Deep Learning Approaches for Detecting and Analyzing Abusive Bangla Comments on Religion in Social Media

Md. Mubtasim Fuad Khan

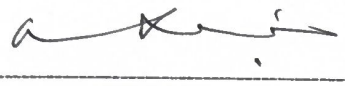
Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

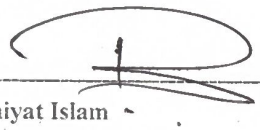
This thesis titled on "**Deep Learning Approaches for Detecting and Analyzing Abusive Bangla Comments on Religion in Social Media**", submitted by **Md. Mubtasim Fuad Khan (ID: 221-35-883)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS



Chairman

Dr. A. H. M. Saifullah Sadi
Professor
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University



Internal Examiner 1

Dr. Rubaiyat Islam
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Dr. Md. Abdul Kader
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Nuruzzaman Faruqui
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md. Mostafiz Khan
Managing Director
Tecognize Solutions Limited

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : ABC

Date of Birth :

Title :

Academic Session :

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

_____
(Student's Signature)

221-35-883

_____
Student ID
Date: 24 December 2025

_____
(Supervisor's Signature)

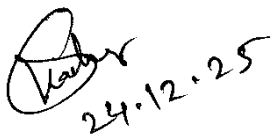
Dr. Md Abdul Kader

_____
Name of Supervisor
Date: 24 December 2025

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR'S DECLARATION

I/We* hereby declare that I/We* have checked this thesis/project* and in my/our* opinion, this thesis/project* is adequate in terms of scope and quality for the award of the degree of *Bachelor of Science/ Master of Science.



(Supervisor's Signature)

Full Name : Dr. Md Abdul Kader

Position : Associate Professor

Date : 24 December 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

(Student's Signature)

Full Name : Md. Mubtasim Fuad Khan

ID Number : 221-35-883

Date : 24 December 2025

Deep Learning Approaches for Detecting and Analyzing Abusive Bangla Comments on Religion in Social Media

Md. Mubtasim Fuad Khan

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Software Engineering)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Almighty Allah because he helped me to achieve the strength, patience, and wisdom to complete this piece of research work successfully.

I would like to thank my supervisor who said that the support, motivation and the tips that he gave me played a significant role in completing this thesis. They have continued to help me and provide useful contributions to this research, making me improve my ideas and achieve research objectives.

I would like to say that I am grateful to the entire faculty of the Department of Software Engineering, Daffodil International University, and their guidance, support, and the academic basis they have provided during my undergraduate career.

I would like to thank my friends and fellow researchers with all my heart, as they have inspired, helped me with technical support, and discussed the issues with my mind, making it possible to overcome all challenges during the work on the project.

Finally, I would like to express my sincere gratitude to my loving parents and family who have remained steadfast in love, support, and prayers to me. They have always believed in me and their faith has been the driving force of all my achievements.

DEDICATION

This dissertation is dedicated to my beloved parents and family which has been my pillar throughout my educational life due to their unconditional love, sacrifice and unbreakable support. My strength has come mostly through their prayers, patience and faith in my abilities.

I would also like to dedicate this work to my teachers and mentors whose support, wisdom, and inspiration have played an important role in nurturing me as both a student and researcher.

Finally, this thesis is devoted to everyone who believes in the use of knowledge, technology, and responsibility to create a safer and more respectful digital society.

ABSTRACT

Increment in hate speech particularly on religious beliefs has been the source of significant concern of concern on the internet social sites. Bangladesh culture and religion have been closely connected, and due to the abuse language against particular religious groups, social cohesion and security is endangered over the Internet. Identification of this harmful material in Bangli is not easy because of the insufficiency of language materials and because of the vague meanings of the contexts. The paper describes a deep learning method, which is automated, to recognize abusive remarks on religion in Bangali. An effective way of learning contextual relations in a text was to include Multi-Head Attention mechanism to a Bi-Directional Long Short-Term Memory (Bi-LSTM) network. The problem of the imbalance of the datasets was dealt with by an adapted Focal Loss that helped to improve detection efficiency. In the collection, there were 24,137 Facebook bangla remarks, and they were classified in the category of Normal and Religious Abuse. The proposed model with data preprocessing and partitioning (72 percent training, 8 percent validation, 20 percent testing) yielded an accuracy of 95 percent, surpassing the traditional machine learning, and other subsequent deep learning models, including CNN-LSTM and regular Bi-LSTM. Comparisons between the models (LSTM, CNN-LSTM, RNN, Logistic Regression, SVM, Random Forest) have been done and it was identified that attention mechanism enhanced the classification of abusive religious content. The results have shown that recurrent neural activity coupled with attention systems is an effective technique of identifying online abuse on religion on low resources languages such as Bengali language (Bangla).

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Spread of hate speech, especially on religious beliefs, has become a major issue on internet social networks. Incidious and offensive words may cause serious social conflict and break the social harmony. Religious abuse is particularly delicate since it often leads to an extreme emotional reaction and even real violent behavior. This form of online harassment has been on the rise in Bangladesh where religion and culture are tied together. This type of content is difficult to detect because of complex meanings and terms of the Bangla language.

This chapter gives an overview of the situation and motive of the research. It analyzes new international and national religious events that formed online debates and increased the need to have automated systems that can eliminate harmful religious comments. The chapter identifies the major goal and research objectives, namely how deep learning and Natural Language Processing (NLP) models can help in developing an effective detection model of Bangla religious.

1.1 Recent Religious Events and their influence on the online discourse

The Various recent events have significantly shaped the landscape of religious discourse at a global level and especially in Bangladesh, often being amplified by the rapid spread of information and misinformation through the social media. These occasions show that effective mechanisms to identify abusive religious text are highly needed because most times, these occurrences lead to heightened tensions, social conflicts and the growth of hate speech on the internet. A understanding of such events provides critical background to the development and application of such systems.

1.1.1 Global Religious Events and the Effects

Over the recent years, a lot of religious conflicts and tensions have been highlighted on online platforms, increasing the hate speech. A good example is the ongoing Israel

Palestine conflict, with online debates often split, and the number of anti-Semitic and Islamophobic comments on the subject has increased. The social media platforms become a platform of stories, featuring strongly expressed opinions by their users which may quickly turn into harassment directed at the Jews or Muslims based on their perceived affiliations. Similarly, the Hindu-Muslim hostilities on the internet frequently appear as conflicts between India and Pakistan, which are usually founded on historical and political matters. The situation in one country may trigger a wave of negative commentaries against the respective religious minority in another, which highlights the connection between the global events and the local online behavior.

Russia-Ukraine conflict is not left behind in the religious terms as the discussion of the Orthodox religion and its various branches is sometimes used to justify or criticize the activities, leading to the use of insults with religious overtones. Religious identities can also be manipulated even in situations that seem secular and this raises the aggregate of religious hate speech on the internet. These global events demonstrate that acts of abuse on religion are not localized in specific regions but rather a global issue that digital communication increases.

1.1.2 Recent events of religion and its consequences in Bangladesh

In Bangladesh, a series of recent events has had a significant impact on the world of religious discourse that is often exacerbated by the rapid spreading of information and misinformation via social media networks. Such events underscore the urgent need to have robust mechanisms of detecting religiously abusive literature as they most of the times lead to escalation of tensions, social incohesion, and an increase in hate speech on the internet. The understanding of such events provides a necessary background to the development and application of these systems.

Another popular problem is the accusations of blasphemy or disrespect of religious authorities/books, and it can easily result in mass violence. One case in point is when a young Hindu known as Akash Das posted an offensive statement about the Quran on Facebook in December 2024. The remark immediately went viral and brought about a lot of unrest and Hindu property and temple attacks [1]. These events demonstrate how one online message (intended or misunderstood) can cause actual consequences, showing

how social media has a powerful and negative effect, in many cases, on the development of the religious conflict.

Following significant political transformations, such as the one in August 2024, it happened that online misinformation and fake reports targeting religious minorities grew in Bangladesh. The news about church attacks and houses of Christians were rapidly spread on social networks, which increased the level of fear and insecurity among these groups of people [2]. Similarly, false information about the Muslim attacks on Hindus was being agitated by right-wing profiles after the resignation of Sheikh Hasina, which further worsened the problem of communal tensions [3]. All these events are evidence of a deliberate use of social media as an instrument to sow hate and instigate violence, often through exploiting already existing religious differences.

According to the Bangladesh Buddhist Hindu Christian Unity Council, 2184 instances of violence against religious minorities occurred between August and December 2024 [4]. This type of cases, which are often preceded or followed by hate campaigns over the Internet, include the destroying of temples, homes, and businesses, as well as the killing or raping of people [5]. Thus, the ease of hate speech dissemination online will simply contribute to mobilizing individuals and organizations to resort to violence, and it is extremely vital to detect and counter abusive content on religion in the name of national security and social cohesion.

In addition, the growing radicalization of religions, as well as the legitimization of extremist views on online platforms, is a threat of the present. The controversial cases of bail of terrorism suspects and the emergence of radical groups of individuals provide an environment in which the digital hate speech can flourish without any restrictions [6], [7]. What makes this even more complicated is the fact that, political instability, religious extremism, and a high level of social media usage interplay to make the situation even more difficult. The combination of these elements allows the spread of offensive remarks on the religious discussions online and thus the necessity to perform the automated detection both badly and urgently.

1.2 Key Factors in Detecting Abusive Content on Religion

In Figure 1.1, several important factors are identified used in detecting abusive religious text based on recent events.

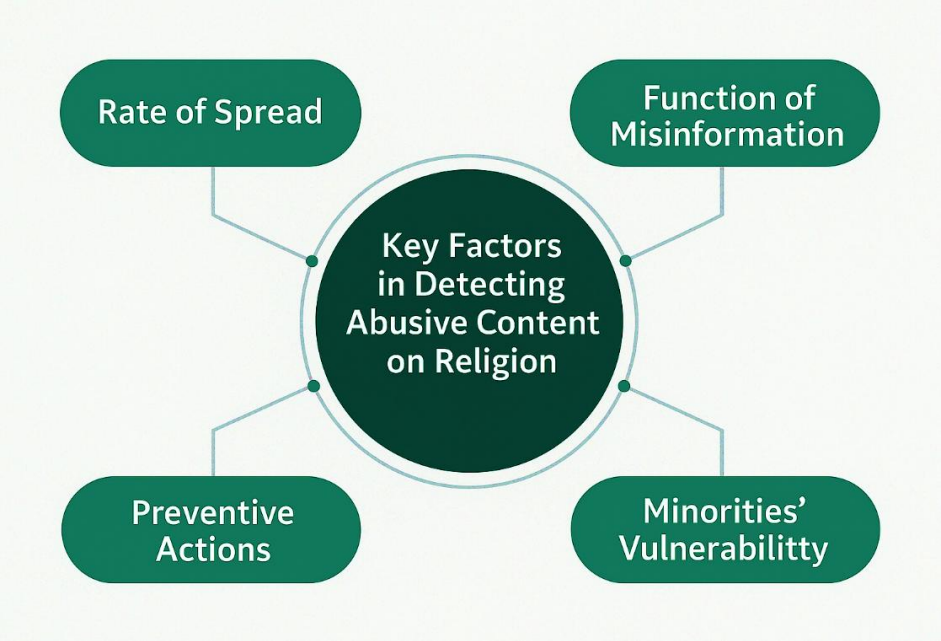


Figure 1.1: Key Elements Influencing Online Abuse Spread on Religion

- **Rate of Spread:** Malicious content is able to go viral within a short period of time, causing violence to reach an unprecedented level in real life.
- **Function of Misinformation:** Fake news and rumors, usually of religious nature, are purported to achieve fear and hatred.
- **Minorities' Vulnerability:** When tension is high, religious minorities are more susceptible to attack, both online and offline.
- **Preventive Actions:** Reacting to content is ineffective, it is important to act proactively and stop damages.

These are the direct concerns of our study, which has narrowed down its focus to the automated detection of such comments, and the aim of providing tools that will help to detect and react to these comments promptly, thereby creating a safer online and offline environment in Bangladesh.

1.3 Research Motivation and Objectives

Conventional methods of content moderation are often challenged when it comes to keeping up with a massive amount of content created by users and also the dynamic nature of harmful language. This has caused a giant interest in the automated solutions on the basis of machine learning and Natural Language Processing (NLP). Although a large number of studies have been done on the identification of hate speech in high-resource languages such as English, there is a major gap in literature regarding low-resource languages such as Bangla and particularly in terms of religious offenses.

This research is aimed to solve this gap by giving a detailed examination in identifying the religious abusive content in the Bangla language. Our work is focused on building and testing the advanced machine learning and deep learning models for the peculiarity of the Bangla language and context. We talk about the efficiency of Bi-directional Long Short-Term Memory (Bi-LSTM) network coupled with Multi-Head Attention mechanism which is a powerful deep learning mechanism that is known for understanding the long-range dependencies and contextual details in sequential data. In addition, we try to test the efficiency of this model with respect to other techniques that are relevant in the matter, learning from the most recent advancements in the areas of sentiment analysis and identification of hate speech.

The objectives of this research are:

- To conduct detailed analysis on the existing study, methodologies, and challenges of identifying anti-religious communication with reference to the Bangla literature.
- To deliver a deep learning-based detection framework, which includes Multi-head Attention and Bi-directional LSTMs approach in detecting the offensive remarks relation to religious topics.
- To validate the effectiveness and strength of the proposed model, test and compare it with the popular performance evaluation metrics, including accuracy, precision, recall, and F1-score.

- To evaluate abusive remarks on religious material using the suggested detection system in order to find language patterns, behavioral trends and important indicators.

1.4 Scope of the Study

The scope of this research includes the detection and analysis of the abusive comments on the religion in the Bangla textual content, particularly the social media platform namely Facebook. The subject is part of the bigger topic of Natural Language Processing (NLP) and Deep Learning with focus on low-resource languages, which have less linguistic and computational resources than English or other high-resource languages.

This research aims at developing a fully automated detection framework which is capable of reliably detecting religiously offensive or damaging comments in Bangla. The areas of research focus are on four areas are Figure 1.2.

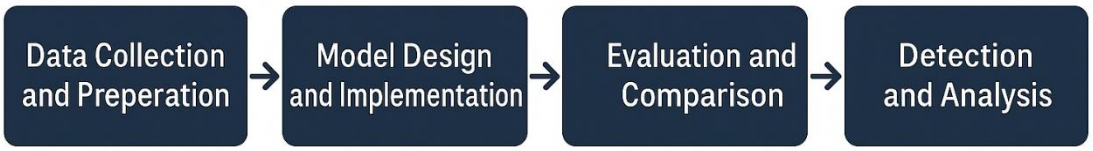


Figure 1.2: Key Research Areas of the Study

- **Data Collection and Preparation:** The data set is only the Bangla comments which is collected from the publicly available social media threads regarding religion. It is a lot of preprocessing, cleaning, labeling, and balancing to make sure that data is of high quality and fair. To make the classification not too difficult and possible to be interpreted easily, the remarks are classified in two groups: normal and religious abuse.

- **Model Design and Implementation:** In the study, a hybrid deep learning architecture is proposed, which is a combination of the bidirectional Long Short-Term Memory (Bi-LSTM) network and Multi Head Attention mechanism. The Bi-LSTM takes sequential dependencies and the meaning of words in context from the past and future in a phrase, and the attention mechanism brings the most influential words, which contribute to abusive context, to the forefront. The proposed method is

intended to go beyond the existing machine learning algorithms and the standard neural networks to understand complex patterns of linguistics in Bangla.

- **Evaluation and Comparison:** Training, validating and testing the model using traditional performance metrics like accuracy, precision, recall and F1-score comes under this scope. To prove the superiority of the suggested method, a comparative analysis of the proposed method with several baseline models like CNN-LSTM, plain LSTM, RNN, SVM, Logistic Regression and Random Forest is performed. The examination focuses more on overall performance as well as class level precision including a focus on the minority class.

- **Detection and Analysis:** This part of the research would be concerned with detecting trends, traits and language of religious offensive Bangla comments. After the model detects abusive contents the study analyses how such comments are created, what keywords or idioms are often used and how the tone of abuse differs from that of normal comments. It also looks at why the model is making a particular prediction, how the attention process draws the attention of important terms and what are the linguistic cues that are indicative of abuse. The point is to get a better scientific insight into the nature of abusive comments on religion in Bangla text, thus, pushing better NLP research for low-resource language (instead of creating real-world moderation solutions).

1.5 Structure of the Report

The remaining part of this report is subdivided into five chapters, and each chapter is dedicated to some specific aspect of the research in a systematic and comprehensive manner.

Chapter 2 provides a comprehensive summary of the past research works on sentiment analysis, hate speech detection, and misinformation classification, particularly, research related to the Bangla language. It reviews existing datasets, models, and techniques that are used in online abuse detection and points out the weakness of existing approaches in the need to take into consideration the linguistic and cultural peculiarities of low-resource languages. This chapter concludes with a brief account of the main research gaps that the given study is aimed at filling.

Chapter 3 is the methodology framework that is used in this research. It tells how the data is collected, labelled and preprocessed, and gives a general overview of the proposed Bi-LSTM with Multi-Head Attention model. In addition, it addresses the concept of tokenization, padding, training, evaluation criteria, and experimental framework used to test the success of a model.

Chapter 4 presents the results of the models that have been conducted and provides an analytic comparison of the effectiveness. Measures such as accuracy, precision, and recall, and F1-score are evaluated with the support of confusion matrixes and performance graphs. The results are compared to the traditional machine learning, deep learning, and transformer-based models to check the effectiveness of the proposed approach.

The fifth chapter is a conclusion of the report, which summarizes the main findings and contributions of the current research. It interprets the results of the research on the identified obstacles and presents the recommendations regarding the further research. The emphasis is placed on the existing attempt to broaden it by means of larger multilingual data, transformer models, and multimodal learning approaches to improve the detection accuracy, flexibility, and responsible AI implementation.

CHAPTER 2

LITERATURE REVIEW

Although the previous chapter focused on the analysis of the body of research related to the identification of abusive, hostile, and deceptive content on the online platforms, the current chapter is devoted to the discussion of the body of research focused on Bangla-language writing. It begins with the review of the influences of human feelings and emotions on the argument and the internet communication, especially the issues of religion and cultural identity. Having reviewed the history and progress of sentiment analysis, it examines studies on hate speech, cyberbullying and detecting misinformation, highlighting the computational approaches, datasets, and language problems involved in these aspects.

In order to summarize the findings of key research papers, a comparative summary (Table 2.1) introduces the results of the techniques, model operation and limitations of various studies. In the chapter, the authors also analyze the current deep learning and transformer-based methods, their advantages and disadvantages of using them in low-resource languages such as Bangla.

Lastly, the chapter addresses critical research gaps in the area, including how to work with cultural nuances, sarcasm, and contextual meanings; how to work with an uneven amount of data; and how to guarantee the model interpretability and flexibility. It then ends by describing the research contributions that the current study aims to make to overcome these limitations and contribute to the further research of automated detection of abusive religious content in online contexts through the proposed model of Bangali religious abuse, based on the Bi-LSTM with Multi-Head Attention.

2.1 Human Sentiment on Religion

Religion has far-reaching implications on the thoughts, emotions, and behaviors of human beings. It has an impact on moral values, sense of community, and social ties and

serves as a unifying and dividing factor of online relationships. Emotions regarding religion are complex, including admiration and devotion, scepticism or hatred- all these mixed feelings are often manifested in social media communication. In the virtual world, these emotions can easily get out of control, the web community can easily escalate feelings and ideas that otherwise would remain intimate or localized.

Alam et al.[8] say that the discussion of religion on social media can go further than emotional polarity (positive or negative), and involve elaborated emotional shades depending on cultural and contextual influences. This makes the analysis of the sentiments in the religious discourse exceptionally challenging, because the algorithms have to identify constructive criticism, neutral remarks, and offensive words. In 2024, Islam et al. [9] emphasized that automated systems should be carefully fine-tuned to prevent religious sensitivities in that words which are offensive in a given culture could mean something different in a different culture.

In 2021, Karim [10] pointed out in 2021 that language on religion has a tendency of being subtle as employed on the drug of abuse through sarcasm, metaphor, or coded language as opposed to blunt attacks. The complexity of this level necessitates models which are able to capture contextual meaning that extends beyond the mere superficial keywords. Kafi et al. [11] further pointed out that idiomatic phrases and allusions to culture that have high emotional appeal are also readily employed by the users of the Bangla social media and this contributes to an increase in complexity in machine comprehension. These linguistic and cultural peculiarities make the definition of the abusive statements about religion in Bangali a rather sensitive and technically problematic activity.

Basically, human emotions about religion are emotionally deep, dependent on the situation and are strongly interconnected with identity. In the case of computational systems, understanding of this sentiment will require not only advanced tools of deep learning but also cultural intelligence the capacity to decode emotion, faith and desire in ways that are socially acceptable and linguistically subtle. It is quite crucial to recognize these dynamics to develop context-sensitive, just, and ethically responsible structures that will enable the identification of religiously abusive language on the internet [8-11].

2.2 Sentiment Analysis and Its Applications

Sentiment analysis, also known as opinion mining, is a computation technique applied to identify and interpret the sentiment, attitude or perspective expressed in a text. It is a necessary sub-discipline of Natural Language Processing (NLP) and Artificial Intelligence (AI) enabling computer programs to read emotions of people in written communication. Sentiment analysis is commonly used in such fields as marketing, politics, entertainment, and social media analytics and helps organizations to measure the attitude of the people, as well as make decisions on the basis of data.

According to a report by Alam et al. [8], the widespread adoption of social media sites has transformed sentiment analysis into an important tool of understanding the behavior of the masses since it is capable of processing large amounts of unstructured text in real-time. Traditional sentiment analysis methods relied heavily on lexicon-based and statistical machine learning systems, including Naive Bayes, Support Vector Machines (SVM) and Logistic Regression, the primary use of which is to give text a positive, negative, or neutral label.

To overcome these limitations, scholars have used a mixture of deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which have contributed to better learning of complex patterns with large bodies of text. Bidirectional LSTM (Bi-LSTM) was developed and enhanced the contextual understanding by processing data in two directions, that is, forward and backward direction, thus making the sentiment prediction more accurate [12].

More so, the transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) and its language-specific analogues such as BanglaBERT have transformed sentiment analysis by introducing the capabilities of contextual embeddings and attention scores to direct a model to focus on important sections of the text [9]. These models have shown themselves to be highly effective to enhance performance in highly linguistic conditions especially those low resource languages as in the case of Bangala language, where context and morphology is highly dynamic [13].

Sentiment analysis can be applied in various ways. In marketing, it is applicable in gauging the customer comments and product reviews. It is a reflection of voter preferences and trends in the state of the political opinion [14]. It helps one identify the fake news in journalism and the media and detects hate speech, analyzing the reaction of the population to events. In addition to these domains, the sentiment analysis can also be used in mental health monitoring, financial forecasting, and social risk assessment, suggesting that it is cross-disciplinary.

Nevertheless, there is not the absence of challenges, and one of them is the inability to distinguish between sarcasm, irony, cultural peculiarities, and ambivalent emotions, which is especially problematic in social networks. According to Alam et al. [8], language and cultural elasticity is a critical question in sentiment analysis research. As a result, it is still significant to use context-sensitive deep learning and transformer-based models that can be improved to ensure the accuracy and the flexibility of sentiment classification systems to another language and other fields.

2.3 Research on Online Abuse, Hate Speech, and Misinformation Detection

It has enabled unhindered communication and start of a mass dialogue but it has also led to a rise in online harassment, hate speech, and misinformation. Such incidences affect profoundly on digital harmony and social well-being particularly in nations such as Bangladesh where religion and politics are inseparable with identity. The pursuit of identifying such malicious content has been a burning research topic in Natural Language Processing (NLP) and Artificial Intelligence (AI), an attempt to ensure that the internet environment is safer and more polite.

2.3.1 Overview of Online Abuse Research Works

Online abuse is a mean or offensive speech targeting individuals or groups of individuals, which is founded on factors like their religion, ethnicity or gender. There are a number of studies which have tried to solve this problem with the help of computational modeling.

In 2020, Islam et al. [15] developed machine learning algorithms to detect cyberbullying on social media, as the damaging posts on social media are identified by classifiers such

as Support Vector Machines (SVM) and Random Forests. On the same note, Vivekananth and Sharma (2025) [16] suggested an NLP classification system that makes use of tokenization, feature extraction, and the best machine learning models to identify bullying language in Twitter text. A detailed overview of machine learning methods of hazardous online behavior provided by Shrestha et al. [17] encompasses CNN, LSTM, and BERT-based hybrid models, and highlights how deep learning is effective in identifying violence online.

Explainability and linguistic adaptation are also issues that have been explored lately. Karim [10] introduced DeepHateExplainer, an interpretable deep learning model, which is used to identify hate speech in Bengali, built using BanglaBERT, mBERT and XLM-RoBERTa. Their results were high-quality and they showed that transparent AI is necessary in very sensitive industries like hate speech regulation. In 2025, Kafi et al. [11] introduced BOISHOMMO, an all-purpose solution to Bangali hate speech classification based on the use of classic machine learning models, including Random Forest and SVM on a multi-label task that addresses annotation bias and cultural diversity.

2.3.2 Overview on Cyberbullying and Hate Speech Detection Works

Most of the negative communication that is displayed on the internet is hate speech and cyberbullying, which are the most harmful and long term. Both go to the use of language to assault, disparage, or discriminate individuals, or groups of individuals, usually basing their argument on their religion, ethnicity, gender or political affiliation. Automatic detection of this content has been identified as one of the major interests of Natural Language Processing (NLP) studies due to its societal implications and language issues.

Initial investigations on the subject were based on machine learning (ML) algorithms. The Support Vector Companies utilized by Islam et al. [15] are in form of the classifier and are able to detect cyberbullying in the social media networks like the Decision Trees and the application of random forests. These models did not necessarily score high in identifying textual or sarcastic connections even though they worked well in simple text classification. Vivekananth and Sharma [16] added more precision to the detection using the assistance of an NLP based classification system, consisting of tokenization, feature

extraction and hyperparameter optimization to classify the tweets in relation to bullying in a more improved manner.

As the deep learning developed, scholars started to investigate those neural structures that could represent semantic relationships. Awal et al. [18] designed an emotion/target co-learning multitask (AngryBERT) model, which also demonstrated a major breakthrough in hate speech recognition. This research indicated the significance of the emotion tone to separate hate based contents. Likewise, Krak et al. [19] integrated the principles of neural network with visual analytics and applied Local Interpretable Model-Agnostic Explanations (LIME) to enhance the level of transparency of the AI predictions.

Besides, there are several efforts, which have led towards advancement in the circumstances of Bangla language. Recently, in 2021, Karim [10] wrote DeepHateExplainer, an explainable transformer model, to support under-resourced Bengali text. It realized significant scores in various categories of hate by integrating the BanglaBERT, mBERT and XLM-RoBERTa and fulfilled the requirement of AI interpretability. Later in 2025, Kafi et al. proposed BOISHOMMO, a complete hate speech detector based on machine learning classifier algorithms, including Random Forest and SVM [11] with the use of the Bangali language. It used ten types of hate in their dataset, which consisted of the social media and internet news, religious was the most prevalent category of hate with an epic 95% F1-score of religious hate speech.

In order to create and enhance the quality of the data and the reflection of culture, Tasnim [20] created BASED (Bangla Adverse Sectarian Expressions Dataset) that is dedicated to hate speech based on religion and community. Transformer models have been shown to work in low-resource environments, with the application of BangaliBERT on the task of multilabel classification of violent and sectarian speech. Ahmed et al [21] then proceeded to use Recurrent Neural networks (RNNs) and Gated Recurrent Units (GRU) to detect and identify political hate articles in Bangali social media and attained a rate of 88.28.

Recent studies also study multilingual adaptability. Shrestha et al. [17] gave an excellent description of the multilingual cyberbullying detection using CNN, LSTM, and hybrid BERT+SVM on a variety of datasets in English, Hindi, Turkish, and Bengali. These

literary sources support the benefits of deep learning and transformer models on the traditional ML, but require huge amounts of labeled data and computation power.

Notwithstanding this development, there are still immense obstacles. The current systems are still grappling with sarcasm, figurative expressions and context interpretation particularly when cultures are subjective issues such as religion or politics. The asymmetry of information is a significant drawback that results to the biased models. Also, there are the ethical concerns, such as the privacy of data, its interpretability and fairness, which should be considered in order to deploy hate speech detection systems with responsibility [11, 17, 18, 20].

2.3.3 Overview on Fake News and Misinformation Detection Works

The misinformation and the fake news spread has become a sharp problem in the world due to the influence of the social media. The appeal to the masses through the false news could influence the manner in which individuals think, create a social conflict, and control political or religious opinion. The concept of machine learning (ML) and deep learning (DL) in detecting and countering disinformation is, thus, noteworthy, as long as information integrity and digital security is involved.

The initial studies were to a large extent based on the classical machine learning algorithms which applied linguistic and statistical attributes in determining the authentic and the fake information. Rahman et al. [22] tried a lot of models, and among them, the Logistic Regression, the Random Forest, Naive Bayes, and CNN were considered and converted to soft and hard voting ensembles. Their ensemble models acquired the high precision of 99.93 percent on the benchmark datasets that represents the advantage of applying a mixture of numerous classifiers.

Among the models formulated to identify fake news in Bangali are a machine learning-based model, and deep learning-based model as one model developed by Das et al. [23]. They have implemented the feature extractors TF-IDF and word embeddings in their system that were trained on Extra Trees, Random Forest, and LSTM algorithms. Their LSTM model got over 86 per cent of the accuracy of Bangali datasets such as the BANS and the BNLPC3, which serves to suggest that the deep learning is highly effective in terms of understanding the context of the limited resource languages.

Alongside Bangla, Ljubi et al. [24] confirmed the comments on the Croatian social media and introduced croBERT, a transformer-based model, that was more efficient than standard algorithms (e.g., SVM and Random Forest) with the accuracy of 89.56. They found that transformer architectures are able to understand semantic and syntactic idiosyncrasy, which can barely be identified by a simple classifier. On the same note, Fajinmi and Joseph [25] investigated the misinformation detection in Afan Oromo through the CNN-LSTM-XLM-RoBERTa network to address the multilingualism and morphologically compound language problems.

Other steps of this direction were also made by Defersha et al. [26] who developed multilingual hostile linguistic materials in Afan Oromo, Tigrigna and Amharic. They tried to enhance the efficiency of multilingual hate and disinformation detection by integrating topic modeling (BERTopic), Word2Vec and deep learning-based transformers. Their results indicated the significance of cross-lingual flexibility that is required in case of the misinformation identification in multi-linguistic and dialectic settings.

In Bangladesh, misinformation is also quite commonly linked to the religious and political biases; hence, it is hard to locate. This has been tackled by Islam et al. [9] who designed transformer based datasets that do not compromise religious sensitivities over the internet. Their design, hatebnBERT, achieved above 98 percent accuracy and the fact that custom transformer models could be designed to suit socio-cultural interests, such as as religious disinformation and incitement.

Despite the achievement of these, there are numerous challenges that are present. Many studies are founded on English data and limit the applicability of cross-linguistic models. Sarcasm, a satire or culture communication that carries the wrong information can barely be determined in the right manner. In addition, it is costly to apply great models in the low resource areas due the massive calculations of the models and the skewed dissemination of the data. Rahman et al. [22] and Das et al. [23] noted that transformers are more appropriate than the traditional models, but they consume a tremendous number of computational resources and tagged data.

2.3.4 Comparative Summary of Related Works

In order to sum up, Table 2.1 is a comparative summary of major research on cyberbullying, hate speech, and misinformation detection in multiple languages and contexts. It includes the briefing of such important aspects as datasets, methodology, achieved results, and identified Limitations. The table depicts that the use of deep learning and transformer-based architectures that are superior to regular models is constant and requires additional computer resources and annotated data.

Table 2.1: Summary of Reviewed Related Studies

Ref No.	Title	Dataset	Methods	Output	Limitations
[8]	Sentiment Analysis in Social Media: How Data Science Impacts Public Opinion Knowledge Integrates Natural Language Processing (NLP) With Artificial Intelligence (AI)	Multi-platform	CNN, LSTM, BERT	BERT improved 8–15%	Poor sarcasm detection
[9]	An Innovative Novel Transformer Model and Datasets for Safeguarding Religious Sensitivities in Online Social Platforms	16,000 Bangla texts	RNN, LSTM, BERT variants	hatebnBERT: 98.8%	Biased annotations
[10]	DeepHateExplainer: Explainable Hate Speech Detection in Under-resourced Bengali Language	Bengali hate data	Explainable BERT	High interpretability	Resource intensive

[11]	BOISHOMMO: Holistic Approach for Bangla Hate Speech	2,499 Bangla comments	RF, SVM, LR	86% Macro-F1	Small dataset
[12]	Comparative Study of Transformer-Based Models and Bi-LSTM for Bangla Sentiment Analysis Using Hybrid	35,000 Bangla texts	Bi-LSTM, BERT, RoBERTa	RoBERTa: 93.6%	High computational cost
[13]	Bangla-Senti: A Large-Scale Corpus for Sentiment Analysis in Bangla and its Applications	536,930 YouTube comments	SVM, NB, CNN, DNN, BiLSTM	CNN: F1 = 0.86	Limited to YouTube
[14]	Political sentiment analysis using natural language processing on social media	14,223 YouTube comments	SVM, LR, RF	SVC: 91.18%	Data imbalance
[15]	Cyberbullying Detection on Social Networks Using Machine Learning Approaches	10,254 Bangla comments	ML + DL models	LSTM: 99.8%	Limited to Bangla
[16]	Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework	47,692 tweets	TF-IDF + ML	LGBM: 83.82%	Limited to Twitter

[17]	Machine Learning for Identifying Harmful Online Behavior: A Cyberbullying Overview	Multi-source	CNN, LSTM, BERT	CNN: 99.86%	Sarcasm/emoji issues
[18]	AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection	Multiple tweet sets	Multitask BERT	90.7%	Dataset dependency
[19]	Method for Neural Network Cyberbullying Detection in Text Content with Visual Analytic	Cyberbullying dataset	NN + Visual Analytics	95.65%	High resource demand
[20]	Mapping Violence: Bangla Sectarian Expression Dataset	13K sentences	BanglaBERT	36.6% (4-class)	Class imbalance
[21]	Automatic Identification of Political Hate Articles from Social Media Using RNNs	1,980 Bangla texts	LSTM, GRU, SVM	88.28% Accuracy	Small dataset
[22]	Fake News Detection: Soft and Hard Voting Ensemble	ISOT dataset	BERT + Ensemble	99.93% Accuracy	High computational demand
[23]	Social Media Bangla Fake News Detection Using Deep and Machine Learning Algorithms	BANS + BNLPC3	ML + LSTM	LSTM: 86%	Unbalanced dataset

[24]	Detecting Disinformation in Croatian Social Media Comments	FRENK 1.1 dataset	SVM, croBERT	89.56%	Language-specific limits
[25]	Machine Learning Approaches for Detecting Fake News in the Afan Oromo	Authentic news data	CNN, LSTM, XLM-R	XLM-R: 95%	Data scarcity
[26]	Topic Words-Based Multilingual Hateful Linguistic Resources	59K cross-language	CNN, LSTM, BERT	CNN+Word 2Vec: 96%	Complex semantics
[27]	AI-Powered Social Media Monitoring: Leveraging NLP for Real-Time Cyberbullying Detection on Twitter	1.2M English tweets	Logistic Regression, NB, MLP, BERT	BERT: 91.7%	Cultural diversity issues
[28]	Machine Learning and Deep Learning-Based Approach to Categorize Bengali Comments on Social Networks	94,000 comments	ML + DL fusion	Hybrid ML: 99.34%	Focused only on Bangla

2.4 Existing Models and Approaches

Online abuse, hate speech, and disinformation detection has been advanced in many different computational models, including machine learning methods to state-of-the-art deep learning and transformer systems. The successive generations of models made a unique contribution to the best text understanding, contextual interpretation, and

performance metrics, but have also revealed the serious drawbacks, especially in the case of low-resource languages, such as Bangla.

2.4.1 Machine Learning Approaches

Initial research mainly used the supervised machine learning (ML) models using manually crafted linguistic and statistical features, including Term Frequency -Inverse Document Frequency (TF-IDF), Bag-of-Words (BoW), and n-grams. The algorithms, which have been used in popular use in classifying text into common categories, include Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), and Random Forests (RF) [15, 16].

These models were explainable and effective with small data but lost semantic associations as well as contextual information. To illustrate the point, Kafi et al. [11] investigated the classification of the Bangali hate speech as the component of the BOISHOMMO system with the help of Random Forest and SVM and obtained the high F1-scores when it comes to certain categories, such as the abusive speech on religion. But the case was worse with a sophisticated language cue, sarcasm, or the switching of languages. On the same note, Ahmed et al [21] applied SVM and GRU to identify the political hate articles in Bengali but they had difficulties in generalizing because of the low and biased quantity of data.

2.4.2 Deep Learning-Based Approaches

The emergence of deep learning (DL) changed how text is classified by enabling models to learn through data by extracting semantic and syntactic relationships in data automatically. Architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) in particular, significantly increased the accuracy of detections. These models identified long-term dependencies and context in text and were therefore effective in detecting emotion-based or implicit hate speech.

To detect fake news in Bangali, Das et al. [23] applied LSTM networks, achieving an accuracy of 86, whereas Uddin et al. [28] evaluated CNN, CLSTM, and RNN, highlighting the advantages of hybrid models. AngryBERT introduced by Awal et al.

[18] is an emotion aware model that incorporates Bi-LSTM with multitask learning, enhancing the detection of abusive and hateful text by jointly learning to model emotions and target sense.

Bidirectional learning has enabled bi-LSTM models to be more effective in text classification. In a study on Bangla sentiment analysis, Ethan [12] compared the Bi-LSTM and transformer models and found out that hybrid optimizers and contextual embeddings lead to an improvement of F1-scores by up to 7%. However, such methods of deep learning often require large computational resources and require substantial resources with high precision annotation, which are scarce in Bangla and other lower-represented languages.

2.4.3 Transformer-Based Approaches

Transformer architectures have brought a novel stage in the study of NLP. Unlike recurrent models, transformers use self-attention mechanisms to obtain the dependencies across full sequences, making it easier to have a deeper contextual understanding. BERT, mBERT (Multilingual BERT), BanglaBERT and XLM-RoBERTa are models that have achieved the state of the art performance in sentiment and hate speech detection.

HatebnBERT by Islam et al. [9] is a specialized transformer that has been trained by data of religious sensitivity, achieving 98.8% accuracy performance on binary classification. Karim [10] utilized different models of transformer into their DeepHateExplainer system on the low resource Bengali language, using BanglaBERT, mBERT and XLM-RoBERTa to enhance explainability. Similarly, Ljubi et al. [24] used croBERT in the detection of disinformation in Croatian, and the results far exceeded traditional ML models.

Transformer-based models have proven to be cross-linguistic. Defersha et al. [26] developed multilingual tools of hate speech on BERTopic and Xlm-R and achieved considerable results on different African languages. However, these models are very resource-intensive and require large amounts of data to fine-tune, and are harder to interpret than traditional ML mechanisms.

2.4.4 Hybrid and Explainable AI Approaches

Recent studies have looked at the hybrid structures in which different types of models or approaches to learning are integrated. Rahman et al. [22] employed ensemble techniques (soft and hard voting) to leverage the advantages of both ML and DL classifiers for identifying fake news, reaching almost flawless performance (99.93% accuracy). Shrestha et al. [17] proposed hybrid models combining CNN + LSTM and BERT + SVM, which would have better generalization when used in a multilingual environment.

To address the problem of transparency, such scholars as Krak et al. [19] introduced visual analytics and LIME (Local Interpretable Model-Agnostic Explanations) to neural networks, which contributed to the improved interpretability in detecting cyberbullying. The value of explainable AI (XAI) has been emphasized by Karim [10] as a way to ensure fairness and trust in the results of hate speech classification, at least in such sensitive areas as religion and politics.

2.5 Identified Research Gaps

Although massive progress is made in sentiment analysis and hate speech detection, and misinformation classification, especially with the use of deep learning and transformer architectures, several significant gaps exist, which restrict the usefulness, fairness, and flexibility of such systems. These loopholes highlight the problems of linguistic heterogeneity, contextual perception, data accuracy, model explainability, and responsiveness to the dynamism of the online dialogues.

2.5.1 Limitations in Handling Linguistic and Cultural Diversity

One of the most important issues which are brought to light in the contemporary literature is limitation of computational models to deal with linguistic and cultural diversity. Most sentiment analysis and hate speech detection models are originally trained and developed with English-language datasets, which comprise rich linguistic content, large corpora and consistent syntax. When applied to low-resource languages, as it is the case with Bangala, the performance drops significantly due to the differences in grammar, morphology, idiomatic phrases, and cultural backgrounds.

As Alam et al. [8] indicated, diversity in languages plays an important role in the display of feelings and views in online communication. The words or phrases that may appear to be neutral in one culture may carry serious emotional or religious content elsewhere. This inconsistency makes direct translation based transfer learning unreliable, because it can change the meaning and misclassify the sentiment or intention. Islam et al. [9] noted that despite the ease that the use of multilingual transformer models such as mBERT and XLM-RoBERTa have enabled the ability to make cross-lingual adaptation, it still faces the challenge of understanding cultural nuances that are vital in interpreting hate speech in Bangla texts.

Kafi et al. [11] observed that the users on the Bangla social media often connect the formal use of the Bangla language, the conversational language, and English code-switching, which leads to lingual complex and informal written forms. The language mixing, coupled with a frequent use of slang and figurative language, makes the process of automatic detection of abusive content much more complicated. Models that are trained with standard text have the problem of not comprehending differences in informal or dialectal text correctly. Similarly, it has been observed by Defersha et al. [26] that when contrasting multilingual hate speech is concerned, there are challenges which emerge when the languages involved are characterized as being rich in their morphology and the cultural affiliations, where the same linguistic structure may have very different interpretation in the different context of social setting.

Besides, culturally sensitive data is further complicated by annotation. Annotators can also be biased and inaccurate due to the influence of personal, religious, or political beliefs on labeling decisions, and this leads to inconsistency and bias in datasets. This especially applies to matters touching on religion in which a single statement can be construed as criticism, sarcasm or insult depending on the annotator. As a result, the dependability of the data set and cross-cultural consistency becomes a significant problem of model effectiveness.

2.5.2 Challenges with Sarcasm and Contextual Understanding

The recognition of sarcasm, irony and meaning, which are context-dependent, is one of the most demanding in sentiment analysis and hate speech as well as content abusiveness.

Sarcastic or ironic statements used can pass across the reverse of what they actually mean unlike direct displays of anger or emotions. Such ambiguity of the computational models rendered by the linguistic phenomenon is that they rely mostly on the lexical indications that can be found at the surface, which causes an enormous number of misclassifications.

Ethan [12] noted that even super-capability deep learning systems such as Bidirectional Long Short-Term Memory (Bi-LSTM) and transformer models such as BERT encounter issues with figurative language, which depends on the tone, intent, or prior discussion. To give an illustration, the expression What a saint! may be used with an admiration or a scorn depending on the context, which the contemporary models cannot be able to interpret without a pragmatic understanding.

Awal et al. [18] addressed this shortcoming by introducing AngryBERT, a multitask learning architecture that concurrently acquires emotion and target semantics. Their study established that the use of emotional tonality helps in identifying subtle cases of hate speech or sarcasm. However, these models still heavily rely on labeled data that is a clear indication of sarcasm or emotion, and that in low resource languages like Bangali is scarce. Similarly, Karim [10] found out that even explainable models such as DeepHateExplainer often fail to detect indirect aggression or hidden animosity, as sarcasm can appear polite on the surface and have a latent aggressive agenda.

Sarcasm and irony create strong challenges in the social media of the Bangali language because the users often express dissent or criticism through humor, idioms, and the alternation between the usage of the Bangali and English language. According to Kafi et al. [11], the speakers of Bengali language apply metaphors, religious symbols, and proverbs that convey meanings which are not directly provided by the words themselves. An example would be, sarcastic comment can indirectly raise the issues of religion and the model may misinterpret this sarcastic comment to be harmless. This demonstrates that one must have a form of understanding of cultural semantics and cultural signals of context to determine the presence of sarcasm and not the patterns of language.

Another interesting obstacle is the problem of contextualization. Most of the models consider text on its own, without considering the context of the conversation, who is speaking, or past communications. Alam et al. [8] emphasized the fact that human

sentiment is very much context-dependent; a statement said by one can be perceived as funny or offensive depending on the speaker, the audience, and the situation. The inability to combine this discourse-level understanding may cause even the most successful models such as BanglaBERT and mBERT to make wrong forecasts.

2.5.3 Data Imbalance and Dataset Limitations

The problem with the research of sentiment analysis, hate speech, and misinformation detection remains that it is impossible to balance the data and cannot be accessed to the particular datasets. Machine learning or deep learning models would greatly depend on large, diverse and well-labeled datasets in order to generalize. In the social media data in the real-life scenario, a minority group of comments is usually abusive/hateful and this leads to the data becoming highly skewed in favor of non-abusive or neutral data. The consequence of this difference is that models assign greater weight to the majority class and lesser weight to sensitivity and recall of the minority (abusive) class which are typically the most sought after target of detection systems.

Tasnim [20] has marked this problem in their BASED (Bangla Adverse Sectarian Expressions Dataset) and revealed that certain forms of hate such as ethno-communal hate or non-denominational hate were less represented compared to more common forms of hate such as religio-communal hate. On the same note, Ahmed et al. [21] were forced to overcome challenges in their dataset of Bengali political hate articles due to low and skewed representations of classes that constrained the scope of the model to generalize in other areas. These variations lead to biased classifiers, which are also highly sensitive to the common class, which are poor at detecting the abuses that are rare or subtle.

Another issue that makes the reliability of a dataset complicated is quality of annotations. In the Kafi et al. [11],it was found that there are changes in the inter-annotator agreement suggested by the changes in the scores of the Kappa across all categories of hate, that imply the subjective and inconsistent nature of human annotators. This stereotype is usually prevalent when the notion of hate or offence between people will be in different definitions due to cultural, political or religious affiliations. Consequently, a particular annotator could label the same remark as hate speech and, conversely, as criticism, which decreases the objectivity and precision of the data.

The small size of the available Bangali datasets also constrain the use of the more complex structures like transformer-based models because they need a lot of training data to attain a high degree of accuracy. Islam et al. [9] and Karim [10] also observed that Bangla is an underserved language in NLP which lacks much high-quality corpus that can be used to define hate speech and abusive uttering on religion. Although such resources as Bangla-Senti [13] and BOISHOMMO [11] provide useful data sets, they are not large enough to reflect the variety of sentiments and dialects peculiarities in various regions and among the users.

Additionally, some of the studies like Das et al. [23] and Uddin et al. [28] are based on the data gathered with help of limited resources e.g., Facebook, Twitter or YouTube and are thus biased by platforms. As an example, the linguistic style and the degree of aggression on Facebook are different compared to the YouTube which usually has short and emotional comments. Deviations of cross platform diversification restrict the stability of trained models in new or unfamiliar social environments.

2.5.4 Computational Complexity and Model Interpretability

The accuracy of sentiment analysis, hate speech detection, and misinformation classification has been greatly enhanced due to the high application of deep learning and transformer-based models. Nonetheless, such improvements are highly computationally expensive and interpretable and this is a significant cost to date particularly in low resource settings and other socially sensitive uses like the identification of religious or political hate speech.

Transformer models (especially, BERT), mBERT, BanglaBERT, and XLM-RoBERTa model consume much data and computation time to train and optimize. As shown by Islam et al [9], though their hatebnBERT model had high accuracy of 98.8% in the binary classification of religiously sensitive Bengali text, it was extremely memory-demanding and also required the use of the GPU, which restricted its applicability to small institutions and by individual researchers. Equally, Ethan [12] discovered that transformer-based frameworks outperformed the Bi-LSTM models in sentiment analysis of Bangali, but require significantly more training time and energy which were issues related to scalability and sustainable deployment.

Training is less expensive than the cost of calculation. To scale large-scale models to a real-time system like social media monitoring or content moderation, there is a need to make inferences on large quantities of text at the time, which incurs increasingly high latency and power consumption. Karim [10] highlighted this to make the transformer models unsustainable in identifying the actual abuse particularly in low-resource conditions where hardware constraints are often constrained. It is also complicated by the fact that fine-tuning and optimization of hyperparameters in the model is an unsolvable problem, the skills and computer capabilities of which demand certain capabilities.

The other significant obstacle is model interpretability as well besides the computational requirements. Most deep learning models are usually described as black boxes since they make correct inputs, but do not give any message on what methods are used to get the inputs. This vagueness poses issues of ethics and accountability in delicate matters such as whether or not there are abusive remarks on religion, e.g. in a system with data that it sees as hate speech but it is not well articulated as to why. To address this problem, Karim [10] designed their DeepHateExplainer that covers explainable AI (XAI) schemes like Layer-wise Relevance Propagation (LRP) to demonstrate and explain how hate detection systems are trained to take decisions. Nevertheless, in the majority of studies that are devoted to transformers, XAI is not fully used.

Krak et al. [19] also improved interpretability by applying visual analytics to detecting cyberbullying with Local Interpretable Model-Agnostic Explanations (LIME) to enable word-level interpretability on the predictions. Their results have shown that the interpretability tools improved the trust in AI systems and detected bias in model results. Nonetheless, the majority of the explainable frameworks possess an extra cost in calculation and therefore, are intricate to expand to the large scales.

Also, Alam et al. [8], and Defersha et al. [26] emphasized the fact that the intricate transformer models can transfer concealed prejudice of training data and proceed with stereotypes and false links, especially with multilingual or religious textual content. In the absence of tools of clear interpretability, it is still a significant problem to locate and address these biases.

2.5.5 Dynamic Nature of Online Content and Model Adaptability

The constantly changing online communication landscape turns out to be a considerable obstacle in creating efficient and effective systems that could help detect the hate speech and false information. The social media language is dynamic and is dictated by trends, new slangs, fluctuating political climate and culture. As users will always have new ways with which to post opinions at one time or another to avoid some form of moderation based on a pre-programmed data set, new trends of the harmful or corrupt words will be located less and less often with time.

Alam et al. [8] have indicated that the human emotions and relationships on the online platforms are dynamic and circumstantial. Even the words, which were perceived not to be negative, may be given overtones when the sociocultural norms change. On the same note, hashtags or terms can come out fast based on a given political or religious event that changes the sense of the online discussions. Such changes can be a challenge to the performance of the statistic models that are dependent on past datasets hence continuous learning and dynamism with time is relevant to maintain the performance.

According to Shrestha et al. [17], cyberbullying and hate speech detection models tend to suffer a temporal decay i.e. with the evolution of online language, they become less accurate. Their results have shown that even the most successful models including CNN- and BERT-based ones deteriorate in a few months once they are put into practice unless they are regularly updated. This reduction has been observed in particular in low-resource languages such as Bengali (Bangla), where digital vocabulary is increasing tremendously via the processes of code-switching, local dialect and social memes. Indicatively, what would be regarded as an amusing joke in a particular context can be used in negativity or political context later and the model would need to be aware of these new associations.

Islam et al. [9] also observed that models, which are trained on religious sensitive dataset, also share the same issues in the adaptation process as the direction and the essence of religious discourse in social media change according to ongoing events. The training data does not reflect this changing nature and hence missed hate contents or false positives will be made. Similarly, Rahman et al. [22] reported that the ensemble-based fake news

detector is typified by time variability as the measure of strategies to deceive the users intensify and apply weaknesses in language or topic in AI systems.

Another challenge to the flexibility is associated with the absence of online education and superior training patterns. Most of the present systems in place presuppose periodical retraining at the beginning as a tedious and resource consuming process. It is not a good strategy that will not allow quick response to the new threats like viral hate-campaigns or planned misinformation. Defersha et al. [26] suggested to dynamically determine the changing topics and refresh the models with the minimum necessary retraining by applying multilingual adaptive learning and topic modelling approaches (including BERTopic). On the same note, Karim [10] indicated the advantage of federated learning in which the models are trained using decentralized sources of data, excluding centralization of sensitive user data, which enhances flexibility and privacy.

2.6 Chapter Summary

This chapter gave a general overview of the available literature that is related to sentiment analysis, hate speech detection, cyberbullying identification, and detecting fake news or misinformation, particularly, the research of the Bangla language and other environments with limited linguistic resources. The presentation was grounded in the designing of computational techniques because of traditional machine learning techniques like Naïve Bayes and SVM, on top of deep learning techniques, like CNN, LSTM, and Bi-LSTM, and lastly on transformer-based ones, like BERT, BanglaBERT, and XLM-RoBERTa.

The articles reviewed also showed significant progress in the area of automatic detection of offensive, hate and deceptive material. Nevertheless, the analysis indicated that gaps in research are still present, such as the fact that the existing models lack the ability to capture linguistic and cultural nuances, have problems with the sense of sarcasm, irony, and contextual meaning, and have the disadvantage due to biased datasets and annotations. In addition to this, severe computational complexity, inferior interpretability and dynamically evolving online interaction still confront implementation of these systems into practice.

Besides, the gathered evidence that even with the increased accuracy level, deep learning and transformer models often demand enormous suffer of computer power and the

masses of annotated data that is typically unattainable in the context of the Bangla NLP field became clear in the comparative summary (Table 2.1). This has resulted in an unending demand of models that are precise, versatile and understandable possessing the added benefit of being productive and culturally sensitive.

By considering the gaps that are identified, it can be seen in the following chapter (Chapter 3: Research Methodology) that the strategy applied to this research is outlined. It includes the description of the creation of the dataset and the preprocessing of the data, the architecture of the proposed Bi-LSTM with Multi-Head Attention model, and the measures to evaluate its performance and the framework applied in the experiment in order to evaluate its performance. The specified methodological framework directly responds to the limitations that have been indicated in the current literature and will target the development of revealing the abusive remarks on religion and manipulation in the text of the social media in Bangladesh.

CHAPTER 3

METHODOLOGY

This chapter explains the methodological approach that was applied in developing and evaluating the proposed system to detect religiously offensive Bangla comments. It begins with data collection procedures, which explain where the social media comments are collected, their selection criteria as well as manual annotation process, explaining how items have been categorized as either Normal or Religious Abuse. The method of preparation that involves cleaning, tokenization, wiping out of stop-words, and normalization of text is also mentioned to show how raw data was transformed into a format that could be used to train models.

The next chapter is on the Bi-LSTM model with Multi-Head Attention in which the chapter shows how the two elements can capture contextual meaning and emphasize on important aspects of semantics. It also talks about the training configuration, including the choice of the parameters, Focal Loss to deal with the imbalance of the classes, and the metrics of performance evaluation. Lastly, the chapter provides the experimental design and the comparison analysis that was conducted against baseline models to prove the legitimacy of the proposed strategy.

3.1 Proposed Methodology

The overall methodology of the given study is depicted in Figure 3.1 that draws attention to all crucial stages of the research process starting with data collection to the model evaluation. The diagram depicts the process of collecting, washing and labeling the raw Bangla Facebook comments. Next is tokenization, padding and separating the dataset into training, validation and testing. The architecture of the proposed Bi-LSTM with Multi-Head Attention model is also presented, and the training process based on optimization methods, the loss of focus, and the weighting of classes are also showcased.

Being a top down description of the entire system pipeline, the diagram allows one to understand how the numerous components interact to form the whole detection system.

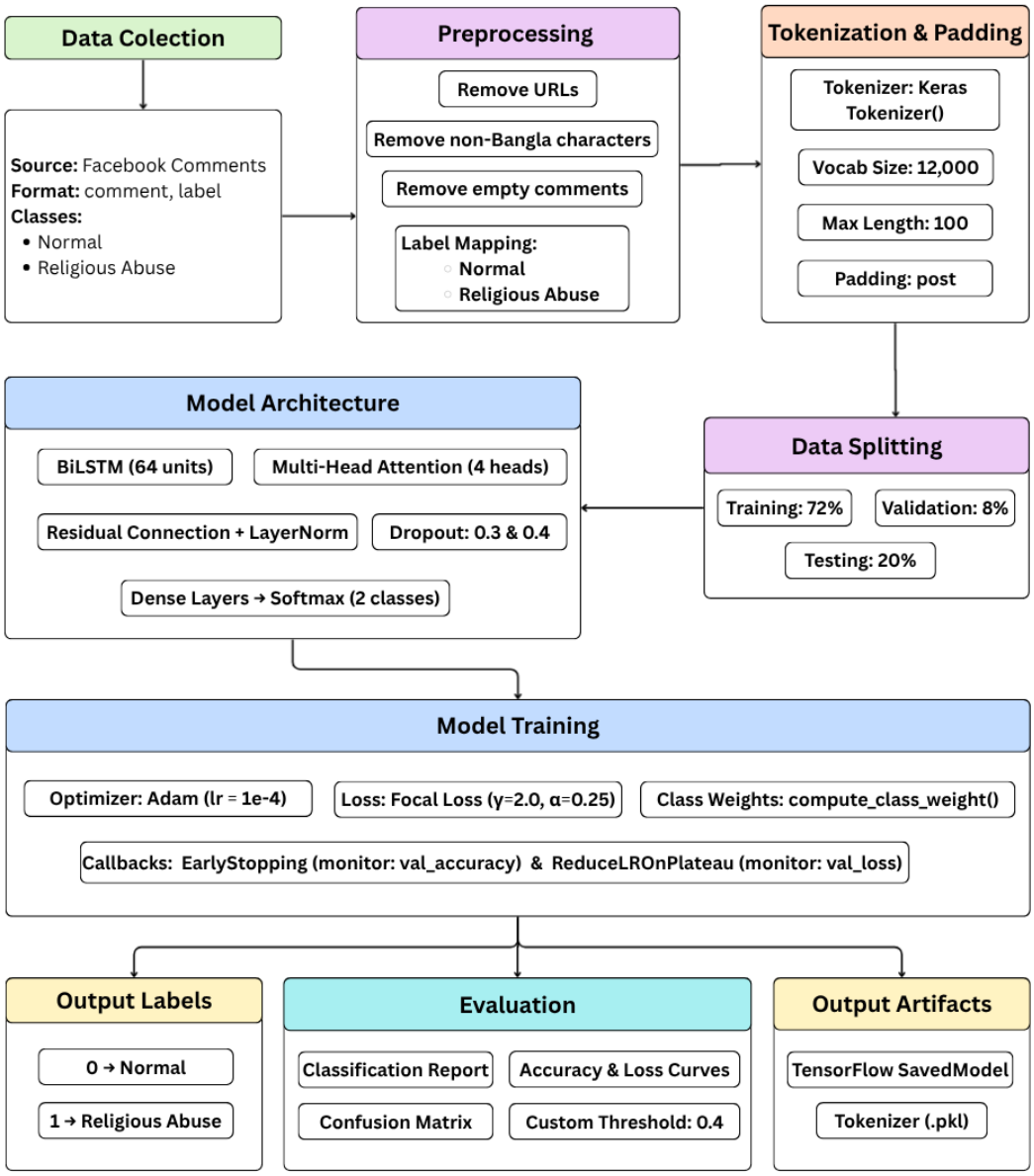


Figure 3.1: Methodology Diagram

The next sections of this chapter provide a detailed explanation of each step depicted in the diagram.

3.2 Data Collection

The data set used in this paper is on the Bangali Facebook comments which have been carefully collected so that it includes both the abusive expressions concerning religion and the non-abusive expressions (regular and religious). The initial sample was constituted of 25,920 comments made on open forums, religious materials and news websites that often host sensitive religious discussions. After cleaning, normalization and elimination of bias, the dataset was reduced to a minimum of 24,137 comments, which included 16,597 and 7,540 as the normal and religious abuse respectively. This improvement ensured the data was linguistically rich and contextually diverse and eliminated noise, unneeded symbols, spam, and redundant records.

Each processing step was performed in Python, with the help of such libraries as Pandas to handle structured data and Regular Expressions (re) to cleanse the text. Letters in non-Bangla, punctuation, URLs, and emojis were removed, and the normalization of all the text was carried out to maintain consistency. The native speakers of the Bangla language were the ones who made these comments and in case of any conflict that arose, they would discuss and have it corrected to ensure correct labelling. The resultant data is thus a neat and clean corpus that can be used to train deep learning models in identifying abuse comments on religion. Figure 3.2 is a word cloud representation that highlights the most frequent and contextually relevant words identified in both classes that highlights the language specifics of the process of social media communication in Bengali.

Figure 3.2: Religious and Abusive Word Cloud

3.2.1 Data Labeling

All the comments within the dataset were categorized as either "normal" and "religious abuse." This binary classification method is the basis of model training and assessment that gives the system an opportunity to differentiate between impartial religious attitudes and abusive or discriminative language. Since the religious issues are sensitive and context-specific, the labeling procedure was thoroughly designed to combine the manual tagging with the automated classification based on rules, to be precise and scalable.

First, the comments were manually categorized by trained annotators who have mastered Bangali as their native language and the contextual and cultural understanding is not misinterpreted. An automated rule-based tagging Python script was also constructed to enhance consistency and remove subjectivity, and relied on predefined keyword dictionaries. These dictionaries were sufficiently broad in their terms (religious phrases (Islamic, Hindu, Christian, Buddhist, and general spiritual references), words related to harassment (personal insults, religious slurs, caste-based epithets and verbal aggressiveness). Religious Abuse was considered as a statement that includes both categories of keywords (religion and harassment) and demonstrates the intent of abuse regarding the issues of religion. In case of a comment whose elements were either religious or related to harassment, the comment was categorized as Normal.

This human-based labeling approach with automatic consistency checks can be used to decrease the number of errors and bias in annotation and produce highly quality dataset suitable to deep learning studies. Figure 4 shows the distribution of the texts in the two binary classes, showing that there is a relative equal proportion of the samples of the two binary classes in the final dataset between the populations of the normal and the religious abuse. Figure 3.3 presents the distribution of data of the dataset.

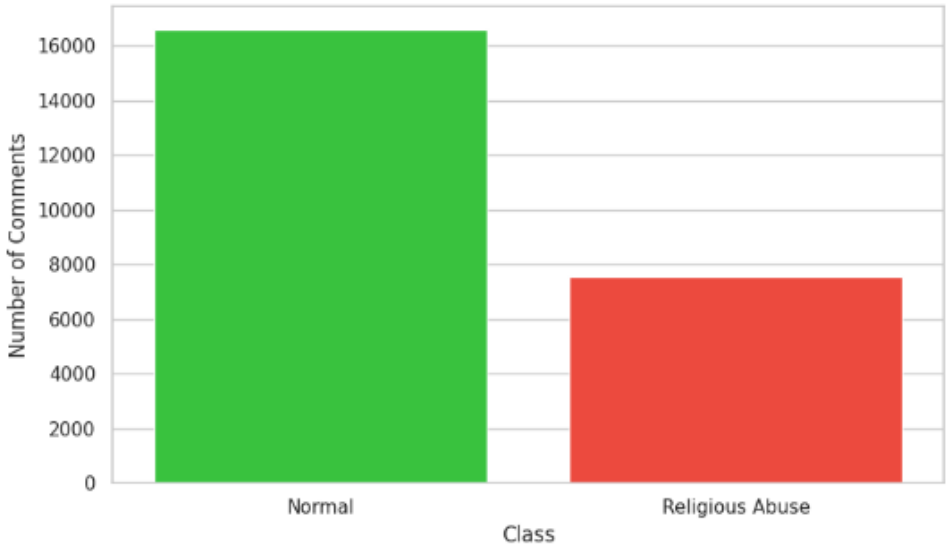


Figure 3.3: Data Distribution of the Dataset

3.2.2 Preprocessing Steps

In order to guarantee the quality and consistency of the textual information the following preprocessing steps were embraced. Figure 3.4 Preprocessing workflow Bangali Religious Comment Data Set.

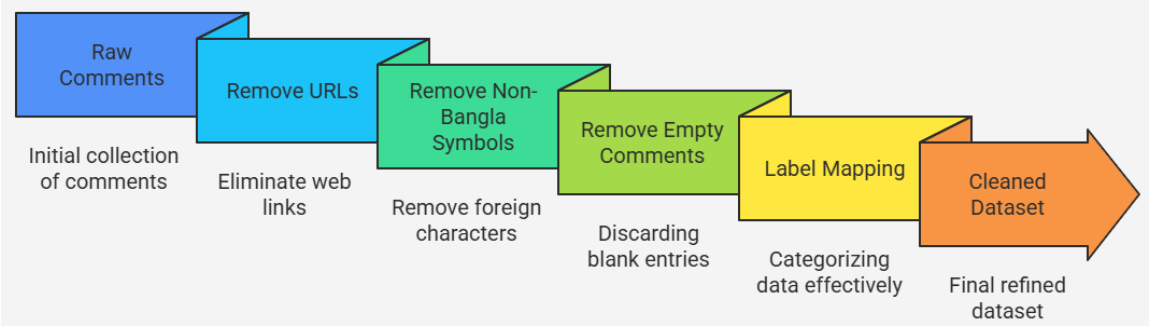


Figure 3.4: Preprocessing Workflow for the Dataset

- **Remove URLs:** The URLs that were in the comments were removed to get rid of extraneous external links and get down to the text itself.
- **Remove non-Bangla symbols:** The models involved removal of the symbols which is not related to the Bangla script so as to ensure that only the features pertaining to the linguistic information is considered by the models.
- **Remove empty comments:** The empty comments that were left empty following the preprocessing stage were removed in the dataset.

- **Label Mapping:** The initial labels were reduced into the numerical ones: 0 became the normal category and 1 the category of religious abuse.

3.3 Tokenization and Padding

The text data was pre-processed, followed by tokenization, and padding to be inputted into the deep learning models. The concept behind tokenization is that a text is turned into a sequence of numerical tokens, and to make the input sequences of equal length, padding is employed. Fig. 3.5 symbolizes Tokenization and Padding Process of Model input.

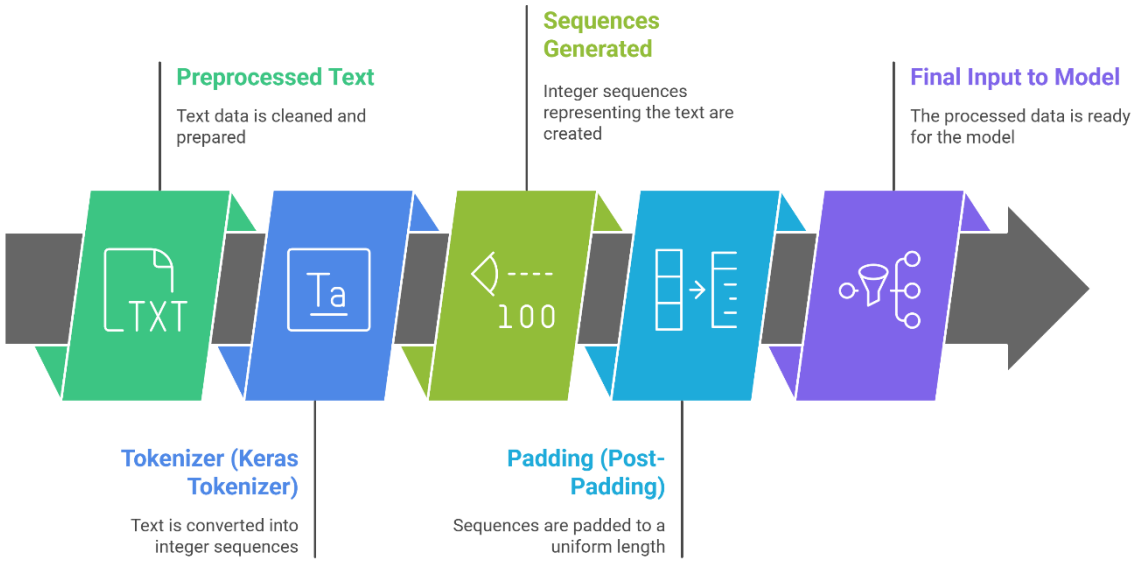


Figure 3.5: Tokenization and Padding Process for Model Input

- **Tokenizer:** In order to encode the text with a set of integers, the keras Tokenizer was employed. Out-Of-Vocabulary (OOV) token also participated in keeping the words which are not featured in the vocabulary.

- **Vocabulary Size:** The definition of the size of the vocabulary was made, i.e., only 12,000 most frequent words were utilized.

- **Maximal Length:** 100 tokens All sequences were made as long as possible.

- **Padding:** The only padding that was used was post-padding that is, adding the zeros at the end of short sequences.

Algorithm: Tokenization and Padding:

- **Step 1 — Initialize Tokenizer**
    - Set vocab_size = 12,000
    - Set oov_token = '<OOV>'
- **Step 2 — Fit Tokenizer**
    - Fit tokenizer to the clean training texts.
- **Step 3 — Convert to Sequences**
    - Put all the comments into a series of tokens, which are integers.
- **Step 4 — Apply Padding**
    - Use post-padding to ensure sequences have length = 100.

3.4 Data Splitting

Preprocessing and padding of the dataset were made along with the division into training, validation and testing sets to evaluate the performance of the model and its capability to generalize. The division of percentage was made as follows:

- **Training:** 72 percent of the information was used to model.
- **Validation:** The training phase consumed 8 percent of the data which was used to validate. This data set will help in tuning of the hyperparameters and prevent overfitting due to monitoring the performance of the model on the new data during the training epochs.
- **Testing:** 2 out of 5 of the dataset was reserved towards the final evaluation of the trained model. The collection presents an unbiased assessment of the model effectiveness on completely new data.

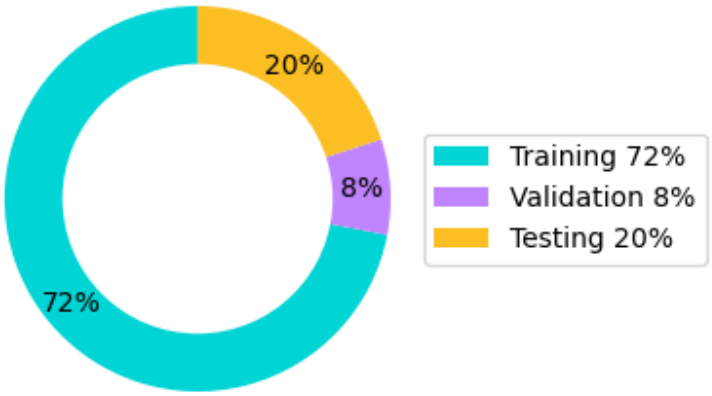


Figure.6: : Donut Chart of Dataset Split

The 72 percent training and 8 percent validation (both of which add to 80 percent total) ratio is a common practice in machine learning, especially with deep learning architecture which requires a substantial quantity of data to be effectively trained. The 20% test set ensures that there is a high measurement of the model generalization. This division has been chosen to make sure that there is sufficient information that the model can use to understand the complex trends without considering a significant portion that would be used to evaluate the data objectively.

Algorithm: Data Splitting

- **Step 1 — Define Split Ratio**
  - Training: 80%
  - Testing: 20%
- **Step 2 — Perform Split**
  - Use train_test_split() with a fixed random_state for reproducibility.

3.5 Model Architecture

The model is founded on the Bi-directional Long Short-Term Memory (Bi-LSTM) network and Multi-Head Attention. The structure is made so as to combine the contextual information and the long distance relations in the sequential text information in a clear form.

- **Input Layer:** Accepts a 100 character sequence with padding.
- **Embedding Layer:** It takes the input token and translates them to dense vectors. At 64 and 12,000 dimensions size and vocab size were added respectively.
- **Bidirectional LSTM Layer:** It is a 64-unit Bi-LSTM layer, which is applied to run embedded sequences. The LSTMs are bidirectional, which means that both forward and backward sequence designs are considered in the bidirectional process of data processing that enables the model to define any past or future dependencies.
- **Dropout (0.3):** The dropout layer has used dropout probability of 0.3 at the end of Bi-LSTM, which prevents overfitting by randomly selected a section of its input units and set them to 0 during the training process.
- **Multi-Head Attention:** MultiHeadAttention of 4 heads and key units of 64. The attention mechanisms also assist the model to concentrate on the various sections of the input chain during the prediction process that increases the power of search of the model to detect pertinent information.
- **Layer Normalization:** This is the next concept that comes after the attention mechanism and it is likely to stabilize the training process and enhance the results.
- **GlobalMaxPooling1D:** It is a layer, which allocates reduced dimension of the result of the preceding layers and chooses the most common of all the time paces to learn more of the significant issues.
- **Dense Layer:** The features are integrated in a dense layer, comprising of 64 units a fully connected and ReLU activation layer.
- **Dropout (0.4):** This is an additional dropout layer whose dropout rate is 0.4 that is applied to the output of the dense layer to more considerably regularize the model.
- **Output Layer:** This is the final dense layer which is 2-units (binary classification) with softmax activation which is intended to provide the probability of one of two sets of labels, which are the normal and the religious abuse.

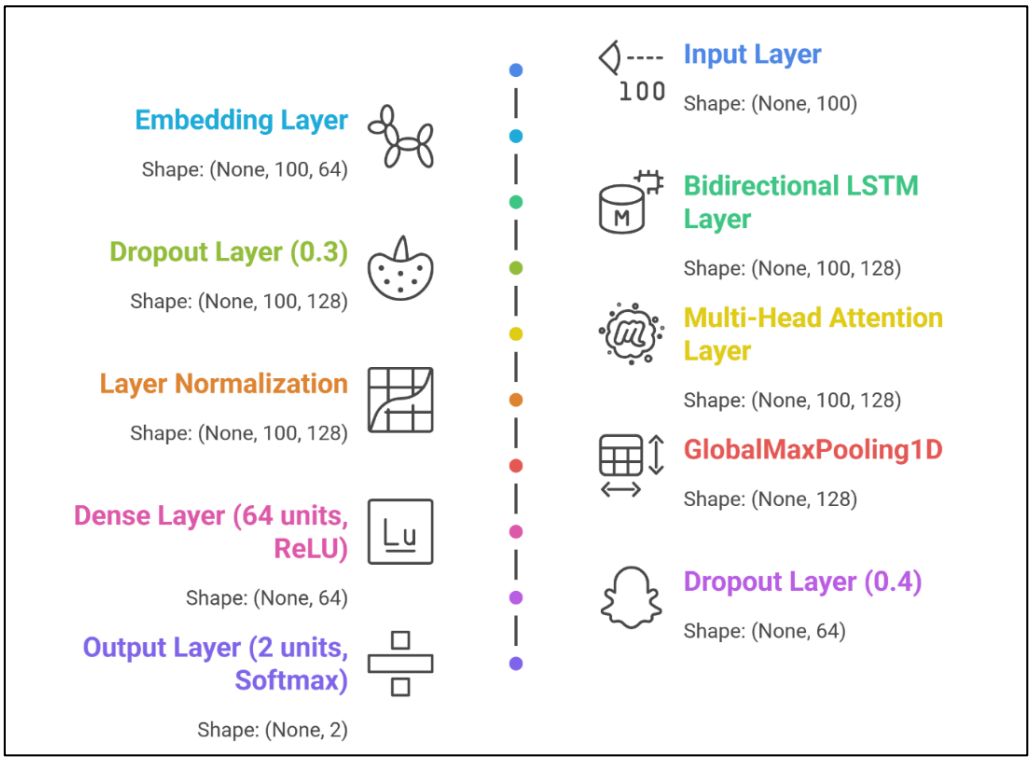


Figure 3.7: : BiLSTM + Multi-Head Attention Model Architecture

Algorithm: Model Architecture

- **Step 1 — Define Input Layer**
  - Shape = (100,) tokens.
- **Step 2 — Embedding Layer**
  - Embedding dimension = 64, vocab_size = 12,000.
- **Step 3 — BiLSTM Layer**
  - Units = 64, return_sequences = True.
- **Step 4 — Dropout Layer**
  - Dropout rate = 0.3.
- **Step 5 — Multi-Head Attention Layer**
  - num_heads = 4, key_dim = 64.
- **Step 6 — Add Residual Connection + Layer Normalization**
- **Step 7 — Global Max Pooling Layer**
  - Reduces sequence to fixed vector.
- **Step 8 — Dense Layer (ReLU)**

    - Units = 64, activation = ReLU.
- **Step 9 — Dropout Layer**
    - Dropout rate = 0.4.
- **Step 10 — Output Layer**
    - Dense(2, softmax).

3.6 Model Training

The model was configured and trained to the following configurations:

- **Optimizer:** To achieve successful gradient descent, an Adam optimizer with a learning rate of 0.0001 was used.
- **Loss Function:** The customized Focal Loss (activation=2.0, alpha=0.25) was used to address the issue of class imbalance in the dataset. Focal loss pulls the weight of examples that are easy but stresses hard misclassified examples, which is particularly beneficial to imbalanced datasets.
- **Measurements:** They used measures that were monitored during the training.
- **Epochs:** The training was done in 20 epochs.
- **Batch Size:** 128 batch size was used.
- **The Class Weights:** compute_class weight(balanced) was used to obtain weights that are negative to the frequencies of classes thus mitigating the impacts of class imbalance further.
- **Callbacks:** EarlyStopping (monitoring val_accuracy with patience=3) and ReduceLROnPlateau (monitoring val-loss with factor=0.2, patience=2) have been used to improve the training and prevent overfitting.

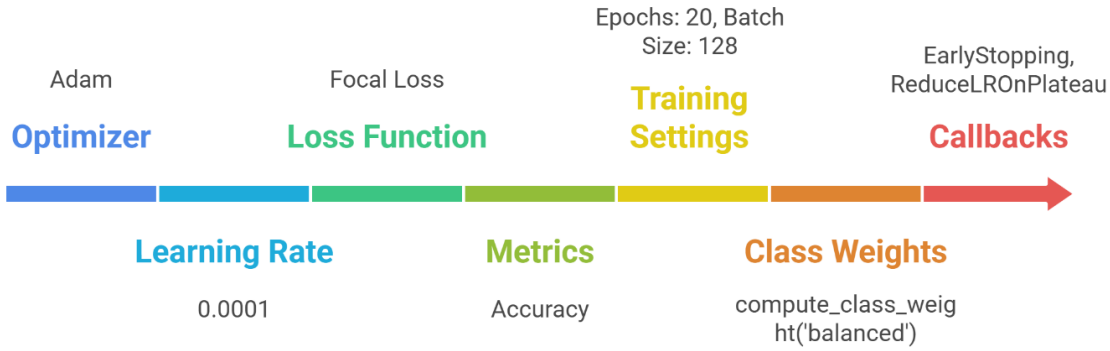


Figure 3.8: Model Training Workflow

Algorithm: Model Training

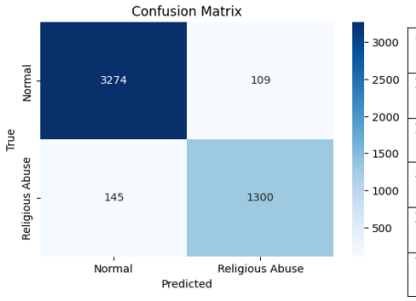
- **Step 1 — Compile Model**
  - Optimizer = Adam (lr=0.0001)
  - Loss = Focal Loss (γ=2.0, α=0.25)
  - Metric = Accuracy
- **Step 2 — Define Callbacks**
  - EarlyStopping (patience=3, monitor='val_accuracy')
  - ReduceLROnPlateau (patience=2, factor=0.2, monitor='val_loss')
- **Step 3 — Train Model**
  - Epochs = 20
  - Batch Size = 128
  - Validation Split = 10%
  - Apply class weights

3.7 Evaluation

The tested model was tested on the unseen test set with a specific threshold of 0.4 to find the commenters ofReligious Abuse. The evaluation metrics were:

- **Classification Report:** Provides the accuracy, recall and F1-score of individual categories.
- **Confusion Matrix:** Shows both the real positives, real negatives, false positives and false negatives in predictions.

- **Accuracy & Loss Graphs:** Plot the accuracy of the training and validation and loss per epoch.



Class	Precision	Recall	F1-Score	Support
Normal	0.96	0.97	0.96	3,383
Religious Abuse	0.92	0.90	0.91	1,445
Accuracy			**0.95**	**4,828**
Macro Avg	0.94	0.93	0.94	4,828
Weighted Avg	0.95	0.95	0.95	4,828

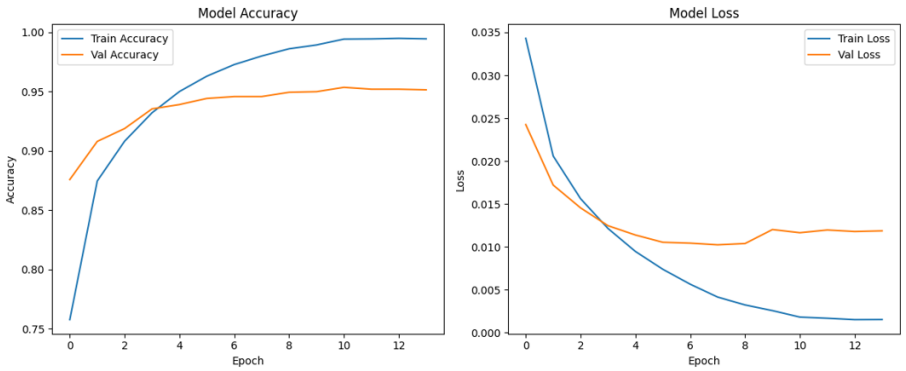


Figure 3.9: Model Evaluation Results

Algorithm: Evaluation

- **Step 1 — Predict on Test Data**
  - Use trained model to predict probabilities for test set.
- **Step 2 — Apply Threshold**
  - If probability ≥ 0.4 → Class 1 (Religious Abuse)
  - Else → Class 0 (Normal)
- **Step 3 — Generate Metrics**
  - Classification report (Precision, Recall, F1-score)
  - Confusion matrix
- **Step 4 — Visualize Performance**
  - Plot accuracy and loss over epochs.

3.8 Chapter Summary

This chapter explained the whole step that is used to develop and evaluate the proposed deep learning model to detect offensive language on religious text in Bangali language.

The methodological workflow was described in the first place and the data collection procedure, labeling procedures and the way the data were preprocessed to prepare the dataset was also described in detail. The chapter then discussed the process of tokenizing and padding and the process of dividing the data into training, validation and testing sets.

The design of the Bi-LSTM with Multi-Head Attention model was explained next, and the key components of it were also described along with the rationale of its selection. The training procedure was also discussed in detail and involved focus loss, class weighting, and optimization callbacks. Finally, the assessment section outlined the measures and tools that were used to assess the model performance. This chapter, all things being said, provides a systematic approach to the understanding of the design, implementation and validation of the system, which preconditions the results to be introduced in the next chapter.

CHAPTER 4

RESULTS AND DISCUSSION

The chapter presents the results of the experiment and performance analysis of the proposed BiLSTM with Multi-Head Attention model to detect religiously abusive comments written in Bangali. It describes the experimental set up including parameters settings, portion of the datasets, and training systems. The chapter breaks down the performance of the model during the training and validation phases focusing on trends in accuracy and loss by epoch. It also illustrates how early termination and changes in the learning rate can be utilized to guarantee the best convergence and avoid overfitting.

The second part of the chapter focuses on the evaluation of the model performance in terms of such measures as accuracy, precision, recall and F1-score along with the support of the confusion matrix and classification report. The efficacy of the proposed method is assessed through a thorough comparison with such baseline models as CNN, GRU and Transformer architectures. The discussion also looks at the ability of the model to cover the imbalance in classes and recognize the sensitive linguistic nuances in Bangla text. Finally, the chapter provides an overview of the benefits of the model, its application in content moderating processes, and potential ways of its improvement in the future.

4.1 Performance Metrics

The accuracy, the precision, the recall, and the F1-Score have been used as the main metrics to test the suggested Bi-LSTM with Multi-Head Attention model. These measures give a relative measure of the capability of the model to define Bangla comments to be either normal or religious abuse. Accuracy measures the total correctness of predictions, Precision and Recall measure the model in its ability to manage erroneous positives and false negatives respectively. F1-score, which is a harmonic mean of Precision and Recall, is a general performance measure, especially useful in imbalance datasets.

4.1.1 Classification Report

The model achieved a total accuracy of 95% which means that it has a great ability to differentiate normal and offensive comments. The F1-score of 0.91 of the Religious Abuse class indicates a balanced tradeoff between the precision and recall which is essential in minimizing false alarms and omissions of abusive utterances.

Table 4.1: Classification Report of Bi-LSTM model with Multi-Head Attention

Class	Precision	Recall	F1-Score	Support
Normal	0.96	0.97	0.96	3,383
Religious Abuse	0.92	0.90	0.91	1,445
Accuracy			**0.95**	**4,828**
Macro Avg	0.94	0.93	0.94	4,828
Weighted Avg	0.95	0.95	0.95	4,828

4.1.2 Confusion Matrix

The confusion table gives a clear account of right and wrong classifications:

- **True Positives (TP):** 3274 (Normal comments identified as Normal)
- **True Negatives (TN):** 1300 (Religious Abuse correctly checked as Religious Abuse)
- **False Positives (FP):** 109 (Normal comments falsely labelled as Religious Abuse)
- **False Negatives (FN):** 145 (Religious Abuse remarks classified as wrongly as Normal)

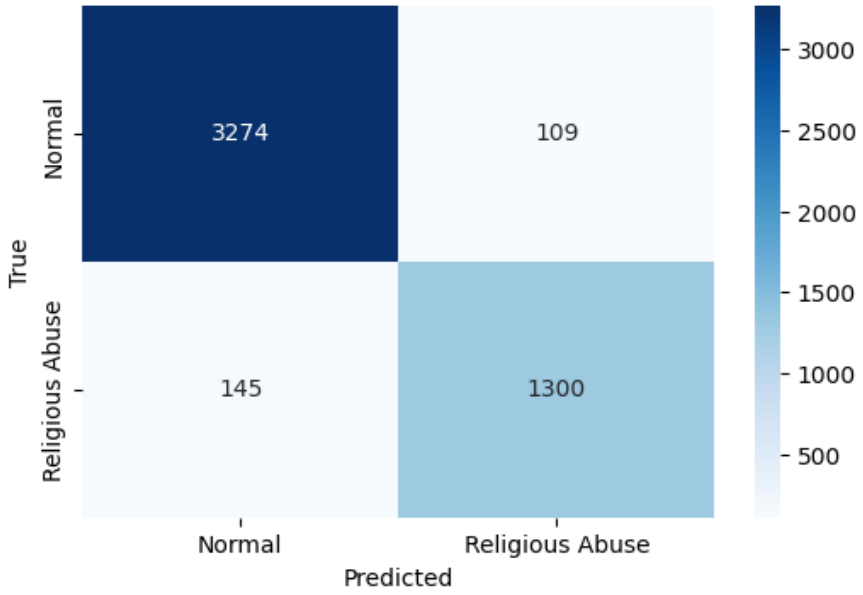


Figure 4.1: Confusion Matrix

These results indicate that the model has high true positive and is incredibly effective in identifying Normal comments. The percentage of false negative is a little higher meaning that still there are an abusive comments that the algorithm fails to pick, although its performance on abusive comments on religion is also quite good. This issue often occurs when the datasets are unbalanced, and the minority group (abusive remarks) is harder to detect.

4.1.3 Experimental Result Analysis

We conduct the detailed analysis using key evaluation measures such as Accuracy, Precision, Recall, and F1-score to determine the effectiveness of the suggested and comparative approaches. These measures can be used to gain a deep understanding of the performance of classification in equal and unequal conditions.

The other important measure in binary classification is True Positive (TP), a positive prediction that the model makes. Similarly, True Negative (TN) is correctly predicted negatives and False Positive (FP) and False Negative (FN) are the errors of classification.

The equations of these metrics are as given below:

$$\textbf{\textit{Accuracy}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{3.1}$$

$$\textbf{\textit{Precision}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.2}$$

$$\textbf{\textit{Recall}} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3.3}$$

$$\textbf{\textit{F1-score}} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{3.4}$$

Whereas Accuracy is used to evaluate the accuracy of the model on average, Precision tries to reduce the number of false positives and Recall shows the ability of the model to detect all positive cases accurately. Since the positive and negative cases of our religious abuse detection dataset are unequally represented, F1-score is specifically appropriate in the said dataset since it balances Precision and Recall.

To reduce the biases of the data splits on the performance results, we use k-fold cross-validation (k = 10) in order to conduct a more stringent evaluation. The dataset is divided into 10 parts, of which 9 are used in the training and 1 in the validation in each iteration and the performance metric is averaged across the parts to make the results reliable and minimize variance in the reported ones.

4.2 Accuracy and Loss Plots

Accuracy and loss charts indicate the model performance during the training and validation of the model per epoch. The more the 'Train Accuracy' curve increases steadily, the more the model is deriving the knowledge on the training set. Also, the curve of the val accuracy increases before it approaches a plateau, which means that the model is getting generalized to new data without significant overfitting. Similarly, the 'Train Loss' continues to reduce, and the 'Val Loss' also reduces and reaches a level, hence demonstrating good generalization.

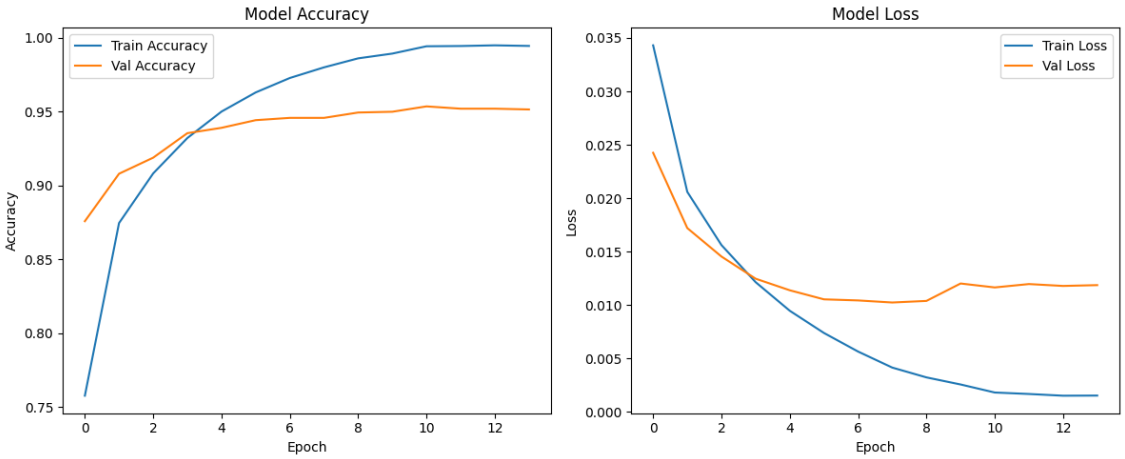


Figure 4.2: Accuracy and Loss Plots

4.3 Comparative Analysis with Other Models

In a bid to give a comprehensive assessment of the model performance, model (Bi-LSTM with Multi-Head Attention) was compared with various other baseline deep learning and machine learning models. These overall performances are presented in Tables 4.2 to 4.8 where Table 4.2 gives the performance of the LSTM model, Table 4.3 gives the CNN-LSTM metrics and Table 4.4 gives the Bi-LSTM (without attention). Table 4.5 demonstrates the findings of the RNN model, and Tables 4.6, 4.7, and 4.8 compare the results of Logistic Regression, SVM, and random forest. These tables demonstrate the way each method copes with the classification of the comments of Normal and Religious Abuse under the same conditions of the experiment which makes it possible to provide the clear comparison of the performance with the proposed model.

Table 4.2: LSTM Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.71	1.00	0.83	3,648
Religious Abuse	0.78	0.04	0.07	1,532
Accuracy			0.71	5,180
Macro Avg	0.75	0.52	0.45	5,180
Weighted Avg	0.73	0.71	0.61	5,180

Table 4.3: CNN_LSTM Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.95	0.94	0.94	3,648
Religious Abuse	0.86	0.87	0.87	1,532
Accuracy			0.92	5,180
Macro Avg	0.90	0.91	0.91	5,180
Weighted Avg	0.92	0.92	0.92	5,180

Table 4.4: BiLSTM Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.94	0.96	0.95	3,648
Religious Abuse	0.90	0.84	0.87	1,532
Accuracy			0.93	5,180
Macro Avg	0.92	0.90	0.91	5,180
Weighted Avg	0.93	0.93	0.93	5,180

Table 4.5: RNN Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.84	0.88	0.86	3,648
Religious Abuse	0.68	0.61	0.64	1,532
Accuracy			0.80	5,180
Macro Avg	0.76	0.75	0.75	5,180
Weighted Avg	0.80	0.80	0.80	5,180

Table 4.6: Logistic Regression Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.82	0.91	0.86	3,648
Religious Abuse	0.71	0.53	0.61	1,532
Accuracy			0.80	5,180
Macro Avg	0.77	0.72	0.74	5,180
Weighted Avg	0.79	0.80	0.79	5,180

Table 4.7: SVM Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.82	0.91	0.86	3,648
Religious Abuse	0.71	0.54	0.61	1,532
Accuracy			0.80	5,180
Macro Avg	0.77	0.72	0.74	5,180
Weighted Avg	0.79	0.80	0.79	5,180

Table 4.8: Random Forest Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.80	0.93	0.86	3,648
Religious Abuse	0.74	0.45	0.56	1,532
Accuracy			0.79	5,180
Macro Avg	0.77	0.69	0.71	5,180
Weighted Avg	0.78	0.79	0.77	5,180

4.4 Comparative Analysis of Classification Reports

The comparison indicates that the Bi-LSTM with Multi-Head Attention model is far better than any other model in detecting religious abusive comments in the Bangla text data, particularly the F1-score of the religious abuse category. Attention mechanism plays a critical role in enhancing the ability of the model to focus on significant parts of the input sequence leading to an improved performance of the model.

Table 4.9: Comparative Table of Classification Reports

Model	Precision (Religious Abuse)	Recall (Religious Abuse)	F1-score (Religious Abuse)	Accuracy
Bi-LSTM with Attention	0.92	0.90	0.91	0.95
CNN_LSTM	0.86	0.87	0.87	0.92
BiLSTM (without Attention)	0.90	0.84	0.87	0.93
RNN	0.68	0.61	0.64	0.80
LSTM	0.78	0.04	0.07	0.71
Logistic Regression	0.71	0.53	0.61	0.80
SVM	0.71	0.54	0.61	0.80
Random Forest	0.74	0.45	0.56	0.79

The results of the deep learning models are usually better than the machine learning ones, and CNNLSTM and BiLSTM without attention also demonstrate spectacular performance. The simple LSTM and RNN models can handle sequential data, but they have more difficulties with the specifics of abusive words, as they have lower F1-scores of the minority class. Traditional machine learning models like the Logistic Regression, the SVM, and the Random Forest show a respectable overall accuracy, but worse recall

and F1-scores on the religious abuse category, which highlights the complexity of the problem in the identification of the minority category without the advanced deep learning models and methods such as focal loss.

4.5 Chapter Summary

In this chapter, the findings of the experiment and performance evaluation of the proposed Bi-LSTM with Multi-Head Attention model were presented. The findings indicated that the model was not only good in classification, which was demonstrated by the actual values such as accuracy, precision, recall, and F1-score. Although the experimental analysis revealed the role of the focal loss and attention systems in improving the minority-class recognition, the classification report and confusion matrix provided possible information about the ability of the model to provide reliable and consistent separation of normal and abusive remarks on religious content.

The comparison analysis revealed that the proposed model was better than the alternative deep learning architecture and the traditional machine learning methods and the accuracy and loss plots confirmed the behavior of stable training and good generalization. Altogether, the results support the effectiveness of the proposed approach and indicate that it is suitable to analyze the derogatory comments on religious text of Bangla social media. This chapter sets the scene of the closing observations of the next chapter and the proposed areas of inquiry.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

This chapter provides summaries of the important findings, contributions and significance of the study. It takes into account the objectives stated in Chapter 1 and explains how the Bi-LSTM with Multi-Head Attention model was utilized to detect offensive statements on the religious material of Bengali text. The model not only performs better than the traditional machine learning and deep learning baselines, but it also maintains its level of reliability and ethical requirements, which proves its high ability to address implicit and context-related abuse.

The future research opportunities have also been outlined in the chapter in order to expand and improve the proposed system. The main directions are to increase the size of data, use larger and multilingual sources, consider transformer-based or hybrid models to achieve better results, and use multimodal data, i.e. images and audio. Moreover, it is also possible to enhance the relevance and influence of this work by creating adaptive real-time systems that react to the changing trends in online language.

5.1 Findings of the Study

The purpose of the research was to develop a robust deep learning framework of detecting and analyzing religiously abusive posts in the Bangla textual data, the general aim of which was to facilitate safer and more thoughtful online communication. With the combination of linguistic analysis, data-driven experimentation, and advanced model design, a number of crucial findings were reached, which, together, serve the purpose of Chapter 1.

The research commenced with the identification of major obstacles to the current research including imbalance in data, cultural and linguistic diversity, recognition of sarcasm, explainable models and adaptable models of Bangali. To overcome these issues, comprehensive data of text of Bangali was collected on the various social media sites and

was then properly classified into two categories; namely, the Normal and Religious Abuse. This data managed to capture the complexity of online Bangla language, with informal expressions, code-switching and references with cultural meaning.

The proposed Bi-LSTM with Multi-Head Attention performed significantly better than the baseline models including traditional machine learning and deep learning techniques. By implementing attention systems, the model would be able to focus on semantically meaningful words and contextual trends and increase its ability to detect implicit and indirect forms of religion-related abuse sarcasm, metaphor, and coded words. The results of evaluation showed strong results on all main measures, including accuracy, precision, recall, and F1-score, and a considerable increase in the recall ability of the minority (abusive) category. This finding is particularly noteworthy, considering that recall is the direct pointer of the model as the means of detecting those cases of abuse that would otherwise remain unnoticed.

5.2 Key Contributions

The paper has important implications to the research of Natural Language Processing (NLP), deep learning, and computational social science, especially to the domain of Bangali text processing and religion identification abuse. The contributions are data production, methodological, model invention, and ethical AI applications.

- **Creation of a Domain-Specific Bangla Abuse Dataset on Religion:** The study is an important contribution because it constructed a curated and annotated Bangla dataset containing all the abuse of religion in internet comments and all the manipulation. The data was gathered by utilizing multiple social media articles, cleaned up, and manually sorted into two, namely, normal and religion-related abuse. The dataset addresses a significant gap in the Bangali NLP resources, which will be the basis of future research on religiously biased or hate speech text classification.

- **Design and Implementation of a Model:** The paper introduces a deep learning framework, which uses Bidirectional Long Short-Term Memory (Bi-LSTM) and Multi-Head Attention model. This hybrid architecture enhances the model to identify forward and backward contextual relationship besides giving importance to

semantically significant words. The architecture helps the model to establish implicit abuse, sarcasm and context-based antagonism that the standard models fail to detect.

- **Better Management of Class Imbalance:** To improve the issue of dataset imbalance that is very prevalent, the research uses strategies of focal loss and data augmentation. These measures enhance the memory of the minority group and see that even minor and infrequent cases of abuse are properly recognized. This is one way to achieve more fair, balanced classification outcomes which is a major consideration of actual content moderation systems.

- **Framework of Extending Research in Low Resource and Multilingual Settings:** Although it is focused on Bangla, the research results lead to the creation of a methodological framework that can be extended to other low-resource languages with similar social and linguistic problems. The suggested methodology of work on the model and the dataset design can be extended to identify political, gender-oriented, or ethnic hate speech in diverse cultural settings and advance the development of inclusive and cross-lingual NLP systems.

5.3 Future Works

Although the present research was able to construct and test a deep learning-based solution to detect and analyze abusive posts in Religious Bangla text, the dynamism of the online communication process, sensitivity to culture, and technical constraints opens many possibilities to future research. To enlarge and enhance the contribution of this work, the following recommendations and research directions are provided:

- **Growth of Dataset Size, Diversity, and Multilingual coverage:** This is because growing the size of a dataset is the most critical step. More generalization needs to be added to future research by incorporating different variants of the Bangla language, code-mixed language texting, and data related to other social network sites. A multilingual dataset, containing multilingual postings (Hindi, Urdu, and Tamil), can be used to identify hate-speech cross-linguistically and better aligns with a bigger South Asian situation.

- **Combination of Transformer-Based, Hybrid, and Adaptive Learning Models:** It is time to leave Bi-LSTM behind and consider transformer models, like BanglaBERT, mBERT, or XLM-RoBERTa, or a combination of recurrent layers, attention, and graph neural networks. Moreover, the adaptive learning technologies like online and federated learning allow changing the model according to the real-time linguistic patterns.

- **Explainable, Fair, and Accountable AI Deployment:** Due to the sensitive nature of the content in religion, future systems should focus more on explainability, fairness, and transparency. The use of explainable AI tools such as LIME and SHAP and consideration of ethical aspects in the collection of data and in the deployment of a model is essential in preventing misclassification, bias, and censorship without compromising trust and responsibility.

Lastly, the general paper presents a comprehensive and focused research of using deep learning to detect offensive bangla remarks on religion. The paper demonstrates that complex contextual modeling could maximally boost both the detection of implicit, subtle, and context-dependent religious maltreatment in online text through the creation of a domain-specific data set and the execution of a robust Bi-LSTM with Multi-Head Attention model and by comparison with a variety of baseline models. The real-life applicability of the system to actual moderation pipelines is validated by the fact that the system is highly accurate, the categorization is balanced, and the system is more sensitive to cases that belong to the minority-class.

In addition to the technical achievements, the process also heavily relies on cultural sensitivity and ethical responsibility, ensuring that the model decisions are in accordance with the social sensitivities of Bangladesh with regard to the religious discourse. These discoveries all point to the significance of incorporating linguistic knowledge, model explainability, and adaptive training processes in handling the complexity and the dynamic nature of online communication. This study provides an important move toward safer, fairer and more inclusive online spaces since it provides a methodological framework that is systematic and the research agenda in the future is outlined clearly.

REFERENCES

[1] "Our lives don't matter': Bangladeshi Hindus under attack after Hasina exit | Religion | Al Jazeera." Accessed: Nov. 15, 2025. [Online]. Available: https://www.aljazeera.com/features/2024/12/12/our-lives-dont-matter-in-post-hasina-bangladesh-hindus-fear-future

[2] "Bangladesh's Religious Minorities Want Peace Amid Country's Turmoil - Christianity Today." Accessed: Nov. 15, 2025. [Online]. Available: https://www.christianitytoday.com/2024/08/bangladesh-protest-christian-hindu-muslim-sheikh-hasina/

[3] "Far-right spreads false claims about Muslim attacks in Bangladesh." Accessed: Nov. 15, 2025. [Online]. Available: https://www.bbc.com/news/articles/cx2n8pzk7gzo

[4] "Country policy and information note: religious minorities and atheists, Bangladesh, June 2025 - GOV.UK." Accessed: Nov. 15, 2025. [Online]. Available: https://www.gov.uk/government/publications/bangladesh-country-policy-and-information-notes/country-policy-and-information-note-religious-minorities-and-atheists-bangladesh-june-2025

[5] "Bangladesh: 'There is no law and order. And Hindus are being targeted again.'" Accessed: Nov. 15, 2025. [Online]. Available: https://www.bbc.com/news/articles/cwy77vgmjlzo

[6] "Bangladesh's Evolving Security Crisis: The Rise of Religious Extremism Amid Political Transition - RSIS." Accessed: Nov. 15, 2025. [Online]. Available: https://rsis.edu.sg/ctta-newsarticle/bangladeshs-evolving-security-crisis-the-rise-of-religious-extremism-amid-political-transition/

[7] "The Upsurge Of Radical And Fundamentalist Islamic Elements In Bangladesh - Expert Speak | ORF." Accessed: Nov. 15, 2025. [Online]. Available: https://www.orfonline.org/expert-speak/the-upsurge-of-radical-and-fundamentalist-islamic-elements-in-bangladesh

[8] S. Alam, M. Sabbir, H. Mrida, ; Md, and A. Rahman, "Sentiment Analysis in Social Media: How Data Science Impacts Public Opinion Knowledge Integrates Natural Language Processing (NLP) With Artificial Intelligence (AI)," *American Journal of*

Scholarly Research and Innovation, vol. 4, no. 01, pp. 63–100, 2025, doi: 10.63125/R3SQ6P80.

[9] M. S. Islam, M. A. T. Rony, M. Ahammad, S. M. N. Alam, and M. S. Rahman, "An Innovative Novel Transformer Model and Datasets for Safeguarding Religious Sensitivities in Online Social Platforms," *Procedia Comput Sci*, vol. 233, pp. 988–997, 2024, doi: 10.1016/J.PROCS.2024.03.288.

[10] M. R. Karim *et al.*, "DeepHateExplainer: Explainable Hate Speech Detection in Under-resourced Bengali Language," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1–13. doi: 10.1109/DSAA53316.2021.9564230.

[11] A. Al Kafi, S. Kumar Banshal, S. Shakib, S. Azam, and T. A. Tabashom, "BOISHOMMO: Holistic Approach for Bangla Hate Speech," *arXiv Preprint*, pp. 1–33, 2025, doi: 10.48550/arXiv.2504.08408.

[12] M. Ethan, "Comparative Study of Transformer-Based Models and Bi-LSTM for Bangla Sentiment Analysis Using Hybrid," 2025. Accessed: Nov. 15, 2025. [Online]. Available: https://www.mendeley.com/catalogue/31769e1d-0354-3e58-b7d9-bd12804572f3/?utm_source=desktop&utm_medium=1.19.8&utm_campaign=open_catalog&userDocumentId=%7B27315610-8f38-44d5-8e40-3f1d4b455a25%7D

[13] M. Shymon Islam, S. Islam Auny, M. Rahman Mou, and M. Masum Hossain, "Bangla-Senti: A Large-Scale Corpus for Sentiment Analysis in Bangla and Its Applications," *Social Science Research Network*, pp. 1–63, 2025, doi: 10.2139/SSRN.5201052.

[14] M. S. Hossain, M. R. Islam, Dr. B. R. Riskhan, M. M. H. HASAN, and R. I. ISLAM, "Political sentiment analysis using natural language processing on social media," *International Journal of Applied Methods in Electronics and Computers*, vol. 12, no. 4, pp. 81–89, 2024, doi: 10.58190/IJAMEC.2024.108.

[15] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020*, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 1–12. doi: 10.1109/CSDE50874.2020.9411601.

[16] P. V. lowast and N. Sharma, “Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework,” *Indian J Sci Technol*, vol. 18, no. 5, pp. 380–389, 2025, doi: 10.17485/IJST/V18I5.1491.

[17] R. Shrestha, R. Dave, R. Shrestha, and R. Dave, “Machine Learning for Identifying Harmful Online Behavior: A Cyberbullying Overview,” *Journal of Computer and Communications*, vol. 13, no. 1, pp. 26–40, 2025, doi: 10.4236/JCC.2025.131003.

[18] M. R. Awal, R. Cao, R. K. W. Lee, and S. Mitrović, “AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection,” *Pacific Asia Conference on Knowledge Discovery and Data Mining*, pp. 1–13, 2021, doi: 10.1007/978-3-030-75762-5_55.

[19] I. Krak, O. Sobko, M. Molchanova, I. Tymofiiev, O. Mazurets, and O. Barmak, “Method for neural network cyberbullying detection in text content with visual analytic,” in *CEUR Workshop Proceedings*, 2025, pp. 298–309. Accessed: Nov. 15, 2025. [Online]. Available: https://cssesw.easyscience.education/cssesw2024/CSSESW2024/paper57.pdf

[20] N. Tasnim *et al.*, “Mapping Violence: Developing an Extensive Framework to Build a Bangla Sectarian Expression Dataset from Social Media Interactions,” *Computing Research Repository*, pp. 1–17, 2024, Accessed: Nov. 15, 2025. [Online]. Available: https://arxiv.org/pdf/2404.11752

[21] S. Ahmed, S. Rakin, K. Urmi, C. K. Nag, and Md. M. Akbar, “Automatic Identification of Political Hate Articles from Social Media using Recurrent Neural Networks,” *arXiv preprint*, pp. 1–8, 2024, doi: 10.48550/ARXIV.2411.04542.

[22] A. Rahman, S. Zaman, S. Parvej, P. C. Shill, M. S. Salim, and D. Das, “Fake News Detection: Exploring the Efficiency of Soft and Hard Voting Ensemble,” *Procedia Comput Sci*, vol. 252, pp. 748–757, 2025, doi: 10.1016/J.PROCS.2025.01.035.

[23] N. R. Das, M. M. Alam, A. P. Polok, M. Raquib, A. Al Mamun, and M. J. Hossen, “Social Media Bangla Fake News Detection Using Deep and Machine Learning Algorithms,” *International Journal of Engineering Trends and Technology*, vol. 72, no. 5, pp. 346–354, 2024, doi: 10.14445/22315381/IJETT-V72I5P135.

[24] I. Ljubi, Z. Grgić, M. Vuković, and G. Gledec, “Detecting Disinformation in Croatian Social Media Comments,” *Future Internet*, vol. 17, no. 4, p. 178, 2025, doi: 10.3390/FI17040178.

[25] M. Benjamin, J. Fajinmi, and O. Joseph, “Machine Learning Approaches for Detecting Fake News in the Afan Oromo,” *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 6, pp. 1–18, 2024, doi: 10.11591/EEI.V13I6.8016.

[26] N. B. Defersha, K. K. Tune, and S. T. Abate, “Topic Words-Based Multilingual Hateful Linguistic Resources Construction for Developing Multilingual Hateful Content Detection Model Using Deep Learning Technique,” *IET Inf Secur*, vol. 2025, no. 1, p. 26, 2025, doi: 10.1049/ise2/6068177.

[27] E. Daisy, “AI-Powered Social Media Monitoring: Leveraging Natural Language Processing for Real-Time Cyberbullying Detection on Twitter,” 2025. Accessed: Nov. 16, 2025. [Online]. Available: https://www.researchgate.net/publication/390877954_AI-Powered_Social_Media_Monitoring_Leveraging_Natural_Language_Processing_for_Real-Time_Cyberbullying_Detection_on_Twitter

[28] K. M. M. Uddin, H. Hamim, M. N. T. Mim, A. Akhter, and M. A. Uddin, “Machine learning and deep learning-based approach to categorize Bengali comments on social networks using fused dataset,” *PLoS One*, vol. 19, no. 10, pp. 1–35, 2024, doi: 10.1371/JOURNAL.PONE.0308862.

Md. Mubtasim Fuad Khan

221-35-883

Quick Submit

Quick Submit

Daffodil International University

Document Details

Submission ID

trn:oid:::1:3450146138

Submission Date

Dec 23, 2025, 10:35 AM GMT+6

Download Date

Dec 23, 2025, 1:41 PM GMT+6

File Name

221-35-883_SWE_Thesis.pdf

File Size

1.5 MB

78 Pages

17,828 Words

98,690 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

221-35-883

ORIGINALITY REPORT

19%	15%	12%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Midlands State University Student Paper	2%
2	arxiv.org Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	ebin.pub Internet Source	<1%
5	umpir.ump.edu.my Internet Source	<1%
6	Submitted to Daffodil International University Student Paper	<1%
7	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufacturing", CRC Press, 2025 Publication	<1%
8	jurnal.itg.ac.id Internet Source	<1%
9	americaspg.com Internet Source	<1%
10	Davy Viriya Chow, Felicia Natania, Oliverio Theophilus Nathanael, Karli Eka Setiawan, Muhammad Fikri Hasani. "Cyberbullying Detection: An Investigation into Natural Language Processing and Machine Learning Techniques", 2023 5th International	<1%

	Conference on Cybernetics and Intelligent System (ICORIS), 2023 Publication	
11	researchinnovationjournal.com Internet Source	<1 %
12	"Advances in Emerging Technologies and Computing Innovations", Springer Science and Business Media LLC, 2025 Publication	<1 %
13	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computational Methods in Science and Technology", CRC Press, 2026 Publication	<1 %
14	ejurnal.seminar-id.com Internet Source	<1 %
15	Arvind Dagur, Sohit Agarwal, Dhirendra Kumar Shukla, Shabir Ali, Sandhya Sharma. "Artificial Intelligence and Sustainable Innovation - Volume 3", CRC Press, 2026 Publication	<1 %
16	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
17	Submitted to School of Oriental & African Studies Student Paper	<1 %
18	Submitted to University of Florida Student Paper	<1 %
19	Submitted to University of Hertfordshire Student Paper	<1 %
20	www.coursehero.com Internet Source	<1 %
	internationalhatestudies.com	

Account Clearence