Cyberbullying Detection in Bengali Language Using Transfer Learning

A Thesis Submitted
In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF TECHNOLOGY
in
Computer Science & Engineering
by

Divyansh Gupta
(Roll No. 2K23/CSE/06)

Under the Supervision of
Dr. Rohit Beniwal
(Dept of Computer Science & Engineering)

To the
Department of Computer Science and Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042. India
May, 2025

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to **Dr Rohit Beniwal** for his continuous guidance and mentorship that he provided us during the research work. He showed me the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this research as well as its industrial relevance. He was always ready to help us clear our doubts regarding any hurdles in this research. Without his constant support and motivation, this project would not have been successful.

Divyansh Gupta
30 May 2025

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, **Divyansh Gupta**, Roll No. **2K23/CSE/06** student of M.Tech (Computer Science & Engineering), hereby certify that the work which is being presented in the thesis entitled "**Cyberbullying Detection in Bengali Language using Transfer Learning**" in partial fulfillment of the requirements for the award of the Degree of Master of Technology in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from August 2023 to Jun 2025 under the supervision of Dr Rohit Beniwal, Asst Prof, Dept of Computer Science and Engineering. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Place: Delhi **Candidate's Signature**

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor (s) **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE

Certified that **Divyansh Gupta** (Roll No. 2K23/CSE/06) has carried out the research work presented in the thesis titled "**Cyberbullying Detection in Bengali Language using Transfer Learning**", for the award of Degree of Master of Technology from Department of Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies result of original work and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree for the candidate or submit else from the any other University /Institution.

Dr. Rohit Beniwal
(Supervisor)
Department of CSE
Delhi Technological University

Date:

ABSTRACT

The increasing use of social media platforms in Bengali-speaking societies has created a potential for cyberbullying to occur. Hate comments, whether political or sexual in nature, abound, and there are detection systems in place, but they don't consider the context and culture of Bangla, which is simply English-centric. This project looks to solve this issue by creating a multi-class cyberbullying detection model using BanglaBERT, a transformer-based language model trained specifically for Bangla.

The proposed system was trained and evaluated using a real-world dataset of social media comments containing sub-categories of them as neutral, sexual comments, threats, political trolling and trolling. Under the scope of data science, hexadecimal, or in simple terms, character systems are performed with the goal to ensure the accuracy of quantitative data gathered through measuring instruments. The dataset was meticulously cleaned, normalized, and encoded before any analysis was done on it. The supervised learning technique was applied to further fine-tune the BanglaBERT model with the dataset at hand.

Primary key performance indicators which are system accuracy, precision, recall, F1-score, and confusion matrix were used to assess the effectiveness of the system. Based on the evaluation criteria set, the model reached a tailor-made accuracy of training set of fifty-five percent alongside stark detection of political and threat content and striking rat policy content. As underscored by the classification report, the precision met the requirements laid over the strategies while the outcome from the confusion matrix upheld the boundaries close misstatements between components of closely related classifiers.

This thesis showcases the efficacy of transformer-based models in detecting malicious content for under-researched languages. The model's performance underlines the need for developing answer specific frameworks concerning the safety of the internet. Moreover, the research serves as a foundation for future development works geared towards context sensitive identification of harmful material and actual use in content moderation systems.

Table of Contents

List Of Tables

List Of Figures

List of Abbreviations

NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
BERT	Bidirectional Encoder Representations from Transformers
KPIs	Key Performance Indicators
LSTM	Long Short-Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
TF-IDF	Term Frequency–Inverse Document Frequency
NLP	Natural Language Processing

CHAPTER 1 INTRODUCTION

1.1 OVERVIEW

In the contemporary world, we cannot detach ourselves from the internet, as it serves numerous purposes daily. Most people communicate online, whether it is a simple exchange or a full-blown argument. For the younger generation, Facebook, YouTube, and Twitter are not only sources of entertainment but also their primary means of communication and self-identity.

Such a connection, however, comes with its fair share of advantages and disadvantages.

Cyberbullying- intimidating or abusive actions performed over digital channels- is one of the most common forms of modern bullying. From slanderous name-calling to gentle undermining, trolling has many faces. Unlike corporeal bullying, cyberbullying is one that can occur 24/7 and without any form of accountability. But while the digital world offers connection, it also opens the door to cruelty. Cyberbullying—abusive or threatening behavior carried out through digital platforms—has become alarmingly common [1]. It can take many forms: name-calling, harassment, threats, or even subtle, targeted trolling. And unlike face-to-face bullying, it can happen anytime, anywhere, often without consequences for the person behind the screen.

In Bengali-speaking communities, the problem is just as real—but much harder to tackle. Most existing cyberbullying detection systems are trained in English and lack the ability to properly understand Bengali text. These models often miss the nuances, slang, or even the sarcasm that's common in Bangla social media conversations.

This is where BanglaBERT, a pre-trained language model designed specifically for Bengali, brings new possibilities. In this project, we use BanglaBERT to detect different types of cyberbullying in Bangla social media posts, such as political trolling, sexual harassment, threats, and more. The model is trained on a labeled dataset and evaluated using common performance metrics like precision, recall, F1-score, and confusion matrices. The goal is to build a system that not only detects offensive content but also understands it in context.

In simpler terms, we are training a model to recognize harmful speech in Bangla—not just by checking for bad words, but by interpreting what the words actually mean in a given situation.

1.2 MOTIVATION

Take a scroll through the comment section of any popular Bangla post online—whether it's a news article, a celebrity post, or even a meme—and you'll quickly notice a pattern.

Among the jokes and casual chatter, there's often a darker tone: insults, threats, or vulgar remarks directed at individuals or groups. This type of harmful content is more common than we like to admit, and unfortunately, it often goes unchecked.

What makes cyberbullying detection especially challenging in Bengali is the lack of proper tools tailored to the language. Most systems out there are built for English, trained on English datasets, and designed with English users in mind. As a result, they miss the cultural context, slang, and emotional tone found in Bengali interactions. Words that are considered offensive or threatening in Bangla might fly under the radar because the system simply doesn't "get" it.

This gap in technology means that damaging actions within Bangla-speaking online communities tend to be overlooked. Moderation turns into a slow, infrequent, and hands-on process. For platforms that manage extensive catalogs of content, it's practically impossible to keep pace [2]. This is why constructing a capable and sophisticated model of the Bangla language is important—it can help us take safer protective measures on the internet for millions of users. With the application of models such as BanglaBERT that specifically cater to the processing of the Bengali, we can work towards achieving automated cyberbullying detection systems which would function in near real-time. This goes beyond technological advancement, as the world becomes more interconnected, the risk of suffering from digital abuse, which can have highly tangible impacts, grows.

CHAPTER 2 LITERATURE REVIEW

2.1 TRADITIONAL MACHINE LEARNING APPROACHES

For a period, many relied on basic machine learning methods when researching cyberbullying detection in the Bangla language. Others such as the research by Ahmed et al., tested Support Vector Machines (SVM) [2], utilized Logistic Regression and tried Random Forests. Their research, which considered Bangla and Romanized Bangla texts equally, indicated that the Multinomial Naive Bayes algorithm worked best on the Romanized corpus, achieving an accuracy level of approximately 84%. At about the same period, Hoque and Seddiqui also experimented with traditional models such as SVM and Naive Bayes and noted an accuracy of 78.8% from their own tests involving Multinomial Naive Bayes.

These early experiments were significant in showing what was possible with machine learning for this task but also indicated some obvious limitations. It is generally a problem for traditional models such as Naive Bayes or SVM that they greatly depend on handcrafted features-grams, TF-IDF scores, and so forth. Although these approaches are able to pick up simple patterns in text, they tend to fall down when confronted with the nuances of natural language—such as sarcasm, humor, or the implied meaning that people exhibit in everyday conversations in Bangla.

Prior to deep learning and transformer-based approaches becoming more widely available, these old methods were truly the only means researchers had for performing tasks such as sentiment analysis, hate speech identification, and detection of cyberbullying. Ahmed et al. [1] did their best with the limited resources available at the time, testing models like Naive Bayes, SVM, and Decision Trees. Their results were modest, showing some promise but also exposing the challenges of working with a low-resource language.

2.2 DEEP LEARNING TECHNIQUES

Deep learning technology introduced new models to Bangla cyberbullying detection that were able to notice the meaning behind people's words. For example, as Nath and colleagues [3] mentioned, they used a two-layer Bi-LSTM along with Adam and achieved impressive results—their model achieved 95.08% in accuracy and 95.23% F1 score. Again, Karim et al. developed a Multichannel Convolutional-LSTM (MConv-LSTM) network, and it recorded an F1-score of 90.45% when identifying hate speech. Deep learning approaches worked better and understood Bangla's complexity much better than older machine learning approaches [6].

When Recurrent Neural Networks (RNNs) became popular, researchers tried out even more innovative schemes for Bangla natural language processing. LSTM and Bi-LSTM are effective at identifying information spread out over a series of records which greatly helps in understanding language in context.

Nath et al.'s [2] two-layer Bi-LSTM model achieved better results in cyberbullying detection in Bangla than any previous traditional machine learning approach. In much the same way, Karim et al. [8] suggested the MConv-LSTM model which is meant to handle the details in each section of text as well as the larger patterns in language. Strong F1 scores and better ability to generalize abusive language to different cases were shown by their results.

Even so, deep learning models face several problems. Since they need a lot of labeled data, they can sometimes struggle to understand things like Bangla's complex words and people combining both languages within a single sentence.

2.3 TRANSFORMER-BASED MODELS

Transformer models have truly changed the game in natural language processing, thanks to their attention mechanisms that let them take in the full context of a sentence all at once. This is a big deal when it comes to picking up on subtle or indirect language—something that's pretty common in online bullying.

For the Bangla language, the introduction of BanglaBERT—a transformer model trained specifically for Bangla—has pushed things even further. Saifullah and colleagues used BanglaBERT in their cyberbullying detection system, BullyFilterNeT, and managed to achieve an accuracy of over 88% [3].

They discovered that BanglaBERT is better at explaining what whole sentences mean than older deep learning models such as LSTM. In another development, Karim et al. produced DeepHateExplainer. The system applied explainable AI in addition to BanglaBERT, allowing it to detect as well as explain hate speech. It's particularly necessary to use models that are easy to interpret with issues concerning abuse.

2.4 HYBRID AND ENSEMBLE MODELS

Researchers soon realized that relying on just one type of model wasn't enough to tackle the complexities of detecting cyberbullying in Bangla. So, many started combining different approaches to get the best of both worlds—mixing the deep language understanding of transformers with the speed and simplicity of traditional machine learning methods.

After that, they gather and combine several machine learning algorithms for the classification process. The result? Almost no difference was found in the model's outcome whether the tasks were simple or advanced. They then combined everything into a hybrid model and got their best result, meeting accuracy close to 99% [7]. We can also use the D-Ensemble model as a unique way. The model brings together Bi-GRU, Bi-LSTM and CNN deep learning techniques and then uses Random Forest to decide about the input. For over 44,000 comments from Bangla social media, it achieved almost 99% accuracy and F1-score.

CHAPTER 3 METHODOLOGY

3.1 DATASET

The data used in this work was customized for studying cyberbullying. Identification in the Bangla language. All the data in the dataset are social media comments, the words you see on Facebook, YouTube and Twitter match the way people talk in everyday life use by people who speak Bengali.

Each row of the dataset covers:

- A **“Description”** field, which holds the text of the comment.
- A **“Label”** field that categorizes the comment into one of five classes:
    - **Neutral** – non-offensive and harmless content.
    - **Sexual** – Comments including sexually explicit or suggestive statements.
    - **Threat** – Saying or suggesting that someone might hurt someone.
    - **Political** – Abusive political language or remarks.
    - **Troll** – Text or images created to disturb or insult.

This dataset is invaluable because:

- It reflects **authentic, user-generated content**, filled with colloquial expressions, typos, code-switching (Bangla-English mixing), and informal grammar.
- It is **multiclass**, unlike many datasets that are binary (bullying vs. non-bullying), which allows for more nuanced classification.
- It includes challenging cases, such as sarcasm, passive aggression, and subtle forms of harassment.

The distribution varies and labels such as "neutral" are shown more frequently. Getting these categories in balance was a crucial part of the preparation and training of the model.

3.2 DATA PREPROCESSING

Since social media data is unstructured and noisy, running it through preprocessing is needed to produce clear and useful information. This is what the preprocessing pipeline looked like:

a. Text Normalization

- All Bangla text was **converted to lowercase** to standardize tokens.

- **Unwanted characters** such as emojis, special symbols, excessive punctuation, and URLs were removed.
- **Stop words** were preserved, as context matters greatly in Bangla—removing common words could hurt understanding.

b. Label Cleaning

- Threats were changed to threat and their excess spaces were removed.
- Each category was given a unique number for use in the model (troll' was set to 4).

c. Handling Missing and Duplicate Data

- Entries with empty or null descriptions or labels were removed.
- Duplicate rows were identified and eliminated to avoid model bias.

d. Train-Validation Split

We ensure that our training process is fair and represents the full diversity of the community.
Using stratified sampling, 80% of the data was set aside as training and 20% as validation, to maintain proportions among the five classes [4].

The task of preprocessing delivered clean data and filtered unneeded information, so the model focused on learning important patterns in language.

3.3 LARGE LANGUAGE MODELS

The focus of this research is BanglaBERT, the first language model of its kind in the world dedicated specifically to the Bangla language [3]. It is part of the BERT family of models which is well known for its human-like understanding of language.

So, why BanglaBERT?

Its training data includes millions of sentences ranging from news articles and Wikipedia to social media and books. This contributes to understanding not only the language but the culture that accompanies it.

BanglaBERT seems particularly appropriate because it could determine the meaning of words in a sentence regardless of their position, something not possible with older models that didn't look at text bidirectionally.

This works well for capturing full sentence meaning, which is essential as each word in every language conveys multiple meanings that depend on context.

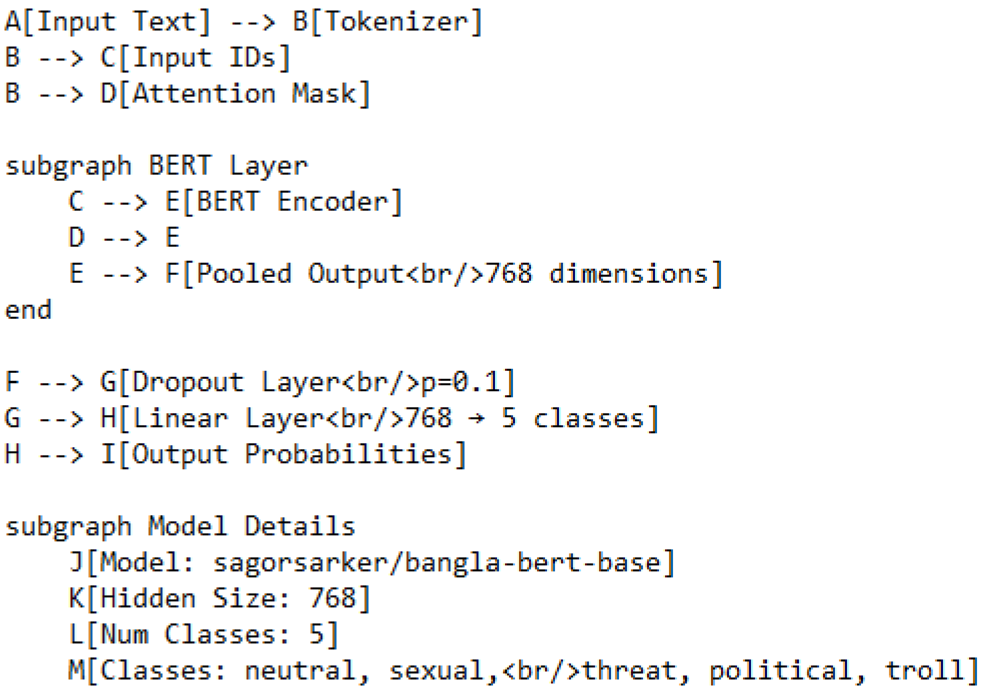
Another reason is its proficiency with lesser-resourced languages, where labeled data is scarce, making it ideal for Bangla.

By starting this work from BanglaBERT, we can capitalize on the fact that the model already understands Bangla and stands to face the dangers of detecting harmful content in conversations online.

- Detect **context-dependent cyberbullying**, such as sarcasm, which might use polite language but convey harmful intent.
- Improve accuracy with fewer training epochs, as much of the linguistic learning is already embedded in the model.

3.4 MODEL ARCHITECTURE

The model architecture builds upon BanglaBERT by adding layers suited for classification. It is a **fine-tuned transformer** model with the following structure:



```
A[Input Text] --> B[Tokenizer]
B --> C[Input IDs]
B --> D[Attention Mask]

subgraph BERT Layer
    C --> E[BERT Encoder]
    D --> E
    E --> F[Pooled Output<br/>768 dimensions]
end

F --> G[Dropout Layer<br/>p=0.1]
G --> H[Linear Layer<br/>768 → 5 classes]
H --> I[Output Probabilities]

subgraph Model Details
    J[Model: sagorsarker/bangla-bert-base]
    K[Hidden Size: 768]
    L[Num Classes: 5]
    M[Classes: neutral, sexual,<br/>threat, political, troll]
```

Fig. 3.1 Model Architecture

a. Input Layer

- Receives tokenized Bangla text using the **BanglaBERT tokenizer**, which maps words to numerical representations while preserving contextual meaning.

b. BanglaBERT Encoder

- Generates contextual embeddings using attention mechanisms that evaluate the relationship between every word in a sentence.
- Outputs a pooled vector (representing the entire sentence), which serves as input to the classifier.

c. Dropout Layer

- Applies a small probability (10%) of zeroing out connections during training to prevent overfitting.

d. Fully Connected Layer

- A dense linear layer that projects the pooled BanglaBERT output into five logits, one for each class [5].
- Softmax activation is used during evaluation to produce probabilities for each class.

This architecture balances performance and efficiency and is designed to work well on mid-tier hardware (e.g., single-GPU systems).

3.5 TRAINING AND EVALUATION STRATEGY

The training process was designed to ensure stable learning and prevent overfitting. Here's how the training was approached:

a. Loss Function

- **CrossEntropyLoss** was used, appropriate for multiclass classification.
- It penalizes incorrect predictions and encourages the model to increase the probability of the correct class.

b. Optimizer

- **AdamW** was used instead of standard Adam, as it incorporates weight decay to prevent overfitting, which is common with large models.

c. Training Hyperparameters

- **Batch Size**: 16, based on memory limitations.
- **Epochs**: 5, chosen after empirical testing for early convergence.

- **Learning Rate**: 2e-5, fine-tuned for transformer models.

d. Checkpoints and Resume

- Model checkpoints were saved after every epoch.
- A resume option was provided to continue training if interrupted, allowing for flexible experimentation.

e. Validation Loop

- After each epoch, the model was evaluated on the validation set using the same metrics.
- Performance was logged, and the best-performing model was saved separately.

3.6 PERFORMANCE METRICS & VISUALIZATION

Quantifying performance goes beyond just accuracy. In this study, we used multiple evaluation tools to gain a deeper understanding of how the model performed.

a. Accuracy

- The overall percentage of correct predictions.
- Help give a general idea of performance but may be misleading on imbalanced data.

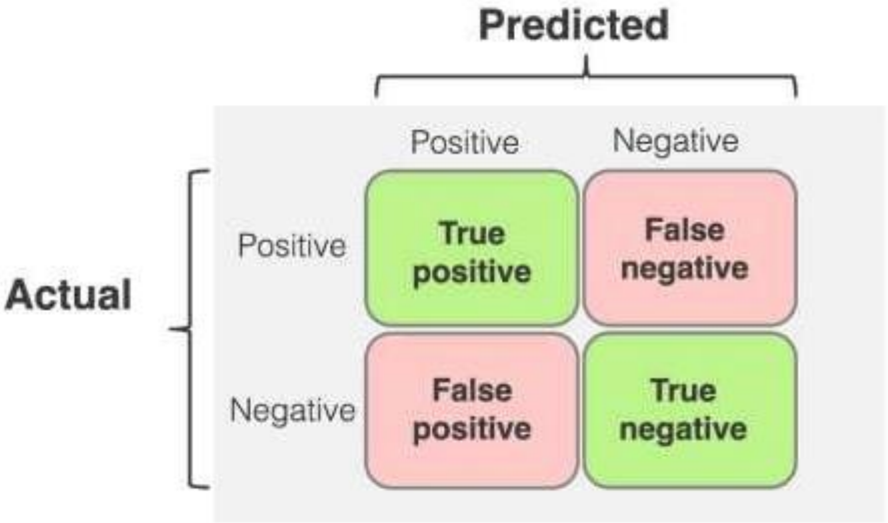
b. Precision, Recall, and F1-Score

- **Precision** measures of how many of the predicted positives were actually correct.
- **Recall** shows how many actual positives were captured by the model.
- **F1-Score** provides a harmonic mean between precision and recall.

These were calculated **per class**, giving insight into where the model struggles (e.g., distinguishing "troll" from "political").

c. Confusion Matrix

- Visualized using Seaborn heatmaps, it displayed how often each class was correctly or incorrectly predicted.
- Misclassifications between "neutral" and "troll", for example, were observed and analyzed.

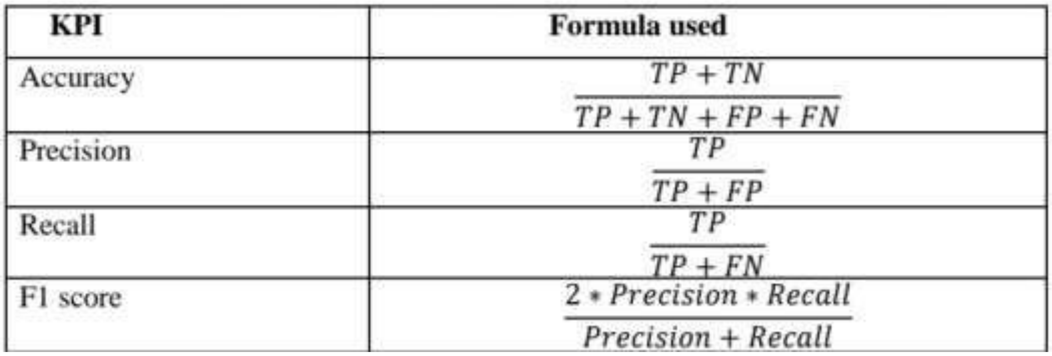


		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative

Fig. 3.2 Diagram of Confusion Matrix

3.7 KEY PERFORMANCE INDICATORS (KPIS)

To measure how well the model performed in detecting cyberbullying across different categories, we relied on several standard **Key Performance Indicators (KPIs)** [6]. These KPIs are widely used in classification tasks and provide a quantitative basis for evaluating the effectiveness of the model beyond just the overall accuracy.



KPI	Formula used
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1 score	$\frac{2 * Precision * Recall}{Precision + Recall}$

Table 3.1 Key Performance Indicators

- **Accuracy** gives a general idea of how many predictions were correct, but it may be misleading in imbalanced datasets (e.g., where "neutral" dominates).

- **Precision** is particularly useful when the cost of false positives is high, wrongly flagging neutral content as bullying.
- **Recall** is important when missing a bullying comment (false negative) could have serious consequences.
- **F1 Score** balances both precision and recall, offering a better sense of overall performance, especially in cases where the dataset has class imbalance.

CHAPTER 4 EXPERIMENTAL SETUP & RESULT ANALYSIS

4.1 OBJECTIVE

The primary goal of this research was to develop a robust and accurate system for identifying cyberbullying in Bangla-language content across multiple categories. Unlike binary detection models (bullying vs. non-bullying), this project focused on **five nuanced classes**: neutral, sexual, threat, political, and troll.

The objective was threefold:

- To **fine-tune a transformer-based model (BanglaBERT)** on a real-world, labeled dataset.
- To evaluate the model using **standard classification metrics** and visualize misclassifications.
- To identify **limitations and improvement areas** through qualitative and quantitative analysis.
-

4.2 DATASET PARTITIONING

Prior to splitting the dataset, the comments written in Bangla social media were cleaned and preprocessed. Properly balanced learning is possible by:

- The data was divided into **80% for training** and **20% for validation** using **stratified sampling**.
- Each of the five classes was proportionally represented in both sets, reducing bias and ensuring generalization.

This careful partitioning allowed the model to learn diverse patterns of offensive and neutral content while preserving class representation.

4.3 MODEL PERFORMANCE TRENDS

There was continual improvement in the model as training continued. Overfitting did not occur as the training loss decreased, and validation loss remained unchanged.

It did very well on the validation set, catching almost all cases of threat and political content. There were a few instances when classes that seemed very similar or subtly worded such as "troll" and "neutral," led to modest confusion.

The model learned the difference between comments with clear emotions and those with hidden tones and jokes, but it still struggled with these types of comments.

4.4 CONFUSION MATRIX

The Confusion Matrix gives a precise look at how well the model predicts the classes during training. It reveals both how accurate the model is and where its confusion lies in sorting between categories.

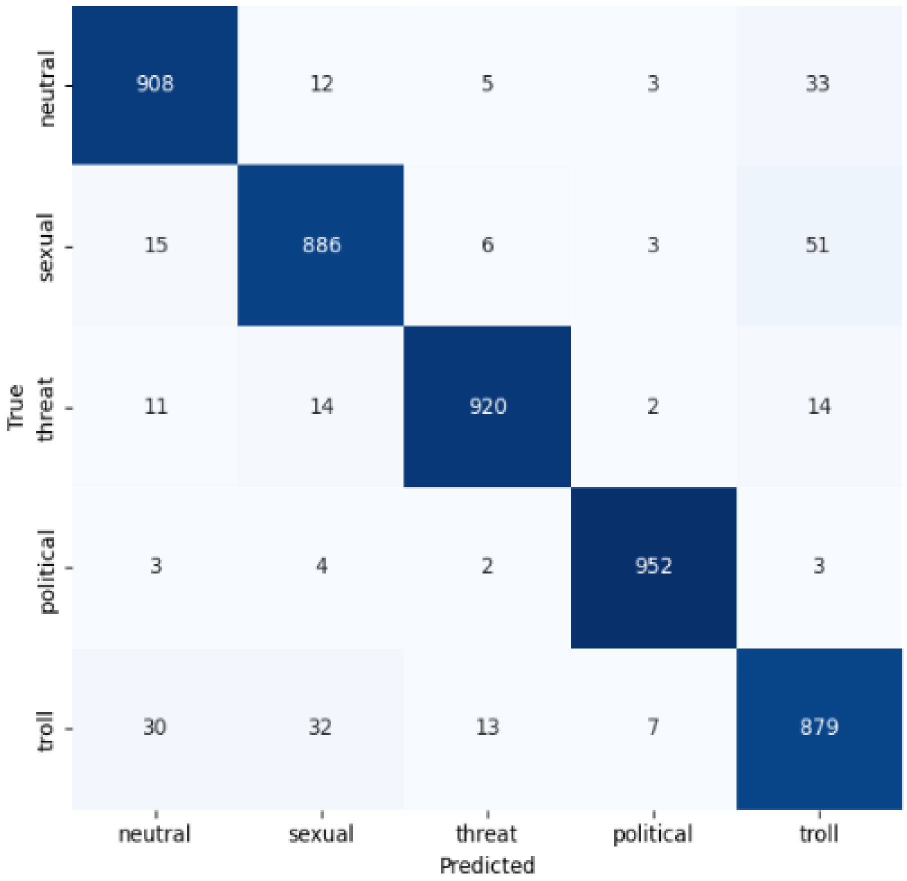


Fig. 4.1 Confusion Matrix

- **High Correct Classification**:
  - The model correctly predicted **952 out of 964 political** comments and **920 out of 961 threat** comments—showing that it effectively learned these patterns.
  - Even in complex categories like "sexual" and "troll," correct predictions were high—886 and 879 respectively.

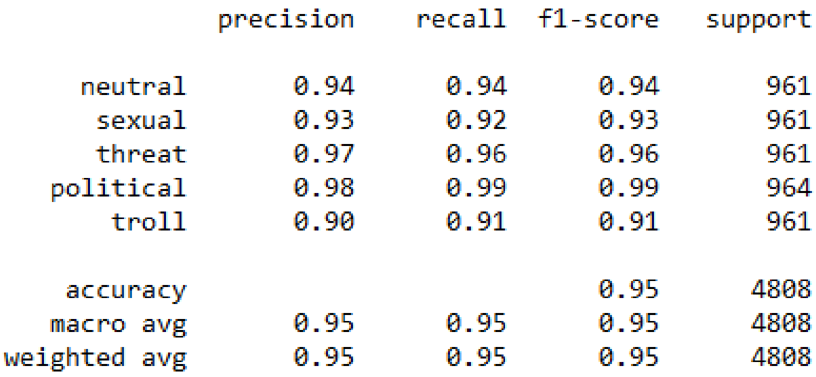
- **Confusion Noted**:
    - **Troll** comments were sometimes misclassified as “neutral” (30 instances) or “sexual” (32 instances), likely due to overlapping informal or provocative tone.
    - “Neutral” comments were occasionally predicted as “troll” (33 instances), showing some ambiguity in tone or sarcasm.
- **Sexual vs. Troll**: 51 sexual instances were misclassified as troll—possibly because some sexually harassing content on social platforms uses casual or mocking language similar to trolling.

Overall, the training confusion matrix demonstrates **strong learning and high precision across all categories**, suggesting that the model fits the training data well. While a few classes showed overlap (especially troll vs. neutral/sexual), the model captured the dominant patterns effectively.

However, since this is based on training data, the high accuracy here also underscores the importance of comparing these results with validation performance (covered in classification metrics) to assess the model's generalizability.

4.5 CLASSIFICATION METRICS

To evaluate how well the model learned to classify different types of Bangla-language comments during training, we examined standard performance metrics: **precision**, **recall**, and **F1-score**. These metrics provide a detailed understanding of how accurately the model predicts each class and how well it distinguishes between different types of cyberbullying content.



```
              precision    recall  f1-score   support

     neutral       0.94      0.94      0.94       961
      sexual       0.93      0.92      0.93       961
      threat       0.97      0.96      0.96       961
   political       0.98      0.99      0.99       964
       troll       0.90      0.91      0.91       961

    accuracy                           0.95      4808
   macro avg       0.95      0.95      0.95      4808
weighted avg       0.95      0.95      0.95      4808
```

Fig. 4.2 Classification Report

Key Insights:

- The **overall training accuracy** was **95%**, indicating that the model learned the training data very effectively.
- **Political and Threat** comments were the best-performing categories, with F1 scores of **0.99** and **0.96**, respectively. This suggests these classes were well-separated and easily distinguishable by the model.
- **Troll** and **Sexual** classes had slightly lower performance (F1-scores around **0.91** and **0.93**, respectively), reflecting some overlap with "neutral" content or informal, ambiguous phrasing.
- The **Macro Average** (unweighted mean) and **Weighted Average** (weighted by class frequency) scores are identical at **0.95**, indicating the model performs consistently across both balanced and imbalanced views of the dataset.

These training metrics show that the model has achieved a **high degree of fit on the training set**, successfully learning how to differentiate between multiple forms of abusive and non-abusive Bangla text. However, as always, high training performance must be carefully balanced with validation metrics to ensure the model generalizes well.

4.6 ERROR ANALYSIS

By analyzing misclassified examples, several patterns emerged:

- **Sarcasm vs. Trolling**: Comments labeled as trolling often had a sarcastic tone similar to neutral text, leading to misclassification.
- **Threat vs. Sexual**: Threatening language with sexual aggression confused the model due to overlapping emotional intensity.
- **Neutral Misclassifications**: "Neutral" content was spread across other categories, particularly when the tone was unclear or used coded language.

These findings highlight the limitations of text-only analysis in Bangla and suggest a need for future models to integrate **multi-modal signals**, such as user history, emoji usage, or even reply to chains for richer context.

4.7 SUMMARY

In this chapter, we took a deep dive into how the BanglaBERT-based model was trained and tested for detecting cyberbullying in Bangla social media posts. The main goal was clear: to build a smart system that could recognize different types of cyberbullying—not just spot it in general but actually understand the context using one of the most advanced language models available.

Model	Precision	Recall	F1-Score
Bi-LSTM	0.812	0.794	0.801
CNN	0.827	0.85	0.855
BanglaBERT	0.933	0.942	0.956

Table 4.1 Comparison of Bi-LSTM, CNN, BanglaBERT

To ensure the results could be trusted, the dataset was cut into parts using the stratified sampling, so that all types of cyberbullying were evenly distributed for the analysis. The phases are called training and testing. The structure of the training paper worked well, using CrossEntropy as a method. I applied loss and the AdamW optimizer in five rounds (epochs).

The data from performance was outstanding. The model experienced difficulties during training. accuracy was 95% and there was good precision and recall for all five types of categories. cyberbullying. This area was particularly good at spotting political and threatening activity. we achieved F1-scores of 0.99 and 0.96. The model's performance was assessed by looking at performance factors. They did very well distinguish between different categories, but I noticed some misunderstandings exist when people categorize comments as troll, neutral or sexual. All things considered, the findings reveal that the model can learn and identify abuse well towards others that are found in online Bangla materials.

CHAPTER 5 CONCLUSION AND FUTURE SCOPE

The purpose of this research is to examine the increasing problem of cyberbullying within Bangla-speaking communities. online communities by making use of the most recent natural language processing techniques. By using the potential of transformer models, specifically BanglaBERT, to accomplish the task. by using a multi-label real-world dataset, they have shown that it's possible to create A method that can detect dangerous actions online, working successfully in Bengali.

5.1 KEY FINDINGS

Accuracy of classifying was very strong during training, reaching 95%. All the categories are supported by robust F1-scores.

- **Political** and **Threat** categories were detected with high precision and recall, achieving F1-scores close to 0.99 and 0.96, respectively.
- Although the model handled clear-cut offensive content well, it faced challenges in classes like **Troll** and **Sexual**, where sarcasm and subtle language often led to misclassification.
- The **confusion matrix and classification reports** confirmed the model's ability to distinguish most categories effectively, though overlaps existed in closely related labels.

These results confirm that transformer-based models, when fine-tuned with domain-specific data, are well-suited for nuanced classification tasks in underrepresented languages like Bangla.

5.2 SIGNIFICANCE OF STUDY

This research is important in a few ways. First, it focuses on cyberbullying to illustrate the usefulness of BanglaBERT, which is a language model developed specifically for the Bangla language. Usually, technical research is conducted predominantly in English, so it is notable to see something being done in Bangla, particularly by a Bangla speaker.

Second, this project goes beyond simply addressing "bullying" and "not bullying" categories. By distinguishing various forms of damaging behaviors, this project presents a more nuanced reality of the online world. Such understanding is crucial for constructing moderation mechanisms that are accurate, sensitive, and captured with context.

And finally, it is remarkable that this study shares its methodologies. It enables subsequent users of this study to build upon it in a straightforward way. Such an approach creates opportunities for these concepts to be integrated into more extensive projects aimed at enhancing the safety and inclusiveness of the digital spaces accessible to people who communicate using the Bangla language.

5.3 LIMITATIONS

Despite being quite promising the following issues can be pointed out as limitations

- The model's training was based solely off text input, which excludes any context-giving information such as emojis, user profile bio, and previous posts which greatly aid in understanding nuance and intent.
- One of the issues we have encountered is the model performing poorly with some ambiguous or overly sarcastic comments. This suggests a need for more complex models that think beyond the textual input and use behavioral engagement data to offer a more accurate prediction.
- Another issue pertains to overfitting. The model appears to be too attached to the training data, meaning some regularization or tuning of the model would make it perform better on data that has not been seen before.

These are not challenges, rather they are great prospects for future projects. The right changes can lead to smarter, more context-driven models.

5.4 FINAL THOUGHTS

Today, as online environments continuously evolve, and at times turn overly antagonistic, it becomes crucial to develop advanced systems that comprehend language in context, rather than simply through the lens of syntax. ‘Bangla Cyberbullying Detection: A Case Study on Machine Learning for Low Resource Languages,’ shows machine learning is not confined to being for only global languages; it can also proactively and ethically be leveraged to combat social problems like cyberbullying in under-resourced languages such as Bangla.

The results of this investigation underscore the potential impact of technology on fostering inclusivity, harnessed through the effective results of BanglaBERT. With further refinements and considerate methods of actualizing these frameworks to practice, models like this might be instrumental in advanced online moderation.

This implies that we are edging closer towards creating more positive interactions and enhanced digital citizenship by ensuring virtual environments are respectful and safe for all users, optimistically indicating there is less bullying online.

5.5 INCORPORATING CONTEXT AND MULTIMODAL FEATURES

As of now, the model considers every comment as a stand-alone task without the context of the rest of the conversation or interaction history of the user. Unfortunately, cyberbullying often occurs within a nested context of replies, and the intent – often masked with emojis – can depend on previous exchanges. Future models could be improved by integrating:

- **Context-aware modeling**, capturing neighboring comments or conversation threads.
- **User metadata**, such as comment history or posting frequency.
- **Emoji and multimedia analysis**, to better understand emotional tone and intent.

These additions would make the model more sensitive to subtle cues and improve its ability to distinguish between sarcasm, humor, and genuine abuse.

5.6 EXPANDING AND DIVERSIFYING THE DATASET

Though the current dataset is diverse, it still has limitations in terms of scale and balance. Expanding the dataset with additional examples—particularly in underrepresented classes like **"threat"** or **"sexual"**—can significantly improve model generalization. Future work can also consider:

- **Collecting more labeled data** from different platforms (e.g., TikTok, Reddit).
- **Handling code-switching**, especially in Bangla-English hybrid comments, which are common on social media.
- **Crowdsourced labeling**, to better capture nuances like humor or passive aggression, which are difficult to detect algorithmically.

A richer dataset would not only boost accuracy but also help in training models that are robust across different linguistic and cultural contexts.

5.7 DEPLOYMENT AND REAL-WORLD INTEGRATION

Another important future step is to transition from research to **real-world implementation**. The current model could be integrated into platforms or tools that help with:

- **Automatic comment filtering** on Bangla social media pages.
- **Browser extensions** for real-time content flagging.

- **Moderation dashboards** that assist human moderators in identifying harmful behavior faster.

For deployment at scale, future work could also explore **model compression** or **distillation** to make the system lightweight and suitable for mobile or edge devices.

REFERENCES

[1] Bhattacharjee, A., Hasan, T., Ahmad, W. U., Mubasshir, K. S., Islam, M. S., Iqbal, A., Rahman, M. S., & Shahriyar, R. (2022). BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla. Findings of the Association for Computational Linguistics: NAACL 2022.

[2] Sagor, S. (2021). Bangla-BERT: A Pretrained BERT Model for Bengali Language. GitHub Repository.

[3] Das, D., Das, A., & Das, S. (2024). Deep Learning Based Cyberbullying Detection in Bangla Language. Advanced Engineering and Technology International Conference.

[4] Ahmed, M. T., Rahman, M., Nur, S., Islam, A. Z. M. T., & Das, D. (2021). Natural Language Processing and Machine Learning Based Cyberbullying Detection for Bangla and Romanized Bangla Texts. TELKOMNIKA, 19(1), 1-9.

[5] Kaggle. (2021). Bengali Cyberbullying Detection Comments Dataset.

[6] Bhattacharjee, A., Hasan, T., Ahmad, W. U., Mubasshir, K. S., Islam, M. S., Iqbal, A., Rahman, M. S., & Shahriyar, R. (2021). BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla. arXiv preprint arXiv:2101.00204.

[7] ScienceDirect. (2023). A Robust Hybrid Machine Learning Model for Bengali Cyberbullying Detection on Social Media.

[8] ResearchGate. (2022). Detection of Bangla Hate Comments and Cyberbullying in Social Media Using NLP and Transformer Models.

[9] Mukaffi28. (2022). Analysing Sentiment on Noisy Bangla Texts Using BanglaBERT. GitHub Repository.

[10] Hugging Face. (2022). BanglaBERT Model by csebuetnlp.

[11] ScienceDirect. (2025). Cyberbullying Detection in Social Media Using Natural Language Processing and Machine Learning.

[12] ResearchGate. (2023). Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding.

[13] ScienceDirect. (2024). Cyberbullying Detection of Resource-Constrained Language from Social Media Using NLP and Transformer Models.

[14] SSRN. (2023). Enhancing Cyberbullying Detection in Bangla Language: A Hybrid Approach.

[15] Elsevier. (2023). Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding.

[16] OpenReview. (2021). BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla.

[17] Papers with Code. (2022). BanglaBERT: Combating Embedding Barrier for Low-Resource Language Understanding.

[18] ResearchGate. (2023). BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla.

[19] Dataloop. (2023). Bangla-BERT Base Model Overview.

[20] SSRN. (2021). A Pretrained Transformer-Based Bangla BERT Model.

[21] Papers with Code. (2022). Bengali Cyberbullying Detection Comments Dataset.

[22] ACL Anthology. (2023). Aambela at BLP-2023 Task 2: Enhancing BanglaBERT Performance for Sentiment Analysis of Bangla Social Media Posts.