত	No Violence	Religio	0.7698

Table 7.1: Overview of Misannotations vs. Model Predictions Across Four Classes

disrupted our model's ability to learn meaningful patterns and correlations between words, leading to poor generalization (see Table 7.2).

Text	Annotation	Model Prediction
এক যে ছিলো শিয়ালে, মোরগ আঁকে দেয়ালে, আপন মনে চাটতে থাকে খেয়ালে।	Non-communal	No Violence
কিসের নিরাপত্তা? যদি নিরাপত্তা এমন ই হত তাহলে হিন্দু সমাজ এত অত্যচার এর সম্মুখীন হতো না। বাংলাদেশের নাগরিক হিসেবে নিরাপত্তা আমাদের অধিকার।	Religio	No Violence
সমস্যা ১০০% সমাধান না হওয়া পর্যন্ত কাউকেই ধন্যবাদ দেওয়া যাবে না! সবই চীন ভারতের বিজনেস!	Non-communal	No Violence
করোনাভাইরাস আশার কারন নির্যাতন।	Non-communal	No Violence

Table 7.2: Misannotated Non-Violent contexts Across Four Classes

Among the 235 texts we reviewed, we found that 44 were incorrectly annotated or lacked context, comprising 18% of the data.

7.2 Pre-trained Model Limitations

7.2.1 BanglaBERT

Our main challenge with the BanglaBERT model was its tendency to misclassify between Religio-communal and Non-communal classes. It is known that word embeddings may reflect or amplify problematic biases from the data they are trained on, for example, gender classes [5]. To look further into this issue, we analyzed the most common words within both the classes. While the Religio-communal data exhibited relevant words like 'কাফের', 'আল্লাহ', 'মুসলিম', 'নাস্তিক','হিন্দু', 'ইসলাম', 'ধ্বংস', 'গজব', the words in Non-communal texts was considerably diverse. Its most common words, such as 'ভাই', 'ভালো', and 'বলে', 'কিছু', 'লজ্জা' lacked significant semantic meaning on their own.

However, the presence of 'মানুষ' (Human) and 'বাংলাদেশ' (Bangladesh) stood out as meaningful most frequent words in the Non-communal class. To investigate further, we compared the cosine similarity between these frequent Non-communal words with the most common Religio-communal words. Remarkably, we found that words like 'ধর্ম' (religion), 'ইসলাম' (Islam), 'আল্লাহ' (Allah), 'নাস্তিক' (atheist), 'মুসলিম' (Muslim), and 'হিন্দু' (religion), কাফের (infidel) had strikingly high cosine similarity with 'বাংলাদেশ' (Bangladesh) or 'মানুষ' (Human).

This high similarity hindered our model's ability to effectively differentiate between the Religio-communal and Non-communal classes. As a result, despite the distinctive patterns in Religio-communal data, the presence of 'মানুষ' (Human) and 'বাংলাদেশ' (Bangladesh) in the Non-communal data posed a significant challenge, leading to misclassifications.

Non-comunal Words	Religio Communal Words	BanglaBERT	BanglaBERT Large	mBERT
মানুষ	কাফের	0.9539	0.9419	0.4839
	নাস্তিক	0.9706	0.9703	0.6049
	ধর্ম	0.9701	0.9873	0.6591
	মুসলিম	0.8825	0.9789	0.5854
	ইসলাম	0.9631	0.9358	0.6146
	আল্লাহ	0.9413	0.9591	0.5444
	হিন্দু	0.9178	0.9678	0.6154
বাংলাদেশ	কাফের	0.8966	0.8978	0.4907
	নাস্তিক	0.9282	0.9233	0.4195
	ধর্ম	0.9429	0.9494	0.4209
	মুসলিম	0.9218	0.9674	0.5267
	ইসলাম	0.9340	0.9848	0.4656
	আল্লাহ	0.9548	0.9843	0.4234
	হিন্দু	0.9365	0.9623	0.5614

Table 7.3: Cosine Similarities: BanglaBERT, BanglaBERT Large and mBERT

7.2.2 BanglaBERT Large

We also encountered the same issue with the BanglaBERT Large model, which struggled to differentiate between Religio-communal and Non-communal texts. To understand this better, we analyzed the cosine similarity issues with this model also and found that the problem was even more pronounced compared to the original BanglaBERT model. The high cosine similarity between Religio-communal and Non-communal words was even more extreme for the BanglaBERT Large model (see Table 7.3). This exacerbated the model's difficulty in accurately classifying the texts. As a result, it performed even poorer in classifying Non-communal texts compared to the BanglaBERT model. F1 scores dropped significantly compared to the best performing fine-tuned BanglaBERT model,with a substantial decline of 0.16 for Non-communal classes (see table 6.9).

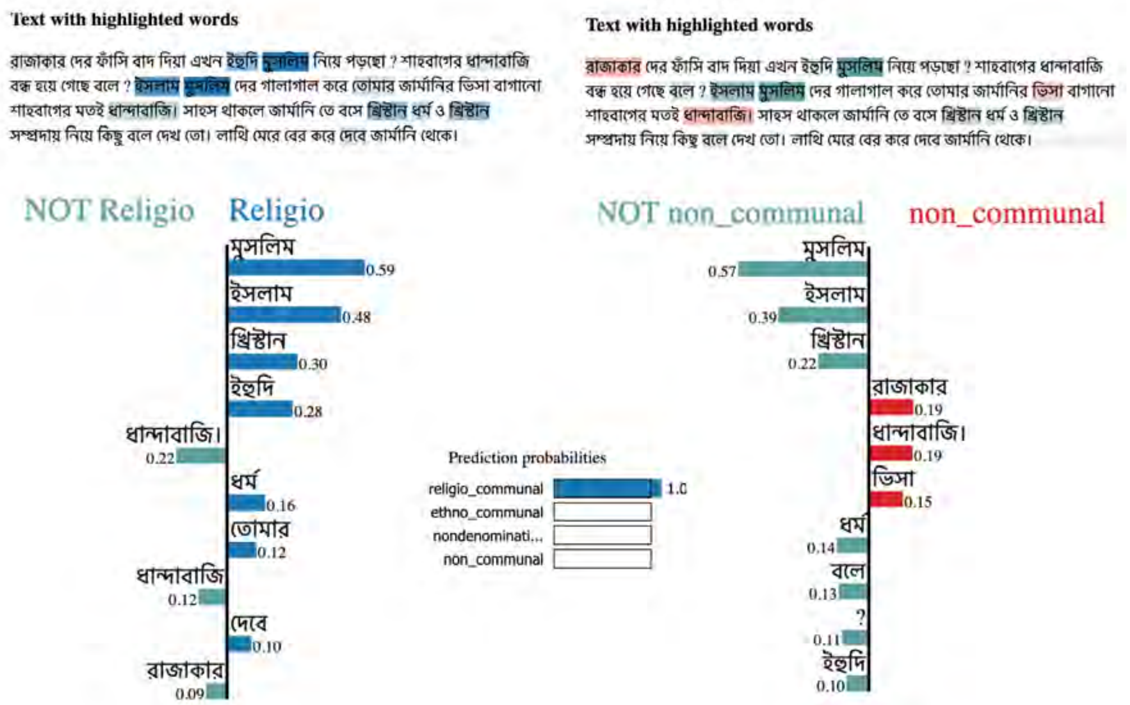
7.2.3 mBERT

However, mBERT had a more balanced cosine similarity score compared to the other ones, as shown by the score of 0.4234 for 'আল্লাহ' and 'বাংলাদেশ' and 0.4209 for 'ধর্ম' and 'বাংলাদেশ'. However its performance in Religio-communal and Non-communal text classification falls short compared to the fine tuned BanglaBERT. The lower F1 scores for those categories (0.42 and 0.39 respectively) suggest that mBERT struggles to capture the subtle contextual nuances crucial for this task. This weakness likely stems from its multilingual training. While mBERT has a broad understanding across languages, it may miss the specific cultural details that are essential for understanding Bengali text. In other words, mBERT shines in situations where clear semantic distinctions exist. This is evident for the Ethno-communal and Non-denominational class, where it performed commendably at F1 scores of 0.61 and

0.67 respectively (see Table 6.8). Its vast multilingual training provides a broad base, but comes at the expense of in-depth understanding of Bengali violent speech patterns. The nuanced context required to differentiate between Religio-communal and Non-communal text might simply be beyond mBERT's grasp. While mBERT excels at balancing semantic similarity, this ability doesn't translate to understanding the specific contexts and connotations that are vital for accurate classification in this domain.

7.3 Our Models Limitation

As previously mentioned, one of the main challenges our models faced was distinguishing between Religio-communal and Non-communal data. The models frequently misclassified non-communal texts as Religio-communal and vice versa. This misclassification arises from the presence of a significant religious context in many Non-communal texts. For example, many of the Non-communal texts often include terms commonly associated with the Religio-communal class, such as "আল্লাহ" (Allah), "মুসলিম" (Muslim), "ধর্ম" (religion), "ইসলাম" (Islam), and "ইহুদি" (Jew). These words were given substantial weight towards Religio-communal class by the model, leading to an erroneous association with the class. This intricate overlap between Religio-communal and Non-communal texts has posed a significant challenge for our model in accurately differentiating between the two classes.



Text with highlighted words

রাজাকার দের ফাঁসি বাদ দিয়া এখন ইহুদি মুসলিম নিয়ে পড়ছো ? শাহবাগের ধান্দাবাজি বন্ধ হয়ে গেছে বলে ? ইসলাম মুসলিম দের গালাগাল করে তোমার জার্মানির ভিসা বাগানো শাহবাগের মতই ধান্দাবাজি। সাহস থাকলে জার্মানি তে বসে খ্রিষ্টান ধর্ম ও খ্রিষ্টান সম্প্রদায় নিয়ে কিছু বলে দেখ তো। লাথি মেরে বের করে দেবে জার্মানি থেকে।

Text with highlighted words

রাজাকার দের ফাঁসি বাদ দিয়া এখন ইহুদি মুসলিম নিয়ে পড়ছো ? শাহবাগের ধান্দাবাজি বন্ধ হয়ে গেছে বলে ? ইসলাম মুসলিম দের গালাগাল করে তোমার জার্মানির ভিসা বাগানো শাহবাগের মতই ধান্দাবাজি। সাহস থাকলে জার্মানি তে বসে খ্রিষ্টান ধর্ম ও খ্রিষ্টান সম্প্রদায় নিয়ে কিছু বলে দেখ তো। লাথি মেরে বের করে দেবে জার্মানি থেকে।

Figure 7.2: LIME Analysis - Example 1

To demonstrate this issue, we present four Non-communal texts that the model incorrectly identified as Religio-communal. We employed LIME (Local Interpretable

Model-agnostic Explanations) analysis to gain insights into these misclassifications. In the first example, terms such as "মুসলিম", "ইসলাম", and "খ্রিস্টান" were assigned high weights in the Religio-communal class, indicating their strong influence on the model's decision. However, these same words had negative weights in the Non-communal class, demonstrating their impact on reducing the likelihood of a Non-communal classification (see Figure 7.3). As a result, the model predicted the text as Religio-communal with a confidence of 1.00.

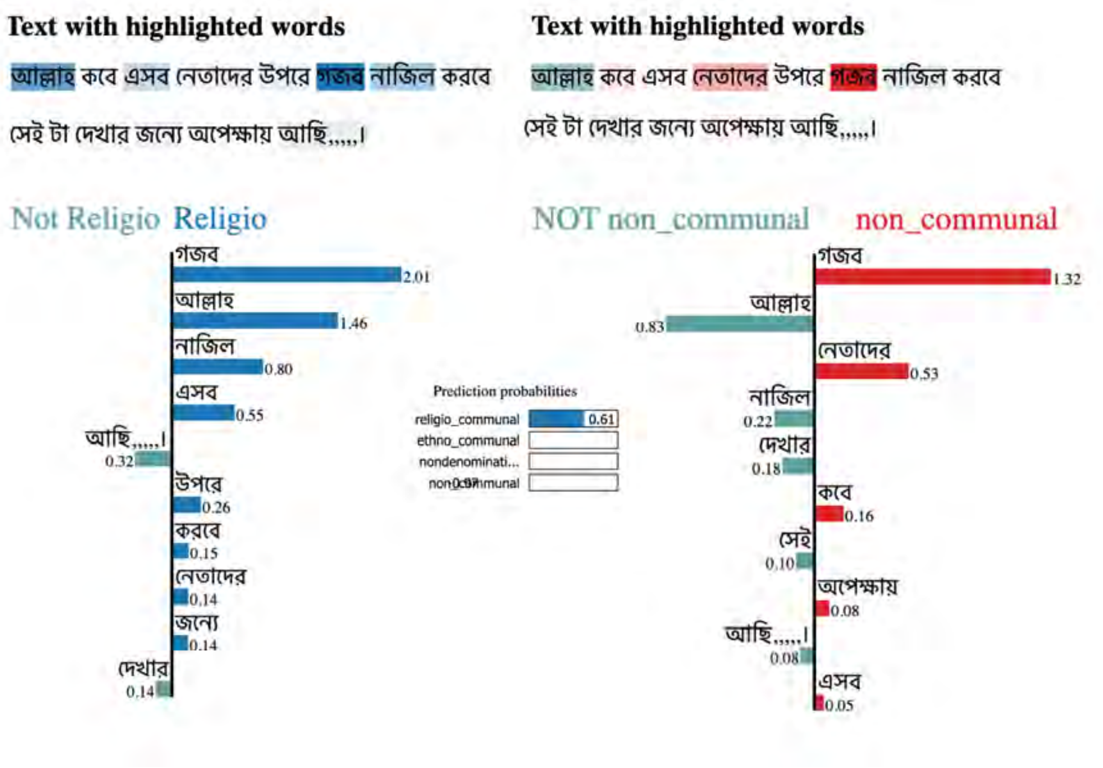
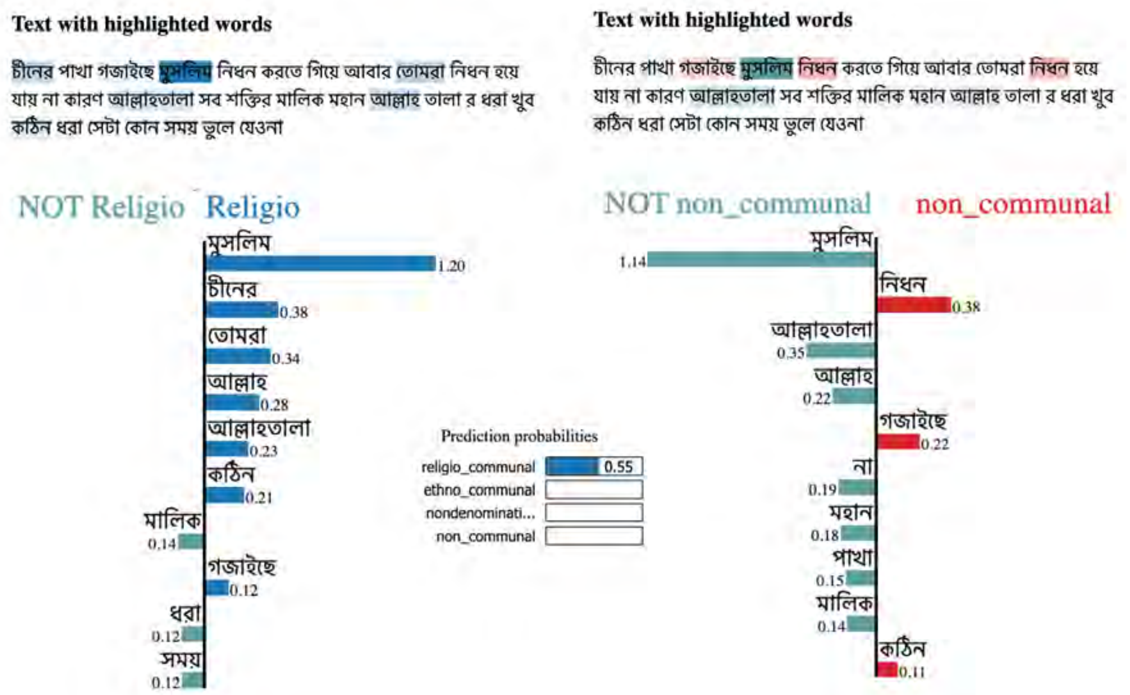


Figure 7.3: LIME Analysis - Example 2

Similarly, in the following examples, words like "মুসলিম", "আল্লাহ", "আল্লাহ'তালা", "নাজিল", "ধর্মান্ধ", "জিহাদ" were assigned high weights in the Religio-communal class while they effected negatively in the Non-communal class, thus misidentifying the text as Religio-communal. This highlights the difficulty our model faces in differentiating between Religio-communal and Non-communal texts due to the overlapping presence of religious terms.



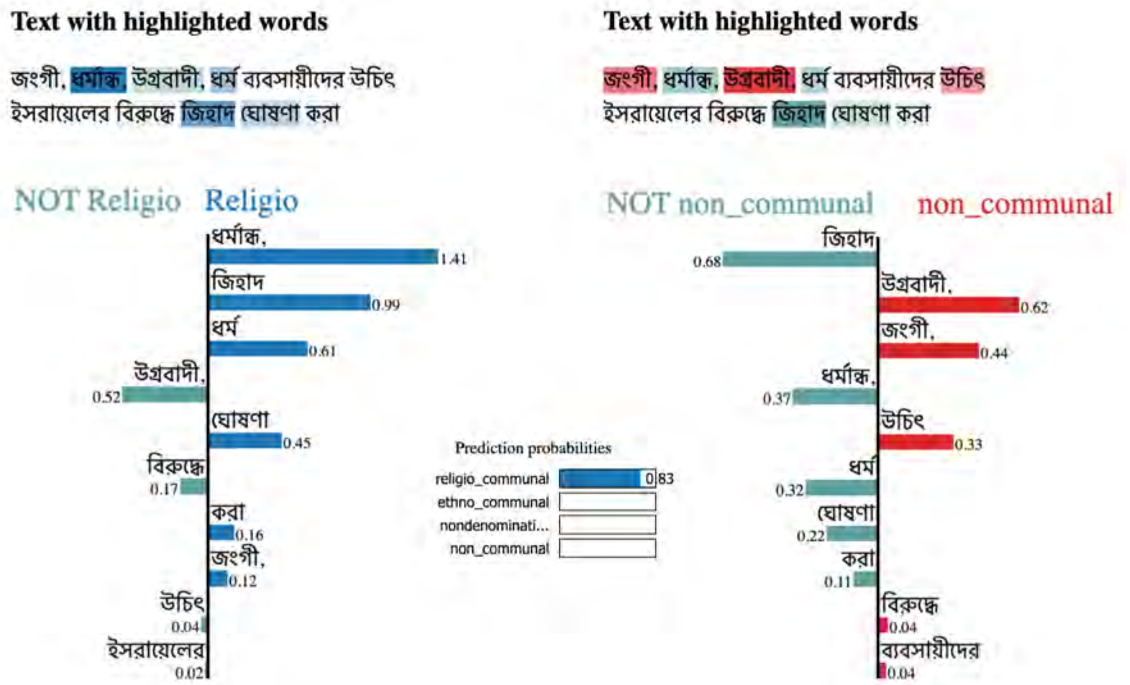
Text with highlighted words

চীনের পাখা গজাইছে মুসলিম নিধন করতে গিয়ে আবার তোমরা নিধন হয়ে যায় না কারণ আল্লাহতালা সব শক্তির মালিক মহান আল্লাহ তালা র ধরা খুব কঠিন ধরা সেটা কোন সময় ভুলে যেওনা

Text with highlighted words

চীনের পাখা গজাইছে মুসলিম নিধন করতে গিয়ে আবার তোমরা নিধন হয়ে যায় না কারণ আল্লাহতালা সব শক্তির মালিক মহান আল্লাহ তালা র ধরা খুব কঠিন ধরা সেটা কোন সময় ভুলে যেওনা

Figure 7.4: LIME Analysis - Example 3



Text with highlighted words

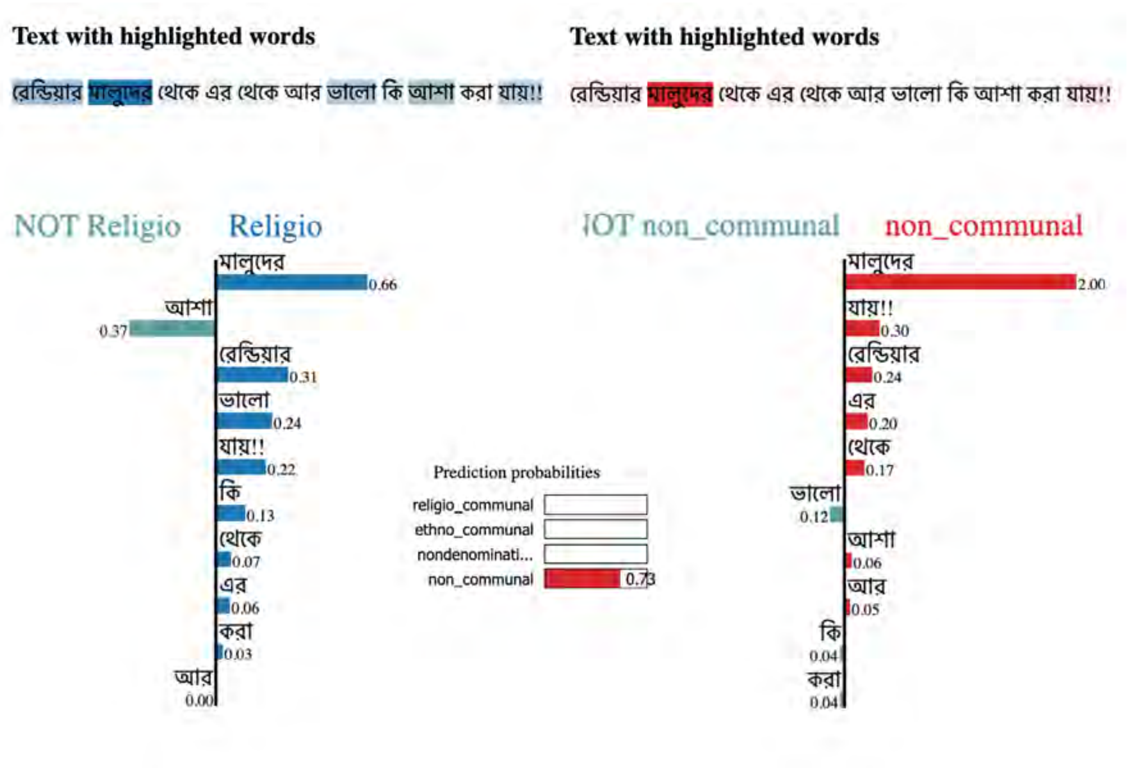
জংগী, ধর্মান্ধ, উগ্রবাদী, ধর্ম ব্যবসায়ীদের উচিৎ ইসরায়েলের বিরুদ্ধে জিহাদ ঘোষণা করা

Text with highlighted words

জংগী, ধর্মান্ধ, উগ্রবাদী, ধর্ম ব্যবসায়ীদের উচিৎ ইসরায়েলের বিরুদ্ধে জিহাদ ঘোষণা করা

Figure 7.5: LIME Analysis - Example 4

Building on this, we extended our analysis to understand the model's confusion in identifying Religio-communal data that it misclassified as Non-communal. By conducting LIME analysis on these misclassified texts, we aimed to uncover the underlying reasons for these errors. The LIME technique allowed us to decompose the model's predictions and examine the contribution of individual words to the final classification. Our findings revealed that the same problem persisted in this context: the model struggled to differentiate and assign appropriate weights to Religio-communal and Non-communal words. In many instances, words typically associated with Religio-communal contexts received high weights in the Religio-communal class, which is expected. However, these words also received high weights in the Non-communal class. Consequently, when the total weights were aggregated, the model therefore categorized Religio-communal texts as Non-communal.



Text with highlighted words

রেন্ডিয়ার মালুদের থেকে এর থেকে আর ভালো কি আশা করা যায়!!

Text with highlighted words

রেন্ডিয়ার মালুদের থেকে এর থেকে আর ভালো কি আশা করা যায়!!

Figure 7.6: Lime Analysis - Example 5

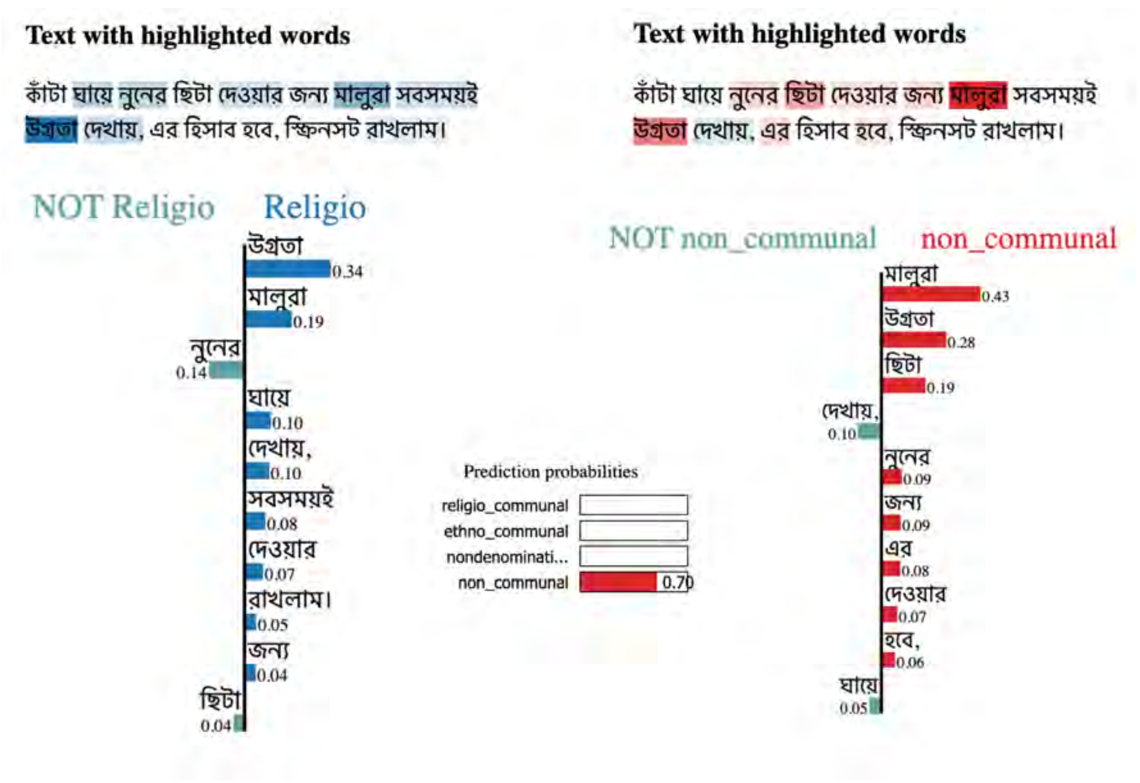


Figure 7.7: Lime Analysis - Example 6

For instance, in Figure 7.6 and Figure 7.7, terms such as “উগ্রতা” (extremism), “মালুর” (a derogatory term for Malayalis), and “মালুদের” (a plural form of the derogatory term for Malayalis) are typically associated with the Religio-communal class. However, the LIME analysis reveals that these words have also been assigned substantial weight in the Non-communal class. This overlap in word weighting suggests that the model’s internal representation lacks sufficient granularity to effectively distinguish between subtle nuances in text. Some prominent words in Religio-communal contexts, might also appear in Non-communal discussions that involve historical, cultural, or social contexts without any communal connotations. The LIME analysis highlighted these specific instances where the model’s interpretive layers did not adequately separate these contexts, leading to misclassification.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

Our research's aim is to use state-of-the-art large language models to effectively analyze and understand the nuances of communal violent speech on social media's Bengali textual content, so that we can proactively detect those content and take necessary actions against them. To achieve this, we analyze a comprehensive dataset of four forms of communal violent speech, categorized by experts into specific categories. We also utilized and fine tuned pre-trained BERT that can accurately classify and predict the thematic category of a given text. In the future, our study will continue to seek more effective methods of detecting communal violence by improving and fine tuning the model further. Our research contributes to the field of NLP by providing a novel and valuable tool for identifying and addressing communal violence propagated through social platforms. We hope that our work will inspire further research on this topic and advance the field of Bengali violent speech detection.

8.2 Future Work

1. Reducing Class Imbalances: Our current research focused on four class of communal violent speech due to the limited availability of data for all sixteen identified class. Future work will aim to expand the dataset to include sufficient examples for all sixteen class, addressing the significant class imbalance we encountered, where Non-communal content comprised one-third of the total data. Additionally, we faced annotation issues, with some annotations being incorrect or lacking context. By gathering more comprehensive, accurate and balanced data, we can ensure accurate annotations and enable more effective training, ultimately improving the model's performance and reliability.
2. Model Improvement Based on Error Analysis: Based on the error analysis we have conducted, future work will focus on addressing specific issues where the model struggles. For example, our current model frequently misclassifies between Religio-communal and Non-communal violence, partly due to the high cosine similarity between related terms in Bangla BERT models. Future improvements will involve refining the model to better distinguish between these categories, potentially through advanced techniques such as contextual embedding and domain-specific fine-tuning.

References

[1] D. L. Horowitz, *The deadly ethnic riot.* Univ of California Press, 2001.

[2] T. Windeatt, "Ensemble mlp classifier design," in *Computational Intelligence Paradigms: Innovative Applications*, L. C. Jain, M. Sato-Ilic, M. Virvou, G. A. Tsihrintzis, V. E. Balas, and C. Abeynayake, Editors. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 133–147, ISBN: 978-3-540-79474-5. DOI: 10.1007/978-3-540-79474-5_6. [Online]. Available: https://doi.org/10.1007/978-3-540-79474-5_6.

[3] A. E. Cano Basave, Y. He, K. Liu, and J. Zhao, "A weakly supervised bayesian model for violence detection in social media," 2013.

[4] S. Agarwal and A. Sureka, "Using knn and svm based one-class classifier for detecting online radicalization on twitter," in *Distributed Computing and Internet Technology*, R. Natarajan, G. Barua, and M. R. Patra, Editors, Cham: Springer International Publishing, 2015, pp. 431–442, ISBN: 978-3-319-14977-6.

[5] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*, 2016. arXiv: 1607.06520 [cs.CL].

[6] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, 2017, pp. 512–515.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[9] D. Agrawal. "Emoji dictionary." (2020), [Online]. Available: https://www.kaggle.com/divyansh22/emoji-dictionary-1 (visited on 11/20/2023).

[10] R. U. Haque, M. F. Mridha, M. A. Hamid, M. Abdullah-Al-Wadud, and M. S. Islam, "Bengali stop word and phrase detection mechanism," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3355–3368, Feb. 2020, ISSN: 2191-4281. DOI: 10.1007/s13369-020-04388-8. [Online]. Available: http://dx.doi.org/10.1007/s13369-020-04388-8.

[11] M. R. Karim, B. R. Chakravarti, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," in *7th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA,2020)*, IEEE, 2020.

[12] I. A. Khandokar, I. Mamun, T. I. A. Chadni, Z. A. Anas, and S. Shatabda, "Event detection and knowledge mining from unlabelled bengali news articles," in *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, IEEE, 2020, pp. 1–6.

[13] A. Kumar, S. Saumya, and J. P. Singh, "Nitp-ai-nlp@ hasoc-fire2020: Fine tuned bert for the hate speech and offensive content identification from social media.," in *FIRE (Working Notes)*, 2020, pp. 266–273.

[14] N. Romim, M. Ahmed, H. Talukder, and M. S. Islam, "Hate speech detection in the bengali language: A dataset and its baseline evaluation," *arXiv preprint arXiv:2012.09686*, 2020. [Online]. Available: https://arxiv.org/abs/2012.09686.

[15] X. Sun, J. Gu, and H. Sun, "Research progress of zero-shot learning," *Applied Intelligence*, vol. 51, no. 6, pp. 3600–3614, Nov. 2020, ISSN: 1573-7497. DOI: 10.1007/s10489-020-02075-7. [Online]. Available: http://dx.doi.org/10.1007/s10489-020-02075-7.

[16] J. Brownlee, *Smote for imbalanced classification with python*, Mar. 2021. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[17] A. K. Das, A. A. Asif, A. Paul, and M. N. Hossain, *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578–591, 2021. DOI: doi:10.1515/jisys-2020-0060. [Online]. Available: https://doi.org/10.1515/jisys-2020-0060.

[18] M. R. Karim, S. K. Dey, T. Islam, *et al.*, "Deephateexplainer: Explainable hate speech detection in under-resourced bengali language," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, 2021, pp. 1–10. DOI: 10.1109/DSAA53316.2021.9564230.

[19] G. Kohli, P. Kaur, and J. Bedi, "ARGUABLY at ComMA@ICON: Detection of multilingual aggressive, gender biased, and communally charged tweets using ensemble and fine-tuned IndicBERT," in *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, NIT Silchar: NLP Association of India (NLPAI), Dec. 2021, pp. 46–52. [Online]. Available: https://aclanthology.org/2021.icon-multigen.7.

[20] R. Kumar, S. Ratan, S. Singh, *et al.*, "ComMA@ICON: Multilingual gender biased and communal language identification task at ICON-2021," in *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, NIT Silchar: NLP Association of India (NLPAI), Dec. 2021, pp. 1–12. [Online]. Available: https://aclanthology.org/2021.icon-multigen.1.

[21] W. Soruar and M. M. Uddin, "Changes in social and religious practices of disputing communities after riot: A case study on communal violence in ramu," in *Journal of Philosophy, Culture and Religion*, vol. 51, 2021, p. 13. DOI: 10.7176/JPCR/51-03.

[22] E. V. Tunyan, T. A. Cao, and C. Y. Ock, "Improving subjective bias detection using bidirectional encoder representations from transformers and bidirectional long short-term memory," *International Journal of Cognitive and Language Sciences*, vol. 15, no. 5, pp. 329–333, 2021, ISSN: eISSN: 1307-6892. [Online]. Available: https://publications.waset.org/vol/173.

[23] A. Akil, N. Sultana, A. Bhattacharjee, and R. Shahriyar, "Banglaparaphrase: A high-quality bangla paraphrase dataset," *arXiv preprint arXiv:2210.05109*, 2022.

[24] A. Bhattacharjee, T. Hasan, W. Ahmad, *et al.*, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98.

[25] R. Botelle, V. Bhavsar, G. Kadra-Scalzo, *et al.*, "Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: An applied evaluative study," *BMJ open*, vol. 12, no. 2, e052911, 2022.

[26] K. Dey, Sumon, M. Cochez, and M. R. Karim, *Bengali Hate Speech Detection Dataset*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5PD07, 2022.

[27] B. D. Dirting, G. A. Chukwudebe, E. C. Nwokorie, and I. I. Ayogu, "Multilabel classification of hate speech severity on social media using bert model," in *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, 2022, pp. 1–5. DOI: 10.1109/NIGERCON54645.2022.9803164.

[28] U. Khan, S. Khan, A. Rizwan, G. Atteia, M. M. Jamjoom, and N. A. Samee, "Aggression detection in social media from textual data using deep learning models," *Applied Sciences*, vol. 12, no. 10, p. 5083, 2022.

[29] G. H. Panchala, V. V. S Sasank, D. R. Harshitha Adidela, P. Yellamma, K. Ashesh, and C. Prasad, "Hate speech & offensive language detection using ml &nlp," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022, pp. 1262–1268. DOI: 10.1109/ICSSIT53264.2022.9716417.

[30] K. Pareek, A. Choudhary, A. Tripathi, K. Mishra, and N. Mittal, "Hate and aggression detection in social media over hindi english language," *International Journal of Software Science and Computational Intelligence (IJSSCI)*, vol. 14, no. 1, pp. 1–20, 2022.

[31] M. J. K. I. Rahman, "Religious nationalism in digitalscape: An analysis of the post-shahbag movement in bangladesh," *Open Journal of Social Sciences*, vol. 10, no. 5, pp. 201–218, 2022.

[32] N. I. Remon, N. H. Tuli, and R. D. Akash, “Bengali hate speech detection in public facebook pages,” in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 2022, pp. 169–173. DOI: 10.1109/ICISET54810.2022.9775900.

[33] D. Antypas and J. Camacho-Collados, *Robust hate speech detection in social media: A cross-dataset empirical evaluation*, 2023. arXiv: 2307.01680 [cs.CL].

[34] T. A. Belal, G. M. Shahariar, and M. H. Kabir, “Interpretable multi labeled bengali toxic comments classification using deep learning,” in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2023, pp. 1–6. DOI: 10.1109/ECCE57851.2023.10101588.

[35] F. Fkih and G. Al-Turaif, “Threat modelling and detection using semantic network for improving social media safety,” *International Journal of Computer Network and Information Security*, vol. 15, no. 1, p. 39, 2023.

[36] S. Kemp. “Digital 2023: Bangladesh.” (2023), [Online]. Available: https://datareportal.com/reports/digital-2023-bangladesh (visited on 09/15/2023).

[37] A. J. Keya, M. M. Kabir, N. J. Shammey, M. F. Mridha, M. R. Islam, and Y. Watanobe, “G-bert: An efficient method for identifying hate speech in bengali texts on social media,” *IEEE Access*, vol. 11, pp. 79 697–79 709, 2023. DOI: 10.1109/ACCESS.2023.3299021.

[38] M. Kim, *Smote: Practical consideration and limitations*, Dec. 2023. [Online]. Available: https://medium.com/@minjukim023/smote-practical-consideration-limitations-f0d926b661a8.

[39] P. Patil, S. Raul, D. Raut, and T. Nagarhalli, “Hate speech detection using deep learning and text analysis,” in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2023, pp. 322–330. DOI: 10.1109/ICICCS56967.2023.10142895.

[40] G. Polat, *Zero-shot learning (zsl) explained*, Oct. 2023. [Online]. Available: https://encord.com/blog/zero-shot-learning-explained/.

[41] S. R. Titli and S. Paul, “Automated bengali abusive text classification: Using deep learning techniques,” in *2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS)*, 2023, pp. 1–6. DOI: 10.1109/ICAECIS58353.2023.10170294.

[42] Jan. 2024. [Online]. Available: https://www.ibm.com/topics/few-shot-learning.

[43] N. Tasnim, S. S. Gupta, F. I. Juee, *et al.*, “Mapping violence: Developing an extensive framework to build a bangla sectarian expression dataset from social media interactions,” 2024. DOI: https://doi.org/10.48550/arXiv.2404.11752.