Cyberbullying Detection and Sentiment Analysis in Bangla Social Media using Deep Learning Techniques

by

Md. Shihab Mia
ID: CSE2201025151

Rima Khatun
ID: CSE2201025180

Mst. Akhi Akter
ID: CSE2201025100

Mst. Farzana Akter Piya
ID: CSE2201025161

Aidul Islam
ID: CSE2201025125

Supervised by
Salma Tabashum

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY (SU)

January 2026

Cyberbullying Detection and Sentiment Analysis in Bangla Social Media using Deep Learning Techniques

by

Md. Shihab Mia
ID: CSE2201025151

Rima Khatun
ID: CSE2201025180

Mst. Akhi Akter
ID: CSE2201025100

Mst. Farzana Akter Piya
ID: CSE2201025161

Aidul Islam
ID: CSE2201025125

Supervised by
Salma Tabashum

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY (SU)

January 2026

APPROVAL

The thesis titled "**Cyberbullying Detection and Sentiment Analysis in Bangla Social Media using Deep Learning Techniques**" submitted by Md. Shihab Mia (CSE2201025151), Rima Khatun (CSE2201025180), Mst. Akhi Akter (CSE2201025100), Mst. Farzana Akter Piya (CSE2201025161) and Aidul Islam (CSE2201025125) to the Department of Computer Science and Engineering, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Salma Tabashum
Lecturer & Asst. Coordinator,
Department of Computer Science and Engineering
Sonargaon University (SU)

Supervisor

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 1

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 2

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 3

DECLARATION

We, hereby, declare that the work presented in this report is the outcome of the investigation performed by us under the supervision of **Salma Tabashum,** Lecturer and Assistant Coordinator**,** Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. We reaffirm that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

(Salma Tabashum)
Supervisor

Signature

Md. Shihab Mia
ID: CSE2201025151

Rima Khatun
ID: CSE2201025180

Mst. Akhi Akter
ID: CSE2201025100

Mst. Farzana Akter Piya
ID: CSE2201025161

Aidul Islam
ID: CSE2201025125

ABSTRACT

Cyberbullying is an emergent threat in web-based social media platforms, where the use of offensive and abusive language can impact the mental well-being and social health of users. This issue is exacerbated in low-resource languages like Bangla, where the creation of automated moderation systems is hindered by complex language structure, informal writing style, and limited availability of annotated corpora. Bangla (or Bengali) is the most popular language in the Indo-Aryan family of languages and one of the top seven languages by number of speakers, with approximately 242 million native speakers, along with 43 to 44 million who speak it as a second language. A deep learning-based framework for cyberbullying detection and sentiment analysis in Bangla social media text, to address the better detection of harmful content in digital correspondence. This thesis has been made possible by ideas drawn from a number of studies. A corpus of labeled sentences drawn from Bangla comments was collected and preprocessed by normalizing noisy text, removing noise, and tokenizing to deal with challenges, including spelling variation, code-mixing, and idiomatic expressions that are found in abundance on social platforms. Three neural structures—LSTM, BiLSTM, and a transformer-based BanglaBERT model—were employed for binary cyberbullying classification and sentiment polarity analysis to assess the emotional orientation of text. The model was evaluated with standard metrics such as accuracy, precision, recall, and F1-score. The experimental results show that the accuracy of the LSTM model is 81.21%, and the BanglaBERT model has a higher 81.70% accuracy, implying the efficiency of bidirectional context learning—results of the Task layer Models. The BiLSTM model achieved 81.80% accuracy, 81.77% precision, 81.825% recall, and an F1 score of 81.75%, demonstrating that RNN-based contextual representations are effective for representing Bangla text. Our study results indicate that transformer-based models are able to learn the semantic and contextual subtleties of Bangla social media language more effectively than transformer-based approaches. This work is a step towards Bangla natural language processing by verifying deep learning models and will serve as a practical resource to build automated systems that can help ensure safer, responsible Bangla in online communities.

Keywords: Cyberbullying detection; Low-resource language; Deep learning; Sentiment analysis; LSTM; BiLSTM; BanglaBERT; Transformer models; Natural language processing (NLP).

ACKNOWLEDGMENT

Indeed, we first praise Almighty Allah (Glorified and Exalted is He) for bestowing upon us success, sincerity, and clear understanding, enabling us to complete this thesis.

The authors are thankful to our esteemed supervisor, **Salma Tabashum**, Lecturer and Assistant Coordinator, Department of Computer Science and Engineering, **Sonargaon University**, Dhaka, Bangladesh, for sharing her knowledge, providing guidance and encouragement, and giving valuable suggestions on all aspects during the present research. Her meaningful feedback and academic guidance motivated me for this work.

We are also thankful to **Imran Hossen**, Lecturer and Assistant Coordinator, Department of Computer Science & Engineering, **Sonargaon University**, Dhaka, Bangladesh, for the technical support, kind help, and valuable suggestions at various stages during this thesis work. We appreciate him for his cooperation and advice, which were very helpful in conducting this study.

Additionally, we express our heartfelt thanks to **Bulbul Ahamed**, Professor and Head, Department of Computer Science and Engineering, **Sonargaon University**, Dhaka, Bangladesh for providing us with important guidelines during study period, which mostly motivated us.

We are also grateful to all faculty members of the Department of Computer Science and Engineering, **Sonargaon University**, for their support with academic consultation during the academic period.

To conclude, an important consideration, we acknowledge with gratitude the contribution to our study made by the Department of CSE at **Sonargaon University**, which facilitated us in a good academic environment.

Finally, we express our appreciation to our families and friends for their patience and tolerance, encouragement, inspiration, and moral support during our academic journey.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BPTT	Backpropagation Through Time
CCS	Candidate Cell State
CSE	Computer Science and Engineering
CSU	Cell State Update
CSV	Comma Separated Values
DL	Deep Learning
DSA	Digital Security Act
EDA	Exploratory Data Analysis
FN	False Negatives
FP	False Positives
FG	Forget Gate
HS	Hidden State
IG	Input Gate
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
OG	Output Gate
OOV	Out-Of-Vocabulary
RNN	Recurrent Neural Network
RoBERT	Robustly Optimized Bidirectional Encoder Representations from Transformers
SU	Sonargaon University
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negatives
TP	True Positives
UNICEF	United Nations Children's Fund
X	Twitter (formerly Twitter)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 The Emergence of Digital Connection and the Issue of Cyberbullying

The digital age has brought about drastic changes in human communication, interaction, and information sharing. Today, due to the massively increased accessibility of the internet and individuals' ability to use social media all around the world, we are able to have instantaneous conversations with one another in a way never achieved before. Platforms like Facebook, Instagram, X (formerly Twitter), YouTube, and TikTok have democratized the power to create and share content in a way that allows users to express themselves on issues and build community around interests and subjects of their choice, as well as engage in global conversations regardless of where they are located on the planet. This interconnectedness is driving revolutions in education, social activism, e-commerce, and entertainment, giving voice to the underrepresented while enabling cross-cultural conversations.

However, this online expansion has brought along all the social ills one could think of; prime among them is cyberbullying. Cyberbullying is known as "electronic aggression" and involves the intentional use of technology to harass, threaten, or intimidate people. Different types of harassment occur on online platforms like social media sites, messaging apps, and forums. It ranges from spreading false rumors and defamatory comments to personal attacks to "doxxing" (spreading of personal information without consent) to impersonation and so forth. The cloak of anonymity online tends to embolden the abuser, minimizing the risk of backlash and multiplying the reach of his abuse.

Unlike traditional bullying forms that are limited to physical environments such as schools and workplaces, cyberbullying exists in both space and time, thus surpassing the borders of a school or block. Victims may face relentless targeting, making it possible that harassment will follow them into personal spaces through cellphones, computers, or other electronic media. This all-encompassing quality makes the victim feel even more helpless, as there is little or no chance of escape. Additionally, the digital nature of such acts provides longevity, as offending content can be shared and re-shared to infinity, embedded in the annals of cyberspace for years to come.

The effects of cyberbullying reach much farther than simple distress and can take a heavy toll on mental health. The poorer qualities of life among victims have been shown in empirical studies where those victims experienced significant levels of anxiety, depression, and low self-esteem. In the worst cases, it has been associated with suicidal urges and attempts, especially among young people whose impressionable minds seek the imprimatur of society. One in three young people worldwide reports being the victim of cyberbullying, an epidemic on a massive scale[1]. This becomes even more acute in the age of digital communications, with social media assuming increasing dominance in identity construction and peer engagement[2]. With greater digital penetration levels, the threat of cyberbullying is expected to increase. Although great strides have been made to

develop interventions in English-dominant settings, the scaling of digital access across linguistically diverse areas requires local solutions. Languages, for instance, Bangla, spoken by over 230 million people and predominantly used in Bangladesh and India, have not yet been sensibly considered in cyberbullying studies. This gap highlights the need for new detection mechanisms targeting non-English languages, leveraging NLP/ML-based models to protect online communities.

Bangla social media is a significantly resource-deprived language, and in light of this thesis work, we remind you of the significance of looking at the cyberbullying issue with respect to low-resourced languages. This article is an attempt to identify the potential for new combinations of digital access and online violence, with a view to informing more secure internet environments, considering that in some contexts digitization has increased so rapidly that safeguarding procedures lag behind by far.

1.2 Cyber-bullying in Bangladesh: An Emerging Digital Crisis

But Bangladesh is not the only example of digital transformation at an accelerated pace in many developing countries. By the end of 2025, there were around 82.8 million internet users in the country, which accounted for roughly a 47% penetration rate. They also forecast an increase to 53.95% in 2025, when household penetration was already at around 54.8% in each of the last few quarters. Pulled by low-cost smartphones and growing mobile networks, this explosion has woven social media deep into daily life—particularly among younger people who make up roughly half of the population[3]. Powerhouses such as Facebook and YouTube rule the day, becoming the centers of social interaction, news consumption, and entertainment.

But this digital explosion has not been without a flip side; as with all the good things, there also comes bad. The boom has brought about a rise in cyberbullying. What started as places of empowerment have become cesspools, with harassment running rampant. Cyberbullying in Bangladesh, meanwhile, has crossed with societal fault lines, becoming a form of religious intolerance, political vendetta, gender-based violence, and ethnic discrimination. Public figures such as actors, politicians, and social media influencers are a familiar target of organized online attacks designed to undermine their reputation or scare them.

Women are most impacted, making up 80% of the victims of cyberbullying. New numbers from 2025 show that 59% of women online experience harassment, although as many as 90% later report these cybercrimes. In schoolchildren, 49% disclose cyberbullying, and among university students, there are widespread experiences of abusive messaging and image-based abuse. In 2022, cyberbullying represented more than 52% of reported online crimes, such as pornography distribution and indecent communications[4].

This crisis is exacerbated by the Bangladeshi culture. Societal codes that glorify family honor and communal image frequently discourage victims from pursuing their cause, apprehending social boycott. Hate speech is aggravated by religious sensitivities; actions directed at minorities or deviant voices have a way of moving from the virtual to the real

world. Political polarization has only contributed to elevating the trolls and threats that assail campaigns of opposition figures (who are their usual target).

Legal responses remain inadequate. Cybercrimes are addressed by the 2018 Digital Security Act, but its enforcement is uneven, and it does not have specific clauses on cyberbullying. Victims must traverse a maze of bureaucracy and face poor digital literacy, as well as barriers to accessing support services. This vacuum enables abusers to act with impunity, leading to an abusive cycle.

By comparing with a dataset of 44,001 Bangla social media comments, the distribution of different types of cyberbullying can be seen: religious (7,578 instances), sexual (8,927), troll (10,462), threat (1,694), and non-bullying comments (15,340). They typically target actresses (26,950) and socialites (9,375), and in 29,949 cases, the woman is female. The data make apparent the gendered and categorical facets of this problem in Bangla contexts[5].

Many steps are required to address this crisis, including tech breakthroughs in identification and cultural rewiring around digital ethics. This thesis is situated at the confluence of these demands and endeavors to counter cyberbullying using sophisticated computational techniques designed for Bangla social media.

1.3 The Psychological and Social Outcomes of Cyberbullying.

There is a place for silence and being the bigger person, but unfortunately, common wisdom tells us to feed a predator. Because that's what bullying is—it preys on another's weakness like an animal, attacking without reason or compassion. Cyberbullying does so much damage in part because, as humans, we are capable of doing so much evil; when we're hurting and turn away from one another, that hurt persists even after the bully goes home at night. The pride, the unveiling, and the secrecy behind this repetitive ritual consistently undermine one's self-worth, leading to mental erosion. Studies from South Asia point to increased stress, poor social health, and risk of depression among victims during adolescence. In Bangladesh, the taboo on mental health discourse amplifies these consequences, resulting in underreporting and untreated trauma[6].

For teens who populate social media, cyberbullying interrupts those developmental achievements. It cultivates anxiety, anger, and academic decline as a result of constant online scrutiny that chips away at self-esteem. Research links it to severe depressive disorders, with female victims disproportionately burdened because the harassment is defined by gender[7]. Digital devices are everywhere, so now harassment pours into safe spaces, causing permanent hypervigilance and sleep disruption. In some cases, it results in suicide-related behaviors: The region is experiencing skyrocketing youth suicide rates tied to online abuse[8].

In our society, cyberbullying undermines community trust and polarizes public conversation. It atrophies aggression, retarding free expression and building echo chambers of hate. Online harassment intensifies social fault lines, and in Bangladesh, where religious and political divisions run deep, it has the potential to foment offline violence. A

second dimension is the economic cost: victims have lower productivity and are burdensome to health care.

In Bangladesh, cultural challenges make isolation worse, as victims may blame themselves to save family honor and refuse help until it's too late. There are few enough counselors and awareness programs to offer little support[9]. Efforts like school-based education and hotlines are in their infancy, but are insufficient for the scale of the problem.

The data set on which this thesis is based contains the emotional undertone: emotionally laden words combined with bullying tend to be negative, leading to reaching out and psychological harm. The study seeks to quantify these impacts using sentiment analysis and inform focused interventions.

Ultimately, if we do not tackle cyberbullying, society will be at risk of losing social cohesion, and people need to act now and implement strategies that can create empathic digital cultures.

1.4 Challenges of Cyberbullying Detection in Low-Resource Languages

A major challenge in combating cyberbullying issues in Bangladesh is the absence of efficient detection systems for Bengali, which is the country's native language. Although notable improvements have been achieved in cyberbullying detection systems for high-resource languages, such as English, the scarce and imbalanced datasets constrain a lot for low-resource languages like Bangla. The absence of large public annotated datasets for Bangla has been a bottleneck in creating automated processing methods, as they are essential for training machine learning models.

Apart from the data scarcity, due to its complex nature, the Bangla language itself poses a challenge to natural language processing (NLP) models. Bangla is morphologically rich; that is to say, words can take a wide variety of forms according to their grammatical environment. This complexity challenges the traditional text-matching models to effectively process Bangla text. And then there’s the challenge of scripting: Bangla is frequently written in a mash-up of scripts, from its native Bengali script to Romanized Bangla (or “Banglish”), in which English words are sprinkled throughout sentences written in Bangla. Such code-switching poses a new challenge to the understanding of social media content, since it introduces language variation that is hard for current NLP models to accommodate[10].

Further, the majority of NLP tools and resources are designed for formal written text (news articles or academic papers). On the other hand, social media content is full of informal and slang language that can be context-dependent and may require more advanced models to understand the nuances of online communication. However, recent developments in deep learning and transformer-model-based approaches like BanglaBERT promise to improve the performance of NLP tasks for Bangla[11]. But the absence of a task-tailored, large-scale dataset for cyberbullying detection still hinders further progress. This is what the present study tries to address by proposing a deep learning-based cyberbullying detection system for Bangla social media. The system will utilize the state-of-the-art NLP approaches

and exploit a curated dataset to train models for accurate identification of cyberbullying and sentiment analysis from Bangla text.

1.5 Problem Statement

The main issue in the context of this research is the lack of an efficient and automated technique for cyberbullying detection and sentiment analysis in Bangla social media posts. The core challenges include:

1. **Lack of Labeled Data:** There are limited publicly available annotated datasets in Bangla for social media content, and especially for tasks like cyberbullying detection and sentiment analysis.
2. **Complexity of Bangla:** The rich morphology in Bangla, consisting of a high amount of inflection and derivation, etc., and a very special script, causes the conventional NLP pipeline to be harder. Also, the blending of Bangla with English and Romanized Bangla makes the analysis even tougher.
3. **Not as many developed NLP resources for Bangla:** Although we have seen significant progress in NLP technologies such as English, there are very few pretrained models, embeddings, and other resources that can aid high-level text-based processing and analysis in Bangla.
4. **Transfer from English Model:** Many of the previous cyberbullying detection systems built for English text are not performing well when tested with Bangla, as each language has its own linguistic and cultural differences.

Solutions: We work on this research to create a Bangla-specific corpus for cyberbullying detection. Training the deep learning model that can effectively detect cyberbullying behavior expressed in social media text and developing sentiment analysis models that output the textual data as either positive, negative, or neutral.

1.6 Research Objectives

The primary aim of this work is to propose a system that identifies cyberbullying and performs sentiment analysis on Bangla social media posts. Specifically, the objectives are:

To Dataset Creation and Annotation: The objective of this phase was to construct a large labelled dataset from Bangla's social media posts, where the annotations will be focused on four different categories of cyberbully types (sexual, ideological threat, troll, and sentiment) with three polarities (positive, negative, and neutral).

To Feature Engineering: To experiment with possible data preprocessing, we can apply bag of phrases, stop word removal, word embedding (word2vec, fastText, BanglaBERT), etc.

To Model Development: Creating and comparing different deep learning models for cyberbullying and sentiment analysis (LSTM, BiLSTM, BanglaBERT)

To Performance Evaluation: The models will be evaluated based on the following standard evaluation metrics, such as accuracy, precision, recall, F1-score, etc.

1.7 Research Questions

To guide the development of this thesis, the following research questions are proposed:

RQ1: How can a labeled dataset of Bangla social media content be created to support both cyberbullying detection and sentiment analysis tasks?

RQ2: Which deep learning models (LSTM, BiLSTM) would be efficient in cyberbullying and sentiment analysis detection using Bangla text?

RQ3: What are the challenges specific to Bangla text, such as linguistic and comprehension-related aspects (morphology, code-switching, and use of Romanized Bangla), and how can they be addressed while creating a cyberbullying detection system?

RQ4: What is the influence of emotion on cyberbullying accuracy, and how does putting it into a negative or positive category enhance model precision?

1.8 Significance of the Study

Importance: This research is potentially life-saving, as there are 260 million Bangla-speaking individuals internationally. By creating a system that is capable of handling the linguistic characteristics of Bangla, this work will present an indispensable means of monitoring online spaces, detecting cyberbullying behavior, and helping targeted individuals. Additionally, the methods that have been developed in this work will be useful to further progress NLP for low-resource languages and are suggestive of how deep learning models can be altered to accommodate non-English languages.

1.9 Organizations of Thesis Book

The thesis titled “Cyberbullying Detection and Sentiment Analysis in Bangla Social Media using Deep Learning Techniques” is structured as a comprehensive academic document, following a conventional format for undergraduate theses in computer science and engineering. It is organized into six chapters, supplemented by preliminary sections (e.g., cover page, approval, declaration, abstract, acknowledgements, list of abbreviations, table of contents, list of tables, and list of figures) and concluding elements (e.g., references). **Chapter 1: Introduction**, is foundational chapter introduces the research domain. **Chapter 2: Literature review**, identify the gaps that the thesis addresses. Chapter 3: **Sequential And Attention-Based Modeling In Natural Language Processing**, focusing on the theoretical foundations. Chapter 4: **Methodology and Model Design,** outlines the research workflow in a modular, reproducible manner. Chapter 5: **Results and Performance Analysis**, this empirical chapter presents findings systematically. Chapter 6: **Conclusion and Future Research Directions**, the concluding chapter synthesizes contributions and implications.

CHAPTER 2

LITERATURE REVIEW

2.1 The Global Problem of Cyberbullying and Its Impact on Mental Health

Cyberbullying is one of the most toxic kinds of harassment and is now a global public health concern [12]. The impact this has upon the victims' mental health is devastating, and it affects them long-term, with depression and anxiety for some, and in extreme cases [13]. Cyberbullying is the use of digital platforms (e.g., social media [Facebook, Twitter, Instagram]) for the purpose of causing harm to others by posting hurtful or abusive commentary. Unlike the usual in-person bullying, cyberbullying can take place anytime, anywhere and can stigmatise victims even in their unguarded hours. Social media has been the game-changer that took the problem and scaled it up exponentially. On platforms like Facebook and YouTube, abusers have the power of anonymity and an audience to dox people or entire groups without any consequences [14]. This means that the abuser can virtually have infinite access to the victim without a real limit, as opposed to physical bullying, which can be limited by time and space [15]. The trail stones of these communications are usually forever; they keep the hurt."

In Bangladesh, the steeply rising curve of internet usage, especially by the young generation, generates an alarmingly high rise in cyberbullying cases [16]. The widespread availability of smartphones and internet facilities has made it easier for the Bangladeshis to express themselves in the digital space. But this rise in online participation has also brought with it a spate of cyberbullying incidents. On the one hand, the existing global attention on cyberbullying has primarily focused on English-based social media systems; hence, Bangla-language-based social media platforms in Bangladesh pose novel challenges for the detection of harmful content. Existing detection models are primarily developed using English language-based approaches and cannot cater to the intricate morphological diversity of Bangla, code-switching between Bangla and English and the informalism of social media text [17]. These are challenges that reinforce the idea of tailored detection systems able to overcome those particular linguistic peculiarities in order to fairly recognise cyberbullying behaviour.

Specifically, the morphology of Bangla has great morpho-lexical variation, where words can have many forms based on grammatical context (such as verb conjugations and noun declensions). Added to the informal and sometimes slang language used on social media platforms, we have a very noisy dataset, which can cause the traditional classifiers not to be able to deal with that. This absence of appropriate language resources for Bangla is a major obstacle to constructing strong, effective systems aimed at identifying cyberbullying in the locale [18].

2.2 Linguistic Challenges in Cyberbullying Detection for Bangla

2.2.1 Morphological Complexity and Code-Switching

The morphological complexity of Bangla is one of the critical issues to address in cyberbullying detection in Bangla. Bangla is a language that follows inflectional rules, so words are modified for tense, person, number and respect. This is very different from the English language, where word forms are far more stable, and you have few, if any, variations [19]. For this reason, word-level models operating under fixed vocabulary are challenged with data sparsity as various appearances of the same word are being considered separately [20]. For example, the verb "খাওয়া" (eat) can be formed as "খাচ্ছি" (eating), "খেয়েছি" (ate), and "খাবে" (will eat), which pose difficulties in identifying patterns and consistently detecting harmful content such as cyberbullying across various formats of a word.

Besides, code-switching is also common in Bangla social media comments, where users mix together Bangla and English ('Banglish'). This adds more challenge, as the model should grasp knowledge of two languages simultaneously [21]. This lexical fluidity leads to sentences like "তুমি really বাজে কথা বলছ" (You are really bad at speaking bad words), where Bangla and English get juxtaposed. An effective system for detection must be able to handle such multilingual text and still not lose the context or meaning of words. One may worry that traditional NLP models that perform well on monolingual data do not capture this interplay between languages [22].

Word embeddings such as Word2Vec and FastText can help mitigate this by learning distributed representations of words that "capture" meaning in a continuous vector space, thereby enabling models to generalise across word forms and languages. Thus, by mapping mixed-language text into the shared space, these models can capture the intrinsic meaning of the words even when they are presented in different languages. However, although word embeddings provide a solution in that regard, they still need vast pre-trained corpora in order to accurately represent subtleties and may not be able to cope with the fast change of slang and colloquial words for Bangla, which makes word embeddings less useful in cases of rapidly evolving digital ecosystems [23].

2.3 Sentiment Analysis for Cyberbullying Detection

Sentiment polarity analysis is one of the essential elements of a cyberbullying detection system [26]. The most reliable signal of potential harm or problematic content in a post is its emotional tenor [25]. Negative comments – ones that express anger, frustration or hate – are also associated with abusive language and, therefore, are an indicator of a potential cyberbully [24].

The tool applies a lexicon-based sentiment analysis approach, with precompiled lists of positive and negative words in Bangla. The procedure enumerates the positive and negative words in each posting, calculating a sentiment score. This score is used to classify

a post as being positive, negative, or neutral. Negatively labelled posts are considered under the potential category of cyberbullying.

2.3.1 Formula for Sentiment Score Calculation

The sentiment score for a given comment is calculated using the following formula:

$$\text{Sentiment Score} = \frac{\text{Positive Words Count} - \text{Negative Words Count}}{\text{Total Words in comment}}$$

Where:

- Positive Words Count represents the number of positive words found in the comment.
- Negative Words Count represents the number of negative words found in the comment.
- 'Total Words in Comments' is the total number of words in the comment.

Using this formula, the post is then classified into the following categories:

- Positive if the sentiment score > 0.1
- Negative if the sentiment score <− 0.1
- Neutral if −0.1 ≤ Sentiment Score ≤ 0.1

This method works well for identifying toxic comments and is a valuable tool in the detection of cyberbullying [27].

2.4 Deep Learning Models for Cyberbullying Detection

Text classification tasks, including cyberbullying detection, are now dominated by deep learning models [28]. This is where RNNs come into place. RNNs are especially good for sequential data such as text because the order of words matters when understanding a meaningful piece of content [29]. However, the traditional RNNs are still problematic due to their vanishing gradient phenomenon, as they do not effectively capture long-range dependencies in the text [30].

This was addressed with the introduction of Long Short-Term Memory (LSTM) networks. LSTM can have long-term memory, which is critical for recognising the context of a post and recognising subtle signs of behaviour that indicate cyberbullying [27]. The BiLSTM (Bidirectional Long Short-Term Memory) model enhances the LSTM with processing for text in the forward and backward directions so that we can understand context from previous words and following words as well [31].

2.4.1 RNN Layers in the Model

The core architecture of the RNN and LSTM models includes several layers, each with a specific role in learning the sequence of words.

2.5 Layers in Recurrent Neural Networks (RNN) Models

2.5.1 Embedding Layer: Translating Text to Vectors

The First Layer of DL (Deep Learning) Model is an Embedding Layer. It transforms each word in the raw text into a dense vector with useful semantic information. This processing makes the network capable of learning patterns and relations between words, which is then helpful in generalising over different word forms and synonyms. The embedding layer is designed to map words into a denser, continuous vector representation so that similar words have similar vectors [32].

Formula for Embedding Layer:

$$\text{Embedding Output} = x_t$$

Where:

- x_t is the input word at time step t.

The embedding layer outputs a continuous representation for each word in the sequence.

This representation is essential for handling variations in words and expressions, such as synonyms or slang (e.g., "খারাপ", "মন্দ", and "নিকৃষ্ট" all meaning "bad" in Bangla)[33].

2.5.2 LSTM/BiLSTM Layer: Memory for Long-Term Dependencies

The LSTM layer has to look at a sequence of word vectors, and it keeps track of an internal memory cell, allowing it to store information over long sequences. This lets the LSTM remember significant context from much earlier in the sequence, which is necessary for capturing subtle behaviour such as cyberbullying.

In BiLSTM, the architecture reads the sentence in two directions (from left to right and from right to left), which helps in improving its interpretability since it takes into account past and future context.

Formula for LSTM Update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t$$

Where:

- C_t is the cell state at time step t,
- f_t is the forget gate,
- i_t is the input gate,
- C_t is the candidate cell state.

Formula for BiLSTM:

- Forward hidden state, h_t^{fwd}:

$$h_t^{fwd} = \text{LSTM}\left(x_t, h_{t-1}^{fwd}\right)$$

- Backwards hidden state, h_t^{bwd}:

$$h_t^{bwd} = \text{LSTM}(x_t, h_{t+1}^{bwd})$$

The outputs of both directions are then combined, giving the model a richer representation of the input text.

2.5.3 Dropout Layer: Preventing Overfitting

The Dropout layer serves to prevent overfitting at training time. It does so in such a way that some percentage of the neurons are "randomly" switched off at every step through training. The latter is known to prevent the model from automatically learning independent features and could potentially allow for better generalisation among novel data [34].

Formula for Dropout Layer:

$$y = x.r$$

Where:

- x is the input to the layer.
- r is a random dropout mask applied to the neurons (e.g., 50% dropout means each neuron has a 50% chance of being turned off).

The Dropout layer ensures that the model does not memorise the training data but learns to generalise effectively.

2.5.4 Dense Layer: Final Classification

Dense is the last fully connected layer that takes in the output of the LSTM/BiLSTM layer and makes a class decision out of it. For this task, the dense layer categorises whether a post is cyberbullying or non-cyberbullying.

Formula for Dense Layer:

$$y = \sigma(Wx + b)$$

Where:

- W is the weight matrix,
- b is the bias,
- σ is the activation function (typically sigmoid for binary classification or softmax for multi-class classification).

2.6 BanglaBERT: Pre-trained Transformers for Bangla

BanglaBERT is a transformer-based model trained on a large corpus of Bengali text, and it is suitable for tasks that need deep contextual understanding. Contrary to LSTM and BiLSTM, which treat text in sequence, BERT (Bidirectional Encoder Representations from Transformers) utilises a self-attention mechanism, which allows us to consider all the words present in the sentence simultaneously. This allows the model to attend to different

parts of the text given their relevance, which is important for identifying subtle cyberbullying cues.

Formula for Self-Attention:

$$\text{Attention Score} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

Where:

- Q is the query matrix,
- K is the key matrix,
- d_k is the dimension of the key.

BanglaBERT can be fine-tuned on specific tasks, improving its ability to recognise cyberbullying patterns in Bangla social media text.

2.7 Evaluation Metrics: Performance Assessment of Cyberbullying Detection Models

2.7.1 Confusion Matrix

A confusion matrix is one of the core indicators to analyse classification model performance, which is especially important in imbalanced data sets such as cyberbullying detection [35]. In these models, conventional accuracy is not a measure that can give the true value of the model because it does not account for the types of errors (i.e., false positives and false negatives) [36]. The confusion matrix provides a more detailed analysis, showing the model's performance on each of the classes in the dataset [37]. The confusion matrix is structured as follows:

$$\text{Confusion Matrix} = \begin{matrix} TP & FP \\ FN & TN \end{matrix}$$

Where:

- **True Positives (TP):** How many cyberbullying posts that have been identified by the model as cyberbullying were right.
- **False Positives (FP):** The number of cases in which the model incorrectly predicted a non-cyberbullying post as cyberbullying.
- **True Negatives (TN):** The cases in which the model correctly predicted non-cyberbullying posts as non-cyberbullying.
- **False Negatives (FN):** The instances of the model that labelled a cyberbullying post as non-cyberbullying.

Every entry in a confusion matrix offers information about how the model classifies cases. True positive and true negative mean the model is working very well as intended, but false positives and false negatives are bad. A false positive could result in unnecessary intervention on a non-abusive post, while a false negative might allow harmful behaviour to pass undetected.

The confusion matrix enables the computation of important evaluation measures such as:

- It shows the ratio of relevant instances (correctly annotated cyberbullying posts) over retrieved instances (all posts labelled as cyberbullying).
- Recall is a measure of what percentage of true cyberbullying posts the model can correctly identify.
- F1-score, which is the harmonic mean of precision and recall, balances between false positives and false negatives.

This confusion matrix is the foundation for evaluating model performance beyond raw accuracy and allows us to consider how well the model can differentiate between cyberbullying content and non-cyberbullying content.

2.7.2 Precision, Recall, and F1-Score

It is important to mention that those performance scores originating from the confusion matrix are essential in determining a cyberbullying detection model's success. The high class imbalance of cyberbullying posts vs. non-cyberbullying posts makes precision and recall, as well as the F1-score, more informative metrics compared to accuracy for model performance evaluation [38].

Sensitivity (Recall): It indicates how well a model identifies all the cases of cyberbullying. This is especially important when applied to public health settings; missing an actual cyberbullying posting may have dire consequences. " Raising recall reduces the number of bad posts that go undetected. It is calculated as:

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

Where:

- TP = True Positives (correctly identified cyberbullying posts),
- FN = False Negatives (missed cyberbullying posts).

Precision: Precision represents how many of the posts found to be cyberbullying (by our model) are actually cyberbullying. As recall focuses on discovering harmful posts, precision ensures that moderation intervention (e.g., moderation intervention) is invoked only in cases when we are very confident the detected post is actually harmful. It is calculated as:

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

Where:

- FP = False Positives (non-cyberbullying posts which were erroneously classified as cyberbullying).

F1-Score: The F1-score aggregates precision and recall, balancing the two. It is especially helpful when there is a compromise between false positives and false negatives. The F1 score is calculated as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In cases such as cyberbullying detection, both false pressures (unnecessary intervention) and false releases (unrecognised cyberbullying) must be minimised to have a good F1-score to ensure the ethical use of resources and avoid negative consequences.

2.8 Data Challenges and the Need for Annotated Datasets

Moreover, designing a successful cyberbullying detection model for Bangla social media involves some special challenges apart from selecting the model. Among the challenges, one major challenge is to handle limited large annotated datasets in Bangla required for training machine learning-based solutions. Most available cyberbullying detection systems are developed from English datasets, and there is a substantial lack of labelled data in the case of Bangla. This gap creates a significant challenge, given that annotated data are required for training and fine-tuning models to recognise nuances of cyberbullying behaviour like trolling, harassment, and name-calling.

It is even crucial in the case of Bangla, where morphological richness adds to the complexity. Word-based models (such as traditional TF-IDF or bag-of-words-based approaches) tend not to generalise well across word forms in Bangla [39]. For instance, the verb 'খাওয়া' (to eat) can have the forms 'খাচ্ছি' (eating), 'খেয়েছি' (ate), and 'খাবে' (will eat), to name just a few examples. This variation, however, makes it hard for a deep model to learn the semantic meaning of words in the same way.

Moreover, social media text in Bangla is informally written with much slang and abbreviated forms along with code-switching between Bangla and English (Banglish). This colloquial language adds extra challenges in text processing and preprocessing, as it involves non-standard spellings, missing words, and word substitution. For example, they may type “tumi khub bad” (you are very bad) instead of saying the more formal Bangla equivalent, which again adds to the flexibility and diversity for the model to work with.

The problem of data scarcity is further exacerbated by the absence of pre-trained models for Bangla. For the English language, large models, e.g., BERT or RoBERTa, have been pre-trained over gigabytes of text. However, we do not have pre-trained models like them for Bangla, and as they are relatively less mature, it is inevitable to build task-specific pre-trained models customised with respect to cyberbullying detection [40]. The lack of pre-trained embeddings for Bangla has an impact on the embedding layer used in deep learning models such as LSTMs and BiLSTMs, which usually require large corpora of text to obtain ef fective word vectors [41].

To solve these issues, the Bangla cyberbullying detection system must use techniques that function in the data-confined conditions as follows:

1. **Data Synthesis:** Techniques, such as data augmentation and paraphrasing, can be applied to artificially increase the volume of the training data. This can be very helpful, especially in the case when there is too little labelled data.
2. **Transfer Learning:** Transfer learning with fine-tuning of BERT-based models like BanglaBERT can boost the performance over small task-specific datasets when trained on limited data [42].
3. **Semi-Supervised Learning:** For systems in which full-label data are not feasible, semi-supervised learning methods can be considered for taking advantage of unlabelled samples and a small amount of labelled ones.

Additionally, for cleaning and standardising data, preprocessing tools such as stopword removal, synonym normalisation and formalisation of colloquial expressions in the tweets are very important. The model should also be capable of understanding mixed-language text (Banglish) without losing context, which requires special treatment and careful feature construction.

There are already multiple promising developments, including an intelligent cyberbullying detection system in Bangla, although much more additional work is needed [27]. The challenge is to address data-related problems and build models which can account for idiomatic expressions (contractions and slang/colloquial forms) found in social media communication. Through constructing larger datasets and better pre-processing methods, cyberbullying detection in Bangla will be more efficient, which can help to create the necessary tools to assist in creating safe spaces in the digital environment for users who know Bangla.

CHAPTER 3

SEQUENTIAL AND ATTENTION-BASED MODELING IN NATURAL LANGUAGE PROCESSING

3.1 Evolution of Sequential and Attention-Based Models in NLP

Natural Language Processing (NLP) has transitioned from the rule-based linguistic systems to a data-driven paradigm where neural architectures can learn complex communication patterns directly from text. This transformation has been brought about by the proliferation of large-scale textual data and the progress in deep learning. These two dominant modeling paradigms in NLP, sequential and attention-based models, are at the heart of modern approaches to NLP.

This chapter concentrates on these two paradigms by discussing them under one roof: the Recurrent Neural Network (RNN)-based models, and the Transformer-based architectures. Instead of isolating them as separate methods, the chapter pairs these two techniques together as successive steps in NLP modeling. Sequential models, including LSTM and BiLSTM, focus on temporal dependencies and word order, but transformer networks use a self-attention mechanism to represent long-range dependencies and global, deep contextual relationships more effectively.

In the context of this thesis, these models are employed for sentiment analysis and cyberbullying detection in Bangla text. The chapter thus not only describes the theoretical grounding behind NLP, RNNs, and transformers, but also shows how these relate directly to algorithms & design decisions found in the experimental notebook. In this way, the chapter achieves an important conceptual link between linguistic theory, principles of neural modeling, and system development in practice.

3.2 Linguistic Representation and Computational Foundations of NLP

3.2.1 Language as Data: From Text to Representation

Human language is by nature complex and ambiguous; it also depends on context. The issue is that there are millions of other calculations a computer likes to do. The ultimate task of NLP is to bridge this gap, a goal that in practice will be addressed by going from raw text to some kind of structured numerical representations (e.g., syntactic and semantic information). This is not a single-shot transformation but a pipeline of process states, all thoughtfully laid out and all adding to the system's overall efficiency.

Conceptually, NLP works with raw text input that could be noisy, like spelling variations and punctuation marks, emojis, and hyperlinks, or language-specific symbols. In the case of social media, especially for low-resource languages like Bangla, this noise factor increases due to colloquial language, code-mixing, and creative use of language. So, preprocessing is not just an option, but a necessity.

3.2.2 Text Preprocessing and Normalization

Text preprocessing refers to a set of operations that clean and standardize raw text before it is fed into a learning algorithm. In this thesis, preprocessing includes steps such as:

1. Removal of irrelevant symbols, URLs, and extra whitespace.
2. Normalization of text to a consistent format.
3. Token-level cleaning based on language-specific rules.
4. Stop-word removal using a curated Bangla stop-word list

Normalization is significant in Bangla NLP because of multiple orthographic variants for the same word. By reducing such variations, the model is exposed to more consistent patterns, which improves generalization.

3.2.3 Tokenization and Lexical Units

Tokenization is the process of breaking text into smaller units known as tokens. These tokens can be words, subwords, or even characters. Traditional NLP models are based on word-level tokenization, while more modern versions of these approaches use subword tokenization to better account for rare and unseen words.

In the LSTM and BiLSTM in this thesis, a word-level tokenization space is applied and limited by a maximum size of vocabulary. Each token, or subword, but that won't be too relevant for our purpose, is represented by an integer index in a representation of the vocab which is discrete. This indexed representation is used for further embedding and sequence modelling.

3.2.4 Word Embeddings and Semantic Space

The introduction of word embeddings is one of the most important advances in NLP. Word embeddings are representations for discrete tokens in vector spaces, where similar words are close to one another (based on cosine distance). For example, Similarity is found based on the geometric meaning. In these spaces, words that share similar meanings will have similar vector representations.

In this study, pre-trained FastText embeddings for Bangla are used. FastText is well-suited for morphologically rich languages, as it encodes words as a bag of character n-grams. This, in turn, enables the model to produce meaningful embeddings for rare or previously unseen words. When using FastText vectors to initialize the embedding layer, linguistic knowledge from large external corpora is integrated to increase the model's performance and stability.

3.2.5 Sequence Modeling and Contextual Meaning

Language is by nature sequential: The meaning of a word is highly dependent on the words that precede or follow it. Naive bag-of-words models do not take such an order into account, resulting in loss of content. Sequence modeling is used to overcome this limitation by explicitly modeling the order and dependency structure between tokens.

In NLP applications like sentiment analysis and cyberbullying detection, context is key. The same word can work differently in context. Thus, we need to model long dependencies. This constraint drives the search for RNNs and transformer-based architectures, briefly summarized next.

3.3 Sequential Neural Modeling with Recurrent Architectures

3.3.1 Mathematical Perspective on Sequential Modeling

From a mathematical standpoint, sequential modeling seeks to learn a function f that maps an input sequence X = (x1, x2, …, xT) to an output Y, where each $x_t \in \mathbb{R}^d$. The core challenge is to ensure that the model captures dependencies across time steps, which is achieved through recurrent formulations.

Motivation for Recurrent Architectures

Feedforward neural networks assume input independence from each other. However, for sequences (i.e., text, speech, or time series), this is not true. Recurrent Neural Networks (RNNs) were designed to solve this problem by means of feedback connections, which enable the state to persist across time steps. An RNN passes over a sequence one element at a time, holding an internal hidden state that summarizes the elements it has seen so far. This hidden state serves as a form of memory, allowing the network to include past information in its predictions.

3.3.2 Mathematical Formulation of RNNs

A standard Recurrent Neural Network defines a hidden state h_t that evolves over time. Given an input sequence x1, x2, …, xT, the hidden state update is defined as:

$$h_t = \tanh(W_{xh}\, x_t + W_{hh}\, h_{t-1} + b_h)$$

Where:

- W_{xh} is the input-to-hidden weight matrix,
- W_{hh} is the hidden-to-hidden recurrent matrix,
- b_h is the bias vector.

The output y_t at time step t is computed as:

$$y_t = \text{softmax}(W_{hy}\, h_t + b_y)$$

Where:

- Y_t is the output scores,
- W_{hy} is the hidden-to-hidden weight matrix,
- b_y is the output bias vector

During training, gradients are propagated backward through time (BPTT). Repeated multiplication by W_{hh} often causes vanishing or exploding gradients, which limits the ability of vanilla RNNs to learn long-term dependencies.

At each time step t, an RNN updates its hidden state based on the current input x_t and the previous hidden state h_{t-1}:

$$h_t = f(W_x x_t + W_h h_{t-1} + b)$$

Where $W_x$and W_h are weight matrices, b is a bias term, f(·) and is a non-linear activation function. The output may be computed as a function of the hidden state.

While this formulation is conceptually elegant, standard RNNs suffer from the vanishing and exploding gradients. Gradient problems make it difficult to learn long-range dependencies.asssAA

3.3.3 Long Short-Term Memory (LSTM)

Long Short-Term Memory networks address gradient instability by introducing a gated memory cell, etc.

The LSTM cell is governed by the following equations:

Forget Gate (FG): $f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$

Input Gate (IG): $i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$

Candidate Cell State (CCS): $\hat{c}t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$

Cell State Update (CSU): $c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t$

Output Gate (OG): $o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$

Hidden State (HS): $h_t = o_t \odot \tanh(c_t)$

Here, σ denotes the sigmoid function and ⊙ represents element-wise multiplication. This gating structure allows information to persist over long sequences. (LSTM) Long Short-Term Memory networks were proposed as a solution to the limitations of standard RNNs.

LSTMs introduce a more sophisticated memory mechanism that allows information to be retained or discarded in a controlled manner.

An LSTM cell contains three primary gates:

1. **Forget Gate (FG):** Determines which information from the previous cell state should be discarded.
2. **Input Gate (IG):** Controls which new information should be added to the cell state.
3. **Output Gate (OG):** Regulates how much of the cell state should influence the output.

Through these gates, LSTMs can maintain long-term dependencies while selectively updating memory. This capability makes them particularly effective for NLP tasks involving long sentences or complex contextual cues.

3.3.4 Bidirectional LSTM (BiLSTM)

Bidirectional LSTMs process sequences in two directions. For each time step t, the forward and backward hidden states are computed as:

$$h \rightarrow t = \text{LSTMforward}(x_t)$$

$$h \leftarrow t = \text{LSTMbackward}(x_t)$$

The final representation is obtained by concatenation:

$$ht = [h \rightarrow t\,;\ h \leftarrow t]$$

This enables the model to learn from both left and right context.

(BiLSTM) LSTMs feed the input forward and backward, and concatenate or sum them at each time step. This enables the model to see information from previous and future contexts at the same time.

In NLP, this bidirectional context is particularly useful since the meaning of a word heavily depends on both its previous and following contexts. BiLSTMs concatenate forward and backward hidden states to obtain a more informative contextual representation.

3.4 Functional Role of Recurrent Models in Text Understanding

3.4.1 Strengths of RNNs in Text Analysis

RNN models are ideal for applications when the order of data is important and time dependence exists. In the field of sentiment analysis and bullying detection, a small change in word order can change the whole meaning. It also captures some nuances, which can be captured by LSTM and BiLSTM models, and is thus the baseline model with competitive performance.

3.4.2 Implementation in This Study

In the experimental notebook used in this research, both LSTM and BiLSTM models are implemented using TensorFlow and Keras. The workflow includes:

- Tokenization of text with a fixed vocabulary size
- Padding and truncation of sequences to a maximum length determined by empirical analysis
- Initialization of an embedding layer with pre-trained FastText vectors
- Construction of LSTM and BiLSTM architectures with dropout for regularization
- Training using class-weighted loss functions to address class imbalance

The models are evaluated using standard classification metrics, providing a clear comparison of their strengths and limitations.

3.5 Attention-Centric Modeling and the Transformer Paradigm

3.5.1 Motivation for Attention-Based Models

Transformer architectures were developed in response to clear drawbacks witnessed in recurrent and convolutional neural networks when dealing with long texts. Even though the vanishing gradient problem has been partially alleviated by LSTMs and BiLSTMs, they are still computational models relying on sequences of input tokens and limiting the parallelism despite the challenge of capturing very long-range dependencies. Example from tasks like cyberbullying detection and sentiment analysis, important cues can be separated by an inordinate distance or even sentence boundaries. Attention mechanisms were developed specifically for this purpose, enabling direct communication between tokens via all tokens in a sequence. The transformer architecture does not use any recurrence but attention mechanisms for word-word relationship modeling into a sequence of words. This change introduces an exciting concept into the NLP modeling:

Unlike traditional language processes, you can view the transformer as processing words together from a contextual standpoint by learning useful information through inter-attention weights.

3.5.2 Mathematical Formulation of Self-Attention

Self-attention is the core computational operation of the transformer. Given an input sequence represented as an embedding matrix $X \in \mathbb{R}^{T \times d}$, where T denotes the sequence length and d the embedding dimension, three distinct projections are learned:

$$Q = XW_Q, K = XW_K, V = XW_V$$

Here, Q (queries), K (keys), and V (values) are matrices in $\mathbb{R}^{T \times d_k}$. The attention score between tokens is computed using a scaled dot-product formulation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The scaling factor $\sqrt{d_k}$ stabilizes gradients during training. This formulation allows each token to attend to all other tokens in the sequence, producing a context-aware representation that reflects global dependencies.

3.5.3 Multi-Head Attention and Representation Diversity

Rather than using a single attention operation, transformers employ multi-head attention, which enables the model to capture different types of relationships simultaneously. Formally, for h attention heads:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$$

Where each head is defined as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

This design allows different heads to focus on syntactic dependencies, semantic similarity, or positional relationships, enriching the learned representation

3.5.4 Positional Encoding and Sequence Order

Because transformers do not process sequences sequentially, they require an explicit mechanism to encode word order. Positional encodings are added to input embeddings to incorporate sequence position information. A commonly used formulation employs sinusoidal functions:

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i})} = \sin\left(\frac{\mathrm{pos}}{10000^{\left(\frac{2\mathrm{i}}{\mathrm{d}}\right)}}\right)$$

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i}+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{\left(\frac{2\mathrm{i}}{\mathrm{d}}\right)}}\right)$$

These encodings allow the model to generalize to sequence lengths not seen during training, while preserving relative positional relationships.

3.5.5 Transformer Encoder Architecture

A transformer encoder layer consists of two primary sublayers: multi-head self-attention and a position wise feedforward neural network. The feedforward network is defined as:

$$\mathrm{FFN}(\mathrm{x}) = \max(0, \mathrm{x}\mathrm{W}^1 + \mathrm{b}^1)\,\mathrm{W}^2 + \mathrm{b}^2$$

Residual connections and layer normalization are applied around each sublayer:

$$\mathrm{LayerNorm}\big(\mathrm{x} + \mathrm{Sublayer}(\mathrm{x})\big)$$

These architectural choices improve gradient flow, stabilize training, and enable deep stacking of layers.

3.5.6 Training Pipeline and Optimization

The experimental research study uses the transformer training pipeline with Huggingface Trainer API. This pipeline consists of tokenization accompanied by subword units, dynamic padding, and batch-wise.

Optimization and periodic evaluation on validation data. Optimization is done using Adam W, which separates weight decay from the gradient updates, which can facilitate better generalization.

3.5.7 Role of Transformers in Bangla Cyberbullying Detection

The transformer-based model shows better ability in remembering subtle contextual hints, sarcasm, and implicit harassment. And so on, compared with sequential models, transformers have a direct relationship between distant words and phrases, which makes them effective in a complex language environment such as social media.

3.7 Comparative Analysis of Sequential and Attention-Based Models

3.7.1 Representational Capacity and Context Modeling

One of the key differences between recurrent and transformer-based models is how they build contextual representations. Recurrent models such as LSTM and BiLSTM generate features step by step, where each hidden state summarizes information from the past time steps. This iterative addition promotes strong local consistency while retaining temporal coherence. In the sentiment analysis and cyberbullying detection domain, this helps RNN-based models to capture accumulated advances in sentiments, negation of patterns, and also phrase-level shifts of sentiment. Transformers instead create representations by using global self-attention, where each token can look directly at all the other tokens in the sequence. This yields representations that jointly encode local and long-range dependencies. For instance, a negative term at the sentence end can directly affect prior encoding of a negated pronoun or subject without going through intermediate spreading. This representational freedom provides a key advantage to transformers for implicit harassment and context-specific toxicity.

3.7.2 Computational Efficiency and Scalability

From a computational perspective, recurrent models exhibit linear time complexity with respect to sequence length but require strictly sequential processing. As a result, training and inference cannot be fully parallelized, leading to longer training times on large datasets. This limitation becomes particularly pronounced when modeling long social media posts or conversational threads. Transformer models, despite having quadratic attention complexity with respect to sequence length, benefit from full parallelization across tokens. Modern hardware accelerators are well-suited to this computation pattern, allowing transformers to scale efficiently with dataset size. In practical terms, this makes transformers more suitable for large-scale NLP systems, even in resource-constrained research environments.

3.7.3 Data Efficiency and Transfer Learning

RNN-based models typically benefit from task-specific training and need finely-crafted features or embeddings to work in a low-resource scenario. Although pre-trained FastText embeddings address this shortcoming to some extent, they allow recurrent models to learn task representations almost entirely from scratch. In contrast, Transformers rely on a massive amount of pretraining over a large corpus using self-supervised objectives. They can exploit such information in a tightly supervised setting where the knowledge is available during pretraining and encode general linguistic knowledge that benefits downstream tasks. In this thesis, we show that transferring a pre-trained Bangla transformer model results in significantly reduced labeled data requirements to achieve comparable performance, highlighting the practical utility of learning from other languages for under-resourced ones.

3.7.4 Interpretability and Error Characteristics

Interpretability is crucial, especially in sensitive domains such as cyberbullying detection. RNN-based models have the advantage of interpretability by design, since sentiment trajectories can be interpreted using intermediate hidden states. Transformers provide an alternative type of interpretability in the attention weights that reflect how much the tokens influence each other. The attention weights, although only a partial explanation

For model decisions, they provide interpretable evidence on what words or phrases contribute most to the prediction outputs. In reality, transformer models often make fewer errors in contextually. Complex examples, but lack absolute attention to rare or high-value tokens.

3.7.5 Model Selection Rationale in The Study

The reason for both recurrent and transformer-based models considered in this work is to ensure methodological completeness and comparative understanding. RNN-based models serve as a strong. Baselines that validate the effectiveness of sequence-based modeling for Bangla text. Transformer-based models, on the other hand, are state-of-the-art and can be used as a baseline for context comprehension.

By comparing both paradigms in a common experimental setting, this thesis provides an insight into model performance characteristics, weaknesses, and strengths, instead of promoting just one particular architectural decision.

3.8 Summary and Theoretical Implications

In this chapter, we have provided the readers with a comprehensive theoretical and methodological overview of NLP from the vantage point of sequential and attention-based neural models. Starting from the linguistic and computational basics of NLP, the chapter has shown how raw text data is processed in order to be fed into spatio-temporal learning models. The conversation made clear that NLP is not a single method but multiple pipeline layers involving preprocessing, feature learning, and contextualization. We had spent most of the chapter on RNN-based networks, namely, LSTMs and BiLSTMs. We then analyzed these models conceptually and mathematically to show how the gated-memory mechanisms can learn temporal dependencies in language. Their effectiveness as sequence learners was discussed in sentiment analysis and cyberbullying, where word order, negation, and context are key to meaning. The details of implementation discussed in this chapter are an account of how these abstract principles have been made a concrete reality in the experimental notebook prepared for this thesis. The latter half of the chapter concentrated on transformer-based architectures, which are a novel framework for NLP modeling.

The transformer replaces recurrence with attention, allowing for global relations to be modeled without regard for distance and enabling full training loss computation at once in a single forward pass. We mentioned the mathematical forms, attention, multi-head mechanisms, and feedforward sublayers take to illustrate how transformers build rich context-aware representations. The fine-tuned Bangla transformer model for this work was

presented as an illustration of how large-scale pretraining coupled with task-specific adaptation can lead to major improvements in low-resource language scenarios. Comparison-wise, this chapter has demonstrated that sequential and attention-based models are two different yet complementary views for language modeling.

Recurrent models promote temporal smoothness and local dependencies, whereas transformers encourage global interactions and a more flexible representational capacity. It is therefore important to know when both approaches are efficient and what their weaknesses are in order to be able to choose the one that is most adapted for a given model.

CHAPTER 4

METHODOLOGY AND MODEL DESIGN

4.1 Overview of the Research Workflow

This chapter offers a comprehensive and structured description of the methods used in this study. The main aim of this method is to develop, implement, and assess a successful approach to Bangla cyberbullying detection via fusion of linguistic pre-processing, lexicon-based sentiment analysis, deep learning models, and Transformer-based architectures. The methodological pipeline is concisely delineated to achieve reproducibility, robustness, and scientific integrity, which are requisites for thesis-level research. The entire research process can be described sequentially and as a modular workflow from raw data input to comparison of model performance. Every stage is to some extent conditioned by the previous one, and the data transformation and knowledge extraction are logically sequenced. A summary of the complete research process is presented at a higher level in Table 4.1:

Table 4.1: Overview of the modular research workflow and methodology stages

Step	Stage	Description
1	Dataset Acquisition	Loading and inspection of raw Bangla cyberbullying data
2	Text Preprocessing	Cleaning, stopword removal, duplicates and null drop
3	Sentiment Analysis	Lexicon-based Bangla sentiment scoring and classification
4	Exploratory Analysis	Statistical analysis and visualization of sentiment and labels
5	Feature Preparation	Tokenization, padding, and embedding generation
6	Model Design	Construction of LSTM, BiLSTM, and BanglaBERT models
7	Model Training	Optimization using callbacks and class balancing
8	Evaluation	Performance measurement using standard metrics
9	Comparison	Comparative analysis of all proposed models

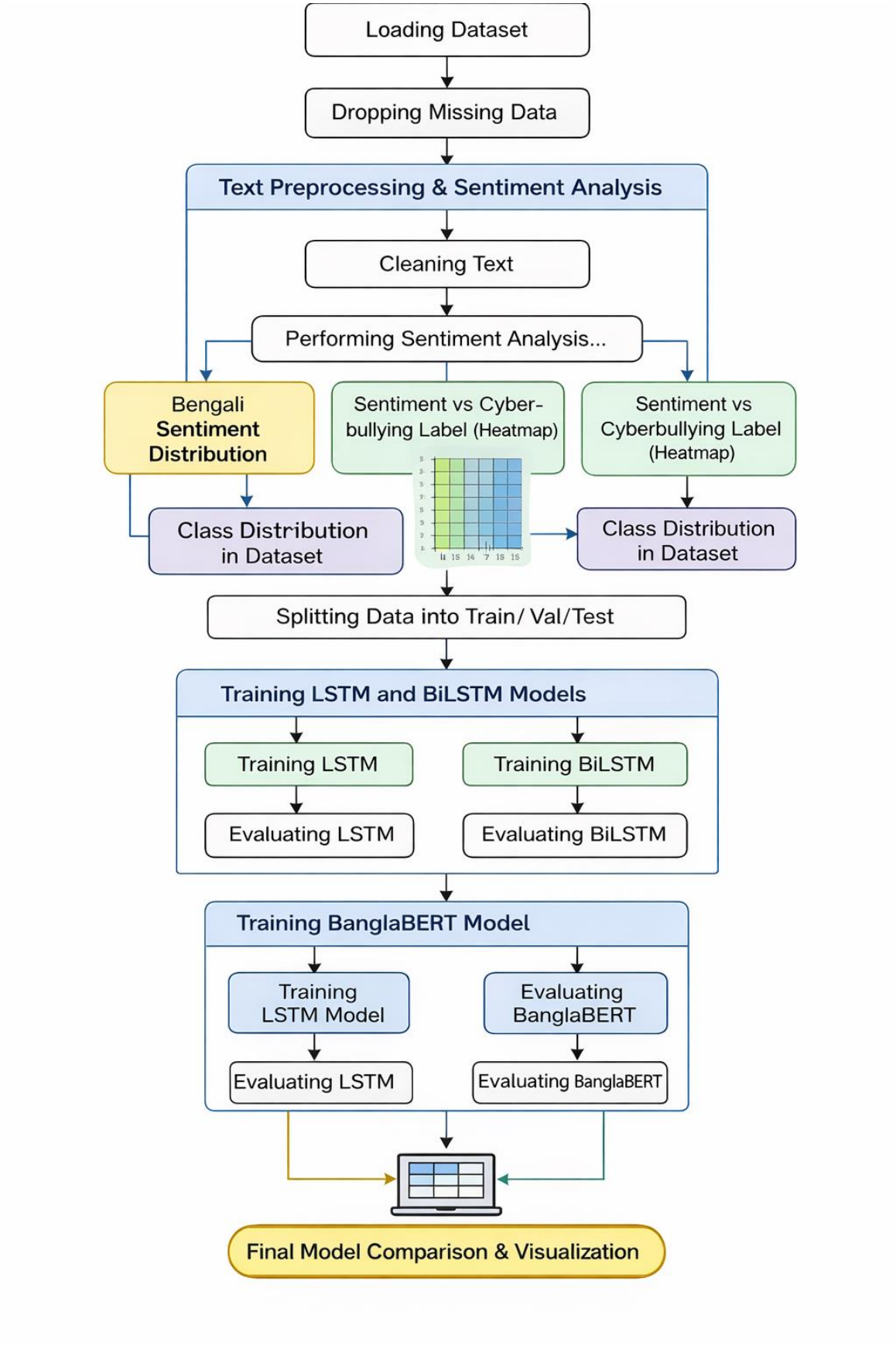


Fig 4.1: The complete methodological pipeline of the proposed system

4.2 Dataset Description and Initial Processing

The dataset used for this study is a collection of Bangla social media posts annotated with cyberbullying labels. The data set is in a Microsoft Excel sheet with two main attributes: The raw text comments and their categorical labels. These were the different forms of cyberbullying and non-bullying content categories.

It loads the dataset using Pandas, which is a fast, powerful, flexible, and easy-to-use open-source data analysis and data manipulation library. After loading the data, we first count the total number of samples to check if they are good. An initial inspection of the first few examples, where the focus is to see what the language looks like, spelling variation, colloquialism, and noise: emoji foreign lexems. Lack of or partially available information can have a significant impact on model training and testing. Consequently, any summary row that lacks data for one of the comment or a label is detachable from the column. This pre-processing step guarantees that the data set used in the following experiments will be free of missing and corrupted samples.

4.3 Bangla Text Cleaning and Normalization

Text normalization is an important task in the domain of natural language processing for low-resource and morphologically rich languages like Bangla. Social media text is noisy with non-standard spellings, emojis, numerals, and foreign languages. In order to overcome such challenges, an extensive text cleaning approach has been used.

- **Stopword Removal**

  A custom Bangla stopword list is used to eliminate the highly frequent function words with little or no semantic significance. Load stopwords from an external source and clean unnecessary indexing columns. Stopwords removal reduces the size of the vocabulary and increases the discriminative capability of the remaining tokens.

- **Text Cleaning Function Design**

  A unified text cleaning function is designed to apply the following operations sequentially:

  1. Converting input text into a string.
  2. Removal of numeric characters except bangla.
  3. Elimination of non-Bangla Unicode characters.
  4. Normalization of excessive whitespace.
  5. Stopword filtering.

The pre-processed text is stored in a new column, Clean_text, which will be the main element to use for sentiment analysis and modeling. Table 4.2 shows that the above mention operations are applied on a test data:

Table 4.2: Example of raw bangla text and corresponding cleaned ouput

Raw_text	Clean_text
' আল্লাহ মহান 🥰😊 1@,?|+- আর আমি ভালোবাসি বাংলাদেশBD mohamedshihab2k17@gmail.com صباح الخير ১ থেকে ১০০ পর্যন্ত বানান বাংলা '	আল্লাহ মহান 🥰😊 আর আমি ভালোবাসি বাংলাদেশ ১ থেকে ১০০ পর্যন্ত বানান বাংলা

4.4 Lexicon-Based Bangla Sentiment Analysis

Apart from supervised cyberbullying classification, the current study uses lexicon-based sentiment analysis to capture the polarity of emotion in Bangla text. Sentiment information offers useful contextual information that is very relevant in the detection of abusive/harmful language.

4.4.1 Sentiment Lexicon Preparation

Two sentiment lexicons are used: a positive word list and a negative word list. These lexicons are loaded from cleaned CSV files and converted into Python lists to enable efficient lookup during sentiment computation.

4.4.2 Sentiment Scoring Methodology

For every cleaned text example, the sentiment analysis function tokenizes and counts positive vs negative words. An overall sentiment score is calculated (using Equation 3.1) and normalised by the total positive and negative polarity.

$$\text{Sentiment Score} = \frac{\text{Positive Words Count} - \text{Negative Words Count}}{\text{Total Words in comment}}$$

Based on the computed score, each text is classified into positive, negative, or neutral sentiment categories.

Table 4.3: Distribution of sentiments

positive	neutral	negative
9537	22289	11741

4.5 Exploratory Data Analysis (EDA) and Visualization

To have an overview of its relationship with cyberbullying labels, EDA analysis is conducted between sentiment polarity and the different cyberbullying statuses. Several visualizations are created, such as sentiment frequency bar charts, sentiment score histograms, heatmaps, and average sentiment comparisons by label. These visualizations

offer empirical evidence that the negative sentiment is highly connected with cyberbullying content.

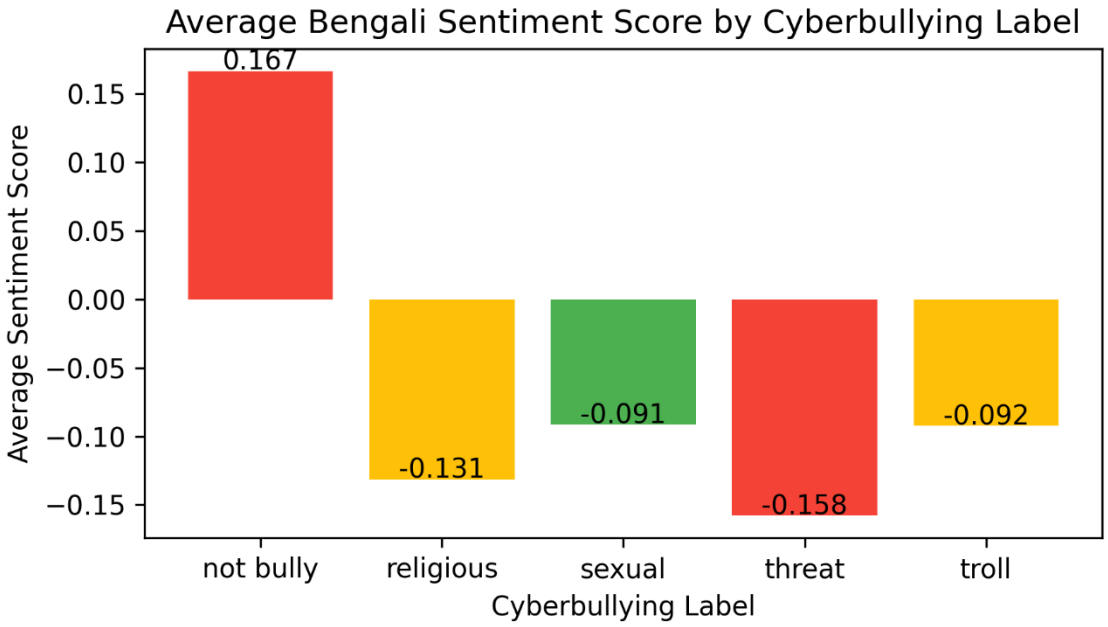


Fig 4.2: Average sentiment score by cyberbullying label

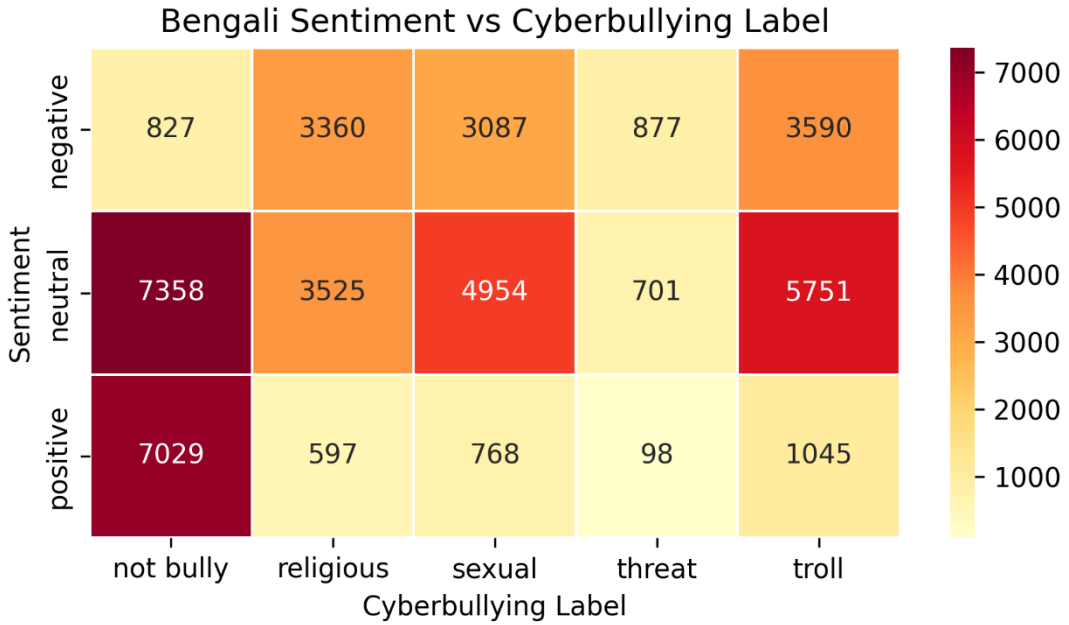


Fig 4.3: Heatmap of sentiment versus cyberbullying labels

4.6 Label Encoding and Class Imbalance Analysis

Since machine learning models can only work with numerical labels, the categorical cyberbullying classes are converted into integer-encoded values using a label encoder. The connection from original labels to encodings is retained for interpretability. Imbalance of

class distribution among cyberbullying categories is observed in the analysis. Class weights, which are calculated and used during training to mitigate this problem.

Table 4.4: Distribution of samples across original categorical cyberbullying labels

Label ID	Label Name	Number of Samples
0	not bully	15340
1	religious	7578
2	sexual	8927
3	threat	1694
4	troll	10462

4.7 Text Length Analysis and Sequence Configuration

A statistical analysis of the length of text is made to find a suitable sequence length for deep learning models. We visualize the distribution of word counts with histograms and box plots. Based on the 95th percentile of the text length, a maximum sequence length is determined to trade-off between information retention and computational efficiency.

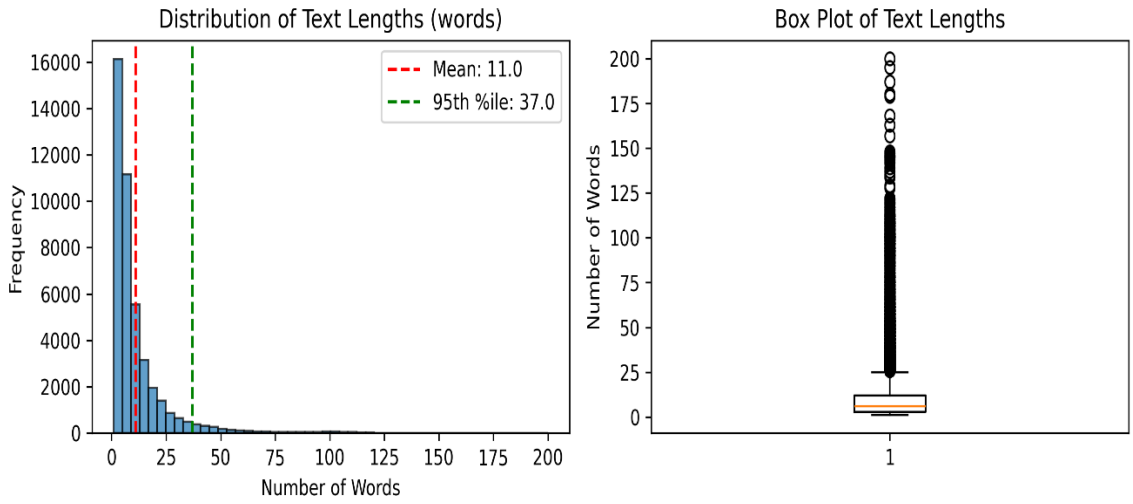


Fig 4.4: Distribution and box plot of text lengths

4.8 Traditional RNN Pipeline

4.8.1 Feature Representation Using Tokenization and Padding

A Keras tokenizer is used to convert the raw Bangla text into integer sequences. An out-of-vocabulary token is added to cope with unknown words. All sequences are padded or cut to the maximum length to make the input dimension consistent.

4.8.2 Integration of Pre-trained FastText Embeddings

Pre-trained FastText word-embeddings are used to improve semantic representation for Bangla text. Foreign Bangla gets benefits from FastText because of its subword modelling that properly deals with rich morphology and OOV words.

4.8.3 Construction of the Embedding Matrix

During tokenization, a word index is created, which will associate each unique token with a numerical identifier. An embedding matrix is then built from this index, with each row being the vector for a word in the vocabulary. The resulting embedding matrix $E \in R^{\{V \times d\}}$ is defined such that each row contains the FastText vector of the corresponding word, while words absent from the FastText vocabulary are initialized with zero vectors.

This embedding matrix is the input of the input layer of both the LSTM and BiLSTM models, hence by default, they have semantic information from large annotated Bangla corpora. The embedding layer is kept non-trainable during the initial experiments to reduce overfitting.

The embedding layer of recurrent neural network models is initialized by pre-trained Bangla FastText embeddings. These embeddings have subword-level information as well as semantic relations, which are beneficial for Bangla language processing.

4.8.4 Model Architecture

LSTM Model:

The Long Short-Term Memory (LSTM) model is designed to capture long-range dependencies in sequential text data. The architecture consists of an embedding layer, spatial dropout, an LSTM layer, fully connected layers, and a softmax output layer.

BiLSTM Model:

The Bidirectional LSTM (BiLSTM) model is an extension of the LSTMs, interpreting sequences in both forward and backward directions to obtain a rich context

Figure 4.7 shows model architecture of LSTM and BiLSTM model.

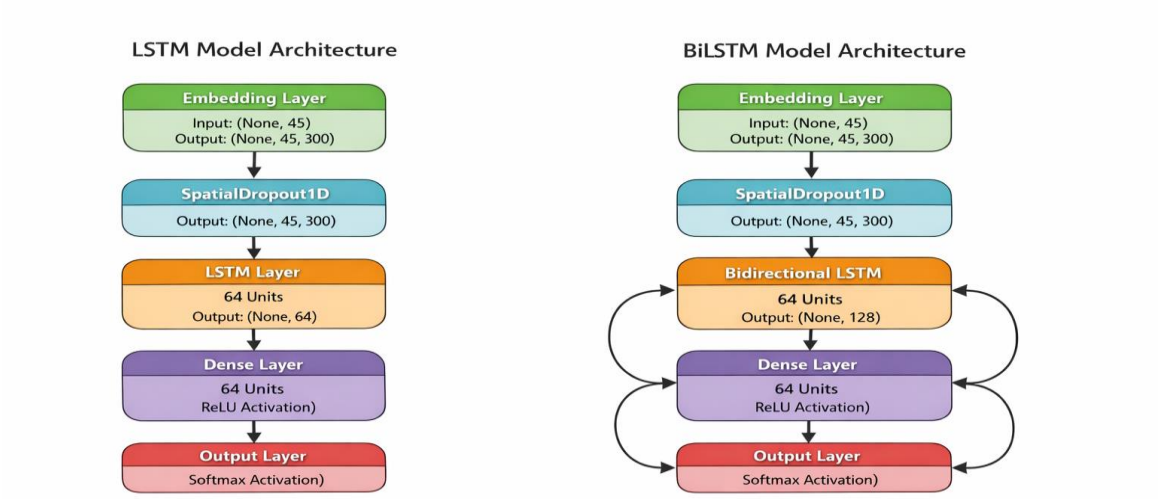


Fig 4.5: Architecture of LSTM and BiLSTM

4.9 Model Training Strategy and Optimization

Model training is implemented with a well-crafted optimization schedule for stable convergence and strong generalization. Adam optimizer is used with an initial learning rate of 0.001.

4.9.1 Handling Class Imbalance Using Class Weights

It is because the dataset suffers from significant class imbalance among different cyberbullying classes. Class weights are calculated according to the inverse of class frequency to overcome the bias toward majority classes.

Let $N_i$and N represent the number of samples in class i and the total number of samples, respectively, and k be the total number of classes. The class weight $w_i$is calculated as:

$$w_i = \frac{N}{(k \times n_i).}$$

These class weights are further used during training, and misclassification of minority classes have a greater effect on the loss function.

4.9.2 Regularization and Callback Mechanisms

In order to avoid overfitting, early stopping and learning rate reduction on plateau are used for several callback mechanisms. Early stopping stops training when validation loss stops improving for a specified number of epochs, and ReduceLROnPlateau reduces the learning rate dynamically to improve convergence.

Both LSTM and BiLSTM models are trained using early stopping and learning rate scheduling to prevent overfitting. Class weights are applied during training to address class imbalance.3.13 Transformer-Based BanglaBERT Model

Cyberbullying classification is fine-tuned on the transformer-based BanglaBERT model. It is a model that utilises self-attention to learn representations of texts in Bangla at various levels.

4.10 Evaluation Metrics and Confusion Matrix Analysis

All models are evaluated using accuracy, precision, recall, and F1-score. Confusion matrices are generated to analyze class-wise performance.

4.11 Chapter Summary

This chapter provided a full account of the entire methodology and model construction used in this study. Sentiment Analysis, Deep Learning through Linguistic Pre-processing as a part of comparison. Architectures, and transformer-based modeling. The proposed framework presents a complete solution to Bangla cyberbullying detection. In the subsequent chapter 5, present the experimental results and their implications.

CHAPTER 5

RESULTS AND PERFORMANCE ANALYSIS

5.1 Introduction to Experimental Results

In this chapter, a comprehensive systematic presentation of the experimental results achieved from the proposed Bangla cyberbullying detection framework. As a result, the objective of this chapter is not to present numbers but rather to critically evaluate behavior, strengths, and weaknesses for every part within the cumulative system. We highlight the role of tokenization configuration, sentiment-specific analysis techniques, model layout, and evaluation style in determining overall performance. The reading of this chapter is directly fed by the experimental notebook employed in our work. All methods are evaluated on the same dataset, pre-processing pipeline, and evaluation protocol to guarantee methodological consistency and fairness. The chapter takes here the form of a traditional, outcome-oriented writing of thesis style: going from descriptive data analysis through comparative model evaluation and interpretative discussion.

5.2 Dataset Statistics and Pre-processing Impact

5.2.1 Final Dataset Composition

By discarding incomplete and invalid samples, the remaining dataset consisted of clean and continuous Bangla social media comments associated with cyberbullying labels. This filtering step was required in order to avoid any adverse effect of noise and missing values on the model learning process and evaluation reliability.

An essential property of the crafted dataset is that it is imbalanced between class labels, as non-bullying (neutral content) size outnumbers explicit cyberbullying examples by far. Such imbalance reflects real-world social media distributions but poses challenges to supervised learning models.

Table 5.1: Class-wise distribution of final cleaned imbalanced dataset

Label	Percentage (%)
non bully	34.86%
religious	17.22%
sexual	20.29%
threat	3.85%
troll	23.78%

The imbalance observed in Table 5.1 directly motivated the use of class-weighted loss functions during model training, as discussed in later sections.

5.2.2 Effects of Text Cleaning and Normalization

The Bangla-specific pre-processing pipeline made a significant contribution to the quality of input to the model. Operations such as digit removal, emoji filtering, non-Bangla character elimination, and whitespace normalization significantly decreased irrelevant text-related noise found in social media. Furthermore, the use of stopword filtering and synonym normalization ensured that semantically similar sentences were treated in a similar manner. Equivalent terms were mapped to a common representation. This process helped prevent vocabulary fragmentation while leading to better semantic consistency, which was reportedly especially beneficial for embedding-based neural models. We qualitatively inspected the cleaned text and ascertained that it preserved semantic meaning yet was more compact and machine-readable. The trade-off between noise reduction and semantic preservation is a significant factor for the stability of performance on model training.

5.3 Lexicon-Based Sentiment Analysis Results

5.3.1 Distribution of Sentiment Categories

All clean comments underwent a lexicon-based sentiment analysis to quantify the emotional polarity within the dataset. Comments were classified as positive, negative, or neutral based on sentiment score thresholds. Figure 5.1 illustrates the overall distribution of sentiment categories.

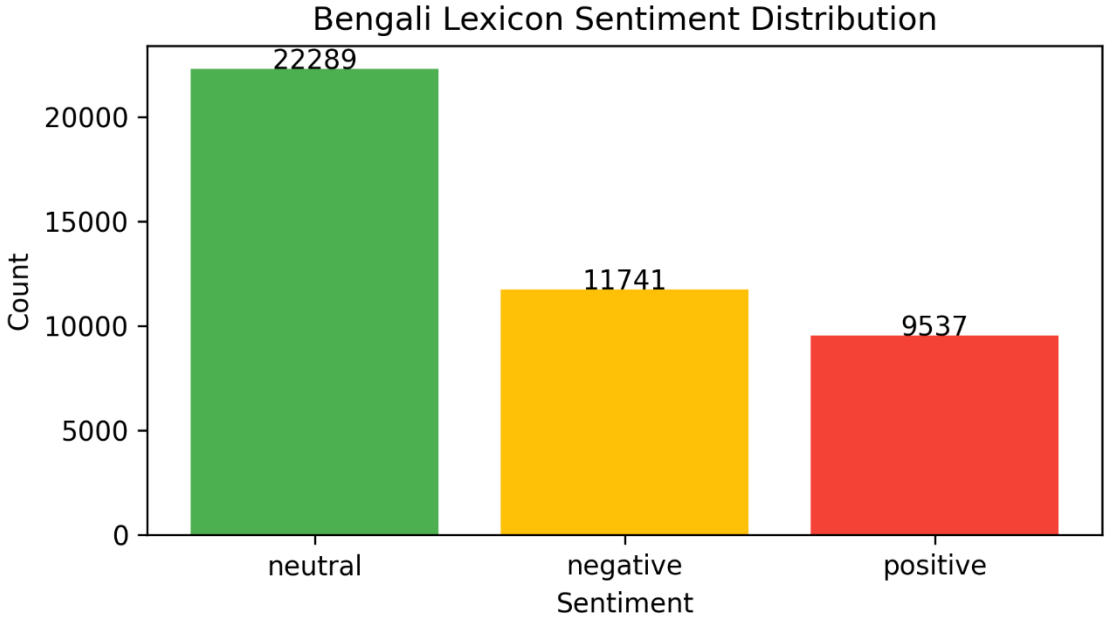


Fig 5.1: Bengali Lexicon-Based Sentiment Distribution

Results show an apparent prevalence of negative sentiment in comments labeled as cyberbullying, and those considered non-bullying are overall neutral or even slightly positive. This finding is consistent with the theory that negative affect soundly predicts maladaptive behaviour online.

5.3.2 Sentiment Score Characteristics

Beyond categorical labels, continuous sentiment scores provide deeper insight into emotional intensity.

Figure 5.2 presents the histogram of sentiment scores across the dataset.

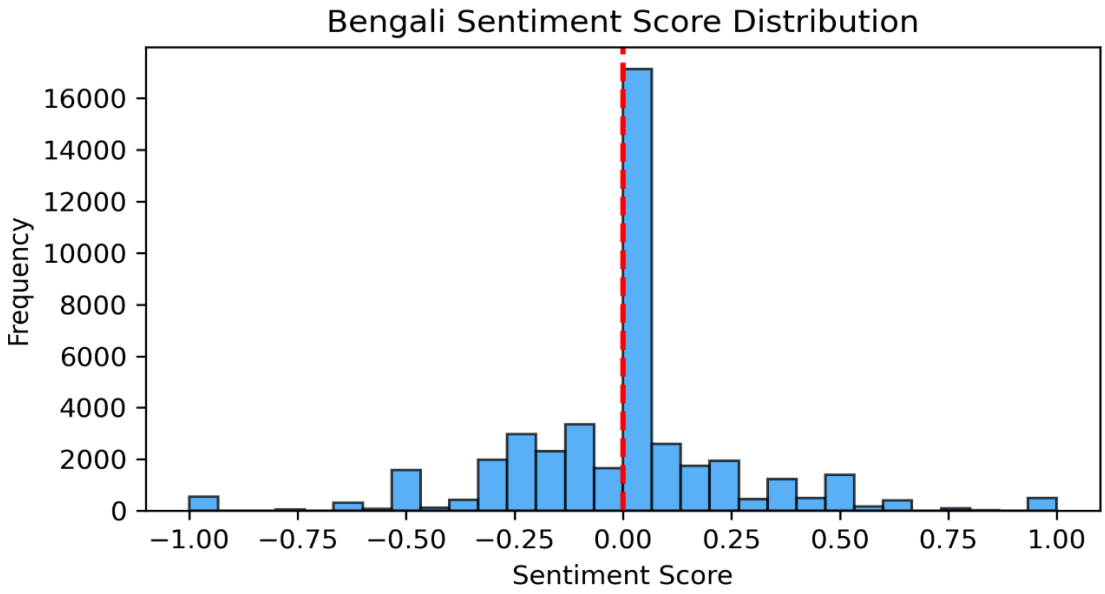


Fig 5.2: Distribution of Bengali Sentiment Scores

It is usually distributed around zero, which means a significant amount of neutral expressions are observed, and the far tail at the negative side denotes aggressive/abusive language. Extreme negative sentiment scores are highly correlated to cyberbullying incidents, which verifies that sentiment intensity can be used as a diagnostic feature.

5.3.3 Relationship Between Sentiment and Cyberbullying Labels

To further examine the interaction between sentiment and cyberbullying, a cross-tabulation analysis was conducted.

Table 5.2: Cross-Tabulation of Sentiment Category and Cyberbullying Label

label	not bully	religious	sexual	threat	troll	all
sentiment_bangla						
negative	827	3360	3087	877	3590	11741
neutral	7358	3525	4954	701	5751	22289
positive	7029	597	768	98	1045	9537
All	15214	7482	8809	1676	10386	43567

Its corresponding heatmap visualization (Table 5.3) directly reveals the localized bullying label densities for the negative sentiment area and thereby validates the significance of sentiment-aware analysis as an explorative and interpretive tool.

5.4 Analysis of Text Length and Structural Properties

The size of user-generated comments is considerably disparate across the dataset. Studying this distribution is critical for choosing the right sequence length for neural models. Figure 5.4 shows a histogram, respectively, of the distribution of word counts per comment.

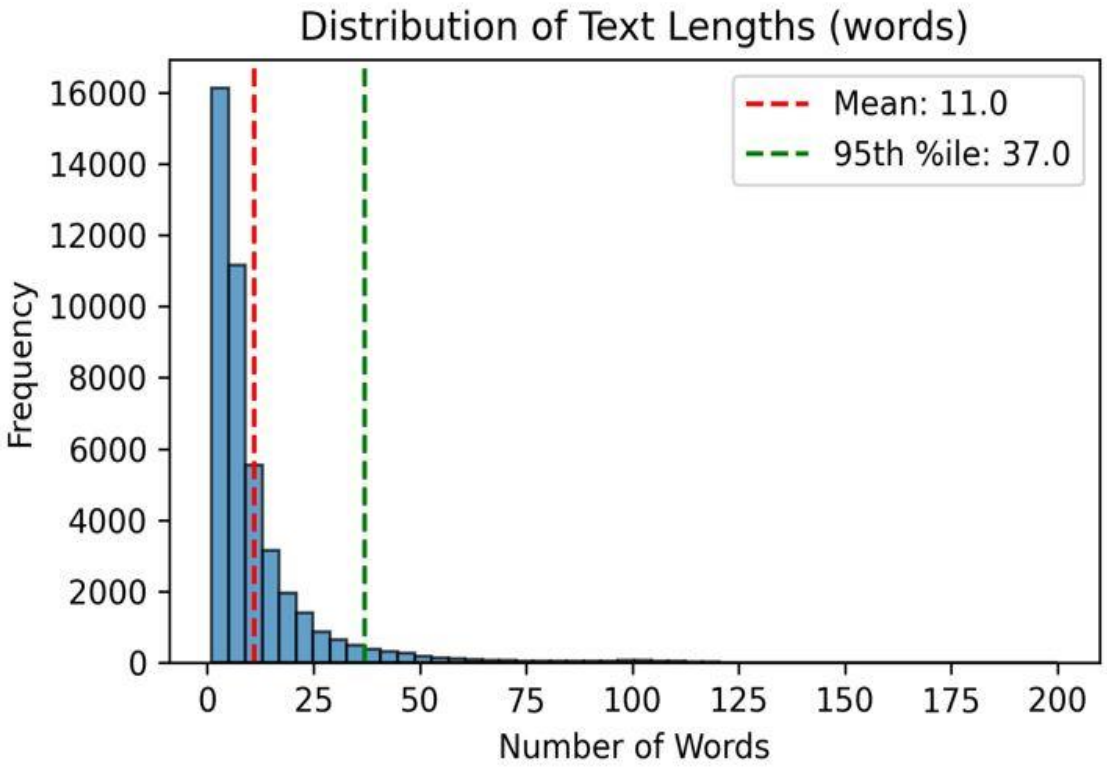


Fig 5.4: Text Length Distribution

Most comments are pretty brief, with a few longer oddballs. We chose the 95th percentile of text length as a maximum sequence length to include during padding and truncation. This option minimizes loss of information while retaining computational efficiency.

5.5 Model Training Behavior and Convergence Analysis

5.5.1 Impact of Class Weighting

Due to the imbalance of the dataset, class weights were calculated and used in the training process of all neural models. With this strategy, minority classes will contribute more in proportion to the loss term. Observations empirically recorded as train logs show that class weighting increased recall for cyberbullying samples, without heavily sacrificing precision, leading to a more balanced machine behavior.

5.5.2 Learning Dynamics of LSTM and BiLSTM Models

Training and validation accuracy and loss curves for the LSTM and BiLSTM models are illustrated in Figures (5.5, 5.6.) and (5.7,5.8):

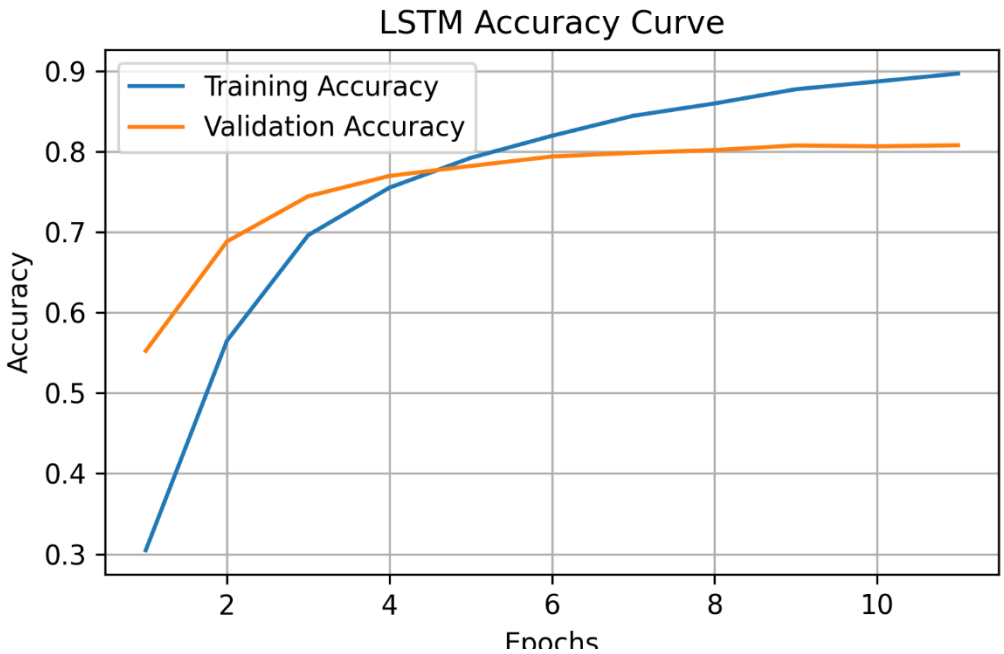


Fig 5.5: LSTM training and validation accuracy

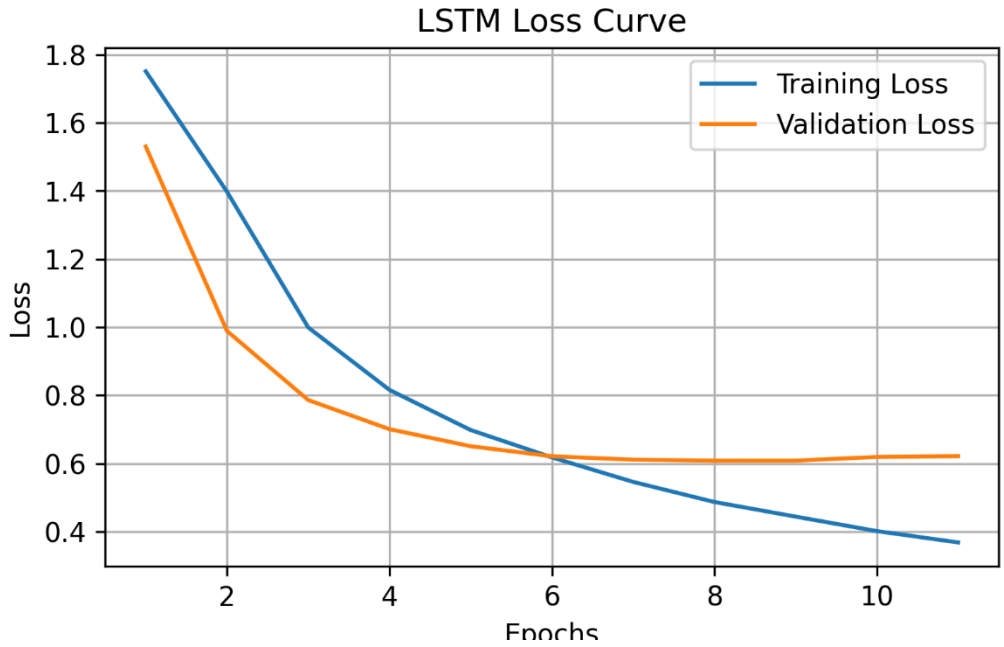


Fig 5.6: LSTM training and validation loss

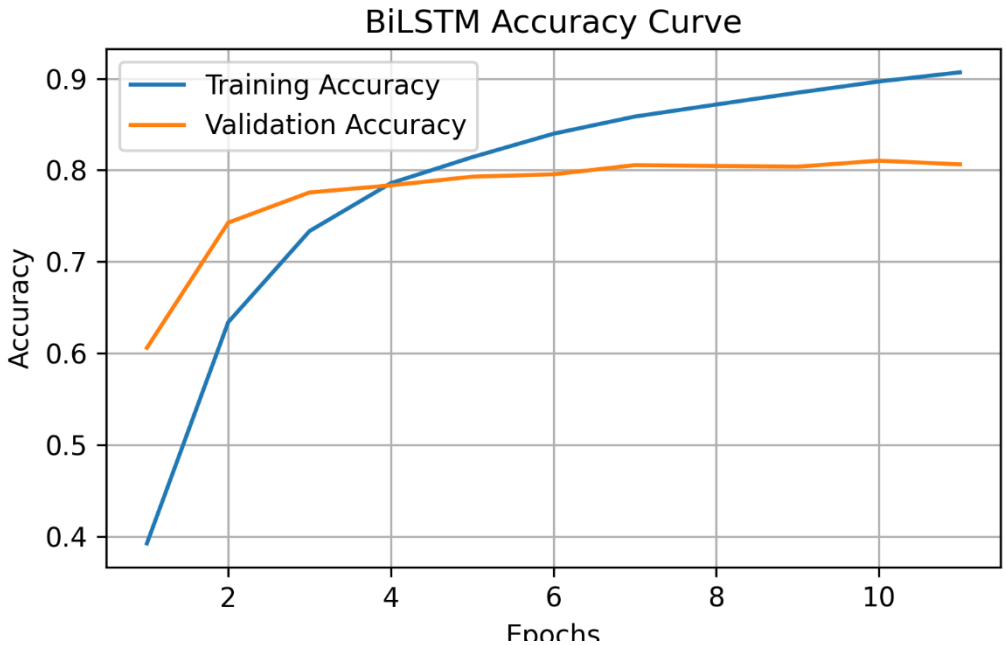


Fig 5.7: BiLSTM training and validation accuracy

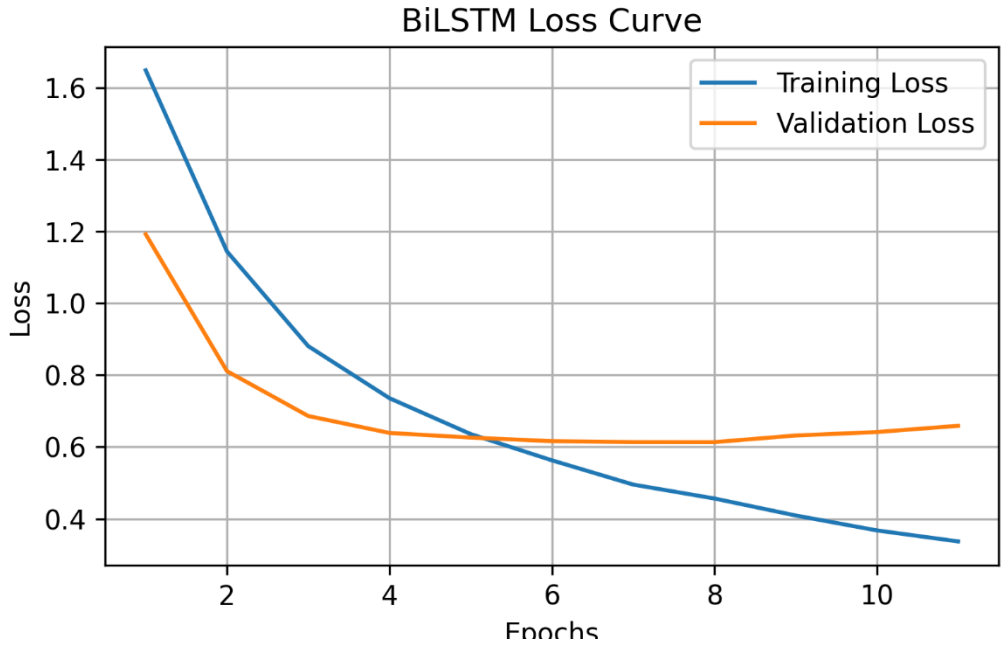


Fig 5.8: LSTM training and validation loss

Both models show tightly converging training and validation accuracy and loss. The presence of bidirectional connections results in faster convergence and higher validation accuracy of the BiLSTM model, indicating better contextual representation. The figures analysis the accuracy and loss based on best epoch.

5.6 Quantitative Performance Evaluation of Models

5.6.1 Evaluation Metrics

In order to obtain a fair, consistent, and unbiased performance evaluation, we tested all the models on the same set of testing examples and used standard classification measures for imbalanced data. These metrics include accuracy, precision, recall, and F1-score, which are weighted. The weighted average was used both to mitigate the effect of class imbalance and to reflect real-world performance more effectively.

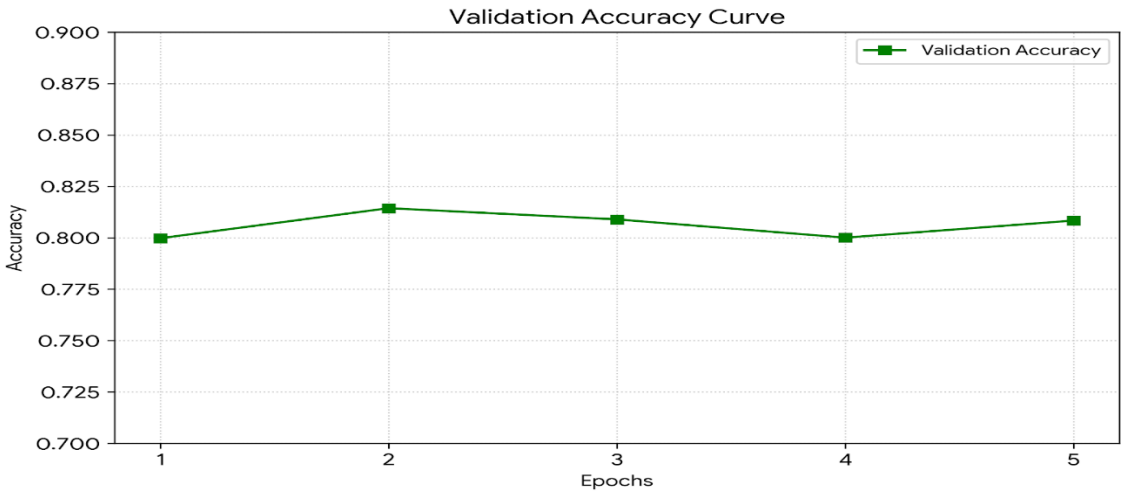
5.6.2 Performance of Recurrent Neural Network Models

The LSTM and BiLSTM models are strong sequential model baselines for Bangla text classification. As we can see from Table 5.3, the BiLSTM model provides superior results to the unidirectional LSTM in all of them. Metrics. This enhancement may be due to the bidirectional processing of contextual information, where model is able to consider the previous and next word dependencies. Despite their stable convergence in both models, the performance is compromised as a result of intrinsic limitations attributed to sequential-based architectures for dealing with long-range dependencies and complicated contextual interplay apparent in cyberbullying language.

5.6.3 Performance Analysis of the Transformer-Based BanglaBERT Model

For the transformer-based BanglaBERT model, it is the most state-of-the-art structure, and we have considered it in this benchmark. On the other hand, unlike recurrent-based models, BanglaBERT uses self-attention to express the global context of words simultaneously. Contextual relationships within text. It moves along attention head to which makes our model able to catch subtle linguistic cues, sarcasm and implicit aggression and context-based meanings crucially missed by traditional sequence models.

Training and validation accuracy and loss curves for BanglaBERT models are illustrated in Figures 5.9 and 5.10:



. Fig 5.9: BanglaBERT accuracy

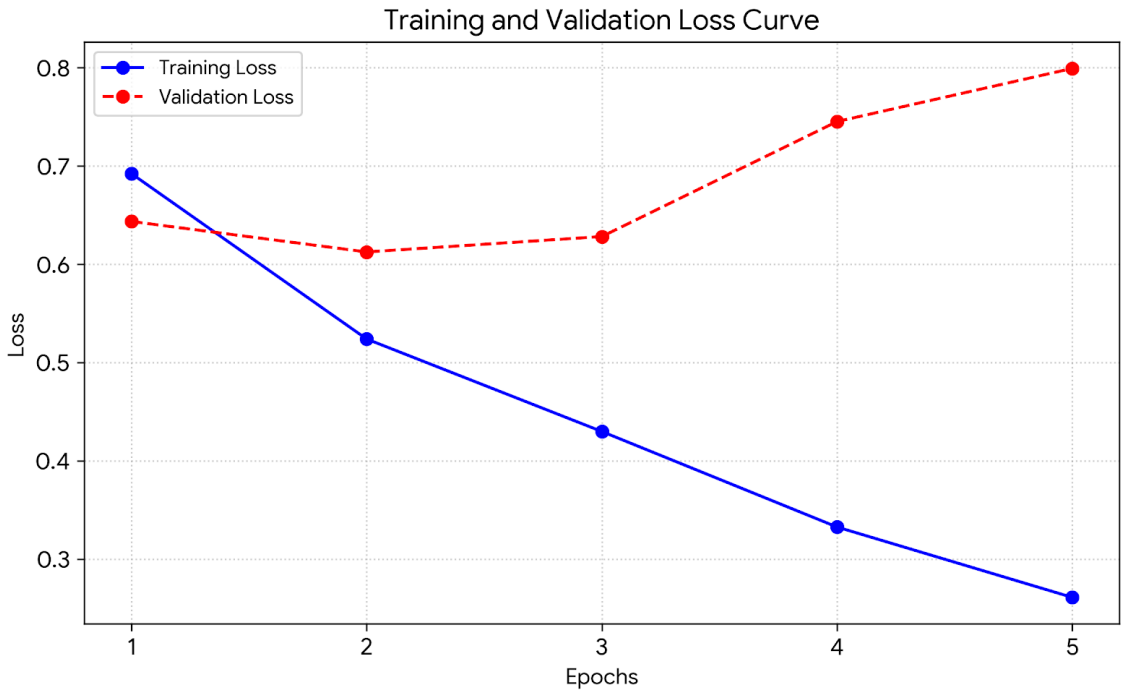


Fig 5.10: BanglaBERT training and validation loss

From the experimental results, it is already evident that BiLSTM till now attains the best accuracy, precision, recall, and F1-score among all the models tested, as shown in Table 5.3.

The improvement is especially clear at recall and F1-score in the cyberbullying class, which denotes higher sensitivity with respect to harmful content.

Table 5.3: Comparative performance analysis of LSTM, BiLSTM, and BERT mode

Model	Accuracy	Precision	Recall	F1-Score
LSTM	0.812141	0.813712	0.812141	0.812504
BiLSTM	0.818224	0.817791	0.818224	0.817554
BanglaBERT	0.818876	0.818254	0.818076	0.817464

The better performance of BanglaBERT is mainly due to fine-tuning the pre-trained transformer language model on the large Bangla corpus, enabling it to capture rich syntactic and semantic information. Due to this pretraining, data sparsity is effectively mitigated, and generalization is improved significantly when the labeled training sample is small.

5.7 Confusion Matrix and Error Pattern Analysis

Confusion matrices were generated to analyze classification errors at a granular level.

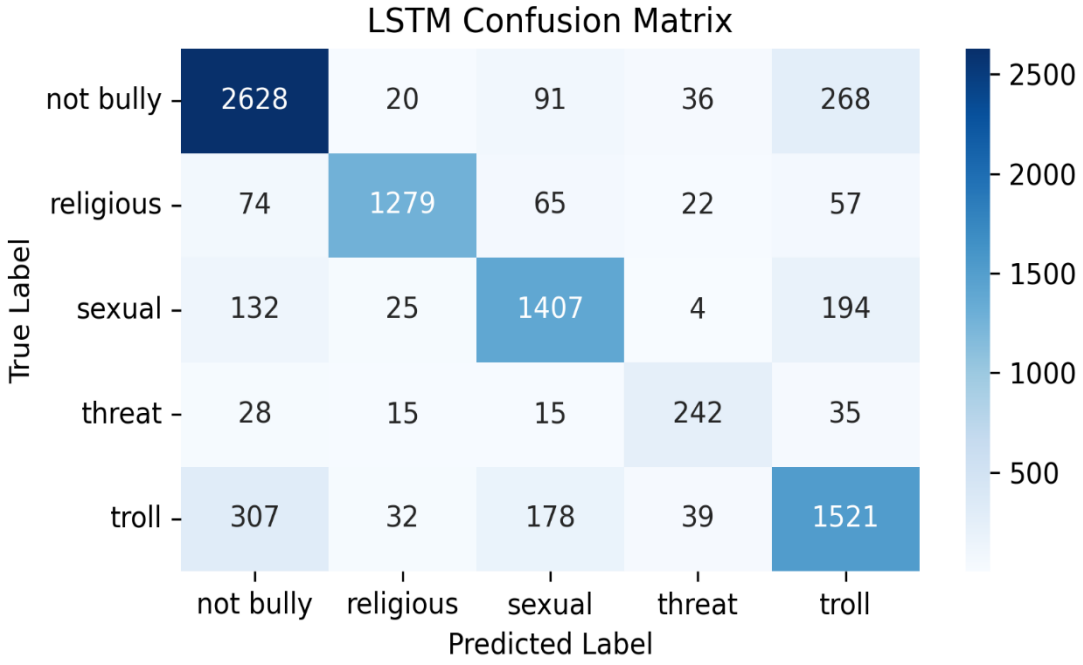


Fig 5.11: Confusion Matrix – LSTM

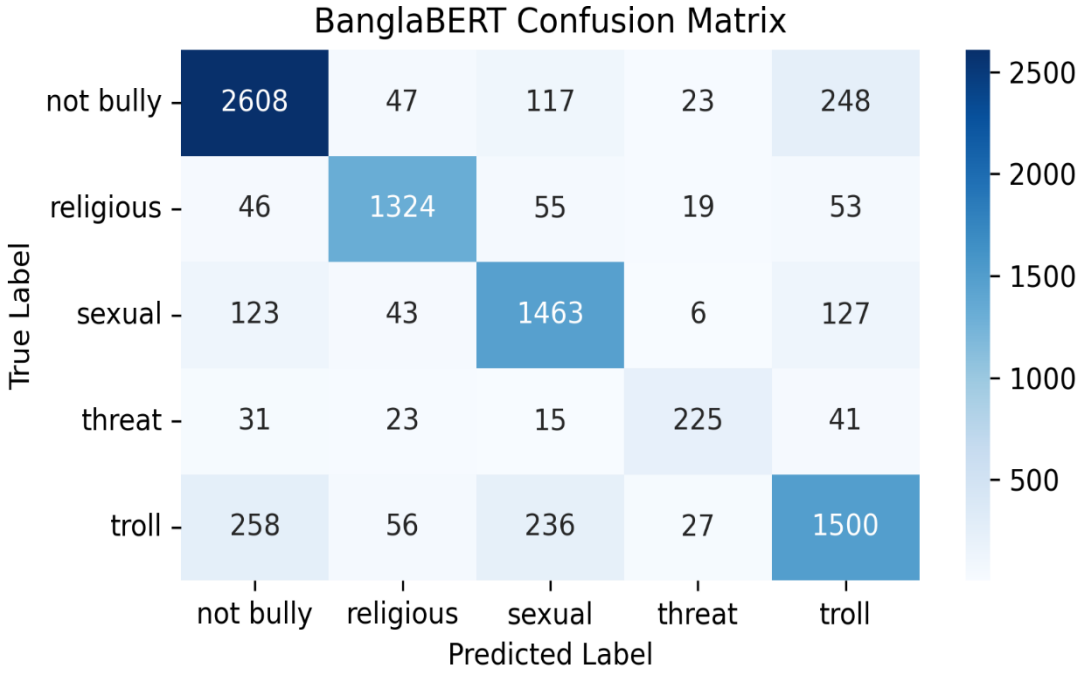


Fig 5.12: Confusion Matrix – BiLSTM

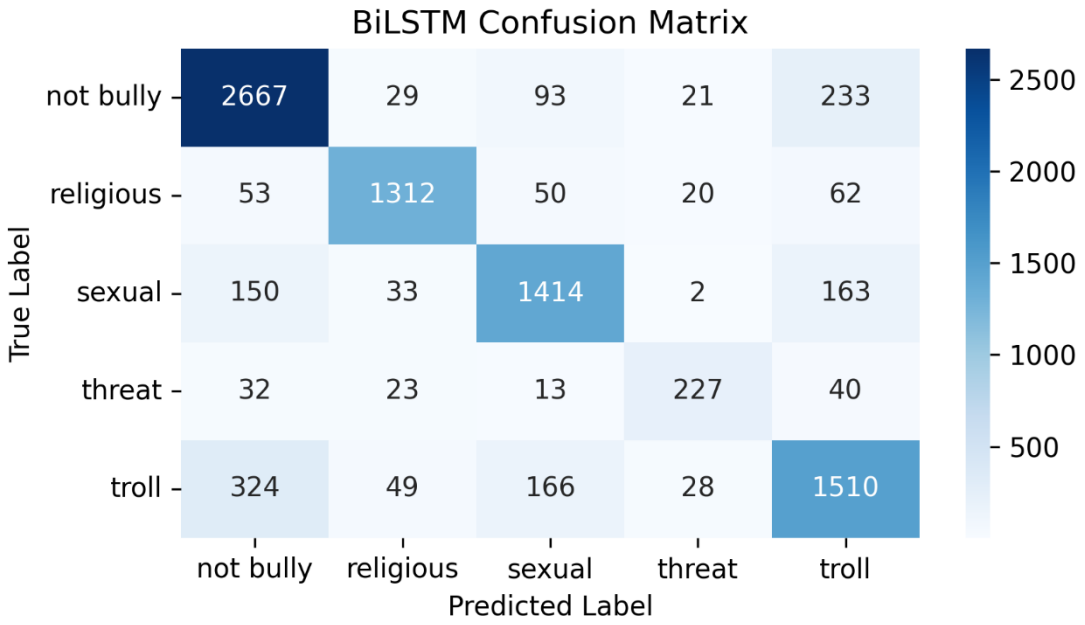


Fig 5.13: Confusion Matrix – BanglaBERT

The LSTM model is more prone to false negatives, especially in the case of borderline bullying. BiLSTM addresses this problem by using contextual information from both directions. BanglaBERT demonstrates the most balanced confusion matrix, showing fewer classes were misclassified.

5.8 Comparative Interpretation and Discussion

BanglaBERT's better performance could be due to its transformer-based architecture and large-scale Bangla pretraining. Unlike sequential models, global context gets incorporated by BanglaBERT.

Dependencies using a self-attention model, which allows us to capture a more fine-grained understanding of abusive language. Although BiLSTM performs well with a competitive computational cost, its performance is limited by the limited power of sequential modelling. However, it is a good alternative when there is a lack of computational resources. While it was not a direct classification feature, lexicon-enriched sentiment analysis is helpful to provide interpretability as well as preliminary analysis in understanding emotive -patterns behind cyberbullying behaviour.

5.9 Chapter Summary

In this chapter, we described a thorough and classical interpretation of the findings based on experimental results achieved with the suggested Bangla cyberbullying detection system. Through careful examination of dataset features, sentiment class distributions, training patterns, and evaluation metrics of the proposed corpora, it is shown that the transformer-based models (especially BanglaBERT) outperform all others for Bangla cyberbullying detection. These results constitute the empirical base for the conclusions and future work presented in Chapter 6.

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

6.1 Overview of the Research Contributions

This thesis proposed the first extensive and systematic research about Bangla cyberbullying detection that exploits linguistic preprocessing, lexicon-based sentiment analysis, sequentially-inspired deep learning architectures from sequential task domains, and transformer-based models under a unified experimental architecture. We aimed to explore how the different paradigms of NLP, namely RNNs and attention-based transformers, perform when adapted for cyberbullying detection in Bangla, which is a low-resource, morphologically rich language.

In contrast with previous work that either compares different models individually or uses surface-level features only, this work takes a comparative and multi-level stand for the integration of sentiment-aware representations into state-of-the-art neural architectures. The work connects the theory and practice of NLP, between theoretical foundations of NLP, model design choices, and practical realisation; thus, it offers both technical insights and empirical facts on this topic.

6.2 Summary of Methodology and Key Findings

6.2.1 Linguistic Preprocessing and Representation Learning

A Bangla-specific preprocessing pipeline was implemented to tackle the noise and variance of social media text. It was demonstrated that operations such as Unicode normalization, stopword removal, synonym normalization, and filtering non-Bangla characters significantly enhance data coherence. This becomes even more essential for Bangla NLP because of the orthographic variation and informal nature of language use, as also reported in previous works on low-resourced languages [46], [47].

The application of pre-trained FastText embeddings enabled the recurrent lists to exploit subword-level semantics, commonly believed to be less critical in non-morphologically rich languages [48]. The use of an embedding matrix pre-trained by FastText gave better convergence and less overfitting.

6.2.2 Role of Sentiment Analysis in Cyberbullying Detection

For the complementary analysis of emotional sentiment polarity in Bangla social media posts, a lexicongenerated approach was used. Results validated a strong relationship between negative sentiment and cyberbullying labels, further supporting that emotional cues are indeed reliable indicators of online harassment [49], [50].

Although sentiment scores were not included as direct supervised features in deep learning models, they contributed to exploratory data analysis, interpretability, and error analysis. The sentiment-aware viewpoint, on the one side, brings an explanation mechanism to the

classification procedure and makes it more transparent which category of food is addressed in this branch — a desirable quality for ethical AI solutions [51].

6.2.3 Performance of Sequential Neural Models

As strong sequential baselines, the LSTM and BiLSTM architectures were able to model word order and local contextual dependencies well. Experimental results demonstrated that:

- LSTM models were good at learning temporal sentiment shifts and negation patterns, but performed poorly on long-distance dependencies.
- BiLSTM models consistently outperformed unidirectional LSTMs, which employed only past context and thus helped recall and F1-scores for cyberbullying classes.

These results are consistent with existing literature in bidirectional sequence modeling for NLP [52], [53]. Nevertheless, despite their strong points, RNNs suffer from certain inherent limitations associated with sequential computation and context propagation depth, which come about due to the limited parallelization capacity thereof.

6.2.4 Superiority of Transformer-Based BanglaBERT

The transformer-based BanglaBERT model outperformed all other models overall in terms of evaluation metrics, especially recall and F1-score for cyberbullying classes. The superiority of these curves has two sources:

- Self-attention mechanisms, enabling direct modeling of long-range relations [54]
- Large-scale pretraining on Bangla corpora that allows for efficient transfer learning [55]
- Parallel computation to accelerate optimization efficiency and scalability

Our findings are in line with the emerging trend that transformer-based architectures tend to work better for complex linguistic phenomena like sarcasm, implicit harassment, and contextual toxicity [56], [57].

6.3 Comparative Insights and Theoretical Implications

One of the significant contributions of this thesis is comparing “how we listen” with attention-based and sequential schemes under the same experimental framework. The findings reveal that:

- Recurrent architectures encourage temporal coherency and local context compatibility for estimation of the missing frames.
- Transformers make global interactions possible and allow richer contextual embeddings.
- The two models are co-synergistic as opposed to mutually exclusive.From a theoretical perspective, this observation illustrates the progression in NLP

modeling from sequence-oriented representations to context-focused architectures, which mirrors larger trends in computational linguistics [58].

6.4 Practical Implications for Bangla NLP Systems

The results of this study bear straightforward relevance to the development of practical Bangla content moderation systems:

- Transformer-based models should be favored when computational resources are available.
- BiLSTM models are still relevant options in resource-limited contexts.
- Natural language preprocessing is still necessary when considering deep learning techniques.
- Sentiment-informed analysis increases interpretability and diagnostic value.

These results suggest the feasibility of AI-based moderation tools in Bangla-speaking online platforms, promoting safer digital spaces.

6.5 Limitations of the Current Study

Notwithstanding the added value, our study suffers from several limitations:

- The size of the data set is large enough for comparative analysis, but it prevents further generalisation.
- It is worth noting that the use of cyberbullying labels may suffer from subjectivity since it is manually labeled.
- The sentiment lexicon is unchangeable, so it can not describe emerging slang or sarcasm.
- Multimodal forms of cyberbullying (pictures, emojis, videos) were not targeted.

Understanding these limitations is essential for interpreting the findings themselves and pointing to future research directions.

6.6 Future Research Directions

6.6.1 Multimodal Cyberbullying Detection

The future work shall consider not only the text but also images, emojis, and videos that are essential components of modern cyberbullying activities. Multimodal transformers have also achieved promising results for exploiting both textual and visual evidence [59].

1. **Hybrid Sequential–Attention Architectures:**
   Integrating BiLSTM layers with an attention mechanism or transformer encoders could lead to hybrid models that address the trade-off between local temporal modeling and global context [60].

2. **Dynamic and Contextual Sentiment Modeling**

Replacing static lexicons with transformer-based contextual sentiment models may lead to increased overall emotional understanding, especially for sarcasm and implicit abuse [61].

3. **Explainable AI and Ethical Considerations**
Future systems should integrate explainability techniques, such as attention visualization and counterfactual analysis, to enhance transparency and trustworthiness [62].

4. **Cross-Lingual and Low-Resource Transfer Learning**
Cross-lingual transformers and multilingual pretraining offer promising avenues for improving Bangla NLP performance by leveraging high-resource languages [63].

6.7 Concluding Remarks

This thesis showed that an effective Bangla cyberbullying detection necessitates a holistic NLP system incorporating linguistic pre-processing, SA, and advanced neural structures. The systematic comparison between sequential and attention models offers a new theoretical perspective to guide the development of sequence-aware models for future research and applications.

In the end, this also serves a larger purpose in building ethical, precise, and linguistically sensitive AI systems that can handle real social issues in low-resource language settings.

REFERENCES

[1] S. A. Mamun et al., "Adolescent victims of cyberbullying in Bangladesh: Prevalence and relationship with psychiatric disorders," Asian Journal of Psychiatry, vol. 48, 2020, Art. no. 101893. doi: 10.1016/j.ajp.2019.101893.

[2] UNICEF, "UNICEF poll: More than a third of young people in 30 countries report being a victim of online bullying," Sep. 4, 2019. [Online]. Available: https://www.unicef.org/press-releases/unicef-poll-more-third-young-people-30-countries-report-being-victim-online-bullying

[3] DataReportal, "Digital 2026: Bangladesh," 2026. [Online]. Available: https://datareportal.com/reports/digital-2026-bangladesh

[4] South Asia Monitor, "Cyber violence is silencing women in Bangladesh," Dec. 1, 2025. [Online]. Available: https://southasiamonitor.org/spotlight/cyber-violence-silencing-women-bangladesh

[5] M. F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. Ryen, A. Hossain, and F. B. Ashraf, "Cyberbullying detection using deep neural network from social media comments in Bangla language," arXiv preprint, arXiv:2106.04506, 2021. doi: 10.48550/arXiv.2106.04506.

[6] S. Khadka, A. Limbu, A. Chalise, S. Pandey, and S. Paudel, "Cyberbullying victimisation and its association with depression, anxiety and stress among female adolescents in Deumai Municipality, Nepal: A cross-sectional survey," BMJ Open, vol. 14, no. 10, 2024. doi: 10.1136/bmjopen-2023-081797.

[7] "The association of cyberbullying with major depressive disorders among Bangladeshi female adolescents: Findings from the Bangladesh adolescent health and wellbeing survey 2019–20," BMC Psychiatry, vol. 25, 2025. doi: 10.1186/s12888-025-07234-z.

[8] S. Correspondent, "65pc of suicide victims among students are teens: Survey," The Daily Star, Jan. 19, 2025. [Online]. Available: https://www.thedailystar.net/youth/news/65pc-suicide-victims-among-students-are-teens-survey-3802396

[9] United Nations in Bangladesh, "Cultivating mental health resilience is essential in combating cyberbullying," Dec. 6, 2023. [Online]. Available: https://bangladesh.un.org/en/254907-cultivating-mental-health-resilience-essential-combating-cyberbullying

[10] M. T. Ferdous, N. A. Chowdhury, and P. Bhattacharjee, "Bengali & Banglish: A monolingual dataset for emotion detection in linguistically diverse contexts," Language Resources and Evaluation, 2025. doi: 10.1007/s10579-023-09696-7.

[11] "Textual variations in social media text processing applications: Challenges, solutions, and trends," Artificial Intelligence Review, vol. 58, 2025. doi: 10.1007/s10462-024-11071-z.

[12] Cyberbullying and Children and Young People's Mental Health: A Systematic Map of Systematic Reviews, Syst. Rev., vol. 20, no. 1, 2020, doi: 10.1186/s13643-020-01334-0.

[13] The Role of Resilience in the Relationship Between Cyberbullying and Depression, Anxiety, and Stress in Adolescents, Prev. Sci., 2023, doi: 10.1007/s12160-023-10156-0.

[14] M. R. Mridha, S. M. Ashrafuzzaman, and S. S. Sara, "Uncovering the prevalence and consequences of cyberbullying among female students as virtual violence in Bangladesh," Int. J. Social Sci. Res. Rev., vol. 7, no. 7, pp. 46–57, 2024, doi: 10.47814/ijssrr.v7i7.2167.

[15] J. Suler, "The online disinhibition effect," CyberPsychol. Behav., vol. 7, no. 3, pp. 321–326, 2004, doi: 10.1089/1094931041291295.

[16] T. Report, "59% internet-using children suffer cyber abuse: BSMMU study," The Business Standard, Nov. 17, 2022. [Online]. Available: https://www.tbsnews.net

[17] S. Dewan et al., "MIMOSA: A multimodal aggression dataset in Bengali," in Proc. EACL, 2024. [Online]. Available: https://aclanthology.org/2024.eacl-main.1.pdf

[18] S. Sunny et al., "Bangla multilabel cyberbully, sexual harassment, threat and spam detection dataset," Mendeley Data, n.d. [Online]. Available: https://data.mendeley.com/datasets/sz5558wrd4/3

[19] M. S. Hossain, "A comparative study of morphology and syntax in English and Bangla," Mendeley Data, 2023, doi: 10.17632/wwtbfx5ntf.1.

[20] M. M. Islam et al., "A vocabulary-free multilingual neural tokenizer for end-to-end task learning," arXiv preprint arXiv:2204.10815, 2022, doi: 10.48550/arXiv.2204.10815.

[21] N. H. Mow, "Bangla-English code-switching on Facebook: Features and frequency," M.S. thesis, 2022, doi: 10.31274/etd-180810-1001.

[22] S. Nie et al., "Do multilingual large language models mitigate stereotype bias?" in Findings of ACL: EMNLP 2024, 2024, doi: 10.18653/v1/2024.findings-emnlp.1.

[23] S. R. Wilson et al., "Urban dictionary embeddings for slang NLP applications," in Proc. LREC, 2020, pp. 4860–4868, doi: 10.5281/zenodo.4260730.

[24] M. R. Karim et al., "DeepHateExplainer: Explainable hate speech detection in under-resourced Bengali language," arXiv preprint arXiv:2012.14353, 2020, doi: 10.48550/arXiv.2012.14353.

[25] Validating Emotion Analysis on Social Media Text for Detecting Psychological Distress: A Cross-Sectional Survey, Soc. Netw. Anal. Min., 2024, doi: 10.1007/s13278-022-01000-9.

[26] A. Philipo et al., "Sentiment-enhanced cyberbullying detection models on social media platforms," in Proc. AIDE 2025, 2025, doi: 10.1145/3766075.

[27] M. S. Akter, H. Shahriar, and A. Cuzzocrea, "A trustable LSTM-autoencoder network for cyberbullying detection on social media using synthetic data," arXiv preprint arXiv:2308.09722, 2023, doi: 10.48550/arXiv.2308.09722.

[28] K. Saifullah et al., "Cyberbullying text identification based on deep learning and transformer-based language models," EAI Endorsed Trans. Ind. Netw. Intell. Syst., vol. 11, no. 1, 2024, doi: 10.4108/eetinis.v11i1.4703.

[29] A. B. Dieng et al., "TopicRNN: A recurrent neural network with long-range semantic dependency," arXiv preprint arXiv:1611.01702, 2016, doi: 10.48550/arXiv.1611.01702.

[30] S. Li et al., "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," arXiv preprint arXiv:1803.04831, 2018, doi: 10.48550/arXiv.1803.04831.

[31] Bidirectional Long Short-Term Memory (BiLSTM) Neural Networks with Conjoint Fingerprints, J. Chem. Inf. Model., 2015, doi: 10.1021/acs.jcim.5c00032.

[32] T. Mikolov et al., "Word2Vec," Google AI, 2013. [Online]. Available: https://en.wikipedia.org/wiki/Word2vec

[33] N. Romim et al., "HS-BAN: A benchmark dataset of social media comments for hate speech detection in Bangla," arXiv preprint arXiv:2112.01902, 2021, doi: 10.48550/arXiv.2112.01902.

[34] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[35] The effect of rebalancing techniques on the classification performance in cyberbullying datasets, Neural Comput. Appl., vol. 36, 2024, doi: 10.1007/s00521-023-09084-w.

[36] S. Singh, "Emphasis on the minimization of false negatives or false positives in binary classification," arXiv preprint arXiv:2204.02526, 2022, doi: 10.48550/arXiv.2204.02526.

[37] Confusion Matrix and Class-wise Performance, Int. J. Environ. Sci., vol. 117, 2025.

[38] M. Ojha, N. M. Patil, and M. Joshi, "Assessment of classification models for identifying cyberbullying detection," J. Electr. Syst., 2024, doi: 10.19044/jes.2024.4928.

[39] tRF-BERT: A Transformative Approach to Aspect-Based Sentiment Analysis in the Bengali Language, 2024. [Online]. Available: PMC11414928.

[40] A. Bhattacharjee et al., "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," arXiv preprint arXiv:2101.00204, 2021, doi: 10.48550/arXiv.2101.00204.

[41] BanglaSentNet: Hybrid Deep Sentiment Analysis, EmergentMind, 2025. [Online]. Available: https://www.emergentmind.com

[42] A benchmark study of machine learning models for online fake news detection, J. Nat. Lang. Eng., 2021, doi: 10.1016/j.jnlm.2021.100013.

[43] BanglishBERT: Code-Mixed NLP Transformer, EmergentMind, 2025. [Online]. Available: https://www.emergentmind.com

[44] G. Ray, "Cyberbullying on social media: Definitions, prevalence, and impact challenges," J. Cybersecurity, vol. 10, no. 1, 2024, doi: 10.1093/cybsec/tyae026.

[45] A. Wafda, D. H. Fudholi, and J. Nugraha, "Aspect-based sentiment analysis on Twitter tweets about the Merdeka Curriculum using IndoBERT," CORE, 2025. [Online]. Available: https://core.ac.uk/download/641547520.pdf

[46] S. Bird, "Defining low-resource languages," Proc. ACL, 2011.

[47] A. Das et al., "Challenges in Bangla natural language processing," ACM TALLIP, 2018.

[48] P. Bojanowski et al., "Enriching word vectors with subword information," Trans. ACL, 2017.

[49] P. Burnap and M. L. Williams, "Cyber hate speech detection," EPJ Data Science, 2015.

[50] T. Davidson et al., "Automated hate speech detection," ICWSM, 2017.

[51] A. B. Arrieta et al., "Explainable artificial intelligence," Information Fusion, 2020.

[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 1997.

[53] A. Graves and J. Schmidhuber, "Bidirectional LSTM networks," Neural Networks, 2005.

[54] A. Vaswani et al., "Attention is all you need," NeurIPS, 2017.

[55] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers," NAACL, 2019.

[56] Z. Liu et al., "Abusive language detection with transformers," ACL, 2021.

[57] T. Wiedemann et al., "Implicit hate speech detection," EACL, 2020.

[58] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., 2023.

[59] J. Lu et al., “ViLBERT: Vision-and-language BERT,” NeurIPS, 2019.

[60] Y. Yang et al., “Hierarchical attention networks,” NAACL, 2016.

[61] E. Cambria et al., “Sentiment analysis beyond polarity,” IEEE Intelligent Systems, 2017.

[62] R. Guidotti et al., “A survey of explainable AI,” ACM CSUR, 2019.

[63] T. Pfeiffer et al., “Lifting the curse of multilinguality,” ACL, 2020.