EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems
Research Article EAI.EU

Cyberbullying Text Identification: A Deep Learning and Transformer-based Language Modeling Approach

Khalid Saifullah[1,*], Muhammad Ibrahim Khan[1], Suhaima Jamal[2], Iqbal H. Sarker[3,*]

[1]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology; Chittagong-4349, Bangladesh

[2]Dept. of Information Technology, Georgia Southern University, Statesboro GA, USA

[3]Centre for Securing Digital Futures, School of Science, Edith Cowan University, Perth, WA-6027, Australia

Abstract

In the contemporary digital age, social media platforms like Facebook, Twitter, and YouTube serve as vital channels for individuals to express ideas and connect with others. Despite fostering increased connectivity, these platforms have inadvertently given rise to negative behaviors, particularly cyberbullying. While extensive research has been conducted on high-resource languages such as English, there is a notable scarcity of resources for low-resource languages like Bengali, Arabic, Tamil, etc., particularly in terms of language modeling. This study addresses this gap by developing a cyberbullying text identification system called BullyFilterNeT tailored for social media texts, considering Bengali as a test case. The **intelligent BullyFilterNeT** system devised overcomes Out-of-Vocabulary (OOV) challenges associated with non-contextual embeddings and addresses the limitations of context-aware feature representations. To facilitate a comprehensive understanding, three non-contextual embedding models GloVe, FastText, and Word2Vec are developed for feature extraction in Bengali. These embedding models are utilized in the classification models, employing three statistical models (SVM, SGD, Libsvm), and four deep learning models (CNN, VDCNN, LSTM, GRU). Additionally, the study employs six transformer-based language models: mBERT, bELECTRA, IndicBERT, XML-RoBERTa, DistilBERT, and BanglaBERT, respectively to overcome the limitations of earlier models. Remarkably, BanglaBERT-based BullyFilterNeT achieves the highest accuracy of 88.04% in our test set, underscoring its effectiveness in cyberbullying text identification in the Bengali language.

Received on 28 December 2023; accepted on 19 February 2024; published on 22 February 2024

Keywords: Cyberbullying; large language modeling; deep learning; transformers models; natural language processing (NLP); fine tuning; OOV; harmful messages

doi:10.4108/eetinis.v11i1.4703

1. Introduction

Cyberbullying text identification refers to the process of detecting and recognizing instances of cyberbullying in written digital communications, such as text messages, emails, social media posts, or other online interactions. The rapid growth of media platforms like Facebook, Twitter, and YouTube has transformed communication, allowing individuals to express opinions on various topics. However, this has also led to the spread of offensive and hateful content. Cyberbullying, a significant issue, can cause psychological distress and undermine respectful conversations. According to research, this kind of conduct happened on Facebook and Twitter quite a bit. Among Bangladesh's 80.83 million Internet users [27], over 90% use Facebook regularly, and the bulk of these users are young, delicate, and in desperate need of protection. Lately, numerous studies have focused on high-resource languages such as English.

*Email: mdkhalidsaifullah@hotmail.com, m.sarker@ecu.edu.au

This emphasis is attributed to the existence of adequately annotated cyberbullying corpora, pre-trained models for text-to-feature extraction, pre-trained cyberbullying identification models, and a suite of finely tuned hyperparameters [9].

Nevertheless, Bengali stands as the seventh most widely spoken language globally, with approximately 245 million people in Bangladesh and two states of India conversing in Bengali [16]. The growing popularity of Bangla on social media is attributed to the widespread acceptance of the Unicode system and the increasing use of the Internet [17]. Consequently, a substantial volume of Bengali bullying texts has proliferated across the unstructured web. Manually identifying these unstructured Bengali texts is impractical and financially burdensome. To address these challenges, the development of a cyberbullying identification system becomes imperative for government policymakers and security agencies. However, the scarcity of annotated corpora related to bullying, the absence of domain-specific pre-trained feature extraction and classification models, and the unavailability of well-tuned hyperparameters for domain-centric tasks pose significant obstacles.

In recent years, considerable research has been dedicated to Bengali text classification [15], authorship attribution [14], sentiment analysis [2], and emotion classification [5]. However, the exploration of Bengali cyberbullying identification from textual data has been relatively limited [1, 3, 4]. Notably, the predominant approaches in existing research involve TF-IDF and non-contextual embedding-based (i.e., GloVe, FastText, Word2Vec) feature extraction, alongside statistical, CNN, and LSTM-based classification models. It is worth noting that the TF-IDF-based feature extractor falls short in capturing semantic meaning-based text features, while non-contextual embeddings like GloVe [26], FastText [6], and Word2Vec [22] struggle to extract context-aware features. To address these shortcomings, our research employs transformer-based language models. These models excel in extracting contextual text features, thus overcoming the limitations associated with traditional classification models. This shift is crucial for advancing the effectiveness of cyberbullying identification in the Bengali language.

In conclusion, to encapsulate the findings of the research, this study centers around the following Research Questions (RQs):

- **RQ1:** How to develop an intelligent cyberbullying text identification model for low-resource language?
- **RQ2:** How can extract the context-aware text features and overcome the limitations of statistical, convolutional, and sequential cyberbullying text classification models?

The noteworthy contributions of this research and potential answers to the Research Questions (ARQs) are outlined as follows:

- **ARQ1:** We develop an intelligent framework for identifying cyberbullying text. This framework systematically gathers a cyberbullying text corpus, extracts features from the text, and ultimately builds the model for textual cyberbullying identification (Section 3).
- **ARQ2:** We implement the transformer-based language models which capture context-aware textual features during the text-to-feature extraction phase and fine-tune the transformer-based language model using the cyberbullying corpus. The fine-tuned model overcomes the limitations of statistical, convolutional, and sequential models (Section 3.2).
- Constructed a cyberbullying text identification corpus comprising 34,433 labeled texts. Within this corpus, 17,901 are categorized as "Bully", and 16,521 are labeled as "Not-Bully". The collection process involved manual gathering from social media, followed by annotation and verification tasks using a manual annotation approach (Section 3.1).
- We trained a total of 12 cyberbullying text identification models, employing a diverse range of methodologies. This includes three statistical models (SVM, Libsvm, SGD), four deep learning models (CNN, LSTM, VDCNN, GRU), and six transformer-based models (BanglaBERT, mBERT, DistilBERT, IndicBERT, XML-RoBERTa, bELECTRA). Through empirical analysis, we identify the top-performing model to detect cyberbullying texts (Section 3.2).

2. Background

The continually evolving landscape of online platforms, including Twitter, Facebook, Reddit, and others, has instigated extensive research into the identification and categorization of undesirable texts in recent years. This research spans diverse domains, addressing aggression classification [23], hate speech detection [11], abuse detection [24], toxicity classification [18], misogyny classification [21], trolling identification [8], cyberbullying detection [25], and offensive text classification [25]. While a substantial body of research has been dedicated to various languages, with a predominant focus on English, this work provides a concise summary of studies addressing violence, hate, offense detection/classification, and related topics in both non-Bengali and Bengali languages. The section includes an overview of studies conducted in English,

Hindi, Arabic, and other languages. Additionally, kumar et al. [19] presents an aggressive language identification dataset featuring three categories covert, and non-aggressive with annotations in both English and Hindi, encompassing 15,000 posts/comments on aggression.

Aroyehun et al. developed deep neural network-based English models employing data augmentation and a pseudo-labeling method. Employing LSTM and CNN-LSTM approaches, their system achieved macro F1-scores of 0.64 and 0.59, respectively. In another study, [28] utilized the TRAC-2 [20] dataset and a bootstrap aggregating-based ensemble with fine-tuned BERT models to detect violence and misogyny. They attained an 80.3% weighted F1 score on the test set of English social media posts. [30] compiled the Offensive Language Identification Dataset (OLID) consisting of 14,000 English tweets. They established a three-layer hierarchical annotation schema to detect, classify, and identify the targets of texts, utilizing SVM, BiLSTM, and CNN for baseline evaluation. CNN outperformed competitors in all three levels, achieving macro F1 values of 0.80, 0.69, and 0.47. Furthermore, Founta et al. [10] provided a dataset comprising 80,000 tweets categorized into hateful, abusive, spam, and normal classes. They employed a comprehensive methodology to address ambiguous categories. Additionally, Davidson et al. [7] created a dataset of 25,000 tweets categorized into hate, offense, and neither. The best macro F1-score of 0.90 was achieved using logistic regression with TF-IDF and n-gram features.

Previous research predominantly focused on high-resource languages, employing transformer-based language models with extensive gold-standard bully identification corpora. However, limited attention has been given to low-resource languages like Bengali and Hindi. Research in low-resource bully identification often relies on non-contextual embedding models, such as GloVe, FastText, and Word2Vec. These embeddings face challenges in overcoming Out-of-Vocabulary (OOV) issues and extracting local and global contextual features specific to low-resource languages. In response to these challenges, this study introduces the **BullyFilterNeT** system. The system systematically develops a Bengali bully identification corpus and empirically evaluates various statistical, deep learning, and transformer-based language models. Ultimately, the top-performing model is selected to address the specific requirements of Bengali bully identification.

3. Methodology

Answer to **RQ1** and the primary objective of this study is to develop an intelligent Bengali cyberbullying text identification system called BullyFilterNeT which can intelligently distinguish between pieces of text as either containing bullying content or not. To fulfill this objective, the research progresses through three steps: (i) Cyberbullying Corpus Development (ii) Cyberbullying Text Identification Models Development (iii) Cyberbullying Text Identification Models Verification and Selection. The abstract view is presented in Figure 1. The subsequent subsections elaborate on each of these steps.

3.1. Cyberbullying Corpus Development

In this research, we have collected the bully and not bully-related texts corpus built in six steps. The overall procedure is presented in Figure 2.

Each of the steps is described in the following.

Data Source Selection. In the realm of data source selection, particularly for social media and blogs, the choice of platforms plays a pivotal role in shaping the nature and quality of the data acquired. Social media platforms like Twitter, Facebook, and Instagram provide real-time and diverse user-generated content, making them valuable sources for understanding public opinions, trends, and sentiments. The informal and conversational nature of social media content can offer insights into current events and user interactions. Similarly, blogs, with their more extended and often reflective narratives, contribute to a deeper understanding of individual perspectives and experiences. However, the selection process should consider the specific objectives of the research, the target audience, and the potential biases inherent in each platform. Striking a balance between the immediacy of social media and the depth of blogs is essential for obtaining a comprehensive and representative dataset for various analyses and applications.

Manually Collection. Manual data collection from social media and blogs involves a meticulous and hands-on approach to gathering information directly from these online platforms. Researchers or data collectors navigate through social media channels such as Twitter, Facebook, and blogs, identifying relevant content based on predefined criteria. This method allows for a more targeted selection of data, ensuring that specific themes, sentiments, or user interactions are captured. The manual collection enables the inclusion of context-rich content that automated tools might overlook, such as nuanced expressions, subtleties, or cultural references. However, this process is resource-intensive and time-consuming, as it requires human reviewers to sift through vast amounts of data. Additionally, ethical considerations, such as user privacy and consent, must be carefully addressed when manually collecting data from social media and blogs. Despite its challenges, manual data collection remains valuable for its ability to provide a nuanced understanding of online content,

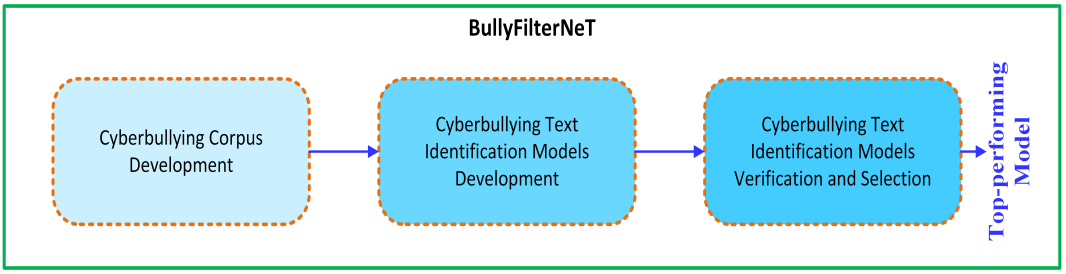

Figure 1. Abstract view of BullyFilterNeT

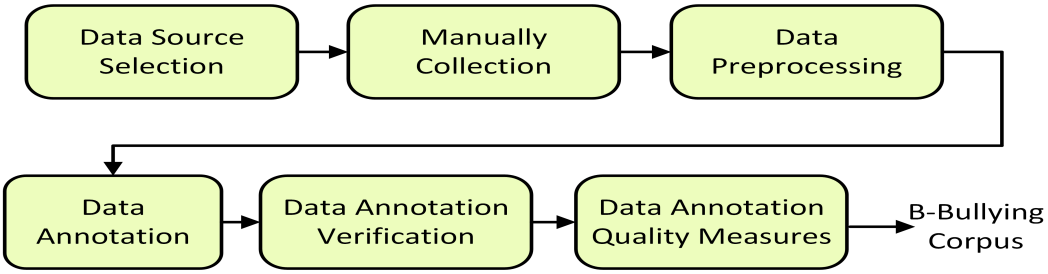

Figure 2. Abstract view of cyberbully corpus development

making it a preferred approach for certain research objectives.

Data Preprocessing. The Bengali full stop is replaced with a newline character. Various forms of whitespace (including non-standard ones) are identified using a regex pattern. These are replaced with a single space to standardize the text. The text is then cleaned of these punctuations using regex substitutions. The text is filtered to retain only Bengali characters and spaces. It achieves this by keeping characters with Unicode values greater than the letter 'z' (which effectively filters out Latin characters) and spaces. Consecutive spaces are reduced to a single space. Finally, the cleaned and processed text is returned. The preprocessing function is to be created specifically to cleanse Bengali text data, rendering it more appropriate for later analysis or training of models.

Data Annotation. Taking into consideration the epistemological issues highlighted by Ross et al. [29], our research employed a tiered approach in the selection of annotators. Two annotators consisted of individuals with diverse academic backgrounds, including undergraduate and postgraduate students. The study established a comprehensive annotation framework to interpret textual content, following rigorous methodological standards. Annotators applied labels like 'Bully' and 'NotBully' based on semantic and pragmatic analysis. Table 1 presents the demographic categorization and epistemic stances of annotators. These measures ensure reliability, and effectiveness, and minimize errors. The process involved two annotators for intersubjective validity.

Data Annotation Verification. Data annotation verification, a crucial step in ensuring the quality and accuracy of labeled datasets, involves the expertise of linguistics professionals who assess and validate annotations based on the opinions of two annotators. In this process, two individuals independently annotate the data, and their annotations are then reviewed by a linguistics expert. The linguist examines the annotations for consistency, coherence, and adherence to predefined guidelines. Any discrepancies or disagreements between the two annotators are carefully scrutinized, and the linguist, drawing on their linguistic expertise, resolves ambiguities and refines the annotations to maintain a high standard of quality. This meticulous verification process helps enhance the reliability of annotated datasets, particularly in linguistically nuanced tasks, ensuring that the labeled data accurately reflects the intended linguistic features or patterns.

Data Annotation Quality Measures. The statistical metric denoted as [63] is utilized to assess the level of agreement in annotations, as represented by the below equation,

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

The sign p(o) is used to represent the observed and hypothetical probability of annotative agreement, denoting them as separate entities. The Kappa coefficient, which measures inter-annotator agreement, yielded a score of 0.87 for the 'Bully' category. This indicates that the lexico-semantic elements associated with this category are notably unambiguous, facilitating a high level of agreement across annotators. In

Table 1. Demographic Categorization and Epistemic Stances of Annotators

Annotator	Academic Level	Research Area	Experience	Gender	Viewed Cyberbullying	Targeted by Cyberbullying
AN-1	Undergraduate	NLP	1 year	Male	Yes	No
AN-2	Undergraduate	NLP	2 years	Female	Yes	Yes
Expert	Senior	NLP, Ethics	10 years	Male	Yes	Yes

contrast, the category labeled 'NotBully' demonstrated a Kappa score of 0.73, suggesting a greater degree of intricacy and the possibility of variations in annotation. Table 2 shows the Kappa Scores for Annotation Classes in 'CyberBullyDetect'.

Table 2. Kappa Scores for Annotation Classes in 'CyberBullyDetect'

Annotation Category	Kappa Score
Bully	0.87
NotBully	0.73
Mean Kappa Score	0.75

Corpus Statistics Following the completion of the crucial six steps, this study successfully establishes a Cyberbullying identification corpus. The statistics summarizing the corpus are detailed in Table 3. This table provides a comprehensive overview of a Bengali corpus categorized into "Bully" and "Not-bully" classes. The corpus consists of a total of 34,422 samples, with 24,108 samples allocated for training and 10,314 for testing. In the training set, there are 12,543 samples labeled as "Bully" and 11,565 samples labeled as "Not-bully." The testing set comprises 5,358 "Bully" samples and 4,956 "Not-bully" samples. These statistics offer a clear distribution of the dataset, indicating the balance or imbalance between the two classes and the total number of samples available for model training and evaluation. Such information is essential for understanding the characteristics of the dataset and guiding the development and assessment of models for Bengali cyberbullying detection.

3.2. Development, Verification, and Selection of the Cyberbullying Text Identification Model

The primary goal of this study is to create a text identification system for low-resource cyberbullying. To achieve this objective, the research is structured into three main steps: (i) Text-to-Feature extraction, (ii) Development of statistical, deep learning, and transformer-based language models, and (iii) Evaluation of models and selection of the top-performing model. The overall methodological procedure is presented in Figure 3. In the following subsequent subsections, each of the steps is described.

Text-to-Feature extraction. In this study, we have employed two types of feature extraction methods, i.e., (1) Non-contextual and (2) Contextual.

(1) Non-contextual The three non-contextual embedding models are used for Bengali text-to-feature extraction purposes, i.e., GloVe [26], FastText [6], and Word2Vec [22]. In the realm of Bengali cyberbullying text analysis, the process of text-to-feature extraction plays a crucial role in uncovering meaningful patterns and representations. This research leverages three prominent word embedding techniques, namely GloVe, FastText, and Word2Vec, to extract informative features from Bengali cyberbullying text. GloVe captures global word co-occurrence statistics, FastText considers sub-word information, and Word2Vec models word embeddings based on contextual similarity. By employing these techniques, the study aims to harness the unique linguistic nuances of Bengali cyberbullying instances, enabling a more nuanced understanding of the language-specific characteristics associated with such content. The comparative analysis of these word embedding methods contributes to the development of a robust text-to-feature extraction pipeline tailored for Bengali cyberbullying detection.

(2) Contextual In answer to the research question **RQ2** this study deployed contextual feature extractors, such as BanglaBERT, XML-RoBERTa, IndicBERT, and ELECTRA, which are advanced language models designed to capture contextual information from text in the Bengali language. These models belong to the family of transformer-based architectures, which have demonstrated remarkable success in natural language processing tasks. BanglaBERT is specifically tailored for Bengali, offering contextualized embeddings by considering the unique linguistic intricacies of the language. XML-RoBERTa extends this idea, emphasizing the importance of contextual embeddings in handling complex structures and multiple languages. IndicBERT, designed for various Indic languages including Bengali,

Table 3. Summary of corpus

Attributes	Values
Total samples	34,422
Total Training Samples	24,108
Bully Training samples	12,543
Not-bully Training Samples	11,565
Total Test Samples	10,314
Bully Testing Samples	5,358
Not-bully Testing Samples	4,956

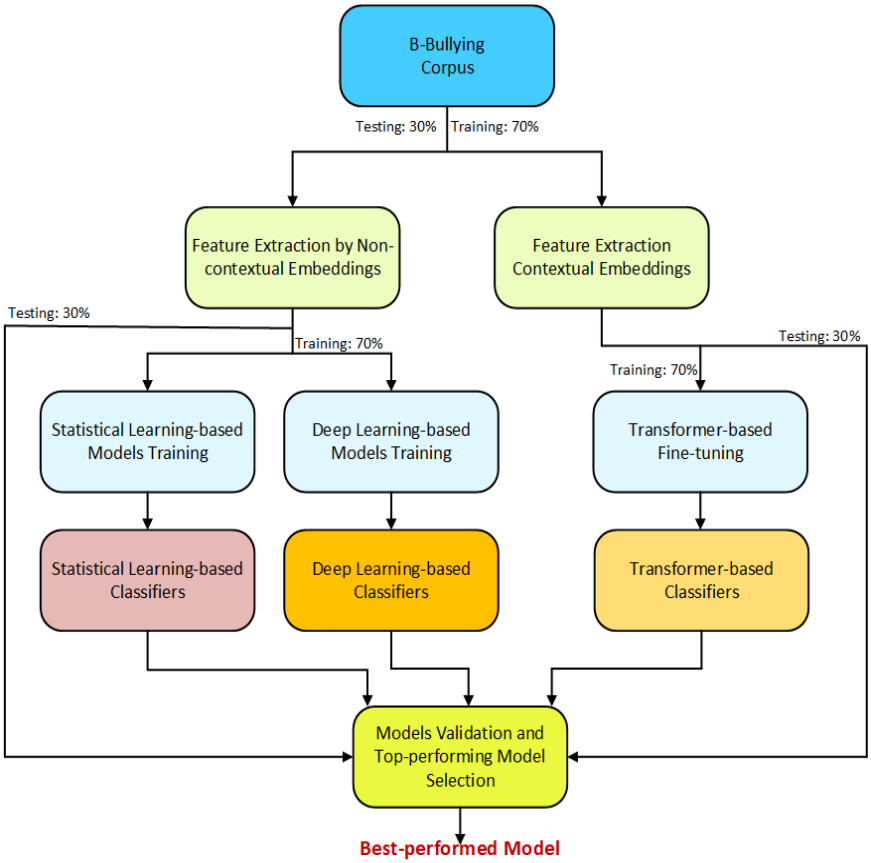

Figure 3. Abstract view of cyberbullying text identification models development and top-performing models selection

focuses on contextualized representations for improved language understanding.

ELCTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) is an approach that introduces an adversarial training strategy to pre-train transformer-based models, enhancing their efficiency in handling contextual features. These contextual feature extractors contribute significantly to tasks like sentiment analysis, named entity recognition, and cyberbullying detection in Bengali text, as they empower models to comprehend the nuanced contextual meanings within the language. Their incorporation in natural language processing pipelines enriches the representation of Bengali text, facilitating more accurate and context-aware language understanding in various applications.

Development of Statistical, Deep Learning, and Transformer-based Language Models. In this study, we have employed two statistical, four deep learning, and four transformer-based language models are verify the Bengali Cyberbullying text identification system.

(1) Statistical Models This study explores the effectiveness of three machine learning (ML) techniques: GPU-based Support Vector Machine (SVM), GPU-based Libsvm and Stochastic Gradient Descent (SGD) for Bengali for Cyberbullies text [13]. The ML-based classifier

models are developed and tuned on a created corpus. SVM and Libsvm share similar parameters, including a sigmoid kernel, tol of 0.00001, and a decision function shape of 'over,' with Libsvm demonstrating faster performance. The SGD classifier employs parameters loss = modified_huber and alpha = 0.001, with the remaining parameters set as defaults.

(2) Deep Learning-based Models **CNN:** Employing a single-layer, multi-kernel Convolutional Neural Network (CNN) architecture, this model investigates Bengali Cyberbullies text performance with three embedding models (FastText, GloVe, & Word2Vec). The distinct kernels are set to 3, 4, and 5, with corresponding filter numbers of 128, 128, and 256. Following the convolution layer, there is a 1D max-pool layer and activation layers [14]. Subsequently, the pooled features are concatenated and subjected to a dropout operation with a threshold value of 0.3.

VDCNN: Introducing the Variable-Size Deep Convolutional Neural Network (VDCNN) architecture for Bengali for Cyberbullies text identification purpose [16]. Unlike the original VDCNN, which operates on character-level embeddings, this adaptation combines VDCNN with different embedding techniques to enhance Bengali text classification performance. By reducing certain convolution operations, it addresses training time and model overfitting issues encountered by the original VDCNN.

LSTM: In this study, a two-layer LSTM is utilized with the following parameters: max sequence length = 256, hidden dimensions = 128, 256, batch size = 12, dropout rates = 0.50, 0.40, loss function = categorical_crossentropy, optimizer = adam, and activation function = softmax [2]. The model is trained for a maximum of 50 epochs on the developed corpus. It is noted that an increased number of sequences negatively impacts classification performance. Additionally, experiments are conducted with max sequence lengths of 1024 and 2048 in this research.

GRU: The two-layer GRU model is configured with the following parameters: hidden states = 128, 128, max sequence length = 512, batch size = 32, epochs = 80, dropout rates = 0.30, 0.25, loss function = categorical_crossentropy, optimizer = adam, and activation functions = tanh, softmax [2]. The last GRU layer is followed by a 1D max-pool layer. Subsequently, the 512 feature values are concatenated for the softmax layer, responsible for generating predictions in the expected category.

(3) Transformer-based Language Models The training module for Transformer-based Language Models, including mBERT, bELECTRA, XML-RoBERTa, IndicBERT, DistilBERT, and BanglaBERT, involves initial pre-training on a large multilingual corpus to learn contextualized representations [15]. Subsequently, these models undergo task-specific hyperparameters adaption, text-to-feature extraction, and fine-tuning to the intricacies of Bengali cyberbullying text identification.

Hyperparameters Adaption: Due to shortage of training samples, we have adapted the transformer-based language models using Eq. 2

$$H_i^O = F_{HPO}(H_i^I)), i = mBERT, ..., BanglaBERT \quad (2)$$

here H_i^O represents the optimised hyperparameters and H_i^I represents the initial hyperparameters of transformer-based language models, i.e., mBERT to BanglaBERT. The function $F_{HPO}(.)$ indicates the hyperparameters adaption function which adapted the maximum sequence length and batch size.

Text-to-feature Extraction: The transformer-based language models are extracted the linguistic features using Eq. 3.

$$T_j^F = F_{TFE}(T_{ji}), j = 1, .., N \quad (3)$$

here T_j^F represent the extracted features of j^{th} sample T_{ji} using the i^{th} model. $F_{TFE(.)}$ indicate the feature extraction function using model i.

Fine-tuning: Fine-tuning is a crucial process in the context of machine learning, especially when working with pre-trained language models like transformer-based models. It involves adjusting the parameters of a pre-trained model on a specific task or domain to enhance its performance. Fine-tuning is particularly valuable when the available labeled data for a specific task is limited. Fine-tuning allows the model to adapt to the specific characteristics and nuances of the target task or dataset, improving its ability to make accurate predictions. The fine-tuning process typically involves feeding the pre-trained model with task-specific data, and updating its weights based on the gradients computed during the training process. This helps the model to specialize in the target task while retaining the knowledge gained during pre-training on a large corpus. These transformer-based language models are fine-tuned using the Eq. 4.

$$\theta_{\text{fine-tuned}} = \arg\min_{\theta} \mathcal{L}_{\text{task-specific}}(\theta) + \lambda \sum_i \|\theta_i - \theta_{\text{pre-trained},i}\|^2 \quad (4)$$

here $\theta_{\text{fine-tuned}}$ represents the fine-tuned model parameters, $\mathcal{L}_{\text{task-specific}}(\theta)$ s the task-specific loss function, $\theta_i - \theta_{\text{pre-trained}}$ denote the parameters of the fine-tuned and

pre-trained models, respectively. The λ is a regularization hyperparameter that controls the balance between the task-specific loss and the regularization term. This equation captures the fine-tuning process where the model is optimized for a task-specific objective while leveraging knowledge gained from pre-training on a large corpus. Regularization helps prevent overfitting during the fine-tuning process.

Evaluation of Models and Selection of the Top-performing Model. In this research, we conducted a comprehensive evaluation of the test set, comprising 10,314 text samples, to assess the performance of various models in the context of Bengali cyberbullying text identification. Three statistical models, namely SVM, SGD, and Libsvm, were employed, along with four deep learning models, including CNN, VDCNN, LSTM, and GRU. Additionally, six state-of-the-art transformer-based language models, mBERT, bELECTRA, XML-RoBERTa, IndicBERT, DistilBERT, and BanglaBERT, were included in the evaluation. The performance metrics, encompassing accuracy and F1-score, for all models, are summarized in Table 4. Notably, BanglaBERT emerged as the top-performing model, achieving the highest accuracy and F1 score among the evaluated models. This underscores the efficacy of transformer-based models, particularly BanglaBERT, in cyberbullying text identification in the Bengali language within the scope of this study.

4. Experimental Results and Discussions

Within this section, we first provide an overview of the experimental setup and the chosen evaluation measures. Subsequently, we delve into the presentation and discussion of the results.

4.1. Experimental Setup and Evaluation Measures

The models were deployed on the Google Colaboratory platform with Python 3 and a Google Cloud Engine backend with GPU capability. This study's computing resources included 12.5GB of RAM and 64GB of disk space. The dataset was analyzed with Python's Pandas (version 1.1.4) and NumPy (version 1.18.5) libraries. The Scikit-Learn package (version 0.22.2) was used to create traditional machine learning models, while Keras (version 2.4.0) and TensorFlow (version 2.3.0) were used to create deep learning models. The ktrain library (version 0.25) was used for models using Transformer architectures. The dataset was partitioned into three sets: training, validation, and test. The training set supported the models' learning phase, whereas the validation set aided in hyperparameter tuning.

The final evaluation was carried out on an unknown test set using a variety of statistical measures, as specified in the following equations: **Precision (p):** It quantifies the proportion of true positive samples within the samples classified as positive.

$$Accuracy = \frac{Numberof CorrectPredictions}{Numberof Samples} \times 100 \quad (5)$$

$$Precision(p) = \frac{Truepositive}{Truepositive + Falsepositive} \quad (6)$$

Recall (r): It calculates the ratio of correctly labeled positive samples to total positive samples.

$$r = \frac{Truepositive}{Truepositive + FalseNegative} \quad (7)$$

Error Rate (e): This is the percentage of misclassified samples.

$$e = \frac{FalsePositive + FalseNegative}{Numberof Samples} \quad (8)$$

Weighted F1-Score: The F1-score is a harmonic mean of Precision and Recall. Because of the dataset's imbalance, a weighted F1-score is produced as follows:

$$F1 = \frac{(Truepositive)}{(Truepositive + 1/2(FalsePositive + FalseNegative))} \quad (9)$$

This section will also measure the weighted average (WA) and macro average (MA) precision, recall, and F1 score. provide a full overview of the findings produced by the various models, with the weighted F1 score serving as the key criterion of evaluation. This part will also include a comparison with existing methodologies, offering insight into the benefits and drawbacks of the proposed paradigm.

4.2. Result Analysis

In this section, we have evaluated the Bengali cyberbullying text identification models with the test dataset. Table 4 presents the performance metrics of various models for Bengali cyberbullying text identification based on a test dataset. The models are evaluated in terms of accuracy and F1 score, providing insights into their classification capabilities. Notably, traditional models such as GloVe combined with SVM, SGD, or Libsvm exhibit reasonable accuracy, ranging from 76.20% to 78.39%, with corresponding F1 scores in the 76.00-79.00 range. Moving to neural network-based architectures, GloVe combined with CNN, VDCNN, LSTM, and GRU achieve higher accuracy, with scores ranging from 79.61% to 84.36%, and F1-scores in the 80.00-84.00 range.

The performance further improves with the integration of transformer-based models. mBERT, XML-RoBERTa, IndicBERT, and DistilBERT consistently

Table 4. Accuracy and F1-score of cyberbullying text identification system based on 10,314 test dataset

Models	Accuracy (%)	F1-score	Precision	Recall
GloVe+SVM	77.73	77.35	76. 64	77.97
GloVe+SGD	76.48	75.00	74.70	75.35
GloVe+Libsvm	78.93	78.71	78.36	79.11
FastText+Libsvm	76.87	75.82	75.03	76.53
Word2Vec+Libsvm	76.20	74.84	73.40	76.18
GloVe+CNN	84.36	83.03	82. 79	83.83
GloVe+VDCNN	82.47	81.88	80. 61	82.96
GloVe+LSTM	81.30	80.10	79. 47	80.88
GloVe+GRU	79.61	79.25	78. 25	79.87
mBERT	86.47	86.21	86. 12	86.22
bELECTRA	85.25	85.12	84.87	85.25
XML-RoBERTa	87.13	86.62	86.62	87.34
IndicBERT	86.75	87.16	86.69	87.45
DistilBERT	85.65	85.96	85.87	86.01
BanglaBERT	**88.04**	**87.85**	**85.80**	**90.0**

demonstrate superior accuracy, ranging from 85.65% to 87.13%, and F1-scores in the 86.00-87.00 range. Notably, BanglaBERT outperforms all other models, achieving the highest accuracy of 88.04% and an F1-score of 88.00%. This underscores the effectiveness of transformer-based models, particularly BanglaBERT, in accurately identifying cyberbullying text in Bengali, showcasing their robust performance on the given test dataset.

BanglaBERT achieves maximum performance in Bengali cyberbullying text identification due to its tailored design for the Bengali language, extensive pre-training on a large corpus, and the ability to generate contextual embeddings. The model's fine-tuning process involves effective parameter tuning, optimizing its performance for the specific task. Leveraging transfer learning, BanglaBERT capitalizes on its general language understanding capabilities, adapting them to the nuances of cyberbullying identification in Bengali. These factors collectively contribute to BanglaBERT's superior accuracy, making it highly effective in discerning and classifying cyberbullying content in the given context.

Based on Table 4 performance, the maximum accuracy of Bengali cyberbullying text identification performance has been obtained from the transformer-based BanglaBERT models. The details of the BanglaBERT model's performance are presented in Table 5.

The table presents a detailed performance analysis of BanglaBERT in the context of cyberbullying text identification, categorizing results into "Bully" and "Not-bully" classes. Precision (p), recall (r), macro average percentage (MA%), and weighted average percentage (WA%) are reported for both categories. For the "Bully" class, the model achieves a precision of 90.00%, recall of 86.00%, and both macro and weighted average percentages of 88.00% for precision and recall. Similarly, for the "Not-bully" class, the precision is 96.00%, recall is 90.00%, and macro and weighted average percentages are 88.00% for both precision and recall. The support column indicates the number of instances in each class, with 5358 instances for "Bully" and 4956 instances for "Not-bully". These metrics collectively demonstrate BanglaBERT's strong performance in accurately identifying both cyberbullying and non-bullying content, with high precision, recall, and consistent average percentages across both categories.

4.3. Error Analysis

Figure 4 presents the confusion matrix of BanglaBERT models for the 10,314 test dataset. In the context of the confusion matrix for Bengali cyberbullying text identification, the terms "error" and "success" can be interpreted as follows:

The model achieved success in correctly identifying instances of cyberbullying (0 for Bully) with a count of 4462. These are instances where the model's prediction aligns with the actual presence of cyberbullying content. The model also demonstrated success in accurately identifying instances of non-cyberbullying content (1 for Not-bully) with a count of 4619. These are instances where the model correctly recognized and classified content as non-offensive. The model made an error in 739 instances by incorrectly predicting cyberbullying (0 for Bully) when the content was, in fact, not offensive. These are instances of false alarms or instances where the model may have been overly sensitive. The model made an error in 494 instances by failing to identify instances of cyberbullying when the content was offensive. These are instances where the model missed detecting actual instances of

Table 5. Statistical summary of BanglaBERT model based on 10,314 test dataset

Category	p(%)	r(%)	MA% (p)	MA% (r)	WA% (p)	WA% (r)	Support
Bully	90.00	86.00	88.00	88.00	88.00	88.00	5358
Not-bully	96.00	90.00	88.00	88.00	88.00	88.00	4956

Table 6. Accuracy comparison of proposed BanglaBERT with existing methods

Method	Accuracy(%)
GloVe+Libsvm [15]	78.39
GloVe+CNN [13]	84.36
GloVe+LSTM [2]	81.30
GloVe+VDCNN [16]	82.47
IndicBERT [12]	86.75
Proposed(BanglaBERT)	**88.04**

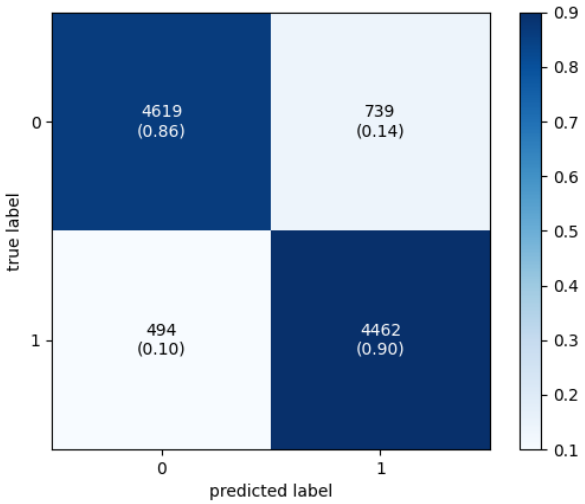

Figure 4. Confusion matrix of BanglaBERT model for 10,314 text dataset

cyberbullying. Understanding these success and error categories provides valuable insights into the model's strengths and weaknesses in differentiating between cyberbullying and non-cyberbullying content.

Comparison with Existing Research. In the absence of a standardized Bengali cyberbullying corpus and established standardization practices, this study employed existing methods along with their associated hyperparameters. The research involved training and validating the test set, and the summarized performance is presented in Table 6.

Various techniques, including GloVe combined with Libsvm [15], GloVe with CNN [13], GloVe with LSTM [2], and GloVe with VDCNN [16], have been previously employed for Bengali cyberbullying text identification. Additionally, IndicBERT [12] represents a transformer-based language model specifically designed for the Bengali language. The proposed model, BanglaBERT, outperforms all these methods, achieving the highest accuracy at 88.04%. This comparison underscores the superior performance of BanglaBERT in the specific task of cyberbullying text identification in Bengali, demonstrating its efficacy in surpassing existing methods.

Overall, our evaluation underscores the superiority of transformer-based models, particularly BanglaBERT, in accurately identifying cyberbullying text in Bengali. BanglaBERT's tailored design, extensive pre-training, and fine-tuning contribute to its exceptional performance, surpassing traditional and other transformer-based models. The model exhibits robust precision and recall, as demonstrated by the confusion matrix analysis. Comparative assessment with existing methods further solidifies BanglaBERT's position as a leading solution for promoting a safer digital environment in the Bengali language. The confusion matrix also provides valuable insights into specific cases where models either succeeded or failed in correctly classifying the given inputs in cyberbullying detection.

5. Conclusion

This study introduces a novel corpus comprising 34,422 samples, with 70% (24,108) designated for training and 30% (12,543) for testing. The corpus undergoes evaluation employing statistical, deep learning, and transformer-based language models. BanglaBERT stands out, achieving the highest accuracy at 88.04%. Deep learning models utilize non-contextual embeddings GloVe, FastText, and Word2Vec yet struggle with Out-of-Vocabulary (OOV) issues. In contrast, transformer-based language models excel in extracting contextual features, mitigating OOV challenges. Statistical models fall short due to limitations in capturing

local and global word and sentence-level semantics. While deep learning models capture local semantics, they lack context awareness. In summary, transformer-based language models prove adept at extracting context-aware features, leading to superior accuracy. While we used Bengali datasets for experiments, this model is also applicable to other low-resource languages such as Arabic, Tamil, etc.

In the future, we will collect other low-resource languages datasets and conduct the relevant experiments. We plan to further refine the large language models (LLMs) using the cyberbully dataset and explore the effects of Bengali to English translation data using the developed model. Overall, this work opens up a promising pathway in cyberbullying research for low-resource languages.

References

[1] Abdhullah-Al-Mamun and Shahin Akhter. Social media bullying detection using machine learning on bangla text. In *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, pages 385–388, 2018.

[2] Sadia Afroze and Mohammed Moshiul Hoque. Sntiemd: Sentiment specific embedding model generation and evaluation for a resource constraint language. In *Intelligent Computing & Optimization*, pages 242–252, Cham, 2023. Springer International Publishing.

[3] Md. Tofael Ahmed, Maqsudur Rahman, Shafayet Nur, Azm Islam, and Dipankar Das. Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–10, 2021.

[4] Arnisha Akhter, Uzzal Kumar Acharjee, Md. Alamin Talukder, Md. Manowarul Islam, and Md Ashraf Uddin. A robust hybrid machine learning model for bengali cyber bullying detection in social media. *Natural Language Processing Journal*, 4:100027, 2023.

[5] Sara Azmin and Kingshuk Dhar. Emotion detection from bangla text corpus using naïve bayes classifier. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–5, 2019.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Tran. ACL*, 5:135–146, June 2017.

[7] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[8] Luis Gerardo Mojica de la Vega and Vincent Ng. Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[9] Amirita Dewani, Mohsin Ali Memon, and Sania Bhatti. Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for roman urdu data. *Journal of Big Data*, 8(1):160, December 2021.

[10] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[11] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*, 2017.

[12] Md. Rajib Hossain and Mohammed Moshiul Hoque. Coberttc: Covid-19 text classification using transformer-based language models. pages 179–186, Cham, 2023. Springer Nature Switzerland.

[13] Md. Rajib Hossain and Mohammed Moshiul Hoque. Toward embedding hyperparameters optimization: Analyzing their impacts on deep leaning-based text classification. In *The Fourth Industrial Revolution and Beyond*, pages 501–512, Singapore, 2023. Springer Nature Singapore.

[14] Md. Rajib Hossain, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Nazmul Siddique, Md. Nazmul Islam, and Iqbal H. Sarker. Authorship classification in a resource constraint language using convolutional neural networks. *IEEE Access*, 9:100319–100338, 2021.

[15] Md. Rajib Hossain, Mohammed Moshiul Hoque, and Nazmul Siddique. Leveraging the meta-embedding for text classification in a resource-constrained language. *Engineering Applications of Artificial Intelligence*, 124:106586, September 2023.

[16] Md. Rajib Hossain, Mohammed Moshiul Hoque, Nazmul Siddique, and Iqbal H. Sarker. Bengali text document categorization based on very deep convolution neural network. *Expert Systems with Applications*, 184:115394, 2021.

[17] Md. Rajib Hossain, Mohammed Moshiul Hoque, Nazmul Siddique, and Iqbal H Sarker. CovTiNet: Covid text identification network using attention-based positional embedding feature fusion. *Neural Computing and Applications*, 35(18):13503–13527, June 2023.

[18] Mladen Karan and Jan Šnajder. Preemptive toxic language detection in wikipedia comments using thread-level context. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 129–134, 2019.

[19] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11, 2018.

[20] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Evaluating aggression identification in social media. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 1–5, 2020.

[21] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. Finding opinion manipulation trolls in news community forums. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 310–314, 2015.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. pages 1–12, 2013.

[23] Nishant Nikhil, Ramit Pahwa, Mehul Kumar Nirala, and Rohan Khilnani. Lstms with attention for aggression detection. *arXiv preprint arXiv:1807.06151*, 2018.

[24] Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pages 363–370, 2019.

[25] John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th international Workshop on Semantic Evaluation*, pages 571–576, 2019.

[26] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, pages 1532–1543, Doha, Qatar, 2014. ACL.

[27] Eric Rice, Robin Petering, Harmony Rhoades, Hailey Winetrobe, Jeremy Goldbach, Aaron Plant, Jorge Montoya, and Timothy Kordic. Cyberbullying perpetration and victimization among middle-school students. *American journal of public health*, 105(3):e66–e72, 2015.

[28] Julian Risch and Ralf Krestel. Bagging bert models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61, 2020.

[29] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

[30] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*, 2019.