Machine & Deep Learning Algorithms for Bengali Cyberbullying Detection

Misbahul Sheikh
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
misbahul.cse.20200204039@aust.edu

Naushin Mamun
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
naushin.cse.20200204010@aust.edu

Nafisa Tasnim Neha
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
nafisa.cse.20200204020@aust.edu

***Abstract*—Social media has become more prevalent and it is now fairly easy to communicate with people online. Social network users have many options to cooperate, interact positively, and exchange information. The same system might create a toxic environment that can create an unpleasant environment for online abuse and bullies. Young adults and celebrities are vulnerable to online abuse more often. That's why cyberbullying should be identified and eliminated from social media because it may significantly lead to psychological as well as emotional suffering. By utilizing Natural Language Processing (NLP), Machine Learning (ML) algorithms like K-Nearest Neighbour (K-NN), Logistic Regression(LR), and Random Forest(RF), as well as Deep Learning Models like LSTM(Long short-term memory), CNN(Convolutional Neural Network), and other models based on Transformers like BERT, we can identify patterns in social media texts used by bullies and create an automated method that can detect abusive texts. In this study, we proposed a reliable machine-learning model for social media cyberbullying detection in the Bengali language. We applied text preprocessing, followed by 7 different feature extraction techniques. Then, we applied 3 ML algorithms on all the feature extraction techniques, 2 DL algorithms, and a transformer-based pre-trained BERT model and evaluated their performances by different performance metrics. Our study found that BERT worked best compared to other algorithms and achieved an accuracy of 89%.**

***Keywords*— Cyberbullying, Bangla bullying detection, Hate speech detection, NLP, Machine learning, Deep learning, BERT.**

I. Introduction

Bullying is a type of aggression that inflicts victims with either immediate or long-term harm or distress. These aggressions include physical such as hitting, and tripping, verbal for example teasing, emotional, and social such as spreading false allegations. As a result, it may pose a severe public health risk. Also, according to the Centers for Disease Control and Prevention (CDC), bullying is defined as any unwelcome hostile behavior(s) by another person or group of people that involves an imbalance of power, and is repeated or is very likely to repeat. In recent times, as a result of advancements in information technology and easy access to digital devices to people, many forms of cyber security threats have surfaced and researchers are trying to mitigate the problems by incorporating different security measures. Moreover, a new form of bullying has emerged in cyberspace dubbed as "cyberbullying." Cyberbullying can be defined as a sort of psychological harassment that takes place through the use of technology including smartphones, blogs, and Social networking sites such as Facebook, YouTube, and Twitter. This type of abuse can happen in different ways like sending threatening personal messages, commenting abusively on social media posts, and spreading rumors by manipulating social media posts. Some people consider cyberbullying as an extension of traditional harassment, however, cyberbullying differs from traditional bullying in several concerning ways. Unlink traditional bullying, a major part of cyberbullying is anonymous. In most of the cases, the perpetrator is unknown to the victim and also to the public. Furthermore, rumors can be spread to a larger audience with remarkable speed. This aspect of cyberbullying makes it incredibly challenging to regulate or deal with.

Over the past decades, Bangladesh has experienced a steady rise in internet usage. However, it is also thought to be contributing to an increase in harassment of women because of society's predominance of patriarchal views and customs as well as the lack of proper legal protection. In a poll conducted by ActionAid Bangladesh, 50% of the women who participated in the study reported experiencing internet harassment. More than 62% of the victims were under the age of 25. It's interesting to note that the victims identified Facebook as the main website where they experienced the most harassment. As a result of these circumstances, 76% of women had mental health issues like anxiety and sadness. About 30% of women were unaware of their options for filing complaints.

Researchers have been trying to build automated systems to solve real-life problems. A lot of research has been done to incorporate the power of AI to develop an automated system that can classify texts to find offensive content on social media platforms in both English and other languages. Also, there are several research that focus on the Bengali language to identify the abusive language because the detection of cyberbullying in the Bengali language is of the utmost importance. Across a large portion of the world, Bengali is the most commonly used language for communication. However, there is a critical need for an enhanced approach that can recognize different forms of offensive Bengali content.

To find and delete offensive Bengali content from social media networks, ML techniques can be very successful. In our study. We proposed an ML technique for effective cyberbullying classification in the Bengali language and to protect users from harassment on social media. Through focusing on the distinct Bengali language context, this study aims to progress the field of Bengali cyberbullying detection. In our study, we explored two machine learning methods, such as Logistic Regression(LR), and Random Forest(RF), 2 deep learning methods such as LSTM(Long short-term memory), and CNN(Convolutional Neural Network) and 1 transformer-based BERT (Bidirectional Encoder Representations from Transformers) model "Bangla-bert-base". We used 7 word embedding techniques such as Term frequency-inverse document frequency (TF-IDF), Bag of Words(Uni-gram, Bi-gram, Tri-gram), Word2Vec, GloVe, and FastText. Findings showed that RF models with the FastText embedding technique performed better than other feature extraction-based models. The deep learning model LSTM gives a better result than any other ML models. Although Bangla-bert-base outperformed all ML, and DL algorithms explored in this work.

II. LITERATURE REVIEW

Multiple studies have been done on detecting cyberbullying using machine learning techniques. Some of the notable works in other languages are discussed here.

The earliest work was done by Reynolds et al. [1]. They collected their data from Formspring.me which was a website focused on queries and answering with a significant number of comments related to bullying. Then labeled the dataset using Amazon's web service is called Mechanical Turk. Later Weka tool was used to train and classify bullying texts. Finally, they successfully achieved 78.5% accuracy by applying the C4.5 algorithm and an instance-based learner.

Hani et al. [2] explored a supervised ML approach to detect a pattern in cyberbullying texts. In their study, they showed that NN (Neural Network) performed better than SVM in abusive text classification with an accuracy of 92.8% compared to SVM (90.3% accuracy).

A study [3] evaluates various machine learning and a transformer-based pre-trained Bangla-Bert model for detecting Cyberbullying from social media platforms. They have used the TF-IDF feature extraction technique and applied 4 ML models such as such as Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and XGBoost (Extreme Gradient Boosting). All the ML algorithms (except BERT) give almost similar accuracy scores ranging from 76% to 79%. However, BERT significantly improves the performance with an accuracy of 90%. The second-best accuracy is demonstrated by SVM (79% accuracy) and the lowest accuracy is achieved by NB (76% accuracy).

Dalvi et al. [4] investigated a software-based approach to identify abusive tweets. They used SVM and NB algorithms and obtained about 71.25% and 52.70% accuracy respectively in identifying bullied texts.

Saha et al. [5] investigated text sentiments from Twitter data using two different techniques named VADER and BERT. Later they surveyed 5 different ML algorithms for detection accuracy and achieved 92% accuracy using BERT.

In one study, methods based on deep learning (DL) as well as ML were utilized to identify abusive texts by Emon et al. [6]. They collected Bengali comments from several social media sites, online blogs, and newspapers and showed that the deep learning technique using Recurrent Neural Networks (RNN) performs better than other ML algorithms by achieving an accuracy score of 82.20%. They also proposed a new stemming technique for the Bengali language.

III. METHODOLOGY

The study consists of 5 parts. First dataset collection and preparation, followed by pre-processing, then feature extraction, and later application of ML algorithms, and finally evaluation of their performances. The whole proposed methodology is shown in Fig. 5 with a block diagram.

A. *Dataset collection and preparation*

The dataset used in this study includes people's opinions from the comments section in social media posts made by celebrities found on Facebook, including television and movie actors, models, sports figures, musicians, and politicians. In total, 44001 comments were collected for this dataset. The dataset contains a total of 5 variables (columns) namely comment, Category, Gender, comment react number, and label, here label is the target variable. From the pie charts, there were about 31.94% of remarks were aimed at males and 68.06% of comments were directed at females as shown in Fig. 1. Moreover, Fig. 2 demonstrates, that comments were categorized into 5 types such as not bully, troll, sexual, religious, and threat, and their percentages are 34.86%, 23.78%, 20.29%, 17.22%, and 3.85% respectively.

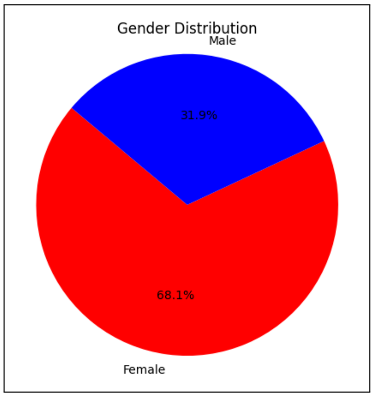


Fig. 1. Gender Comment Distribution

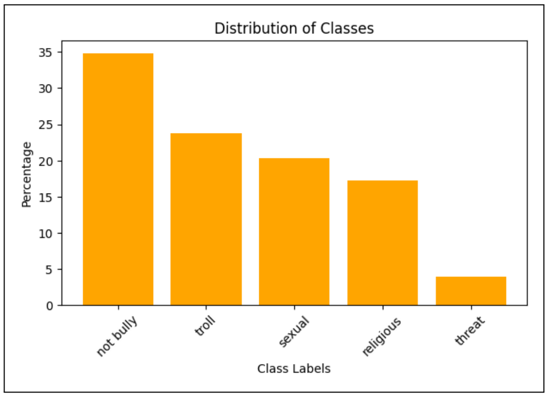


Fig. 2. Distribution of Classes

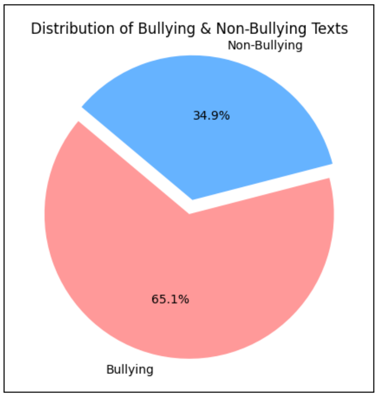
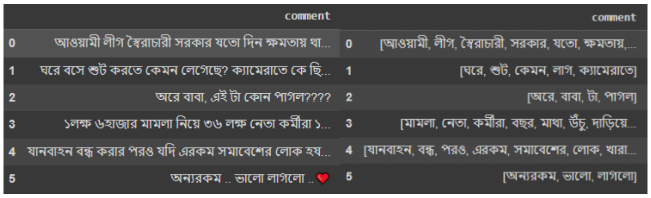


Fig. 4. Binary Class Distribution

B. *Text preprocessing*

Making sure that the data can be understood by machines is a vital part of data analysis. That's why, before applying any kind of ML, and DL algorithm data needs to be processed through filtering and tokenization. Moreover, normal texts contain many kinds of symbols and words other than our desired language, so filtering is needed. In our research, by filtering techniques, we removed all kinds of digits, website links, punctuation, and symbols, signs, emoticons, any characters other than Bengali. Followed by the elimination of Bengali Stopwords. Tokenization is dividing each text into smaller words based on a delimiter (space). These words or smaller units are called tokens. We also apply lemmatization in the comment column. Lemmatizer reduce words to their base or root form thus normalizing the dataset. A sample from the dataset after filtering is depicted in Fig 3.



	comment		comment
0	আওয়ামী লীগ স্বৈরাচারী সরকার যতো দিন ক্ষমতায় থা...	0	[আওয়ামী, লীগ, স্বৈরাচারী, সরকার, যতো, ক্ষমতায়,...
1	ঘরে বসে শুট করতে কেমন লেগেছে? ক্যামেরাতে কে ছি...	1	[ঘরে, শুট, কেমন, লাগ, ক্যামেরাতে]
2	অরে বাবা, এই টা কোন পাগল????	2	[অরে, বাবা, টা, পাগল]
3	১লক্ষ ৬হাজার মামলা নিয়ে ৩৬ লক্ষ নেতা কর্মীরা ১...	3	[মামলা, নেতা, কর্মীরা, বছর, মাথা, উঁচু, দাড়িয়ে...
4	যানবাহন বন্ধ করার পরও যদি এরকম সমাবেশের লোক হয...	4	[যানবাহন, বন্ধ, পরও, এরকম, সমাবেশের, লোক, খারা...
5	অন্যরকম .. ভালো লাগলো ..❤	5	[অন্যরকম, ভালো, লাগলো]

Fig. 3. Preprocessed Texts

C. *Conversion to binary class from multi-class*

For our experiment, we converted our dataset from multi-class to binary class considering troll, sexual, religious, and threat texts as bullying text, on the other hand, not-bully comments were kept unchanged. For this purpose troll, sexual, religious, and threat texts are represented as 1 (bully) and neutral as 0 (non-bully), at this point, the dataset contains 28661 bullying texts (1) and 15340 non-bullying texts (0). See the binary class distribution in Fig 4.

D. *Feature extraction*

Feature extraction is implemented because text data needs to be handled to train ML models. These methods are employed to represent the words numerically. There are several techniques used by researchers. In our study, we have used 7 different word embedding techniques such as Term frequency-inverse document frequency (TF-IDF), Bag of Words(Uni-gram, Bi-gram, Tri-gram), Word2Vec, GloVe, and FastText.

*1) **TF-IDF**:* Term frequency-Inverse document frequency vectorizer (TF-IDF) is a vectorization technique that converts text data into vectors. From raw texts, it creates a matrix of features. Term Frequency (TF): TF counts the number of times a term appears in a document. TF is calculated using the equation 1.

$$\text{TF} = \frac{\text{num. of occurrences of a term } t \text{ in the document}}{\text{total words in the document}} \tag{1}$$

Inverse Document Frequency (IDF): This is a weight that is an indicator of how frequently a term is used in a document. Its score decreases with increased usage across a document and scales up the less frequent words using equation 2.

$$\text{IDF} = \log\left(\frac{N}{DFt}\right) \tag{2}$$

Here, N is the total number of text documents and DFt is the number of texts that use the term t. TF-IDF: It is the product of IDF and TF as shown in equation 3.

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \tag{3}$$

*2) **Bag of words**:*

- Uni-gram, represents individual words in a text corpus without considering neighboring words.
- BI-gram, involves pairs of consecutive words in a text, capturing some contextual information.
- TRI-gram, considers sequences of three consecutive words, providing even more context than BI-gram.

*3) **Word2Vec**:* Word2Vec is a popular technique in natural language processing (NLP) used for learning word embeddings, which are numerical representations of words in a continuous vector space. The key idea behind Word2Vec is to learn distributed representations of words based on their contextual usage in a corpus of text.

$$\max \prod_{t=1}^{T} \prod_{-c \leq j \leq c, j \neq 0} P(w_{t+j}|w_t) \tag{4}$$

where:

- T is the total number of words in the corpus.
- c is the context window size.
- w_t represents the current word at position t.
- w_{t+j} represents the context word at position $t+j$ within the window around w_t.
- $P(w_{t+j}|w_t)$ is the conditional probability of observing context word w_{t+j} given the current word w_t.

*4) **GloVe**:* GloVe is a word embedding technique designed to capture global statistical information about word co-occurrences in a corpus. Unlike methods like Word2Vec which focus on local context (e.g., nearby words in a sentence), GloVe constructs word vectors by leveraging the global statistical information of how frequently words co-occur across the entire corpus. This approach allows GloVe to generate embeddings that encode semantic relationships and analogies between words effectively.

$$\sum_{i,j=1}^{V} f(P_{ij})(\mathbf{w}_i^T \mathbf{w}_j + b_i + b_j - \log P_{ij})^2 \tag{5}$$

where:

- $\mathbf{w}_i$ and $\mathbf{w}_j$ are word vectors,
- b_i and b_j are bias terms,
- P_{ij} is the co-occurrence probability of word j given word i,
- f is a weighting function used to emphasize informative co-occurrences.

*5) **FastText**:* FastText represents words as bags of character n-grams, where each word is represented as a sum of the vector representations of its character n-grams. This allows it to generate embeddings not just for words present in the training corpus, but also for unseen words that share common character n-grams with the known words.

$$\mathbf{v}_w = \sum_{g \in \mathrm{G}(w)} \mathbf{z}_g \tag{6}$$

where:

- $\mathrm{G}(w)$ denotes the set of character n-grams (including the word itself) of w.
- $\mathbf{z}_g$ represents the vector representation of character n-gram g.

E. Bangla-Bert

Bidirectional Encoder Representations from Transformers, which is termed as BERT, is a DL model that is based on Transformers. Every output element in a transformer is connected to every input, and attention-based dynamic weighting determines the relative importance of each element. The difference between BERT and previous language models is that BERT can simultaneously read texts in both directions contrary to other language models which could read text inputs in one direction only. BERT is a full language model that uses an embedding method as one of its constituent parts, not just an embedding technique.

Here, bangla-bert-base is a pre-trained model for the bengali language utilizing mask language modeling which is detailed in BERT. Two primary sources were used to download Corpus which are Open Super-large Crawled Aggregated coRpus (OSCAR) and Bengali Wikipedia. Google BERT's code was used to train bangla-bert and the latest model contains 12 layers, 768 hidden layers, and 110 million parameters in its architecture.

We used the "simple transformer" NLP library to implement BERT and "Classification Model" as simple transformer model for binary classification.

F. Classification

We tried 3 different ML classifiers (K-NN, LR, and RF), 2 DL classifiers (LSTM, and CNN) and 1 BERT pre-trained model called bangla-bert-base to train our dataset and classify Bengali cyberbullying texts.

Later we evaluated and compared their effectiveness considering different performance indicators.

G. Performance metrics

Evaluation metrics in machine learning are crucial for assessing the performance of a model. They help quantify how well a model is performing based on the predictions it makes compared to actual values.

*1) **Accuracy**:* This metric measures the proportion of correct predictions among the total number of predictions made. It's suitable when the classes in the dataset are balanced.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

*2) **Precision**:* Precision measures the proportion of true positive predictions (correctly predicted positive instances) among all positive predictions made. It focuses on the accuracy of positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

*3) **Recall**:* Measures the proportion of true positive predictions among all actual positive instances. Focuses on how well the model can find all positive instances.

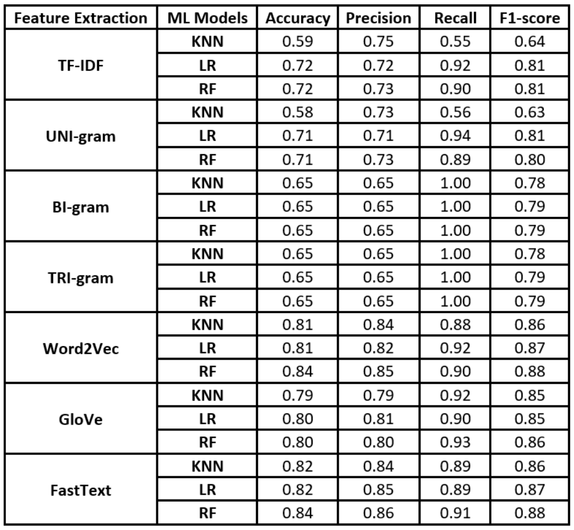
$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

*4) **F1-score**:* A single metric that combines precision and recall using the harmonic mean. F-measure with equal importance to precision and recall is denoted as F1-score.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

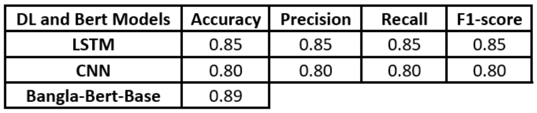
IV. RESULT ANALYSIS

The performance of our cyberbullying detection models was evaluated using key metrics such as accuracy, precision, recall, and f1-score.



Feature Extraction	ML Models	Accuracy	Precision	Recall	F1-score
TF-IDF	KNN	0.59	0.75	0.55	0.64
	LR	0.72	0.72	0.92	0.81
	RF	0.72	0.73	0.90	0.81
UNI-gram	KNN	0.58	0.73	0.56	0.63
	LR	0.71	0.71	0.94	0.81
	RF	0.71	0.73	0.89	0.80
BI-gram	KNN	0.65	0.65	1.00	0.78
	LR	0.65	0.65	1.00	0.79
	RF	0.65	0.65	1.00	0.79
TRI-gram	KNN	0.65	0.65	1.00	0.78
	LR	0.65	0.65	1.00	0.79
	RF	0.65	0.65	1.00	0.79
Word2Vec	KNN	0.81	0.84	0.88	0.86
	LR	0.81	0.82	0.92	0.87
	RF	0.84	0.85	0.90	0.88
GloVe	KNN	0.79	0.79	0.92	0.85
	LR	0.80	0.81	0.90	0.85
	RF	0.80	0.80	0.93	0.86
FastText	KNN	0.82	0.84	0.89	0.86
	LR	0.82	0.85	0.89	0.87
	RF	0.84	0.86	0.91	0.88

Fig. 5. ML Result Table



DL and Bert Models	Accuracy	Precision	Recall	F1-score
LSTM	0.85	0.85	0.85	0.85
CNN	0.80	0.80	0.80	0.80
Bangla-Bert-Base	0.89			

Fig. 6. DL and Bert Result Table

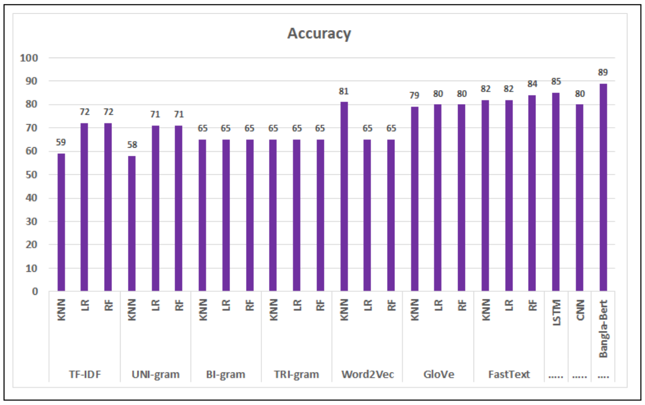


Fig. 7. Accuracy Graph (%)

Our study found that Bangla-BERT worked best compared to other algorithms and achieved an accuracy of 89%. RF model with the FastText embedding technique performed better than other feature extraction-based models. RF with FastText embedding technique gives an accuracy of 84%. Its precision, recall, and F1 scores are 0.86, 0.91, and 0.88 respectively. LSTM gives better results than CNN and other ML models. LSTM gives an accuracy of 85%. Models with BI-gram and TRI-gram provide almost the same results.

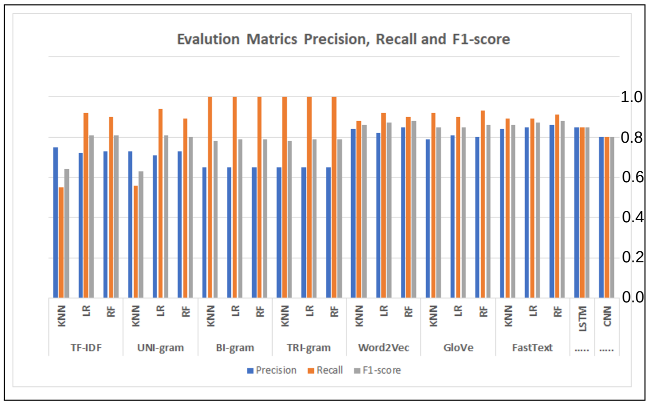


Fig. 8. Classifiers Result Graph

V. CONCLUSION AND FUTURE WORKS

The study demonstrates that advanced models, especially the Bangla-BERT, significantly enhance the detection of Bengali cyberbullying, achieving the highest accuracy of 89%, followed by the Random Forest model with FastText embeddings and the Long Short-Term Memory model. Future research should focus on expanding the dataset to include more diverse and varied texts, exploring other state-of-the-art transformer-based models, developing real-time detection systems for social media platforms, integrating multilingual capabilities, and incorporating user feedback mechanisms to continuously improve the model's accuracy and relevance.

REFERENCES

[1] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011, 2011, vol. 2, pp. 241–244, doi: 10.1109/ICMLA.2011.152.

[2] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, pp. 703–707, 2019, doi: 10.14569/ijacsa.2019.0100587.

[3] Subrata Saha, Md. Shamimul Islam, Md. Mahbub Alam, Md. Motinur Rahman, Md. Ziaul Hasan Majumder, Md. Shah Alam and M. Khalid Hossain, "Bengali Cyberbullying Detection in Social Media Using Machine Learning Algorithms," 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI), 979-8-3503-9431-3/23/$31.00 ©2023 IEEE, DOI: 10.1109/STI59863.2023.10464740

[4] R. R. Dalvi et al., "Detecting A Twitter Cyberbullying Using Machine Learning," Proc. Int. Conf. Intell. Comput. Control Syst. ICICCS 2020, pp. 297–301, May 2020, doi: 10.1109/ICICCS48265.2020.9120893.

[5] S. Saha, M. I. H. Showrov, M. M. Rahman, and M. Z. H. Majumder, "VADER vs. BERT: A Comparative Performance Analysis for Sentiment on Coronavirus Outbreak," 2023, pp. 371–385, doi: 10.1007/978-3-031-34619-4 30/COVER.

[6] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das, and T. Mittra, "A Deep Learning Approach to Detect Abusive Bengali Text," Jun. 2019, doi: 10.1109/ICSCC.2019.8843606.