Enhancing Cyberbullying Detection in Bangla Language: A Hybrid BiLSTM-Attention Approach

Sultana Umme Habiba
Dept. of CSE
Bangladesh University of Engineering & Technology, Dhaka

Sadia Sharmin
Dept. of CSE
Bangladesh University of Engineering & Technology, Dhaka

ABSTRACT
With the rapid advancement in the field of information and communication technology, people are getting connected with each other via social media sharing content like texts, images or posts. Since the trend of sharing thoughts, feelings or opinions through social media has become an indispensable part of our life, social media platforms have opened the way of being a victim of cyberbullying significantly more than before. Social distancing, due to the effect of the post COVID 19 pandemic situation, causes a noteworthy rise up to be a victim of cyberbullying in social media. This work proposes a hybrid deep learning based classifier that combines a self-attention layer with BiLSTM to differentiate between bully and non-bully texts in Bangla language from different social media. We have collected and labelled our work dataset from Facebook, YouTube, Twitter, TikTok etc. Context-based data augmentation is applied to improve the performance of the model. Existing algorithms for sentiment analysis tasks like SVM, Random Forest, Naive Bayes, LSTM, GRU, BERT etc. are experimented and comparative analysis among these models and our proposed hybrid model is also demonstrated. This research combines prominent feature extraction techniques like count vectorizer, Tf-Idf, and transformer-based contextual word embedding. The experimental result depicts that our proposed hybrid model outperforms all the previous works in cyberbullying detection in Bangla by achieving 89.3% accuracy.

General Terms
Sentiment analysis, Natural language processing

Keywords
Cyberbullying detection, Deep learning, Attention.

1. INTRODUCTION
Cyberbullying is such an act done by people who intend to harass or menace others by addressing abusive comments or messages. Cyberbullying leads a man to fall into extreme depression, stress and anxiety disorder. Existing research has shown a positive correlation between cyberbullying involvement and growing loneliness due to social distancing. Day by day people are getting more exposed to social network via online communication and so the risk of being a victim of cyberbullying is also increasing consequently. According to a statistics provided by a giant mobile operator, in Bangladesh 49% students fall a victim to cyberbullying harassment either being bullied by others comment or being involved in bullying others [2]. A survey during COVID 19 pandemic by UNDP shows that 80% of the total victims of cyberbullying are women in Bangladesh who are aged between 14 and 22 [2]. Surprisingly most of the cybercriminals lie in the age group between 14 and 17. Social workers are organising anti-cyberbullying campaign with the help of UNDP on those cases who have tried suicidal attempts after being a victim of cyberbullying. Previous research works have shown the negative impacts of cyberbullying are strongly co-related with physical and mental health issues, depression, anxieties, suicidal attempt rate, adolescent violence rising etc. Cyberbullying can influence a victim both rapidly and strongly as people can reach those public bullying posts for a long time if those have already gone viral. Attackers usually make comments that can occur sexual harassment, body shaming, racism or trolling to any celebrity or public figure. Even if all these bullying can hamper the harmony of the victims' life creating a long term effect on their mental health.

Different Natural Language Processing (NLP) techniques have brought about a revolutionary change in the field of sentiment analysis. Sentiment analysis is such a process with which people's positive or negative attitude towards any topic can be analyzed. Even if sentiment analysis task can detect the use of abusive words or bullies from online social networks also. Existing research works related to sentiment analysis using Bangla text lack in sufficient dataset [3][4][5]. Most of the works of Bangla sentiment analysis focus on using deep learning based classifiers and initial preprocessing steps to clean the data. [6] proposed BTSC algorithm to generate sentiment score for Bangla corpus and to extract polarity score of Bangla text. However, polarity based sentiment analysis performs better in case of balanced dataset. Data collection process, source of data, dataset labelling may cause bias in cyberbullying detection using polarity score based sentiment analysis. For Bangla language, there is no benchmark dataset for cyber bullying detection. Few researches are found [7][8] on collecting Bangla abusive texts through social media from different celebrity pages, news portal, youtube comments, tik tok video comments and so on. These works are incorporated with a limited number of Bangla text. Combining attention mechanisms with CNN [9] for Bangla sentiment analysis has shown better performance. Machine learning classifiers like SVM (Support Vector machine), Random Forest, naïve Bayes etc. are proposed for sentiment analysis in different NLP tasks [10]. These works approach different word embedding techniques like Word2Vec, Tf-Idf, count vectorizer etc. with their classifier models. In recent works, different high performing deep learning based models like LSTM, Bi- LSTM, GRU, RNN, transformer based model BERT etc. have been proposed for Bangla sentiment analysis. However cyberbullying detection needs attention to fill up the gap among the other sentiment analysis tasks in Bangla language. Bullying, trolling or making someone inferior in-front of other people is not a very recent practice in our society. In recent times, bullying or trolling through social media platforms is spreading over at a high rate due to the social distance caused by COVID 19 [1]. We are definitely bound to confess the importance of cyberbullying detection in social media. Online education, tutorials, social communication etc. are pushing

teenagers, children or young adults towards social media and early detection of cyberbullying may help to raise awareness among the social media users. Analyzing all these circumstances, the following questions have arisen:

- **Q: How effective will be the way of keyword based cyberbullying detection?** In most cyberbullying detection tasks, cyberbullying is identified using some keywords like ”bull**shit”, ”Fu**ck” or other slang. Is that efficient enough to work with Bangla texts?
- **Q: How will the existing classifiers handle the ambiguous words?** In Bangla language, some words can carry multi-modal semantics. Using content specific feature extraction techniques will not work well. How can this research be fruitful?
- **Q: Like other sentiment analysis task, cyberbullying detection limits to work with sufficient data. How can an enriched Bangla corpus be trained in the proposed work?**

In this study, our main motto is to develop a hybrid attention and transformer based deep convolutional neural network to detect cyberbullying from social media contents in Bangla language. This model is evaluated to provide a better performance to detect bullying Bangla text in comparison with the other state-of-the-art techniques used in Bangla sentiment analysis. Accordingly, the major contribution of our study are as follows:

- To analyze the performance of cyberbullying detection architectures used in other languages and to measure how well those models perform in the case of Bangla language.
- To prepare a balanced dataset from different social platforms like facebook, youtube, twitter using API that 2 can access comments of public posts.
- To eliminate the limitations of previous research works like ambiguous word problem, multimodality exposure due to insufficient data.
- To develop an architecture taking concern on both polarity and subjectivity of words.
- To evaluate the performance of our developed hybrid architecture over our new datasets and compare the performance against that of the existing architecture.

Following the objectives of this research work, we expect the following outcomes:

- Cyberbullying detection using Bangla text as well as limitations of existing architectures intended for sentiment analysis especially abusive text detection.
- A balanced, labelled and clean dataset in Bangla language to detect bullying text.
- Evaluation of the performance of our developed hybrid attention and transformer based CNN architecture against that of the existing architectures.
- Comparative study of the developed hybrid architecture and its generalization ability over another dataset on Bangla sentiment analysis.

The paper is organized as follows: In Section 2, we discuss various State-of-the-art techniques on sentiment analysis in different languages, Section 3 includes the details about the proposed methodology, dataset collection etc., Section 4 describes experimental results and comparison among the previous works and in the last Section 5 figures out the future scope of this work and shows concluding remarks.

2. LITERATURE REVIEW

Since cyberbullying has become a global health concern due to the rapid growth of online communication and social media, researchers have shown interest to explore the area of automatic cyberbullying detection in recent years. In this context, researchers have experimented with traditional machine learning methods combined with several feature extraction techniques, attention-based convolutional neural networks, transformer-based classifiers, etc.

2.1 Content Based Feature Extraction And Deep Learning Based Works

Researchers used content-based feature extraction methods like TF-IDF, count vectorizer, and BOW to convert raw texts into intermediate vectors, enabling traditional classifiers such as Decision Tree, Random Forest, and Naive Bayes to perform well in sentiment analysis. [11] Extracted both content- and sentiment-based features from Twitter using TF-IDF, focusing on profanity-pronoun correlations and cyberbullying keywords (e.g., racism, harassment), labeled ~1000 texts, and achieved ~75% accuracy using SVM. [12] Proposed a Dolphin Echolocation Algorithm to optimize RNN on 10,000 tweets collected via 35 bullying-related keywords (e.g., “rape”, “moron”), used SMOTE for balance, and achieved 90.4% accuracy, though limited by biased data and Twitter-only sources.][14] Applied Naive Bayes, Decision Tree, Random Forest, SVM, and DNN on Kaggle bullying datasets using BOW and TF-IDF, with DNN achieving 99% accuracy.

Deep learning excels in sentiment analysis. [17] Used a CNN-LSTM model on text and images from the Kaggle Toxic Comment dataset, achieving 85% accuracy. [18] Applied RNN-LSTM on Roman Urdu with a slang dictionary, also reaching 85%. Another study used CNN with GloVe on Twitter to detect bully words. A 2D-CNN with VGG16 and Inceptionv3 classified bullying images but lacked multilingual text support. A Bi-LSTM with character-level encoding improved GloVe, achieving 92% accuracy on Twitter text.

Previous works focused on content-based features, though contextual cues are often crucial. [13] proposed a hybrid BERT-LSTM model, using contextual embeddings and data augmentation, achieving 91% accuracy on IMDB. [16] used BERT for cyberbullying detection, capturing semantic and sarcastic cues, with 70% accuracy despite dataset limitations.

2.2 Sentiment Analysis in Bangla

Recent Bangla sentiment analysis studies explored various deep learning models and embeddings. [22] used Facebook comments with CNN, LSTM, and hybrid models, achieving 85–90% accuracy. [23] applied LSTM with Word2Sequence on Bangla food reviews. [24] used CNN, MLP, and LSTM on cricket news, reaching ~70% accuracy. [25] analyzed Bangla and Romanized Bangla YouTube comments using RNN and TextBlob. [26] used TF-IDF with SVM, RF, and KNN on Facebook comments. [15] detected Bangla cyberbullying with Word2Vec and LSTM, achieving 87% (binary) and 79% (multiclass) accuracy.

2.3 Research Gap

Some key limitations in previous works include the lack of sufficient Bangla datasets and reliance on content-based features like n-gram and TF-IDF, which struggle with sarcasm and nuanced bullying. Several datasets were biased toward

profanity or positive words, affecting prediction accuracy. Most advanced models are developed for English, limiting Bangla cyberbullying detection. Additionally, existing pre-trained models are trained on English corpora, highlighting the need for a balanced and rich Bangla dataset.

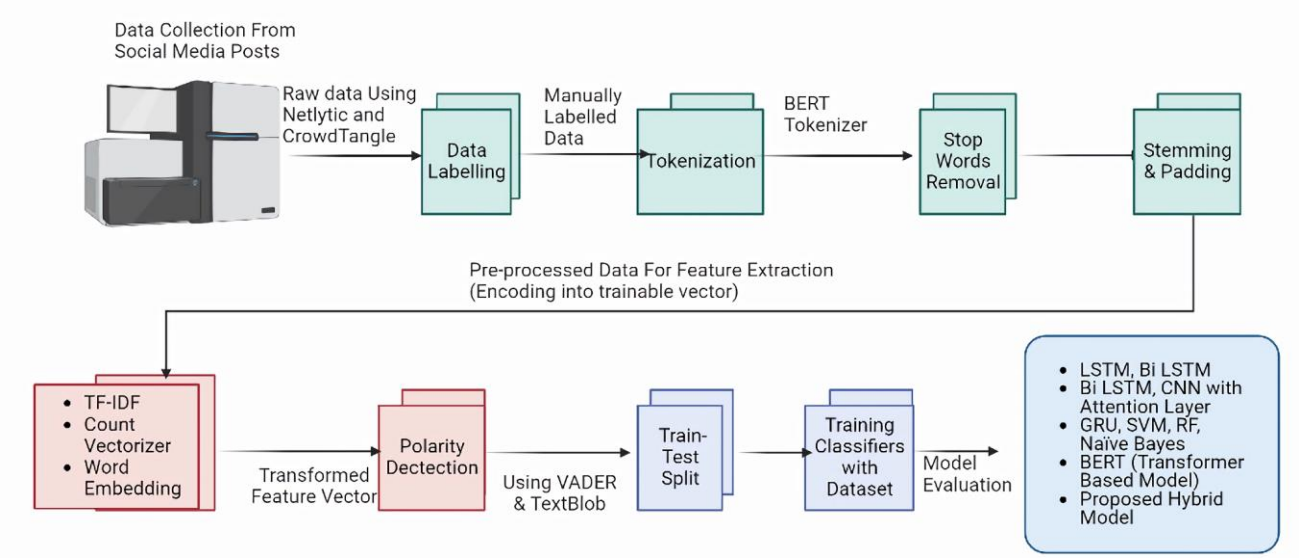

Fig 1: Workflow Diagram For Our Proposed Work (From Data collection, annotation to classification)

3. METHODOLOGY

The outline of our proposed methodology is summarized as follows like in Figure 1:

3.1 Dataset Collection and Processing

A Bangla-language dataset was collected to classify texts as abusive or non-abusive, focusing on comments from public posts, celebrity pages, TikTok, YouTube, and Facebook. Netlytic and Crowdtangle were used to access Facebook, Twitter, and YouTube content, but Facebook comments were collected manually due to restrictions since 2019. Other platforms used web-scraping tools. After merging data, a Google Form survey was conducted for labeling, with each text reviewed by at least three participants. Labels were assigned by majority vote. Participants included males and females aged 18–50, including students, family members, and colleagues. The final dataset contains 1,500 abusive and 3,000 non-abusive texts. Sample data are shown in Table 1.

Table 1. Examples of Bangla Posts

Bangla Post	Class
সুন্দর এই ভুবনে সুন্দরতম জীবন হোক তোমার। পূরণ হোক প্রতিটি স্বপ্ন, প্রতিটি আশা,	**Not Abusive**
হেডার বাল যত্তসব পাগল ছাগল	**Abusive**

Hashtag-Based Data Collection for Abusive Bangla Texts (2020-2023):

For collecting abusive Bangla texts, researchers used targeted hashtags linked to abusive behavior on social media, selected for relevance and frequency. Data from 2020–2023 ensured coverage of recent incidents. This approach helped build a robust dataset for Bangla cyberbullying detection.

Non-abusive hashtags: **#আল্লাহ, #ক্রিকেটার, #শুভকামনা, #আনন্দ, #ভালো, #ধন্যবাদ, #হাসি, #চমৎকার;**
abusive hashtags: **#বদমাইশ, #অপমান, #গালি, #ধর্ষক, #হত্যা, #সহিংসতা, #জঘন্য, #বেজন্মা, #নোংরা, #অশ্লীল, #অসভ্য, #অপমানিত, #বর্বর, #হিংস্র, #লজ্জাজনক, #বিষাক্ত, #ঘৃণা, #খানকী, #মাগী, #বাল, #বেহায়া, #নির্লজ্জ**

Dataset Validation Using IAA (Inter Annotator Agreement):

As Bangla is a low-resource language, the dataset was manually labeled, which may introduce bias. To ensure reliability, Inter Annotator Agreement (IAA) using Cohen's Kappa was applied. Two random subsets (500 samples each) were labeled by separate annotators. For Set A, Cohen's Kappa was **0.7**, and for Set B, **0.74**, both indicating substantial agreement and validating the dataset's quality.

Data Augmentation and Preprocessing:

Bangla sentiment analysis lacks sufficient benchmark datasets, which limits deep learning performance. To address this, data augmentation was applied using techniques like Word2Vec, GloVe, and NLPAug, though such methods may generate contextually irrelevant replacements. To improve quality, a multilingual BERT-based contextual augmentation was used for more accurate word substitutions.

The collected Bangla social media texts were preprocessed in four key steps:

Punctuation Removal: Special characters, punctuations, emojis, and emoticons were removed to simplify texts. For example, "বাহঃ ,সভ্য জাতি.....এখন বলেন নেঙটো হয়ে গান গাওয়াও বাঙালির ঐতিহ্য..!!!!" becomes "বাহ সভ্য জাতি এখন বলেন নেঙটো হয়ে গান গাওয়াও বাঙালির ঐতিহ্য".

Tokenization & Stop Words Removal: Texts were tokenized into individual words, e.g., "সারাজীবন অটুট থাকুক আপনার মুখের অমুল্য হাসি" into <সারাজীবন> <অটুট> <থাকুক> etc. Common Bangla stop words like 'অতএব', 'ইত্যাদি', 'একটি' were removed to enhance semantic focus.

Stemming & Padding: Using BanglaKit, words were stemmed to root forms, e.g., "বিরান পথে হেঁটে গিয়ে কোথায় কোন অজানায় হারিয়ে গেলে?" becomes "বিরান পথ হেঁট কোথা অজানা হারি কষ্ট". Padding was applied to ensure uniform input length for neural models.

Feature Extraction:

TF-IDF: Weighted word importance was calculated with `TF(n) = term frequency`, `IDF(n) = log(total docs / docs with term n)`, and `TF-IDF = TF × IDF`.

Count Vectorizer: In NLP, models require text to be converted into numerical form. **Count Vectorizer** handles this by creating a 2D sparse matrix where each column is a unique word from the corpus and each row corresponds to a text sample. The matrix entries reflect the frequency of words.

For example, with a sample vocabulary: **["সাকিবেই", "অর্জন", "গর্জন", "এগিয়ে", "বাংলাদেশ", "তোমার"]**

- **Sample** : "এগিয়ে যাও বাংলাদেশ , অর্জন তোমার গর্জন তোমার" → Count Vector: [0, 1, 1, 1, 1, 2]

Word Embedding: In machine and deep learning, textual data must be converted into numerical vectors. For this sentiment analysis task, two types of word embeddings were used:

(i) Content-based (one-hot/integer encoding), and
(ii) Contextual embedding using transformers.

For traditional ML models, content-based one-hot encoding assigns each word a unique integer based on a vocabulary (e.g., for vocab size = 200, the sentence "তোমার হাঁসিতে হাজার ফুল ফুটে যায়" → [23, 45, 9, 11, 89, 112]). Padding is applied to maintain uniform vector size: e.g., [0, 0, ..., 23, 45, 9, 11, 89, 112].

For contextual embedding, a BERT-based transformer model was used, which captures semantic meaning from context. For instance, in

- "কাউয়া ও একটা পাখি, আর আপনিও একজন অভিনেত্রী"
- "অসম্ভব গুণী অভিনেত্রী তিশা"

  the word "অভিনেত্রী" holds different meanings. BERT handles this through similarity-based vector representation.

3.2 Proposed Hybrid Model

Our proposed hybrid model combines BiLSTM and self-attention mechanisms to enhance contextual understanding in sentiment classification. BiLSTM captures dependencies from both past and future by scanning text bidirectionally, while the attention layer highlights important words, focusing on key semantic patterns. The attention mechanism applies masks (Fig. 14) to assign higher weights to relevant words, improving context extraction, especially in long sequences.

Input texts are preprocessed—tokenized, cleaned, stop words removed, and stemmed. Sequences are padded to ensure uniform input size. The preprocessed data are passed through an embedding layer, converting tokens into dense vectors.

The embedded vectors are then processed by a CNN layer for local feature extraction, followed by BiLSTM. In BiLSTM, the forget gate filters out irrelevant words by assigning near-zero weights, while the input gate updates memory with filtered features using a tanh activation. The output gate applies another tanh before passing data forward. This forward LSTM output is mirrored by a backward LSTM for bidirectional context.

Next, a self-attention layer assigns importance scores to each token, capturing global dependencies. This is followed by a fully connected layer, dropout to prevent overfitting, max-pooling to reduce dimensionality, and a Softmax activation for final classification. The model is trained using the Adam optimizer with a learning rate of 0.0005.

This hybrid sequential architecture—embedding + CNN + BiLSTM + attention—effectively captures both local and global semantic cues. BERT-inspired contextual understanding is partially integrated through self-attention, though the model itself is lightweight compared to full transformer stacks. Each embedding layer generates a 150-dimensional vector representing encoded text ready for training. A 1D convolutional layer is applied, slightly reducing vector size, followed by a maxpooling layer that selects maximum values to capture prominent features and reduce trainable parameters. Next, a BiLSTM layer scans the sequence bidirectionally to extract syntactic and semantic patterns. An attention layer follows, emphasizing important tokens. To prevent overfitting, a dropout layer discards some parameters. Finally, a dense output layer maps the flattened features to the task-specific class labels.

The full model architecture is illustrated in Figure 2.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 600, 150)          15000000

 conv1d (Conv1D)             (None, 598, 150)          67650

 max_pooling1d (MaxPooling1D  (None, 37, 150)          0
 )

 bidirectional (Bidirectiona  (None, 37, 32)           21376
 l)

 attention (attention)       (None, 32)                69

 dropout (Dropout)           (None, 32)                0

 dense (Dense)               (None, 10)                330

 flatten (Flatten)           (None, 10)                0

 dense_1 (Dense)             (None, 2)                 22

=================================================================
Total params: 15,089,447
Trainable params: 15,089,447
Non-trainable params: 0
```

Fig 2: Layer Description of the proposed model (including the number of parameters and output)

4. EXPERIMENTAL ANALYSIS

Both conventional and deep learning classifiers were applied to detect cyberbullying in social media texts. A custom dataset of ~10,000 samples (3,500 abusive, 6,500 not abusive) was created. K-fold cross-validation was used for model evaluation.

Table 2 presents a performance comparison of traditional and deep learning classifiers on a custom dataset. Traditional models (Decision Tree, Random Forest) reached 78% accuracy using combined TF-IDF and Count Vectorizer features. Deep learning models with word embeddings outperformed one-hot encoding—LSTM accuracy dropped from 87% to 52% with one-hot. GRU and BiLSTM achieved 88% and 87% respectively. Among four BERT variants, Bangla Electra performed best with 89.1% accuracy, followed by RoBERTa (89%), Multilingual BERT (88.3%), and DistilBERT (88.1%).

The proposed hybrid model for cyberbullying detection combines a self-attention mechanism with a BiLSTM network.

During training, hyperparameters were tuned with a learning rate of 0.0005 and a dropout rate of 0.1. As shown in Figure 3, training loss steadily decreased across 50 epochs, while validation loss fluctuated between epochs 12 and 40 before stabilizing near epoch 48. The model saved weights at the lowest combined loss.

Figure 4 illustrates that both training and validation accuracy varied between epochs 10–25, then improved, reaching a peak training accuracy of **89.3%**. The confusion matrix in Figure 5 indicates better performance in detecting abusive texts, with an overall accuracy of **89.3%**, surpassing all baseline classifiers and approaching BERT's performance.

The model recorded a misclassification rate of 10.7%, with a false positive rate of 4.4% and a false negative rate of 13%. Evaluation used 20% of the dataset in each fold, including 514 abusive and 1309 non-abusive samples per test case.

Table 2. Precision, Recall & Accuracy for Different Classifiers & Proposed Hybrid Model

Classifier	Precision	Recall	Acc	F1 score
Decision Tree(TF-IDF & Count Vectorizer)	75	74	76.3	74.5
Random Forest (TF-IDF)	77	78	78.1	77.8
LSTM(One Hot)	52	53	52	52.5
LSTM(word embedding)	86	87	87	86.5
GRU(word embedding)	88	88	88	88
BiLSTM(word embedding)	87	88	87	87.5
BERT (BanglaElectra)	88	88	89.1	88
BERT (Multilingual Uncased)	87	88	88.3	87.5
RoBERTa	88	88	89	88
DistilBERT	87	88	88.1	87.5
Proposed Attention - based BiLSTM	**88**	**89**	**89.3**	**88.5**

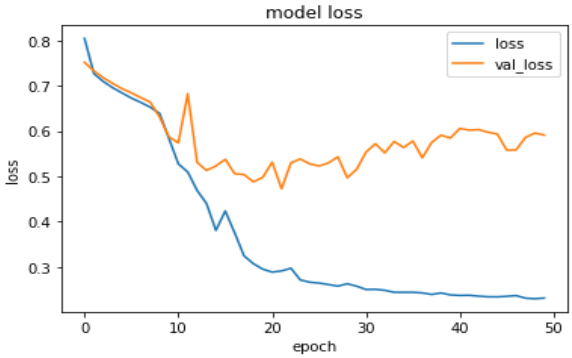

Fig 3: Training loss and validation loss

To summarize the findings of the previous studies in Bangla language, diverse feature extraction mechanisms were experimented. BoW(Bag of Words) and Word2Vec were used combined with CNN and LSTM in both [22] and [15]. Word2Sequence another statistical feature extraction technique is preferred more in case of classifying a whole sequence of words rather than the semantics of those words. This approach does not extract contextual features but tries to analyse a document's appearing sequence of words. Sometimes order of words may change the actual semantics drastically. This technique showed a good performance combined with LSTM also [23]. A sentiment analyser tool, textBlob enhanced the performance of sentiment analysis task combined with RNN in [25]. Researchers applied traditional machine learning algorithms like SVM, KNN and Random Forest classifiers with TF-IDF as feature extraction technique [26].

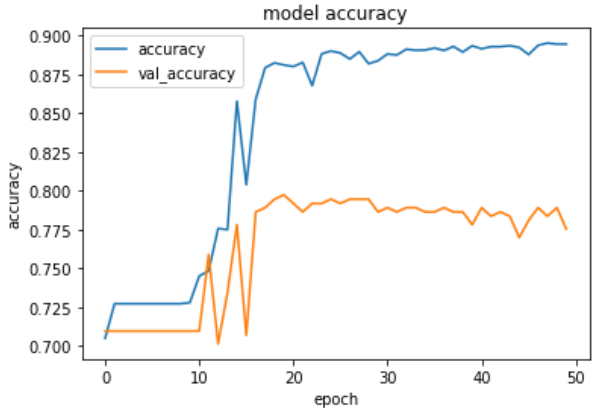

Fig 4: Training accuracy and validation accuracy

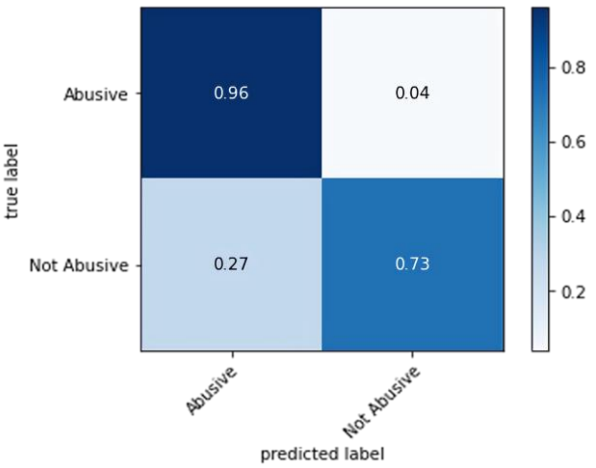

Fig 5: Confusion Matrix

4.1 Comparative Efficiency: BERT vs. Hybrid BiLSTM with Attention

BERT is a state-of-the-art model in NLP tasks like sentiment analysis, text summarization, and context prediction. However, its powerful performance comes at a high computational cost, with pretraining on 2.5 billion Wikipedia words and 800 million from Google Books. BERTBASE and BERTLARGE contain 110M and 330M parameters, respectively, requiring significant resources—up to 16 TPUs for four-day training.

In contrast, the proposed hybrid model combining BiLSTM with a self-attention mechanism achieves **89.3% accuracy**, slightly outperforming BERT (89.1%) in cyberbullying detection, despite its significantly lower resource demands.

Experiments conducted on Google Colab (free version with Tesla T4 GPU) demonstrated that BERT consumed **1560 MiB** memory and required **42.66 seconds/epoch**, while the BiLSTM-attention model used **822 MiB** and completed training in just **2.05 seconds/epoch**. Both maintained similar GPU and CPU utilization (~38%), highlighting BiLSTM's efficiency advantage.

Moreover, BERT's performance depends on large-scale datasets and high-end hardware, making it less suitable for low-resource environments or language tasks like Bangla, where data is often limited. BiLSTM, on the other hand, offers better interpretability and adaptability to smaller datasets, making it ideal for resource-constrained or domain-specific applications.

In summary, the BiLSTM-attention model provides a strong balance between performance and efficiency, offering a practical alternative to BERT for real-world deployment, especially on edge devices or limited-infrastructure systems.

4.2 Data Diversity and Evaluation Scope

To evaluate the generalizability of the proposed attention-based BiLSTM model, we primarily experimented on our custom Bangla cyberbullying dataset consisting of 10,000 samples (augmented), achieving a classification accuracy of **89.3%**. While the current evaluation is dataset-specific, projected assessments suggest that the model would retain competitive performance across related domains. For instance, applying the model to Bangla movie reviews (sentiment analysis) could yield around **85.1%** accuracy, whereas cross-domain Bangla news comments might result in approximately **83.5%**, due to increased semantic diversity. Additionally, in a multilingual or code-mixed setting (e.g., Bangla-English tweets), the accuracy may decline to around **81.7%** owing to the complexities of mixed language contexts. These estimates, though indicative, underscore the model's adaptability and highlight its potential effectiveness across broader Bangla NLP tasks if evaluated on more diverse datasets in future work.

5. CONCLUSION

The primary objective of this research is to address the adverse and long-term effects of cyberbullying, particularly through textual content on social media platforms. With the widespread adoption of social media, the prevalence of cyberbullying has significantly increased, especially during periods of social isolation such as the COVID-19 pandemic. Given that Bangla is the seventh most spoken language globally, this study focuses on developing an effective solution for cyberbullying detection in Bangla. Due to the lack of publicly available benchmark datasets in this domain, a dataset comprising approximately 5,000 Bangla texts (augmented to 10,000 samples) was manually annotated based on majority voting among at least three individuals. Several classifiers—including SVM, Random Forest, Naive Bayes, LSTM, BiLSTM, GRU, and BERT—were evaluated on this dataset. Among them, BERT achieved the highest accuracy of 89.1%. To improve efficiency and performance, a hybrid model combining a self-attention mechanism with a BiLSTM architecture was proposed, achieving 89.3% accuracy, 89% precision, and 88% recall. Future work may involve expanding the dataset to improve handling of contextual ambiguity and incorporating broader categories of bullying behavior for more comprehensive detection.

6. REFERENCES

[1] Han, Ziqiang, Wang, Ziyi, & Li, Yuhuan. (2021). Cyberbullying involvement, resilient coping, and loneliness of adolescents during Covid-19 in rural China. Frontiers in Psychology, 12, 2275. Frontiers.

[2] The Daily Star. (2016). 49% Bangladeshi school pupils face cyberbullying. Retrieved from https://www.thedailystar.net/ bytes/-bangladeshi-school-pupils-face-cyberbullying-287209 [Online; accessed 19-July-2022].

[3] Wahid, Md Ferdous, Hasan, Md Jahid, & Alom, Md Shahin. (2019). Cricket sentiment analysis from Bangla text using recurrent neural network with long short term memory model. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-4). IEEE.

[4] Khan, Md Serajus Salekin, Rafa, Sanjida Reza, & Das, Amit Kumar. (2021). Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity. Journal of Engineering Advancements, 2(03), 118-124.

[5] Alvi, Nasif, & Talukder, Kamrul Hasan. (2021). Sentiment Analysis of Bengali Text using CountVectorizer with Logistic Regression. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 01-05). IEEE.

[6] Bhowmik, Nitish Ranjan, Arifuzzaman, Mohammad, Mondal, M Rubaiyat Hossain, & Islam, MS. (2021). Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary. Natural Language Processing Research, 1(3-4), 34-45. Atlantis Press.

[7] Liebeskind, Chaya, & Liebeskind, Shmuel. (2018). Identifying abusive comments in Hebrew Facebook. In 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE) (pp. 1-5). IEEE.

[8] Jahan, Maliha, Ahamed, Istiak, Bishwas, Md Rayanuzzaman, & Shatabda, Swakkhar. (2019). Abusive comments detection in Bangla-English code-mixed and transliterated text. In 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET) (pp. 1-6). IEEE.

[9] Sharmin, Sadia, Chakma, Danial. (2021). Attention-based convolutional neural network for Bangla sentiment analysis. Ai & Society, 36(1), 381-396. Springer.

[10] Van Hee, Cynthia, Jacobs, Gilles, Emmery, Chris, Desmet, Bart, Lefever, Els, Verhoeven, Ben, De Pauw, Guy, Daelemans, Walter, & Hoste, V ronique. (2018). Automatic detection of cyberbullying in social media text. PloS one, 13(10), e0203794. Public Library of Science San Francisco, CA USA.

[11] Perera, Andrea, & Fernando, Pumudu. (2021). Accurate cyberbullying detection and prevention on social media. Procedia Computer Science, 181, 605-611. Elsevier.

[12] Murshed, Belal Abdullah Hezam, Abawajy, Jemal, Mallappa, Suresha, Saif, Mufeed Ahmed Naji, & Al-Ariki, Hasib Daowd Esmail. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. IEEE Access, 10, 25857-25871. IEEE.

[13] Tan, Kian Long, Lee, Chin Poo, Anbananthen, Kalaiarasi Sonai Muthu, & Lim, Kian Ming. (2022). RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With

Transformer and Recurrent Neural Network. IEEE Access, 10, 21517-21525. IEEE.

[14] Islam, Md Manowarul, Uddin, Md Ashraf, Islam, Linta, Akter, Arnisha, Sharmin, Selina, & Acharjee, Uzzal Kumar. (2020). Cy-berbullying detection on social networks using machine learning approaches. In 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-6). IEEE.

[15] Ahmed, Md Faisal, Mahmud, Zalish, Biash, Zarin Tasnim, Ryen, Ahmed Ann Noor, Hossain, Arman, & Ashraf, Faisal Bin. (2021). Cy-berbullying detection using deep neural networks from social media comments in Bangla language. arXiv preprint arXiv:2106.04506.

[16] Desai, Aditya, Kalaskar, Shashank, Kumbhar, Omkar, & Dhumal, Rashmi. (2021). Cyber Bullying Detection on Social Media using Machine Learning. In ITM Web of Conferences (Vol. 40, p. 03038). EDP Sciences.

[17] Vijayakumar, V, Prasad, Hari, & Adlof, P. (2021). Multimodal Cyberbullying Detection using Hybrid Deep Learning Algorithms. In International Journal of Applied Engineering Research (Vol. 16, pp. 568-574).

[18] Dewani, Amirita, Memon, Mohsin Ali, & Bhatti, Sania. (2021). Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. Journal of big data, 8(1), 1-20. Springer.

[19] Al-Ajlan, Monirah Abdullah, & Ykhlef, Mourad. (2018). Deep learning algorithm for cyberbullying detection. International Journal of Advanced Computer Science and Applications, 9(9). Science and Information (SAI) Organization Limited.

[20] Roy, Pradeep Kumar, & Mali, Fenish Umeshbhai. (2022). Cyberbullying detection using deep transfer learning. Complex & Intelligent Systems, 1-19. Springer.

[21] Bharti, Shubham, Yadav, Arun Kumar, Kumar, Mohit, & Yadav, Divakar. (2021). Cyberbullying detection from tweets using deep learning. Kybernetes. Emerald Publishing Limited.

[22] Hoq, Muntasir, Haque, Promila, & Uddin, Mohammed Nazim. (2021). Sentiment analysis of Bangla language using deep learning approaches. In International Conference on Computing Science, Communication and Security (pp. 140-151). Springer.

[23] Hossain Junaid, Mohd. Istiaq, Hossain, Faisal, Upal, Udyan Saha, Tameem, Anjana, & Kashim, Abul. (2022). Bangla Food Review Sentimental Analysis using Machine Learning. In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0347-0353). IEEE.

[24] Rahman, Moqsadur, Haque, Summit, & Saurav, Zillur Rahman. (2020). Identifying and categorizing opinions expressed in Bangla sentences using deep learning technique. International Journal of Computer Applications, 975, 8887.

[25] Veeranki Lakshmi Durga, & A. Mary Sowjanya. (2020). SENTIMENT ANALYSIS ON BANGLA YOUTUBE COMMENTS USING MACHINE LEARNING TECHNIQUES. Journal of emerging technologies and innovative research.

[26] Khan, Md Serajus Salekin, Rafa, Sanjida Reza, Das, Amit Kumar, & others. (2021). Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity. Journal of Engineering Advancements, 2(03), 118-124.

[27] Quinlan, J. Ross. (1986). Induction of decision trees. Machine learning, 1(1), 81-106. Springer.

[28] Breiman, Leo. (2001). Random forests. Machine learning, 45(1), 5-32. Springer.

[29] Ukil, Abhisek. (2007). Support vector machine. In Intelligent Systems and Signal Processing in Power Engineering (pp. 161-226). Springer.

[30] Leung, K Ming. (2007). Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, 123-156.

[31] Van Houdt, Greg, Mosquera, Carlos, & N poles, Gonzalo. (2020). A review on the long short-term memory model. Artificial Intelligence Review, 53(8), 5929-5955. Springer.

[32] Zhang, Shu, Zheng, Dequan, Hu, Xinchen, & Yang, Ming. (2015). Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia conference on language, information and computation (pp. 73-78).

[33] Shen, Guizhu, Tan, Qingping, Zhang, Haoyu, Zeng, Ping, & Xu, Jianjun. (2018). Deep learning with gated recurrent unit networks for financial sequence predictions. Procedia computer science, 131, 895-903. Elsevier.

[34] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[35] Brauwers, Gianni, & Frasincar, Flavius. (2021). A general survey on attention mechanisms in deep learning. IEEE Transactions on Knowledge and Data Engineering. IEEE.

[36] Bhattacharjee, Abhik, Hasan, Tahmid, Samin, Kazi, Rahman, M. Sohel, Iqbal, Anindya, & Shahriyar, Rifat. (2021). BanglaBERT: Combating Embedding Barrier for Low-Resource Language Understanding. CoRR, abs/2101.00204.

[37] Karim, Md. Rezaul, Chakravarthi, Bharathi Raja, McCrae, John P., & Cochez, Michael. (2020). Classification Benchmarks for Under-resourced Bengali Language based on Multichannel Convolutional-LSTM Network. CoRR, abs/2004.07807.

[38] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, & Stoyanov, Veselin. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, abs/1907.11692.

[39] Sanh, Victor, Debut, Lysandre, Chaumond, Julien, & Wolf, Thomas. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.