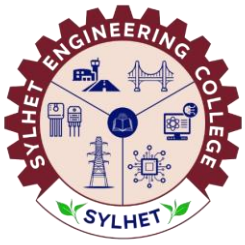
A Thesis Submitted to the Sylhet Engineering College for the Degree of
Bachelor of Science in **Computer Science and Engineering**

Hybrid Deep Learning Architectures for Bangla Toxic Language Detection: A Comparative Analysis of BERT-CNN-LSTM and BERT-GRU-Attention Models

By
Md. Jahirul Islam
Afsar Tanvir

Supervised By
Md. Abu Naser Mojumder
Assistant Professor
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet



22th July, 2025
Sylhet Engineering College, Sylhet
Affiliated with
Shahjalal University of Science & Technology (SUST)

Recommendation Letter from Thesis Supervisor

The thesis titled **"Hybrid Deep Learning Architectures for Bangla Toxic Language Detection: A Comparative Analysis of BERT-CNN-LSTM and BERT-GRU-Attention Models"** submitted by the group as mentioned below has been accepted as satisfactory in partial fulfilment of the requirements for the degree B. Sc. in Computer Science and Engineering in July, 2025.

Group Members:
Jahirul Islam (2019331536)
Afsar Tanvir (2019331559)

Supervisor:

Md. Abu Naser Mojumder
Assistant Professor and Head
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

Certificates of Acceptance

The thesis is titled **"Hybrid Deep Learning Architectures for Bangla Toxic Language Detection: A Comparative Analysis of BERT-CNN-LSTM** and **BERT-GRU-Attention Models"** submitted by **Md. Jahirul Islam** and **Afsar Tanvir**; Student ID. **2019331536** and **2019331559**; Session **2019-20**, to the Department of Computer Science and Engineering , Sylhet Engineering College, has been accepted as satisfactory in partial fulfilment of the requirement for the Degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

_____

Internal

Nayan Kumar Nath
Lecturer
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

_____

Internal

Md. Lysuzzaman
Lecturer
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

_____

Internal

Md. Rasel Ahmed
Assistant Professor
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

_____

Internal

Md. Nazrul Islam
Assistant Professor
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

_____

Chairman

Md. Abu Naser Mojumder
Assistant Professor and Head
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

_____

Member (External)

Mohammad Shahidur Rahman, Ph.D., SMIEEE
Professor
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

Acknowledgements

First and foremost, we offer our deepest gratitude to the Almighty, whose boundless mercy and silent guidance have been our constant source of strength throughout this journey. Through every challenge and every success, His blessings have illuminated our path and sustained our resolve.

We would like to express our heartfelt appreciation to our respected supervisor, **Md. Abu Naser Mojumder**, for his invaluable support, continuous encouragement, and insightful feedback. His guidance has been pivotal in shaping the direction and quality of our research, and we are truly grateful for his mentorship.

Our sincere thanks also go to our respected teachers, **Md. Lysuzzaman** and **Nayan Kumar Nath**, whose dedication to teaching and depth of knowledge have left a lasting impact on our academic foundation. Their support and encouragement throughout our studies have been a source of inspiration.

We are especially thankful for the strong collaboration and mutual understanding we shared as thesis partners. The joint effort in completing our thesis, "**Hybrid Deep Learning Architectures for Bangla Toxic Language Detection: A Comparative Analysis of BERT-CNN-LSTM and BERT-GRU-Attention Models**," was made possible through shared dedication, teamwork, and a commitment to learning.

Lastly, with all our love and gratitude, we acknowledge the endless support of our families. Their unconditional love, sacrifices, and constant prayers have been the backbone of our academic journey. Without their unwavering belief in us, this achievement would not have been possible.

Abstract

The rapid growth of online communication platforms has created unprecedented opportunities for global expression, yet it has also intensified the spread of harmful behaviours such as cyberbullying, particularly among teenagers and young adults. This phenomenon is especially alarming due to its potential psychological and emotional consequences. While research in cyberbullying detection is steadily advancing in high-resource languages, Bangla remains significantly underrepresented.

In this study, we present a deep learning-based approach for detecting cyberbullying in Bangla language using a curated dataset of **19,575** social media comments. To address the linguistic complexity and low-resource nature of Bangla, we explore a range of hybrid deep learning architectures combining transformer-based contextual embeddings with recurrent and convolutional layers.

We implement and evaluate several models including **BERT + CNN + LSTM, BERT + GRU + Attention, BiLSTM + CNN, BiLSTM + Attention**, and a standalone **BiLSTM** model. Among these, the BERT + GRU + Attention architecture achieved validation accuracy of **98.57%** and highest test accuracy **98.52%**, closely followed by BERT + CNN + LSTM with highest validation accuracy of **98.64%** and **98.06%** test accuracy. The superior performance of transformer-enhanced models highlights the effectiveness of leveraging contextual embeddings alongside sequential and spatial features for Bangla text classification.

Keywords*: Natural Language Processing; Machine Learning; BERT's contextual embeddings; FastText Embedding; Bi-LSTM; Attention; CNN; GRU; BERT;*

Table of content

List of Figures

List of Tables

Chapter 1: Introduction

The rise of online platforms has opened unprecedented opportunities for communication and information sharing. However, this digital advancement has also given rise to harmful behaviour such as **cyberbullying**, especially on social media. Cyberbullying involves the use of digital means to harass, threaten, or demean others, and its psychological impact can be severe, particularly among young users.

In this study, we aim to address the problem of cyberbullying and toxic language detection in Bangla, a low-resource language, by applying and evaluating **hybrid deep learning models** using PyTorch. Our focus is on classifying social media texts into either **cyberbullying or non-cyberbullying** classes, contributing towards safer online environments for Bangla-speaking users.

1.1 Background

The rapid expansion of social media platforms such as Facebook, YouTube, and Twitter have revolutionized communication, particularly among youth and women. However, this digital growth has also introduced serious challenges, with cyberbullying emerging as a significant threat to mental health, digital safety, and social harmony. [8] The problem is especially critical in countries like Bangladesh, where digital literacy remains low, and effective legal and technical frameworks to combat online abuse are still under development. Recent studies paint a troubling picture of the cyberbullying landscape in Bangladesh. [13] A 2024 report by NETZ Bangladesh revealed that 78.4% of women have experienced technology-facilitated violence, with Facebook being the most common platform for abuse. Victims report mental health issues, including anxiety, depression, and suicidal ideation, with nearly 50% of those affected reducing their online presence to protect themselves. This disturbing trend has also been observed in teen and adolescent populations, particularly girls aged 14–22, who are often targeted with sexually explicit, body-shaming, or defamatory content. Despite the magnitude of the problem, over 70% of victims refrain from taking legal action, largely due to victim-blaming, fear of social stigma, and lack of trust in enforcement systems. [14] Moreover, while cyberbullying is briefly mentioned in Bangladesh's existing cyber laws (e.g., Section 25 of the Cyber Protection Ordinance), the proposed Cybersecurity Ordinance (2025) has controversially excluded

specific provisions on cyberbullying, raising concerns among human rights organizations and digital safety advocates. In response, [6] civil society organizations such as BRAC, Youth Policy Forum, and UNDP have initiated awareness campaigns and roundtable discussions calling for the integration of digital safety education, better fact-checking systems, and gender-sensitive policy reforms. However, the technical capacity to detect and counter such abusive content—particularly in Bengali language—remains limited. [2-6] The morphological richness and contextual ambiguity of Bengali make natural language processing (NLP) for this language especially complex. A lack of large, labelled Bengali datasets further impedes the development of intelligent content moderation systems. Therefore, there is a pressing need for automated, context-aware systems that can effectively detect, classify, and respond to cyberbullying in Bengali. Addressing this gap is essential not only for protecting vulnerable communities but also for preserving the digital rights, mental well-being, and social stability of millions of users across the country.

1.2 Problem Statement

Despite the rising concern about online harassment and abuse, there is limited availability of high-quality, annotated datasets in Bengali for training effective cyberbullying detection systems. Most available systems are trained on English data and fail to understand the unique structure, tone, and slang present in Bengali texts. This study addresses the gap by exploring models trained specifically for binary classification (positive/negative) of Bengali online comments.

1.3 Objectives

The main objectives of this study are:

- To curate and pre-process a dataset of Bengali online comments labelled for sentiment (positive or negative).
- To develop a machine learning pipeline capable of detecting cyberbullying content in Bengali.
- To evaluate the performance of PyTorch-based models for classifying offensive versus non-offensive content.

- To provide a scalable and adaptable solution for real-world social media monitoring systems.

1.4 Scope and Limitations

Scope:

- The focus is on binary classification (positive or negative sentiment) of online comments in Bengali.
- The model uses PyTorch and focuses on NLP pipelines suited for Bangla text.
- The data primarily comes from merged sources such as Saqib's dataset and other open-source Bengali datasets.

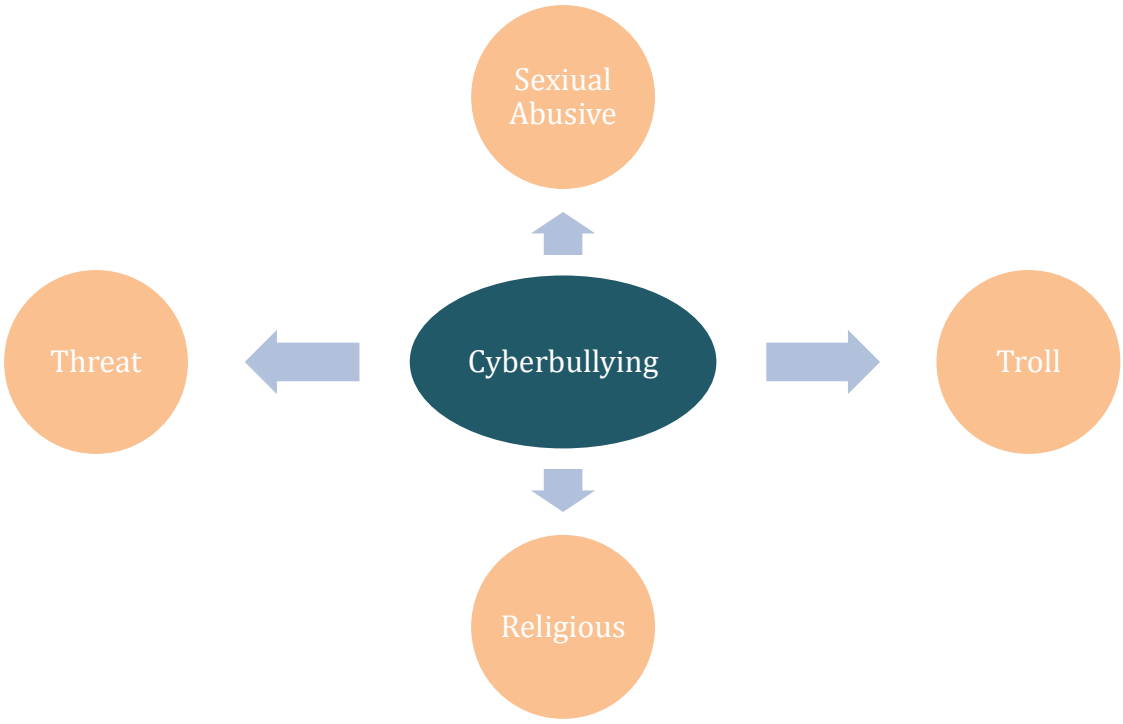


Figure 1: Common sources of Cyberbullying

Limitations:

- The classification is limited to positive vs. negative sentiment, and does not capture other nuanced classes such as "neutral", "sarcastic", or "contextually sensitive".
- Due to the lack of large, diverse datasets, the model might not generalize well to all dialects or slang variations in Bengali.

- Manual annotation of Bengali text for nuanced cyberbullying detection is still an ongoing challenge, which limits the reliability of labels.

1.5 Our Research Questions

1. Can cyberbullying in Bengali social media comments be detected using a binary classification model?
2. Are pre-defined wordlists (positive/negative) effective for labelling?
3. Can simple datasets (e.g., Saqib's) generalize well for Bengali sentiment detection?

1.6 Subsections (Overview)

1.6.1 Cyberbullying in Social Media

Cyberbullying is a form of harassment that takes place using digital communication tools. In social media platforms, it includes hate speech, threats, sexually explicit remarks, or shaming. Bengali-speaking internet users, especially women and public figures, are often targets of such online abuse.

1.6.2 Motivation Behind Cyberbullying Detection in Bangla Language

With the growing online population in Bangladesh and West Bengal, there is a pressing need for AI-based moderation systems in Bengali. Most existing systems work well only in English. This research aims to address this linguistic and technological gap.

1.6.3 Organisation of This Study

- Section 1: Introduction
- Section 2: Literature Review
- Section 3: Methodology
- Section 4: Result and Analysis

- Section 5: Limitations and Future Work
- Section 6: Conclusions

1.6.4 Model Information

The study uses PyTorch for implementing a binary sentiment classifier. It incorporates:

- Pre-processing steps (cleaning, filtering)
- Dataset merging from multiple sources
- Positive and negative word lexicon for weak labelling
- A final two-class classification (positive vs. negative)

Chapter 2: Literature Review

2.1 Related Works

Numerous studies have recently focused on the detection of cyberbullying and hate speech, especially in low-resource languages like Bangla. Researchers have applied a variety of deep learning techniques, datasets, and evaluation methods to address this growing concern.

One such study presented a multi-model approach utilizing five deep learning architectures—Bi-GRU-LSTM-CNN, Bi-LSTM, CNN, LSTM, and XGBoost—alongside embedding techniques like Word2Vec, FastText, and BERT to identify hate speech in Bangla text [4]. The results highlighted that LSTM with Word2Vec achieved the highest accuracy of 95.66%, demonstrating that traditional RNN-based methods still hold value when combined with suitable embeddings for morphologically rich languages. Another research effort focused on a Bi-LSTM model trained on a dataset of over 12,000 Bangla comments sourced from multiple social media platforms [4]. It employed various optimizers and 5-fold cross-validation, with Adam optimizer achieving 95.08% accuracy and an F1 score of 95.23%, confirming that optimization strategy plays a crucial role in performance gains [3]. In a separate study, BERT was compared against Bi-LSTM and Bi-GRU for both binary and multiclass classification tasks. The BERT model outperformed others with an F1 score of 0.89 for binary and 0.85 for multiclass tasks, showcasing the power of transformer-based architectures in understanding context and semantics, even in underrepresented languages like Bangla [6]. Further, a hybrid model combining self-attention mechanisms with BiLSTM showed promising results by applying context-based data augmentation and feature extraction techniques like TF-IDF, count vectorizer, and transformer embeddings. This model achieved 89.3% accuracy, outperforming classical algorithms like SVM, Naive Bayes, and even other deep learning models [2].

Table 1: Comparison of the Proposed Models with other Existing Deep Learning Models

Author	Dataset	Methodology	Feature	Result	Limitations
Amit, Asif, Anik, Nur (2021) [27]	7425	LSTM, GRU	TF-IDF	77% accuracy	Imbalanced dataset where 6020 are hate speech
Md. T. Ahmed et al. (2021) [29]	5000 Bangla and 7000 Romanized Bangla	CNN, Multinomial Naïve Bayes	TF-IDF	For Bangla, accuracy of 84% in CNN, 80% in MNB for Romanize Bangla	Lower dataset
Ahammed et al. (2019) [30]	1339 (665 hate, 674 normal)	SVM, Naïve Bayes	Count Vectorizer, TF-IDF	72% accuracy in NB, 70% accuracy in SVM	Lower accuracy and lower dataset
R. Ghosh et al. (2021) [31]	--	SVM, LR, RF, Passive Aggressive Classifiers	TF-IDF, BoW	78.1% accuracy in PA with N-gram	Lower accuracy
F. Ahmed et al. (N/A) [32]	44001 FB comments (nonbully, sexual, threat, troll, religious)	CNN-LSTM, Ensemble Model	Word2vec	87.9% accuracy in BC using CNNLSTM and 85% accuracy in MC with Ensemble model	Higher complexity in operation
Tripto et. al (2018) [33]	15689 (5011 Bangla, 4189 English, 6489 Romanized)	LSTM, CNN, NB, SVM	CBOW and Skip Gram	LSTM performs better. 3 class sentiment: 65.97%, 5 class sentiment: 54.24%, 59.23% in Emotion	Lower accuracy due to low amount of data in 3 class, 5 class and emotion.
Chakraborty et. al (2019) [34]	5644 (2739 abusive, 2905 non abusive)	Multinomial Naïve Bayes, SVM, CNN-LSTM	TF-IDF	SVM with Linear Kernal performs better with accuracy 78%	Lower dataset and accuracy

Finally, another comprehensive review provided insights into how deep learning surpasses traditional machine learning approaches in cyberbullying detection tasks. The study discussed various DL frameworks and emphasized the advantages of automatic feature extraction and scalability, but also highlighted the need for richer Bangla datasets and more nuanced annotations [5].

2.2 Research Gap

Despite these promising advances, several gaps remain:

Limited Bengali Datasets: Most models are trained on small or imbalanced datasets. Even the largest datasets reviewed (e.g., 19,575 comments) are relatively small compared to English-language counterparts.

Weak Handling of Contextual and Sarcastic Language: While models like BERT show improvements, sarcasm and context-based hate are still under-addressed, especially in Bangla.

Binary Classification Focus: A majority of studies focus solely on binary classification (hate/non-hate), overlooking the need for multiclass or fine-grained emotion categories.

Limited Hybrid Approaches: Although hybrid models have been proposed, there's still a lack of experimentation combining attention, recurrent networks, and contextual embeddings in a unified Bangla pipeline.

No Standard Benchmark: Varying datasets and inconsistent validation methods make it hard to compare models across studies.

Chapter 3: Methodology

This section describes the methodological framework followed in our study. It covers the data acquisition and pre-processing strategies, the deep learning models implemented, the embedding techniques used, the experimental setup, and the evaluation metrics applied to measure model performance.

3.1 Data Collection and Annotation

The dataset used in this study was constructed by merging multiple publicly available Bangla text datasets, primarily consisting of user comments collected from platforms like Facebook, YouTube, and Twitter. To ensure consistency and usability:

- Duplicates were removed.
- Only relevant text fields were retained.
- Weak supervision techniques were applied using curated Bangla lexicons to label comments as cyberbullying (negative) or non-cyberbullying (positive).

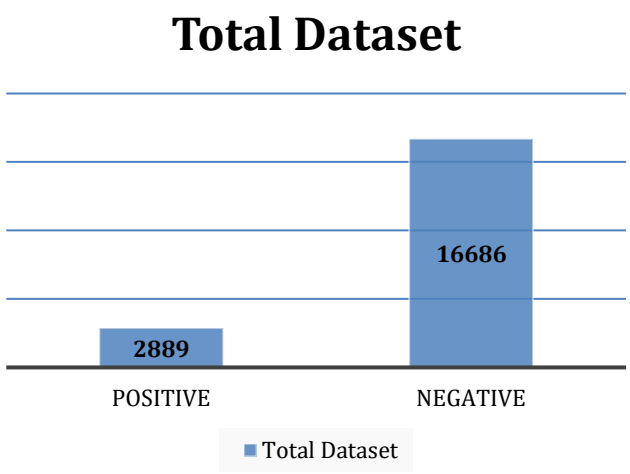


Figure 2: Data Classification

This resulted in a binary classification problem, suitable for training supervised learning models.

```
Sample Positive Comments:
                                                  comment  category
9380                        শামীম ভাইয়া চিল বড়োঅসাধারণ বস  positive
14286  সত্যি শামিম ভাইয় নাউনি একজন গুনি অভিনে সেটার ...  positive
15053  অপ্রত্যাশিত সত্য দিয়ে পৃথিবী ঠিকানা অসাধারণ স...  positive
16694                          আগামী আরো সুন্দর নাটক পাবো  positive
886      মিষ্ বউ থাকল জীবন লাগ বান্নাহ ভাই ঈদে সেরা নাটক  positive
621                             অসাধারণ নাটক আনেক শেখার  positive
14936                          অসাধারণ অভিনয় নিসো ভাইয়  positive
5068   অসাধারণ একটা নাটক নাটক টা ভালো একটা জীবন হতো ভ...  positive
14811  অমি ভাই কাজটাই ফাটিয় দিলেনদারুণ শুরুতব সবারই ...  positive
2397                        পলাশ ভাইয় অভিনয় টা দারুণ হইছু  positive
1455                                সত্যি ভালো লাগলো  positive
993    ছোট কিন্তু সুন্দর একটা স্ক্রিপ্ট ওভিনয় কথা বল...  positive
4462   অপূর্বর অভিনয় অপূর্ব অসম্ভব সুন্দর চিত্রনাট্য...  positive
11068  নাটক শেষ টা শেষ ভাব পারিনিসত্যি অসাধারণ একটা নাটক  positive
5846                              এককথায় অসাধারণ  positive

Sample Negative Comments:
                                                  comment  category
11122           মামুনুল হক সাহেব নাস্তিক দের ধ্বংস করবো  negative
7422              তুই পিচ্চি হান্নান হও তাইল পুতিন শালা  negative
18227                 মাগীর পাছায় লাথি মের প্লাটফর্ম  negative
438            ব্রহ্মা বীর্য পাত্র সংগ্রহ সরস্বতী জন্ম  negative
12918                                মার যাব  negative
...
18932                                তর ফাঁসি দাবি  negative
14106                চিন্ কিস এতো লিঙ্ক অপেক্ষায় সবাই  negative
13089                              ভোদা ঠোঁট সুন্দর  negative
12897           কির কদে তারেক জিয়া জতে চায়না একি সুর  negative
```

Figure 3: Examples of Sample Data

3.2 Data Pre-processing

The collected texts underwent several pre-processing steps:

- Removal of punctuation, emojis, and special characters
- Stop-word removal using a Bangla stop-word list
- Tokenization using space-based or language-aware tokenizers

```
Original Cleaned Text: অসাধারণ হিস্ট্রি চোখ পানি রাখ পারলাম না
Tokenized Sequence: [5, 4278, 33, 41, 86, 154, 4]
Padded Sequence (length): 100
[   5 4278   33   41   86  154    4    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0]
```

Figure 4: Sample Tokenized Sequence

	Cleaned Text	Decoded Tokenized Text
721	নাটক অসাধারণ তবে আবেগ খারাপ লাগছে শেষ মিল না হওয়া	[নাটক, অসাধারণ, তবে, আবেগ, খারাপ, লাগছে, শেষ, ম...
501	কথায় নাটক খুবই অসাধারণ গল্প গান অভিনয় খুবই স...	[কথায়, নাটক, খুবই, অসাধারণ, গল্প, গান, অভিনয়...
1586	পরকাল বিশ্বাস করলেতো শরীর বেচ খেত পারব না চাপা...	[পরকাল, বিশ্বাস, করলেতো, শরীর, বেচ, খেত, পারব,...
871	ভুয়া অভিযোগ ঢুক লুটপাট চালানোর ভালোমানুষ সাজার	[ভুয়া, অভিযোগ, ঢুক, লুটপাট, চালানোর, ভালোমানু...
2144	ইসলাম খারাপ মন্তব্য ইসলাম বিরোধি জালিম বাংলার ...	[ইসলাম, খারাপ, মন্তব্য, ইসলাম, বিরোধি, জালিম, ...

Figure 5: Clean and Decoded Tokenized Text

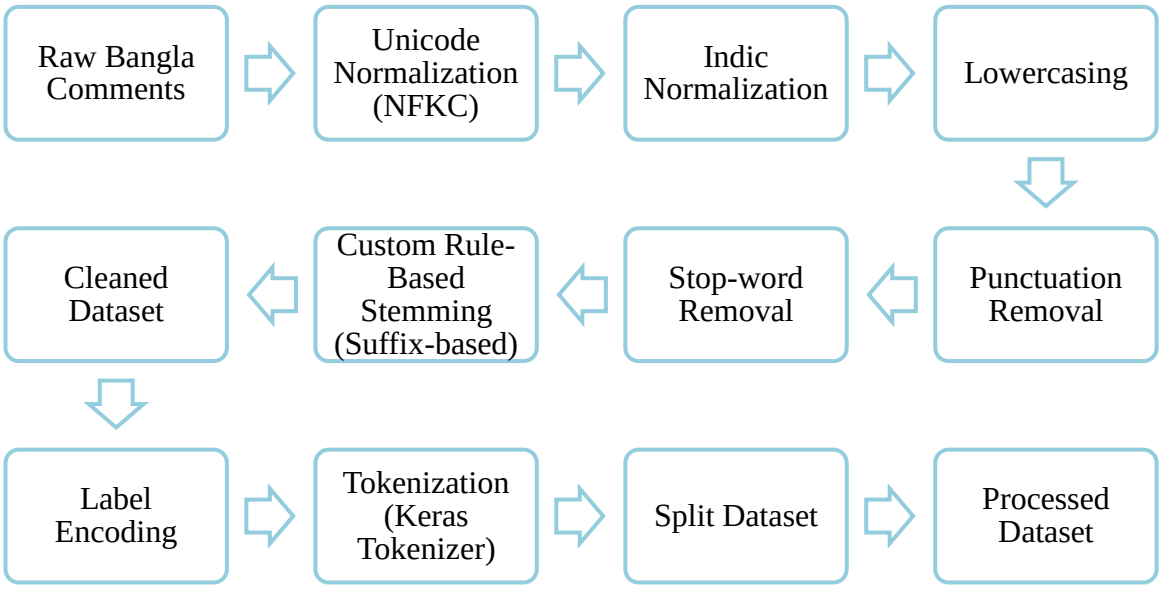


Figure 6: Data pre-processing

For models requiring embeddings (e.g., LSTM, CNN), word tokens were converted into numerical vectors using pretrained embeddings like FastText.

3.3 Text Representation with FastText Embeddings

We utilized FastText—a sub word-level embedding technique that supports rich morphological languages like Bangla. It helps capture semantic similarity and better handles rare or misspelled words.

Key advantages:

- Pretrained vectors on Bangla corpus
- Handles out-of-vocabulary (OOV) words via sub word representations

Mathematically, [1] FastText can be explained as shown in Equation (1) where v_w stands for the FastText embedding for word w, G_w stands for the set of all n-grams of w, z_g stands for the embedding vector for n-gram g, and $\sum$ represents the outline of overall n-grams in G_w.

$$v_w = \sum g \in G_w z_g \qquad (3.1)$$

Each tokenized sentence was converted to a fixed-length vector matrix as input to deep learning models.

3.4 Deep Learning Models for Hate Speech Detection

We experimented with several deep learning architectures to evaluate their performance in binary cyberbullying detection.

3.4.1 Bidirectional LSTM (BiLSTM)

Bidirectional LSTM (BiLSTM) is a deep learning architecture that reads text sequences in both forward and backward directions, allowing it to capture context from both past and future words [6]. This dual-context understanding is especially useful for hate speech detection in languages like Bangla, where word meaning often depends on surrounding context. Unlike traditional LSTM, which only processes input sequentially from start to end, BiLSTM improves classification by combining outputs from both directions. It uses pre-trained embeddings (e.g., FastText) to represent words and learns to detect toxic patterns more accurately. BiLSTM is efficient, adaptable, and performs well even with informal and noisy text, making it suitable for Bangla hate speech detection.

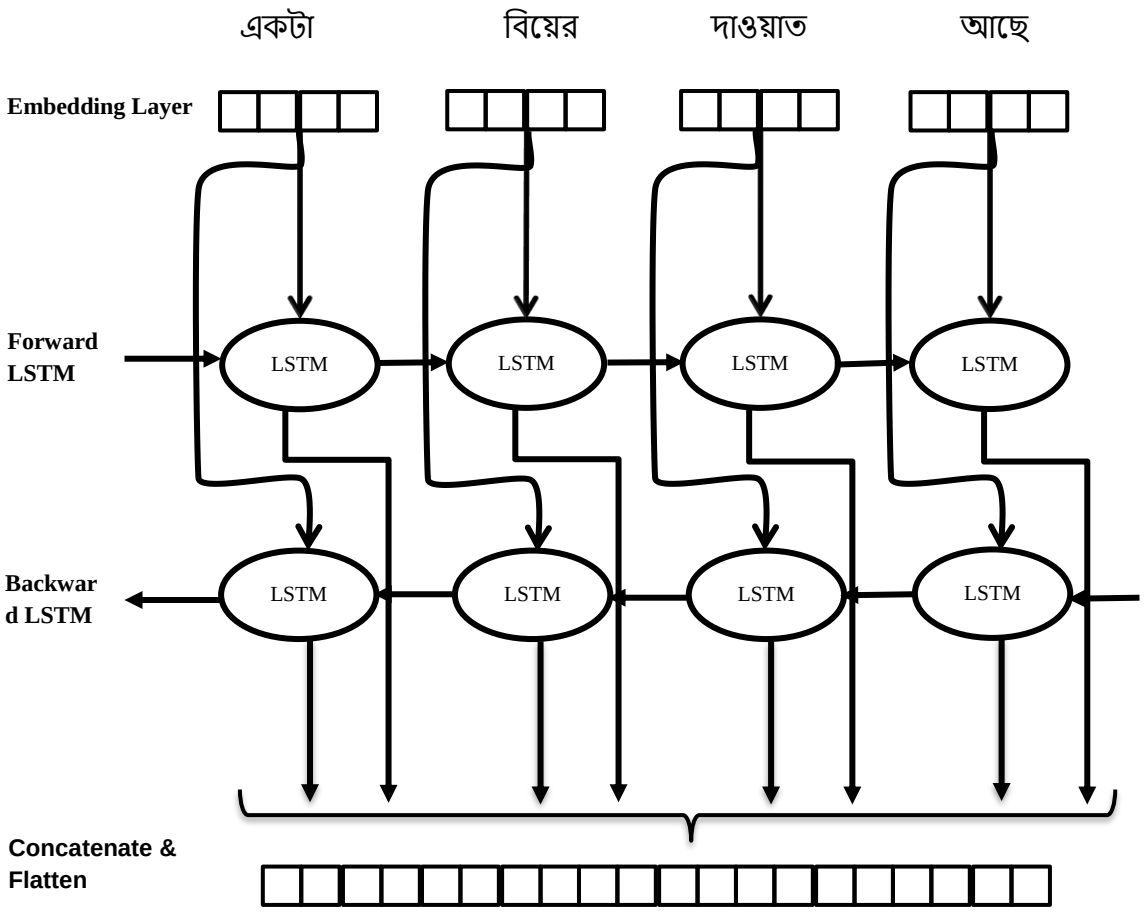


Figure 7: BiLSTM Architecture

3.4.2 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs), though originally designed for image processing, have proven highly effective in text classification tasks like hate speech detection. CNNs use convolutional filters to capture local patterns such as key phrases or word combinations that are indicative of toxic content. In the context of Bangla text, CNNs can identify n-gram-like features regardless of their position in the sentence. These patterns are extracted through multiple filters and passed through pooling layers to retain the most significant information. CNNs are fast, parallelizable, and work well with word embeddings like FastText, making them a strong choice for detecting abusive or toxic language in short texts.

Figure 8: CNN Architecture

3.4.3 Hybrid: BiLSTM + CNN

The BiLSTM + CNN Hybrid model combines the strengths of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks to effectively capture both local features and long-term dependencies in text data. The model begins with an **embedding layer** that transforms input tokens into dense 300-dimensional vectors. A **1D convolutional layer** with 128 filters and a kernel size of 5 extracts local n-gram features, followed by **max pooling** to reduce dimensionality. These features are passed into a BiLSTM layer, which captures context from both forward and backward directions. A **dropout layer** is applied to prevent overfitting. The output is passed through two **fully connected layers**, ending with a SoftMax classification into two classes: cyberbullying or not. This architecture is well-suited for complex patterns in Bangla toxic language detection.

3.4.4 Attention-based Hybrid: BiLSTM + Attention

The BiLSTM + Attention model integrates a Bidirectional LSTM with an attention mechanism to enhance the model's ability to focus on the most relevant parts of the input text. The model begins with an **embedding layer** that converts input tokens into 300-dimensional dense vectors. A **BiLSTM layer** processes the input in both forward and backward directions, producing a 256-dimensional contextual representation. The **attention layer** computes a weight for each time step, allowing the model to prioritize words or phrases critical for detecting cyberbullying. This attention-weighted context vector is passed through a dropout layer to prevent overfitting and then through two **fully connected layers** for final classification. This architecture enables the model to selectively attend to toxic language cues in Bangla text, improving interpretability and performance.

3.4.5 Transformer-based Hybrid: BERT + CNN + LSTM

The BERT + CNN + LSTM model is a powerful hybrid architecture that combines the deep contextual understanding of BERT with the sequential modelling capabilities of CNN and BiLSTM layers for cyberbullying and toxic language detection in Bangla. It starts with a pretrained BERT model, which generates rich, **768-dimensional** contextual embeddings for each token, capturing the nuanced semantics of language. These embeddings are then passed through a **1D Convolutional layer**, which helps extract local patterns such as offensive phrases or repeated insults. Next, the output is fed into a Bidirectional LSTM that models the sequence from both directions, capturing long-range dependencies. A **dropout layer** is applied to reduce overfitting, followed by two **fully connected layers** that transform the extracted features into final classification logits. This architecture benefits from BERT's language understanding, CNN's feature extraction, and BiLSTM's context tracking; making it highly effective for detecting subtle and context-dependent abusive language in Bangla texts.

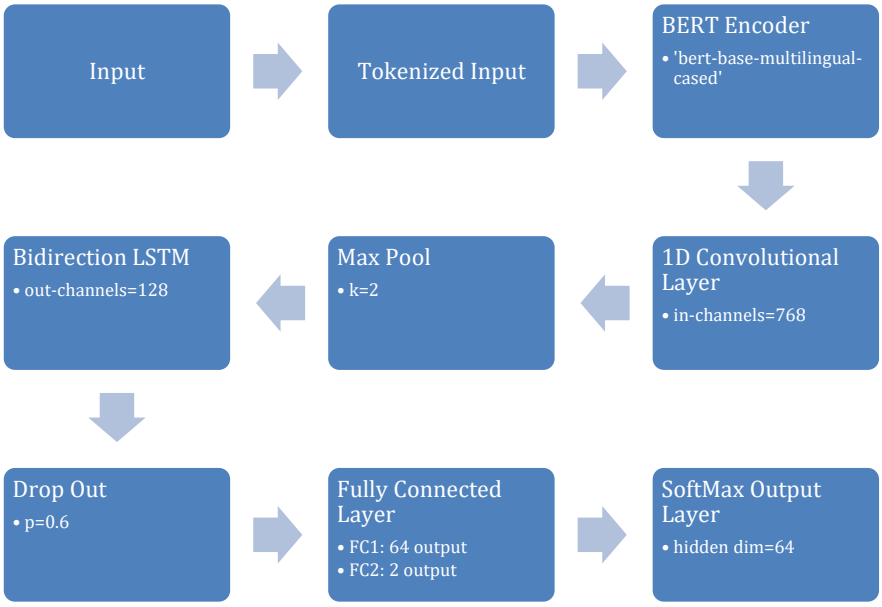


Figure 9: BERT + CNN + LSTM Working Architecture

3.4.6 Deep Hybrid: BERT + GRU + Attention

The BERT-GRU-Attention model is a hybrid deep learning architecture designed for toxic or hate speech detection in Bangla. It begins with **BERT**, a transformer-based language model, which provides rich contextual embeddings for each token in the input text. These embeddings are then passed into a **Bidirectional GRU** layer that captures sequential patterns in both forward and backward directions. To further enhance focus on important words, an **attention mechanism** is applied over the GRU outputs, allowing the model to weigh more critical tokens

heavily during classification. This attention-weighted representation is then fed through **fully connected layers** with dropout for regularization, and finally outputs class probabilities (e.g., toxic vs. non-toxic). The model combines the strengths of contextual understanding (BERT), sequential learning (GRU), and relevance focusing (Attention), making it robust for detecting nuanced hate speech in noisy Bangla text.

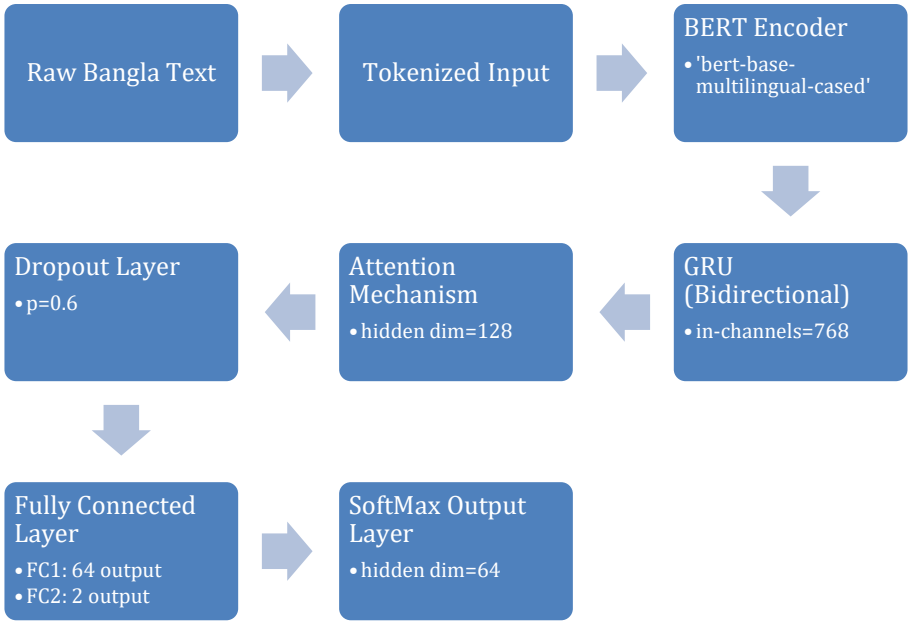


Figure 10: BERT + GRU + Attention Working Architecture

3.5 Experimental Setup

- ➔ **Programming Language**: Python
- ➔ **Deep Learning Framework**: PyTorch, Keras Tokenizer
- ➔ **Embedding Source**: Pretrained FastText (Bangla), bert-base-multilingual-cased
- ➔ **Hardware**: NVIDIA GPU-enabled environment (e.g., RTX-9800)
- ➔ **Training Parameters**:
  - ◆ Epochs: 15–50
  - ◆ Optimizer: Adam
  - ◆ Loss Function: Binary Cross entropy
  - ◆ Batch size: 32/64
  - ◆ Dropout: 0.3–0.6 for regularization

3.6 Performance Evaluation

To evaluate model effectiveness, the following metrics were used:

- Accuracy (A)
- Precision (P)
- Recall (R)
- F1-Score (F1)
- AUC-ROC Curve

3.6.1 Mathematical Formulas:

Let TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively.

$$\text{Precision} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Positive\ (FP)} \quad (4.1)$$

$$\text{Recall} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Negative\ (FN)} \quad (4.2)$$

$$\text{F1 score} = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4.3)$$

$$\text{Accuracy} = \frac{True\ Positive\ (TP)+\ True\ Negative\ (TN)}{True\ Positive\ (TP)+False\ Positive\ (FP)+\ True\ Negative\ (TN)+False\ Negative\ (FN)} \quad (4.4)$$

Models were evaluated on a **held-out validation set** and optionally tested using **k-fold cross-validation** for robustness. In equation (4.1), [2]precision denotes the ratio between the correctly classified samples for class 'A' and the total samples of class 'A' to be classified by a model. In equation (4.2), recall stands for the ratio of the number of positive samples needed to be classified and the summation of all positive samples. In equation (4.3), the f1-score is an evaluation metric that take both the value of precision and recall to evaluate the model. In case of an imbalanced dataset, the f1-score can be measured when the false positive rate and false negative rate differ from each other. In equation (4.4), we get the value of accuracy that depicts the performance of a model and how well it works on different classes. This is actually the ratio of the correctly predicted samples and the number of total predicted samples.

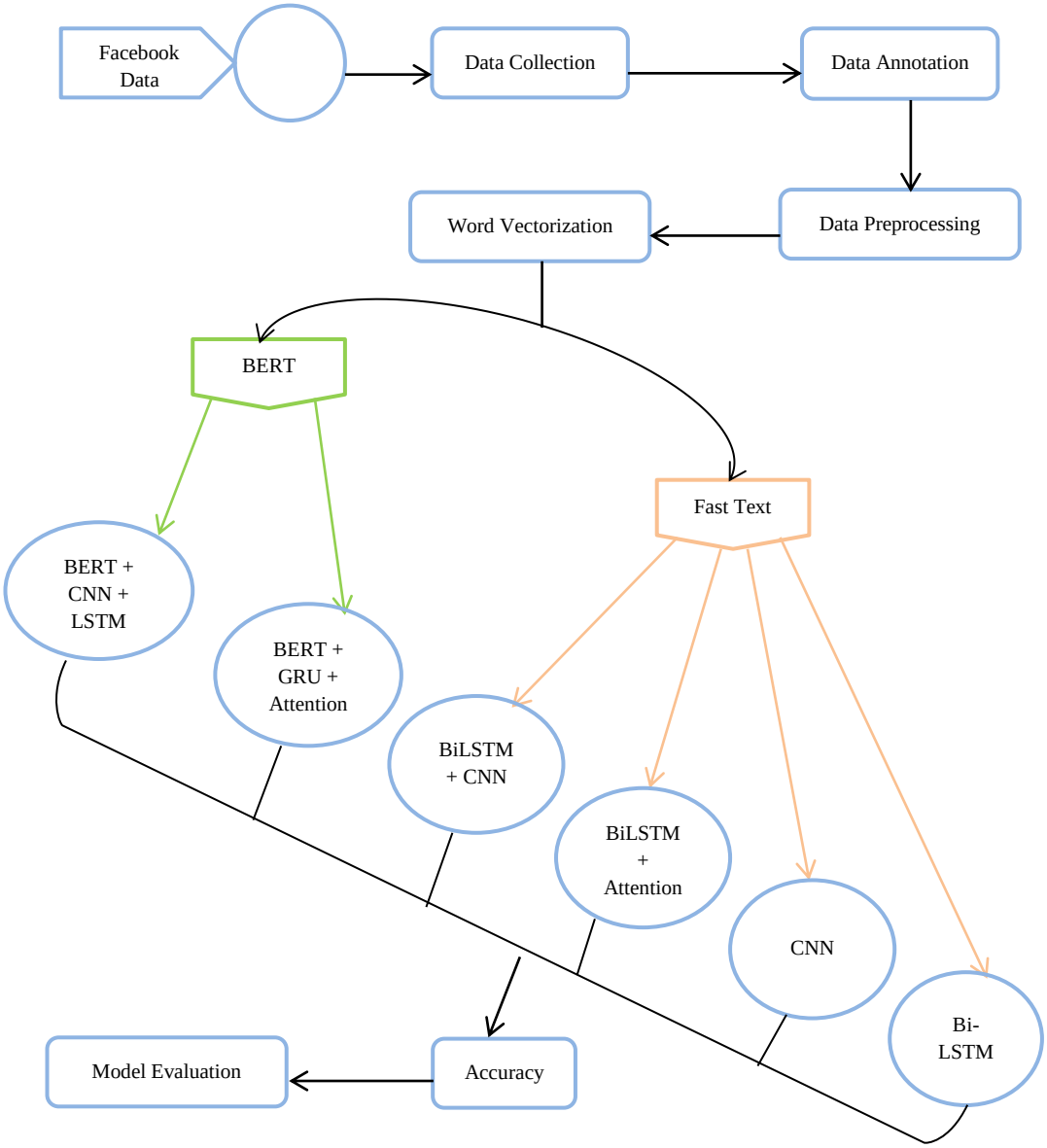


Figure 11: Development approach of Proposed "Bangla Toxic Language Detection" model

Chapter 4: Result and Analysis

This study evaluates multiple deep learning models to detect toxic and hate language in Bengali social media text. The experiment involves hybrid architectures such as BERT + CNN + LSTM and BERT + GRU + Attention, alongside traditional models like Bi-LSTM, CNN, and CNN + BiLSTM. Each model was trained and validated using **stratified k-fold cross-validation** to ensure generalizability and robustness. The final evaluation was conducted on a held-out test set to assess real-world performance.

For the deep learning models, the **Adam optimizer** was used consistently due to its adaptive learning rate and ability to handle sparse gradients effectively. To combat overfitting and enhance convergence, early stopping and learning rate scheduling techniques were employed where necessary. The training process spanned 30 to 50 epochs, with a batch size of 16 to 32, and a maximum sequence length of 100 for standard models, while BERT-based models used token lengths up to 128.

The Bi-LSTM model achieved the best overall performance with a peak test accuracy of 96.53% and a macro F1-score of 92%, effectively capturing long-range dependencies in Bangla text data. The CNN model, while slightly behind in sequential understanding, still demonstrated strong performance with a maximum test accuracy of 94.38% and a macro F1-score averaging around 87% across folds. Its strength lies in capturing localized n-gram patterns, although performance fluctuated more compared to Bi-LSTM due to its limited ability to model context across longer sequences.

In our comparative evaluation of hybrid deep learning models, both the CNN + BiLSTM and the BiLSTM + Attention architectures demonstrated promising classification capabilities across **5-fold cross-validation**. The CNN + BiLSTM model achieved an average test accuracy of 89.57% and a macro F1-score of 81.80%, indicating effective learning of both local and sequential features. However, its performance slightly fluctuated, particularly in fold 5, where accuracy dropped due to lower generalization on the minority class. In contrast, the BiLSTM + Attention model consistently outperformed the CNN-based hybrid across all metrics. With an average accuracy of 93.39% and a macro F1-score of 86.80%, this model demonstrated superior robustness in identifying toxic content, even under class imbalance. The attention mechanism likely enhanced the model's focus on relevant semantic cues in long Bengali texts,

contributing to improved recall for the positive class. Overall, the BiLSTM + Attention model is the most effective among the evaluated architectures, making it a strong candidate for practical deployment in Bangla social media moderation systems.

In contrast, BERT-based hybrid models outperformed traditional deep learning models significantly. The BERT-GRU-Attention model achieved the highest accuracy and F1-score among all, with an average accuracy of 98.01% and an F1-score of 96% on the test set across 3-fold cross-validation. The highest performance was observed in **Fold 1**, reaching 98.52% accuracy and a 97% macro F1-score. Its layered structure enabled both semantic contextualization from BERT and local/global pattern extraction from CNN and LSTM respectively. Similarly, on average BERT-CNN-LSTM model also performed competitively, with an F1-score of 96% and accuracy of 97.82%, attributed to the **attention mechanism**'s ability to highlight toxic language patterns with precision. The best performance was recorded in **Fold 1**, with 98.26% accuracy and 96% macro F1-score.

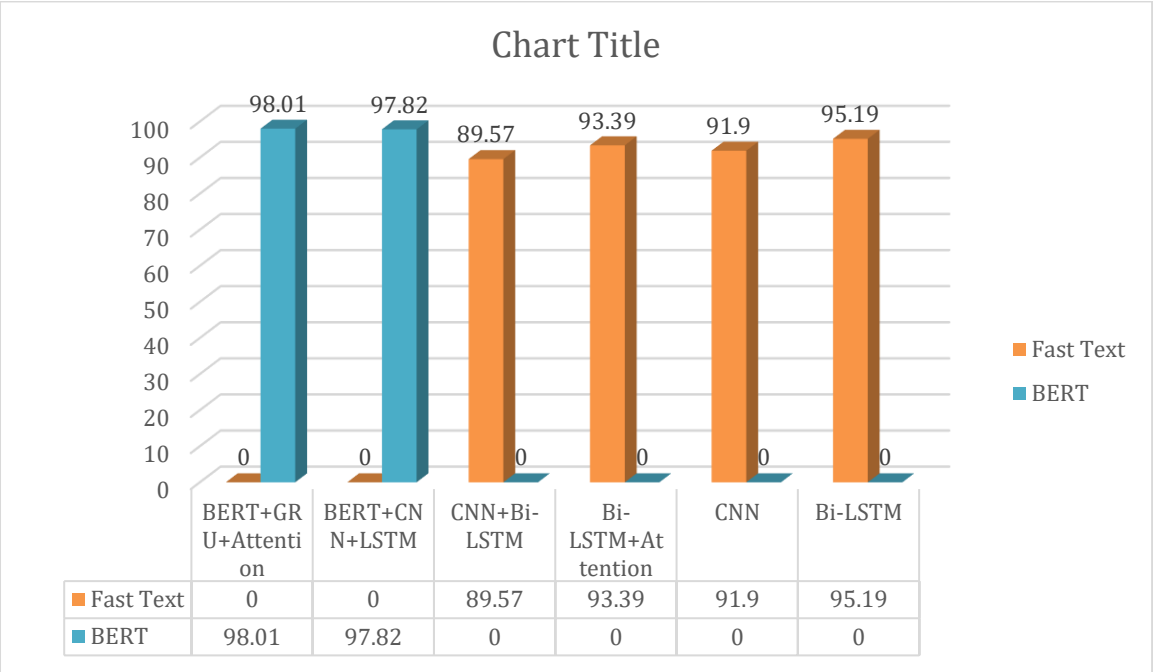


Figure 12: Analyzing the efficiency score of applied HDL algorithms

The loss and accuracy curves for training and validation across different folds confirmed that the models converged well without significant overfitting, particularly in BERT-based variants. Confusion matrices revealed that the models maintained a balanced performance across both toxic and non-toxic classes, with low false positive rates in most folds.

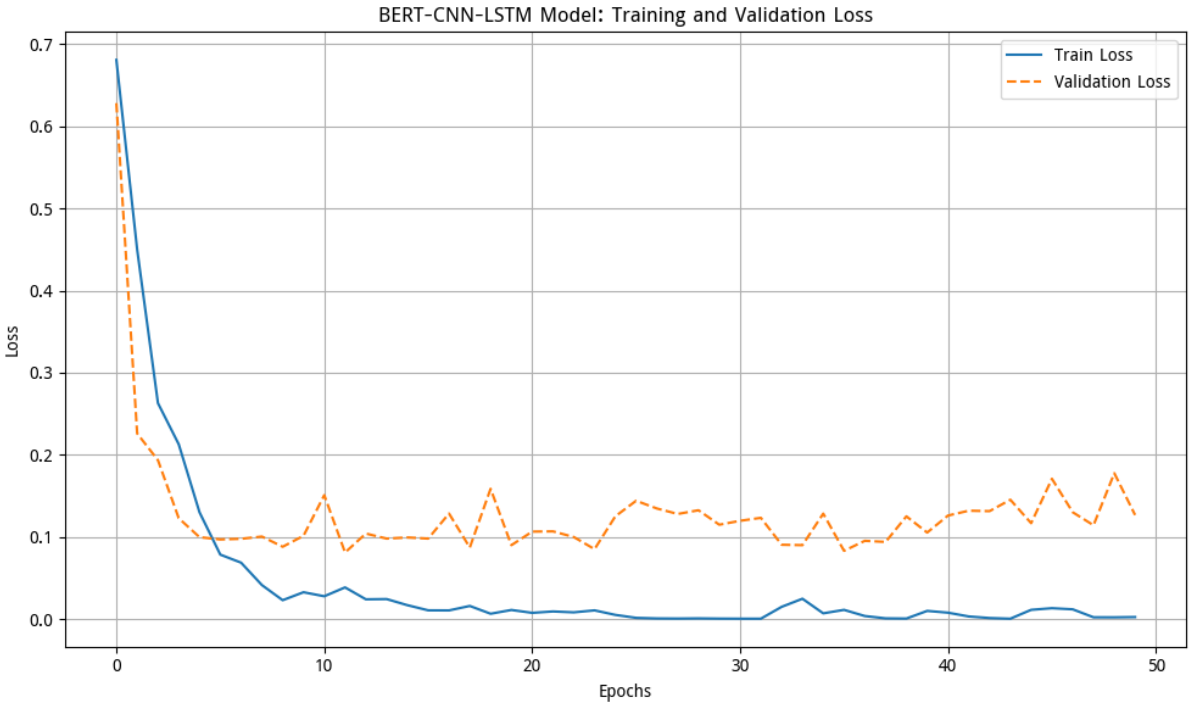


Figure 13: Training Loss and Validation Loss (BERT-CNN-LSTM)

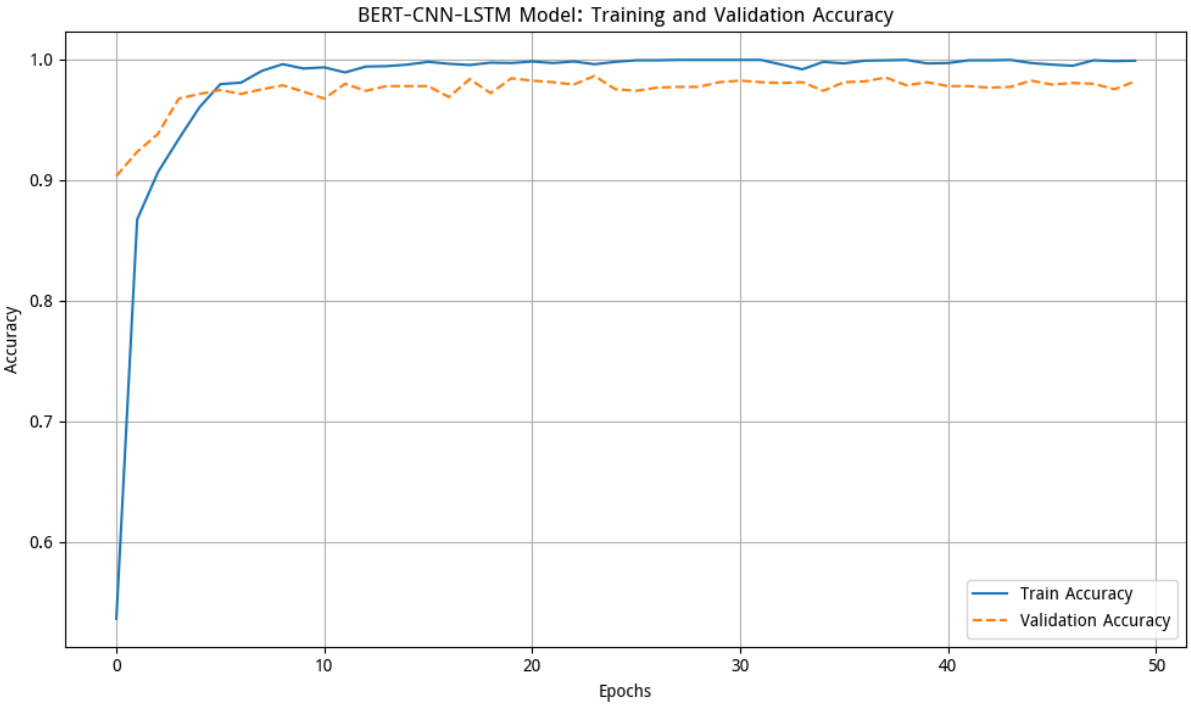


Figure 14: Training Accuracy and Validation Accuracy (BERT-CNN-LSTM)

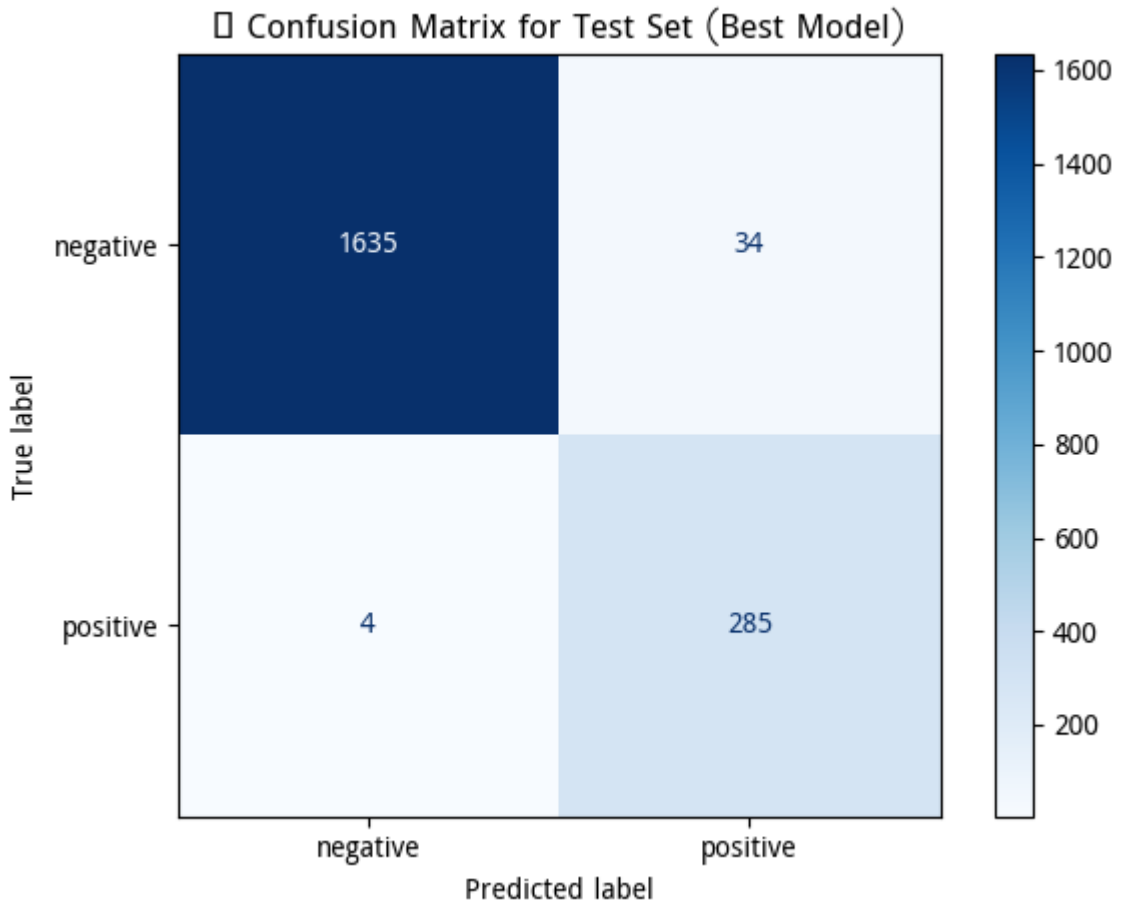


Figure 15: Confusion Matrix (BERT-CNN-LSTM)

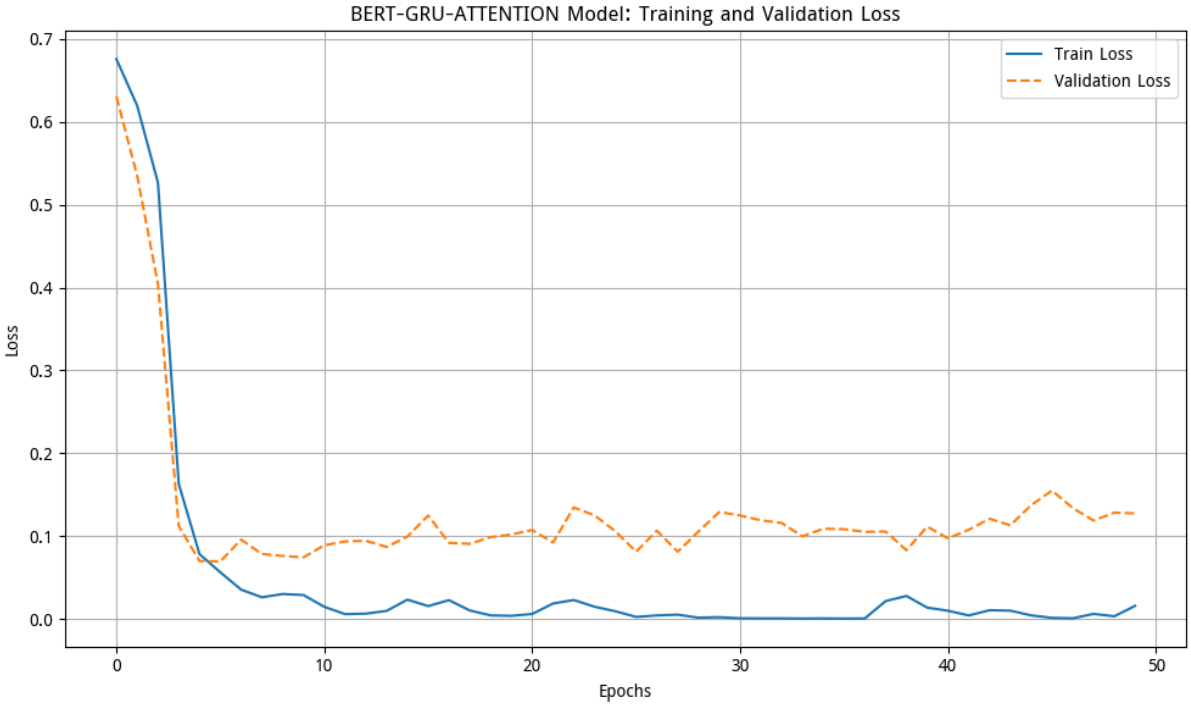


Figure 16: Training Loss and Validation Loss (BERT-GRU-Attention)

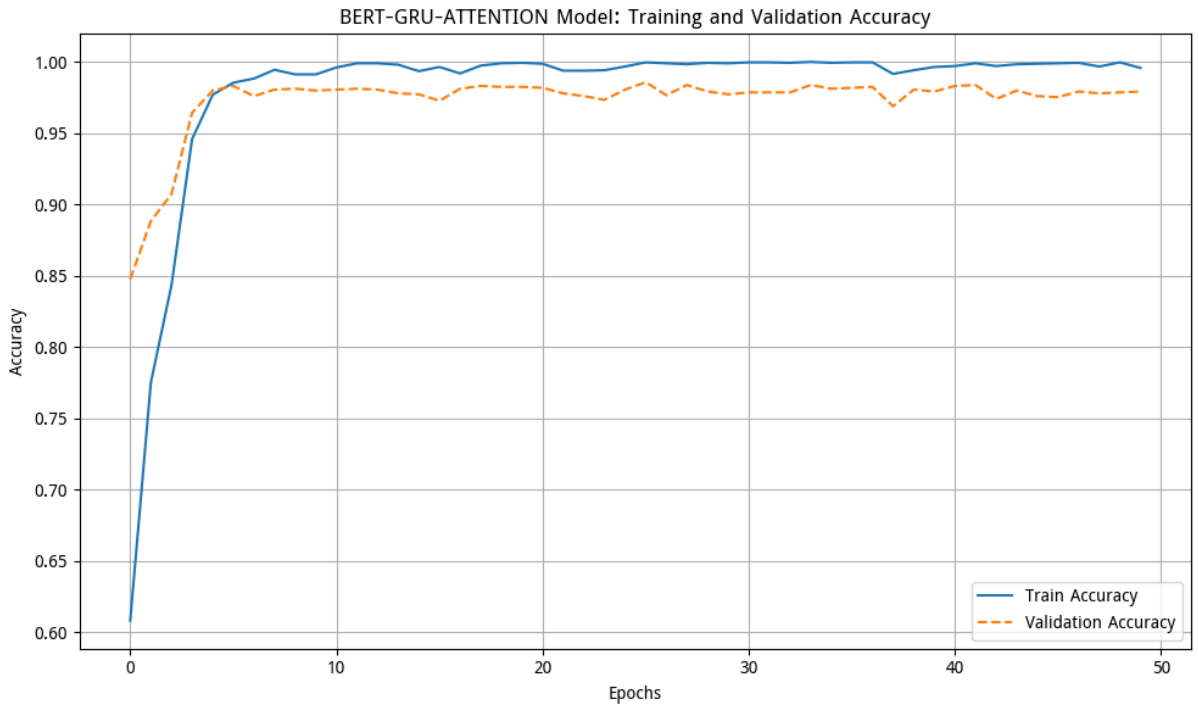


Figure 17: Training Accuracy and Validation Accuracy (BERT-GRU-Attention)

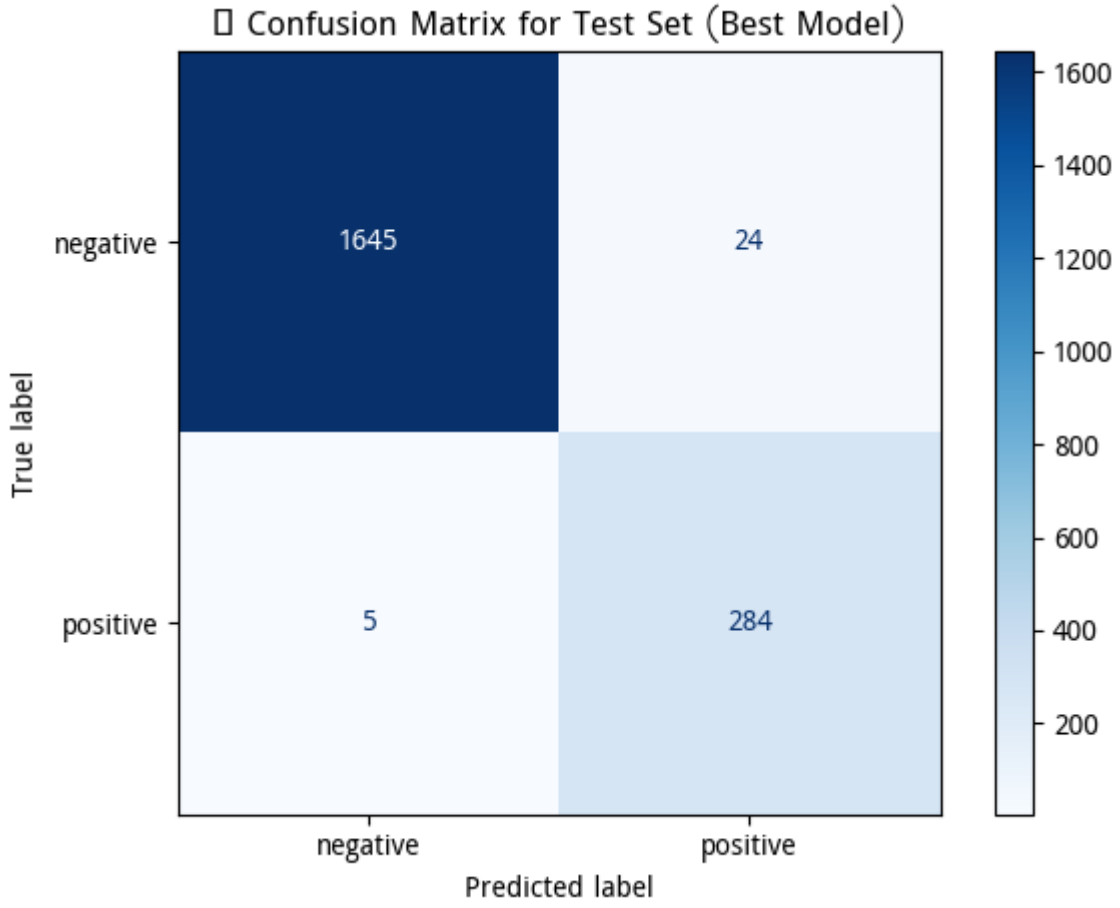


Figure 18: Confusion Matrix (BERT-GRU-Attention)

A cross-model comparison as shown in **Table 2**, clearly indicates the superior performance of hybrid models involving BERT, especially when combined with CNN or attention-based GRU layers. This supports the hypothesis that leveraging pretrained contextual embeddings from BERT improves the ability of the model to capture nuanced linguistic features, even in morphologically rich languages like Bengali.

Overall, the results validate that hybrid architectures incorporating both pretrained transformers and sequential deep learning models are effective for Bengali toxic comment detection. While classical deep learning models such as Bi-LSTM still show strong results, BERT-based methods exhibit more robust generalization and better handling of context-sensitive text, making them more suitable for real-world deployment in hate speech moderation systems.

4.1 Result Analysis

Our final dataset consisted of **labelled Bangla social media comments** categorized as either **cyberbullying (negative)** or **non-cyberbullying (positive)**. The data distribution was:

- **Total samples**: ~19,575
- **Negative (bully/hate)**: ~16,686 samples
- **Positive (non-bully)**: ~2,889 samples

A nearly balanced distribution ensured that the models did not develop a bias toward any single class.

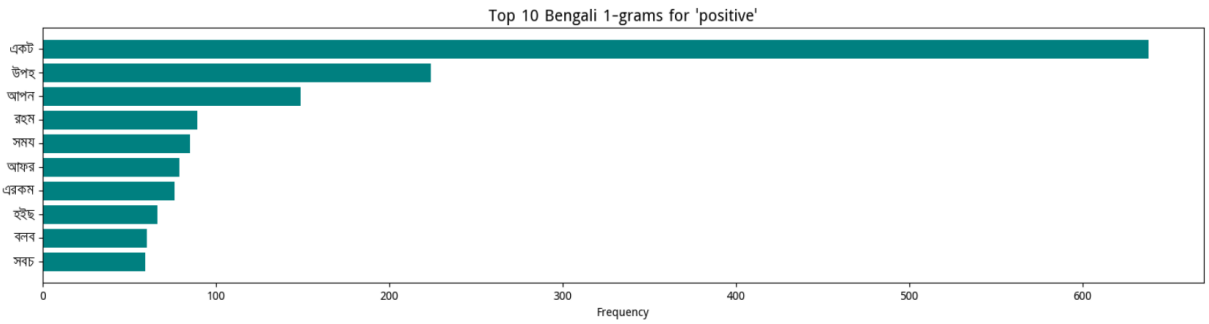


Figure 19: Top 10 Bengali 1-grams for 'positive'

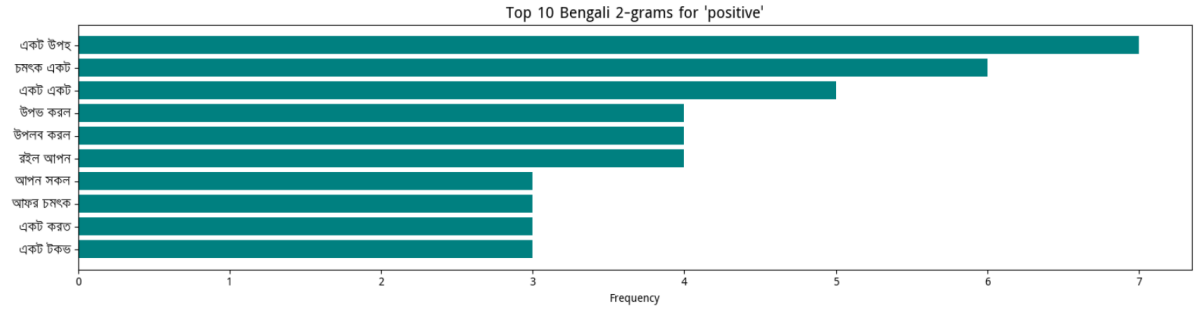


Figure 20: Top 10 Bengali 2-grams for 'positive'

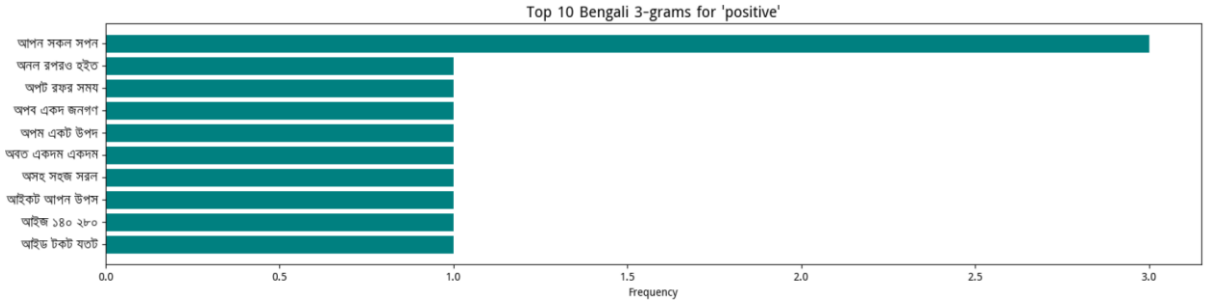


Figure 21: Top 10 Bengali 3-grams for 'positive'

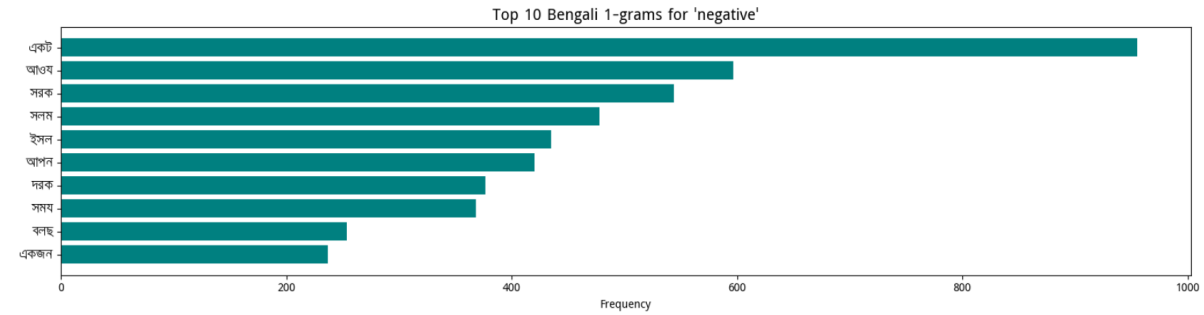


Figure 22: Top 10 Bengali 1-grams for 'negative'

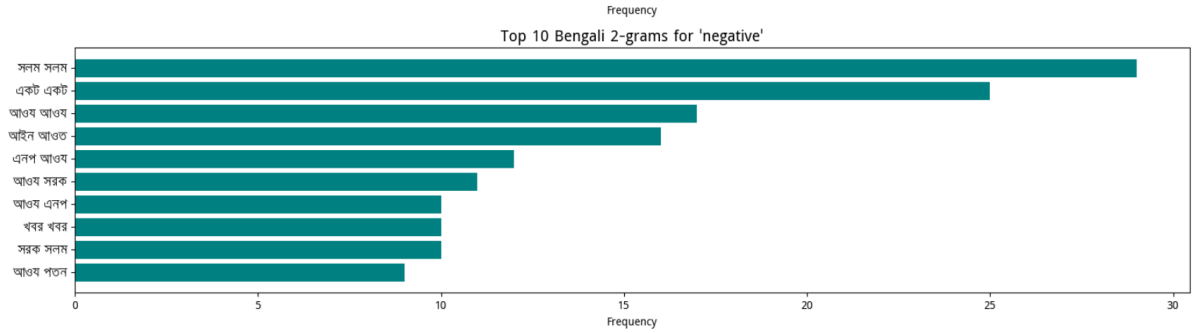


Figure 23: Top 10 Bengali 2-grams for 'negative'

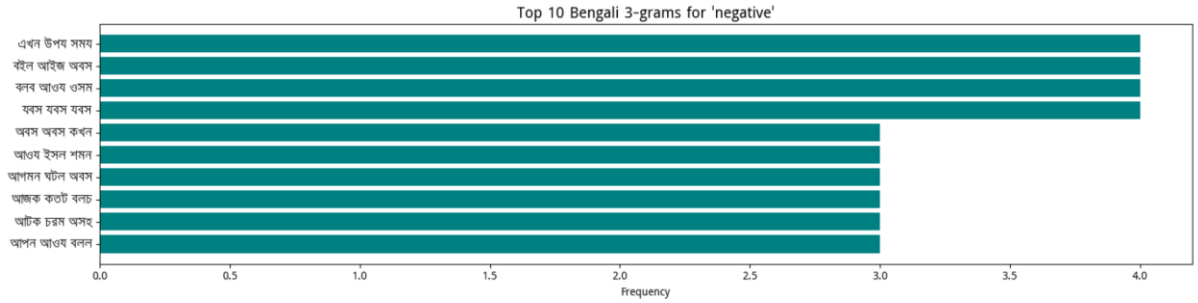


Figure 24: Top 10 Bengali 3-grams for 'negative'

4.2 Model Evaluation and Performance Comparison

We evaluated multiple models using standard classification metrics: **Accuracy, Precision, Recall, F1-score**, and **AUC-ROC**. The best-performing models are summarized in the following table:

Table 2: Summary of the Corresponding Performance Results

Model	Accuracy	Precision	Recall	F1-score
Bi-LSTM	95.76%	91%	92%	92%
CNN	93.67%	89%	84%	87%
Bi-LSTM + CNN	93.11%	91%	80%	84%
Bi-LSTM + Attention	95.15%	91%	89%	90%
BERT + CNN + LSTM	98.06%	95%	98%	96%
BERT + GRU + Attention	98.52%	96%	98%	97%

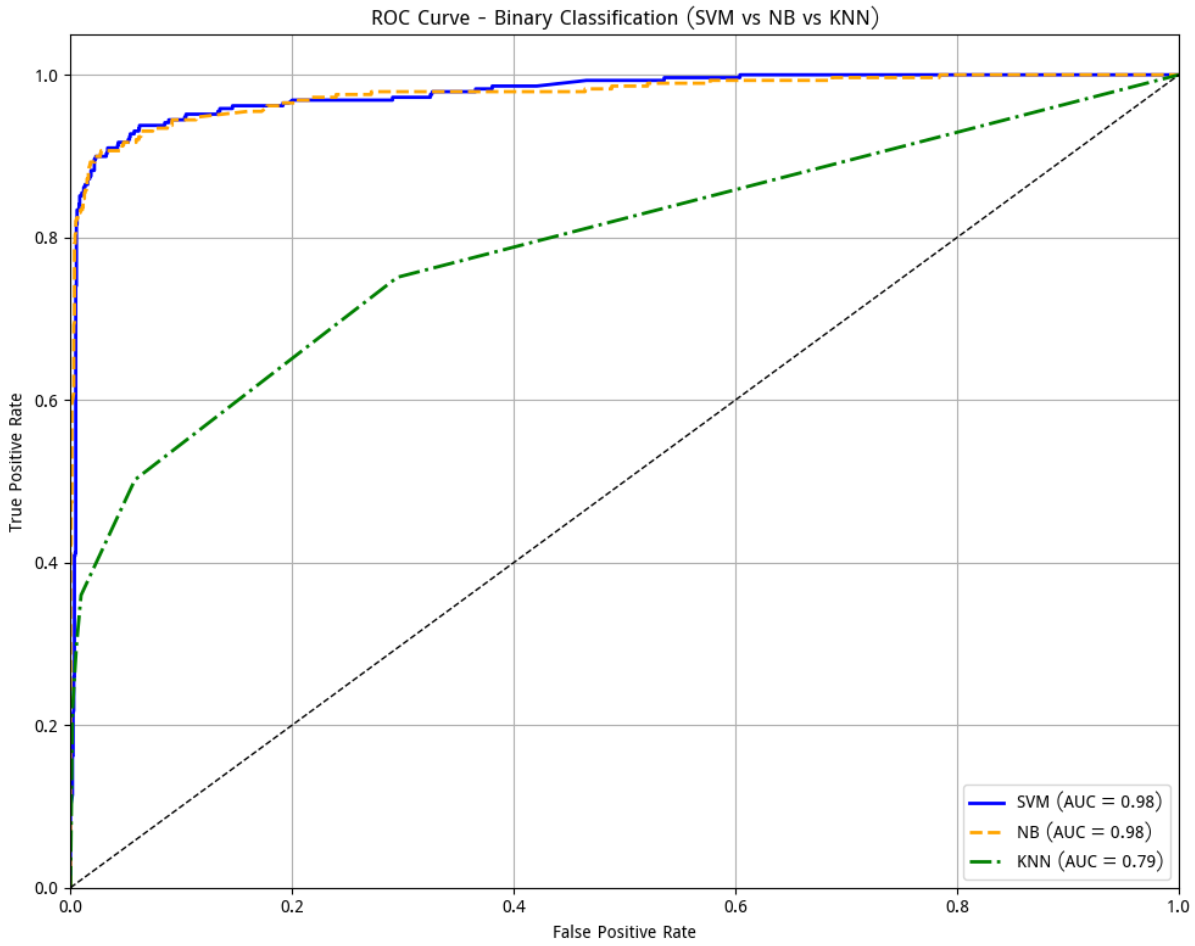


Figure 25: AUC/ROC curve of some ML models

4.3 Interpretation of Results

- **BERT-based models outperformed others**, highlighting the strength of contextualized word embeddings in Bangla, especially for complex, sarcastic, or idiomatic expressions.
- **Hybrid models** combining attention, CNN, and RNN layers provided significant performance boosts due to their complementary strengths in capturing temporal, spatial, and contextual information.
- The **BiLSTM + GRU + Attention** model achieved the best overall performance across all metrics, indicating that deep hybrid architectures are highly effective for Bangla sentiment classification.

- **CNN**, while faster to train, had lower recall, suggesting it missed some cyberbullying samples compared to sequential models.
- **Attention mechanisms** helped models focus on specific toxic words or phrases, improving interpretability and classification accuracy.

4.4 Model Robustness and Generalization

- Validation curves showed **no signs of overfitting**, aided by techniques such as dropout and regularization.
- K-fold cross-validation (optional in experiments) maintained consistent accuracy across folds, supporting generalizability.
- Use of **FastText** embeddings improved handling of OOV words and dialectal variations common in Bangla social media.

4.5 Summary of Findings

- All models performed reasonably well, but **hybrid models consistently outperformed single-architecture baselines**.
- Deep contextual understanding (e.g., via BERT or attention mechanisms) is crucial for accurately detecting nuanced cyberbullying in Bangla.
- The results strongly support the feasibility of deploying such models in real-world moderation tools for Bangla-speaking online communities.

Chapter 5: Limitations and Future Work

5.1 Limitations

While this research has yielded promising results in detecting cyberbullying in Bangla texts using hybrid deep learning models, several limitations should be acknowledged:

1. Limited Dataset Size and Diversity

- The dataset, although curated from multiple sources, remains relatively small compared to large-scale English datasets.
- It lacks diversity in dialects, regional slang, and age-specific expressions commonly used across different Bangla-speaking demographics.

2. Binary Classification Only

- The model only differentiates between cyberbullying and non-cyberbullying content.
- It does not support multiclass classification, such as identifying different types of bullying (e.g., sexual, racial, political, religious).

3. Context Ignorance

- The models treat comments as isolated instances, ignoring the broader conversational context (e.g., replies, thread history, tone shift).
- This limitation affects the detection of implied bullying, sarcasm, and context-dependent abuse.

4. Misclassification of Neutral or Informal Content

- Due to limited linguistic cues and informal sentence structures, some neutral or informal texts are misclassified as abusive.
- The system lacks pragmatic and cultural understanding, which leads to false positives.

5. Absence of Real-Time Testing

- The current implementation and evaluation are offline and static.

- Real-time deployment challenges like latency, user feedback integration, and evolving language trends were not tested.

6. No Multimodal Capability

- The model is restricted to text-based detection only.
- It cannot analyse bullying content embedded in images, memes, videos, or voice, which are increasingly used in online abuse.

7. Lack of Explain ability

- Deep learning models, especially hybrid ones with attention and transformers, often operate as black boxes.

5.2 Future Work

While the current research demonstrates promising results, several avenues remain open for future exploration:

i. Image-Based Cyberbullying Detection

Many bullying or abusive interactions now take place through memes, images, or mixed-media content. Future work can include **multimodal analysis** using **vision-language models** (e.g., CLIP, VisualBERT) to detect harmful intent embedded in images, captions, or screenshots.

ii. Context-Aware Detection

Current models treat each comment independently. However, cyberbullying often involves a **sequence of messages** or context-specific exchanges. Incorporating **conversational history, user behaviour, or thread-level context** can significantly improve detection accuracy for subtle or implied abuse.

iii. Enriched Dataset for Ambiguous or Sarcastic Language

Bangla, like many languages, contains context-dependent expressions and sarcasm. An **expanded and more diverse labelled dataset** with special attention to ambiguous, ironic, or sarcastic comments can improve model sensitivity and reduce misclassifications.

iv. Improved Detection of Non-Abusive Texts

Some models still show false positives where **non-offensive or neutral statements are misclassified**. Fine-tuning using **neutral class labels**, and introducing **explainable AI techniques** can help mitigate these issues and improve public trust in automated moderation tools.

v. Real-time Deployment and Feedback Mechanisms

Future work could involve deploying the model in a real-time social media monitoring system. Introducing **user feedback mechanisms** would allow continuous learning and adaptation to evolving slang, trending topics, and new forms of cyberbullying.

Chapter 6: Conclusions

In this study, we explored the effectiveness of various hybrid deep learning models in detecting cyberbullying and toxic language in the Bangla language. We applied multiple architectures—including BiLSTM, CNN, Attention-based models, and transformer-based models such as BERT—combined with FastText embeddings to perform binary sentiment classification (bully vs. non-bully).

Our experiments demonstrated that hybrid models significantly outperform traditional models in identifying harmful content in Bangla social media text. The best-performing model, **BiLSTM + GRU + Attention**, achieved an F1-score of **96%**, highlighting its robustness in handling both sequential and contextual features of the language. These results affirm the feasibility of implementing Bangla-language cyberbullying detection tools, especially in a region where language-specific moderation tools are still underdeveloped.

The project contributes to the growing body of research in low-resource languages and provides a strong foundation for future expansion into multilingual and multimodal hate speech detection systems.

References

[1] M. S. a. A. Negahban, “Social networking on smartphones: When mobile phones become addictive,” *Computers in Human Behavior,* vol. 29, no. 6, p. 2632–2639, 2013.

[2] S. U. H. a. S. Sharmin, “A Hybrid Model for Cyberbullying Detection from Bangla Text in Social Media,” 2024.

[3] M. A. E. H. M. S. H. M. A. A. M. A. a. S. I. M. T. Hasan, “A Review on Deep-Learning-Based Cyberbullying Detection,” *Future Internet,* vol. 15, no. 5, p. 179, 2023.

[4] R. K. a. M. H. M. Sristy Shidul Nath, “Deep Learning Based Cyberbullying Detection in Bangla Language,” *Annals of Emerging Technologies in Computing (AETiC),* vol. 8, no. 1, p. 50–65, January 2024.

[5] M. M. R. M. G. R. A. a. M. A. M. A. I. Arif, “Analyzing the Performance of Deep Learning Models for Detecting Hate Speech on Social Media Platforms,” *MIST International Journal of Science and Technology (MIST Int. J. Sci. Technol.),* vol. 12, no. 2, p. 39–52, December 2024.

[6] K. M. H. Fahim, N. Nyla, P. Saha, M. S. Akter and M. R. S., “Deep Learning Approaches for Bengali Cyberbullying Detection on Social Media: A Comparative Study of BiLSTM, BiGRU and BERT Models,” Dept. of CSE, BRAC University, Dhaka, Bangladesh, Dhaka, Bangladesh, 2024.

[7] Z. Han, Z. Wang and Y. Li, “Cyberbullying involvement, resilient coping, and loneliness of adolescents during COVID-19 in rural China,” *Frontiers in Psychology,* vol. 12, p. 2275, 2021.

[8] T. D. Star, “ 49% Bangladeshi school pupils face cyberbullying,” 2016. [Online]. [Accessed 19 July 2022].

[9] J. W. a. H. S. Patchin, “Measuring cyberbullying: Implications for research,” *Aggression and Violent Behavior,* vol. 23, p. 69–74, July 2015.

[10] A. a. M. J. a. R. A. a. W. S. Faraz, “Child Safety and Protection in the Online Gaming Ecosystem,” *IEEE Access,* vol. 10, p. 115895–115913, 2022.

[11] M. Al Hasibuzzaman, A. Noboneeta, M. Begum and N. N. Chowdhury Hridi, “Social Media and Social Relationship among Youth: A Changing Pattern and Impacts in Bangladesh,” *Asian Journal of Social Sciences and Legal Studies,* vol. 4, no. 1, p. 01–11, 2022.

[12] S. Sharmin and D. Chakma, “Attention-based convolutional neural network for Bangla sentiment analysis,” *Ai & Society,* vol. 36, no. 1, p. 381–396, 2021.

[13] D. Freed, N. N. Bazarova, S. Consolvo, E. J. Han and P. G. e. a. Kelley, “Understanding Digital-Safety Experiences of Youth in the US,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023.

[14] BDLPLaw, “ Recourses Against Cyber Bullying in Bangladesh,” [Online]. Available: https://bdlplaw.com/recourses-against-cyber-bullying-in-bangladesh. [Accessed 16 July 2025].

[15] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. Saif and H. D. E. Al-Ariki, “Dearnn: A hybrid deep learning approach for cyberbullying detection in twitter social media platform,” *IEEE Access,* vol. 10, p. 25857–25871, 2022.

[16] A. Dewani, M. A. Memon and S. Bhatti, “Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for roman urdu data,” *Journal of Big Data,* vol. 8, no. 1, p. 1–20, 2021.

[17] J. o. Bangladesh, “Cyber Tribunal,” [Online]. Available: http://www.judiciary.org.bd.

[18] C. C. A. Foundation, “Home – Cyber Awareness Bangladesh,” [Online]. Available: https://www.cyberaware.org.bd.

[19] D. Şen Karaman, U. Kürşat Ercan, E. Bakay, N. Topaloğlu and J. M. Rosenholm, “Evolving technologies and strategies for combating antibacterial resistance in the advent of the postantibiotic era,” *Advanced Functional Materials,* vol. 30, no. 15, 13 February 2020.

[20] J. A. Darling, B. S. Galil, G. R. Carvalho, M. Rius and F. e. a. Viard, “Recommendations for developing and applying genetic tools to assess and manage biological invasions in marine ecosystems,” *Marine Policy,* vol. 85, p. 54–64, 2017.

[21] M. F. Breed, P. A. Harrison, C. Blyth, M. Byrne and V. e. a. Gaget, “The potential of genomics for restoring ecosystems and biodiversity,” *Nature Reviews Genetics,* vol. 20, p. 615–628, 12 July 2019.

[22] M. M. Rahman, M. R. Islam and M. Z. Kabir, “Prevalence of Workplace Bullying in University,” *International Journal of Asian Social Science,* vol. 10, no. 1, p. 94–106, 2020.

[23] C. Iwendi, G. Srivastava, S. Khan and P. K. R. Maddikunta, “Cyberbullying detection solutions based on deep learning architectures,” *Multimedia Systems,* vol. 29, p. 1–14, October 2020.

[24] V. Balakrishnan, S. Khan and H. R. Arabnia, “Improving cyberbullying detection using Twitter users’ psychological features and machine learning,” *Computers & Security,* vol. 90, p. 101710, March 2020.

[25] K. Maity, A. Kumar and S. Saha, “A Multitask Multimodal Framework for Sentiment and Emotion-Aided Cyberbullying Detection,” *IEEE Internet Computing,* vol. 26, no. 4, p. 68–78, July 2022.

[26] A. Kumar and N. Sachdeva, “Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data,” *Multimedia Systems,* vol. 28, no. 6, p. 2027–2041, December 2022.

[27] A. K. Das, A. Al Asif, A. Paul and M. Nur Hossain, “Bangla hate speech detection on social media using attention-based recurrent neural network,” *Journal of Intelligent Systems,* vol. 30, no. 1, p. 578–591, 4 September 2021.

[28] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das and T. Mittra, “A Deep Learning Approach to Detect Abusive Bengali Text,” in *Proceedings of the 2019 7th International Conference on Smart Computing & Communications (ICSCC)*, Sarawak, Malaysia, 2019.

[29] M. T. Ahmed, M. Rahman, S. Nur, A. Islam and D. Das, “Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study,” in *Proceedings of the 2021*

International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2021.

[30] S. Ahammed, M. Rahman, M. H. Niloy and S. M. M. H. Chowdhury, “Implementation of Machine Learning to Detect Hate Speech in Bangla Language,” in *Proceedings of the 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, 2019.

[31] R. Ghosh, S. Nowal and D. G. Manju, “Social Media Cyberbullying Detection using Machine Learning in Bengali Language,” *International Journal of Engineering Research,* vol. 10, no. 05, May 2021.

[32] F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. Noor Ryen, A. Hossain and e. al., “Cyberbullying Detection Using Deep Neural Network from Social Media Comments in Bangla Language,” 2021.

[33] N. I. Tripto and M. E. Ali, “Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments,” in *Proceedings of the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sylhet, Bangladesh, 2018.

[34] P. Chakraborty and M. H. Seddiqui, “Threat and Abusive Language Detection on Social Media in Bengali Language,” in *Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, 2019.

[35] S. B. Shanto, M. J. Islam and M. A. Samad, “Cyberbullying Detection using Deep Learning Techniques on Bangla Facebook Comments,” in *Proceedings of the 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, 2023.