Natural Language Processing Journal 4 (2023) 100027

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

A robust hybrid machine learning model for Bengali cyber bullying detection in social media

Arnisha Akhter[a], Uzzal Kumar Acharjee[a,*], Md. Alamin Talukder[a], Md. Manowarul Islam[a,*], Md Ashraf Uddin[b]

[a] Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh
[b] School of Information Technology, Deakin University, Waurn Ponds Campus, Geelong, Australia

ARTICLE INFO

Keywords:
Cyberbully
Bengali
Bullying
Machine learning
Harassment
Language
Social media

ABSTRACT

Social networking platforms give users countless opportunities to share information, collaborate, and communicate positively. The same platform can be extended to a fabricated and poisonous atmosphere that gives an impersonal, harmful platform for online misuse and assault. Cyberstalking is when someone uses an internet system to ridicule, torment, insult, criticize, slander, and discredit a victim while never seeing them. With the growth of social networks, Facebook has become the online arena for bullying. Since the effects could result in a widespread contagion, it is vital to have models and mechanisms in place for the automatic identification and removal of internet cyberbullying data. This paper presents a robust hybrid ML model for cyberbullying detection in the Bengali language on social media. The Bengalibullying proposal involves an effective text preprocessing to make the Bengali text data into a useful text format, feature extraction using the TfidfVectorizer (TFID) to get the beneficial information of text data and resampling by Instance Hardness Threshold (IHT) procedure to balance the dataset to avoid overfitting or underfitting problems. In our experiment, we used the publicly available Bangla text dataset (44,001 comments) and got the highest performance ever published works on it. The model achieved the most elevated accuracy rate of 98.57% and 98.82% in binary and multilabel classification to detect cyberbullying on social media in the Bengali language. Our best performance findings are more effective than any previous effort in identifying and categorizing bullying in the Bengali language. As a result, we might use our model to correctly classify Bengali bullying in online bullying detection systems, protecting people from being the targets of social bullying.

1. Introduction

Social media platforms serve as dynamic mediums for exchanging vital facts, opportunities, stories, and encouraging ideas. However, one of the significant challenges associated with social media is the prevalence of threatening and abusive language, commonly known as cyberbullying (Chakraborty and Seddiqui, 2019). Among these platforms, Facebook stands out as the most popular platform for human connection, boasting approximately 2.27 billion active users. Through various features provided by Facebook, users engage in conversations, debates, and share their thoughts with others and communities. Unfortunately, these interactions often devolve into contempt for opposing viewpoints, leading to insulting or hateful comments targeted at specific individuals, groups, cultures, or religions (Ishmam and Sharmin, 2019). Such communication can be construed as abusive or threatening when it employs sexist or racist slurs, criticizes or attacks specific cultures or religious beliefs, incites criminal activity, and more.

The prevalence of cyberbullying is on the rise as social media platforms continue to expand (Kee et al., 2022). Within social networks, individuals frequently experience harassment from strangers and unauthorized users (Das et al., 2019). Recent surveys have indicated that women constitute 73.71% of cyberbullying victims (Akter et al., 2017). This egregious violation of human rights manifests in various forms, including trolling, harassment, revenge porn, stalking, and cybersex. Women are particularly targeted by unwanted and often explicit sexual solicitations and defamatory online communications from anonymous and unreliable sources (Powell et al., 2017). Instances of fraud, sexual content, sexist remarks, and lewd propositions have become commonplace on social media, alongside the circulation of fabricated and edited images of naked women (McGlotten, 2013). Tragically, some individuals have even taken their own lives due to hostile remarks and online bullying. In certain cases, offensive posts have incited individuals to commit crimes and engage in other wrongful activities (Emon et al., 2019).

* Corresponding authors.
E-mail addresses: arnisha@cse.jnu.ac.bd (A. Akhter), uzzal@cse.jnu.ac.bd (U.K. Acharjee), alamintalukder.cse.jnu@gmail.com (M.A. Talukder), manowar@cse.jnu.ac.bd (M.M. Islam), ashraf.uddin@deakin.edu.au (M.A. Uddin).

https://doi.org/10.1016/j.nlp.2023.100027
Received 9 April 2023; Received in revised form 12 July 2023; Accepted 13 July 2023

Cyberstalking and cyberbullying have both physical and mental impacts on individuals. Abusers take advantage of the anonymity provided by social media platforms, allowing their cruel behavior to go unchecked. Furthermore, as harassment becomes more frequent over time, the situation worsens (Dalvi et al., 2020). People utilize social media platforms such as Facebook, Twitter, Instagram, and others to express their thoughts, sentiments, and emotions. As social media continues to grow, online harassment, blackmail, and other forms of cyber oppression are also escalating rapidly. These incidents often occur as a result of individuals sharing offensive photographs, making derogatory remarks, and sending abusive messages (Hussain et al., 2018). Teenagers exposed to abusive language experience devastating consequences that may even lead to suicide. Bullying victims are reported to have a 2–9 times higher risk of suicide compared to non-victims (Shireen et al., 2014).

Numerous studies have focused on text categorization in English and other languages to identify offensive messages, comments, or photos (Mahmud et al., 2023; Ahmed et al., 2022a; Emon et al., 2022; Shakambhari et al., 2022; Raj et al., 2022; Kumar and Sachdeva, 2022; Islam et al., 2020). Bengali is widely spoken as a primary language of communication worldwide, ranking seventh among the top 100 most spoken languages globally. With over 210 million Bengali speakers, primarily residing in the Indian states of West Bengal, Assam, and Tripura, as well as significant immigrant populations in the United Kingdom, the United States, and the Middle East (Sen et al., 2022), the detection of cyberbullying in the Bengali language is of utmost importance. While several papers have been published on cyberbullying detection in Bengali and English languages (Mahmud et al., 2023; Ahmed et al., 2022a; Emon et al., 2022; Shakambhari et al., 2022; Raj et al., 2022; Kumar and Sachdeva, 2022; Islam et al., 2020), there is a limited number of studies specifically addressing Bengali language bullying detection (Emon et al., 2022; Aurpa et al., 2022; Ahmed et al., 2022a; Shakambhari et al., 2022). Therefore, this study presents an opportunity to contribute to the detection of Bengali cyberbullying in order to safeguard individuals from such abuse. Despite the existence of different approaches in the literature for detecting cyberbullying in the Bengali language (Ahmed et al., 2022a; Mahmud et al., 2023; Emon et al., 2022; Aurpa et al., 2022), their performance rates in bullying detection are not satisfactory enough to protect users from Bengali cyberbullying. Consequently, there is a significant need for an improved strategy capable of identifying various types of abusive Bengali content. Machine learning (ML) approaches can be highly effective in identifying and removing abusive Bengali content from social media platforms.

To address the classification of cyberbullying in the Bengali language and protect users from bullying on social media, we propose a hybrid ML approach for efficient cyberbullying detection. Our proposed methodology encompasses text preprocessing to prepare the dataset, feature extraction to derive important information from the text data, resampling techniques to balance the data and mitigate overfitting or underfitting issues, and the application of several ML algorithms such as Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Multilayer Perceptron (MLP). The performance of these models will be evaluated to identify the best approach for accurately detecting Bengali cyberbullying on social media platforms.

In light of the above, this study aims to contribute to the field of cyberbullying detection by focusing on the specific context of the Bengali language. By developing an effective ML-based approach, we hope to mitigate the detrimental effects of cyberbullying and promote a safer online environment for all users.

The key contributions of this paper are as follows:

- We proposed a robust hybrid ML model for cyberbullying detection in the Bengali language on social media.
- The procedure involves effective text preprocessing to make the Bengali text data into a useful text format, feature extraction using the TfidfVectorizer (TFID) to get the beneficial information of text data and resampling by Instance Hardness Threshold (IHT) procedure to balance the dataset to avoid overfitting or underfitting problems.
- Finally, apply ML algorithms to evaluate the performance and find out the best ML model to detect Bengalibullying and provide a comparative analysis to prove the effectiveness of our ML model.

1.1. Research questions

This study seeks to address the following research questions:

1. What is the prevalence of cyberbullying in the Bengali language on social media platforms, particularly Facebook?
2. What are the physical and mental impacts of cyberbullying on individuals, and how do these impacts differ based on gender?
3. What are the existing approaches and models for cyberbullying detection in Bengali, and what are their limitations in terms of performance?
4. How can a hybrid machine learning (ML) approach be developed to improve the detection and classification of cyberbullying in Bengali?
5. How does the proposed ML model compare to existing approaches in terms of performance and effectiveness?

These research questions will guide the study and provide a comprehensive understanding of the prevalence and impact of cyberbullying in the Bengali language, the limitations of existing approaches, and the potential of a hybrid ML approach to enhance cyberbullying detection.

The remainder of the essay is structured as follows. In Section 2, we examine the relevant articles for identifying Bengalibully. We outline our approach and several ML algorithms in Section 3. Section 4 presents the experimental findings. After that, in Section 5, we give the findings and our analysis, and finally, in Section 6, we draw to a close by discussing how this study may be improved in the future.

2. Related works

Several studies have been conducted in the field of cyberbullying detection, particularly focusing on the Bengali language. These studies provide a strong justification for the current research, demonstrating the need for effective methods to address cyberbullying in social media platforms.

Dalvi et al. (2020) developed an ML model for detecting and preventing bullying on Twitter. Their model utilized Support Vector Machine (SVM) and Naive Bayes (NB) classifiers to identify bullied tweets and postings. SVM achieved an accuracy level of 71.25%, while NB achieved 52.70% accuracy in identifying true positives.

Emon et al. (2019) proposed a methodology for identifying abusive content in open comment pages of social media websites. Their study focused on improving online forum safety by detecting various forms of abusive Bengali text. They explored ML and deep learning (DL) models, with Recurrent Neural Network (RNN) achieving the highest accuracy rate of 82.20%.

Ahmed et al. (2021b) presented a binary and multilabel classifier model for detecting bully expressions on Facebook pages. Their study analyzed 44,001 consumer reviews from popular public Facebook pages, categorized into non-bully, sexual, threat, troll, and religious categories. Their NN + Ensemble technique achieved an accuracy of 87.91% for binary classification and 85% accuracy for multilabel classification.

Chakraborty and Seddiqui (2019) developed an ML model utilizing Natural Language Processing (NLP) methods to identify bullying in the Bengali language on Facebook. They focused on spotting Unicode emoticons and Unicode Bengali symbols as input to detect offensive content. Their SVM model achieved an accuracy rate of 78%.

Ishmam and Sharmin (2019) created a Gated Recurrent Unit (GRU) model for categorizing user reviews on Facebook pages. Their study collected 5126 Bengali opinions and categorized them into hate speech, racial and religious intolerance, incitement, political commentary, and religious commentary. The GRU model achieved an accuracy of 70.10% in hate speech recognition.

Akhter et al. (2018) advocated using ML algorithms for detecting online bullying in Bangla text. They trained various ML-based classification algorithms using a collection of Bangla text from social media sites and achieved a detection rate of 97% using the support vector machine (SVM) model.

Ahmed et al. (2021c) developed an ML and DL model for identifying cyberbullying in texts written in Bangla and Romanized Bangla. Their study created three datasets from social media: one for Bangla, one for Romanized Bangla, and a combined dataset. The ML algorithm Multinomial Naive Bayes (MNB) achieved an accuracy rate of 80% in the combined dataset.

Mahmud et al. (2023) proposed a methodology to recognize abusive language in Bangla using ML algorithms. They used annotated translated Bengali corpora and achieved a 97% accuracy rate using logistic regression (LR) to identify Bengalibullying.

Hussain et al. (2020) conducted an investigation assessment on the identification of fake news in Bangla from social media. They utilized ML algorithms MNB and SVM, as well as feature extraction tools, to identify fake news in Bangla. SVM achieved an accuracy rate of 96.64% using the linear kernel, outperforming MNB's accuracy of 93.32%.

Emon et al. (2022) proposed a technique for identifying cyberbullying on social media in the Bengali language. They employed various transformer models, including Bangla BERT, Bengali DistilBERT, and XLM-RoBERTa, on a dataset of 44,001 Bangla comments from Facebook. The XLM-RoBERTa model achieved the highest accuracy rate of 85% and F1-score of 86% among the models.

Aurpa et al. (2022) focused on recognizing offensive comments in Bengali on Facebook. They utilized transformer-based deep neural network models, including BERT (Bidirectional Encoder Representations from Transformers) and ELECTRA, on a dataset of 44,001 comments. The BERT model achieved an accuracy rate of 85.00%, while the ELECTRA model achieved 84.92% accuracy.

Chakravarthi (2022) proposed a customized deep network model leveraging multilingual data to identify and encourage positivity in comments. Their model, using a combination of embedding from T5-Sentence, achieved macro F1 scores of 75% for English, 62% for Tamil, and 67% for Malayalam using the CNN model.

Romim et al. (2021) offered a hate speech (HS) dataset in Bengali called HS-BAN, comprising over 50,000 annotated remarks. They explored linguistic traits and neural network-based techniques to create a standard hate speech detection mechanism for Bengali. Their benchmark showed that word embedding models trained on informal texts outperformed those trained on formal texts, with a Bi-LSTM model achieving an F1-score of 86.78% on top of FastText casual word embedding.

These studies collectively provide strong justification for the current research on cyberbullying detection in Bengali text. They demonstrate the effectiveness of various ML, DL, NLP, and transformer-based models in detecting and categorizing abusive and bullying content.

The summary of the previous studies is shown in Table 1 on cyberbullying detection, abusive content identification, and hate speech recognition in various languages, including Bengali, have contributed valuable insights. However, these studies have certain limitations. Some achieved lower accuracy rates, highlighting the need for improvement, while others focused on specific languages and text types, limiting generalizability. Specific ML algorithms and techniques were often relied upon, calling for exploration of a wider range of approaches. Limited dataset categories and insufficient justification for classification decisions were also observed. To address these gaps, our study aims to develop a robust hybrid machine-learning approach that encompasses Bengali languages, text types, and abusive content categories. We will explore various ML algorithms, provide detailed justifications for classifications, and contribute to the advancement of cyberbullying detection and prevention strategies.

3. Methodology

In this section, we describe our proposed methodology as well as the various machine algorithms used in the scheme. To begin, we will explain how the proposal works. The machine learning models are then briefly described.

Fig. 1 shows the workflow of our proposal. The proposal has six significant parts namely Bengalibullying text data collection, text preprocessing to prepare the data, feature extraction to transform it into numerical information of text data, resampling to balance the data, utilizing k-fold cross-validation to split the data, applying ML algorithms to build the models and finally evaluating the performance using various performance metrics. we describe the workflow of each part briefly as follows:

3.1. Data collection

In our experiment, we use a public Bengali cyberbully text dataset (Ahmed et al., 2021a). The data utilized in this study include opinions from the interactivity section under postings made by actors, musicians, politicians, athletes, and other public figures on the Facebook platform. There have been 44,001 responses in the count. Our analysis shows that 31.9% of remarks are directed at male victims and 68.1% are directed at female victims. In addition, 21.31% of comments seek victims who were social figures, 5.98% of target politicians, 4.68% of target athletes, 6.78% of target singers, and 61.25% of target actors. The distribution of this dataset is depicted in Figs. 2 and 3.

Following is a description of the dataset's labels.

- Non-bully: Remarks lack a personal attack intent against the target of the message.
- Sexual: Remarks that harass or transmit sexually offensive sentiments against another person.
- Threat: Responses posted by users that make threats to hurt or kill someone else belong to the threat class.
- Troll: Purposefully hurt another individual with their comments.
- Religious: Opinions that use derogatory language or incite hatred against different religious communities.

3.2. Data preprocessing

Data preprocessing is the method by which raw data is prepared for the model's fit. It is one of the most important steps while developing a model. In our text preprocessing, we handle null values replacing with mode as its categorical feature, all the web links, punctuation marks and special sign. Then we perform tokenized the text and remove all Bengali stop words as well as all the frequent words. Later we applied to lemmatize to retain the actual words of text. Furthermore, we handle the banglish words which is a mix of Bengali and English words using detecting the Bangla words followed by separating the Bengali and English words and converting the English to Bengali words and joining the processed tokens. This preprocessing aids in lowering the input dimension, preserving valuable information, and preparing the data for model construction with minimal complexity (Denny and Spirling, 2018). The original and processed text data is shown in the following Fig. 4.

Table 1
Summary of previous studies.

Research issue	Technologies/Models used	Results	Limitations
Detecting and preventing bullying on Twitter (Dalvi et al., 2020)	SVM, NB	SVM: 71.25% accuracy, NB: 52.70% accuracy	Limited accuracy and dataset coverage
Identifying abusive content in open comments pages of social media websites (Emon et al., 2019)	DL (RNN)	RNN: 82.20% accuracy	Limited to Bengali text
Binary and multilabel classification of cyberbullying in consumer reviews on Facebook (Ahmed et al., 2021b)	Hybrid neural network, ensemble technique	Binary: 85% accuracy, Multilabel: 87.91% accuracy	Limited to specific categories and dataset
Hate speech recognition on Facebook pages in Bengali (Ishmam and Sharmin, 2019)	DL (GRU)	GRU: 70.10% accuracy	Limited accuracy, dataset size, and categories
ML model utilizing NLP methods for identifying bullying in the Bengali language (Chakraborty and Seddiqui, 2019)	SVM with a linear kernel	78% accuracy	Limited to Bengali language and specific techniques
Identifying cyberbullying in Bangla text using ML algorithms and user data (Akhter et al., 2018)	SVM and other ML algorithms	SVM: 97% detection rate	Limited to Bangla text and specific algorithms
Model for cyberbullying identification in Bangla and Romanized Bangla texts (Ahmed et al., 2021c)	CNN, MNB	CNN: 84% accuracy (Bangla), MNB: 84% accuracy (Romanized Bangla)	Limited to specific datasets and algorithms
Methodology for recognizing abusive language in Bangla (Mahmud et al., 2023)	ML (LR)	LR: 97% accuracy	Limited to Bengali language and specific algorithms
Identification of fake news in Bangla from social media (Hussain et al., 2020)	MNB, SVM	MNB: 93.32% accuracy, SVM: 96.64% accuracy	Limited to fake news detection and specific algorithms
Cyberbullying identification on social media using transformer models (Emon et al., 2022)	Transformer models (Bangla BERT, Bengali DistilBERT, XLM-RoBERTa)	XLM-RoBERTa: 85% accuracy	Limited to transformer models and specific dataset
Efficient classification of offensive comments in Bengali on Facebook (Aurpa et al., 2022)	Transformer-based models (BERT, ELECTRA)	BERT: 85.00% accuracy, ELECTRA: 84.92% accuracy	Limited to offensive comments and specific models
Identification and encouragement of positivity in comments using multilingual data (Chakravarthi, 2022)	CNN	CNN: 75% accuracy (English), 62% accuracy (Tamil), 67% accuracy (Malayalam)	Limited to specific languages and sentiment analysis
Standard hate speech detection mechanism for Bengali (Romim et al., 2021)	Bi-LSTM model with FastText word embedding	Bi-LSTM: 86.78% F1-score	Limited to hate speech detection and Bengali language

3.3. Feature extraction

Textual information is transformed into numerical information using feature extraction. A fairly simple technique to use in natural language processing to better grasp the context is called feature extraction (Bird et al., 2009). It is a key step in natural language processing since no machine learning algorithm, not even computers can comprehend a text. Text can be converted into vectors with the aid of the text vectorization technique TfidVectorizer (TF-IDF) vectorizer, which is a well-liked method for conventional machine learning algorithms (Sharifani et al., 2022). In our proposal, we take text data so we need to extract information from the text to find insight into the data and build models. We apply TfidVectorizer, a feature extraction process from text data, to perform our extraction process. It creates a matrix of TF-IDF features from a set of raw texts. Because it considers the importance of the words as well as their frequency, TF-IDF is superior to Count Vectorizers in terms of word analysis (Ahmed et al., 2022b).

3.4. Resampling the data

Resampling is a technique that involves taking samples from the source information repeatedly (El-Amir and Hamdy, 2020). Due to having a large volume of text data, we adopted the undersampling technique to produce a balanced dataset that matches reality and is most effective at identifying Bengalibullying. It is a technique for leveling unequal data that consists of minimizing the size of the majority class while keeping all of the data from the minority class. This method uses initially unbalanced text collections to extract more precise data. It has better analysis run times and lower storage needs. With less data, we may acquire insightful information faster and with less storage space (Singh et al., 2022).

The undersampling is carried out using an innovative technique known as Instance Hardness Threshold (IHT), which is an undersampling strategy for reducing class imbalance by minimizing majority class (Le et al., 2018; Jin et al., 2021). IHT is a special approach that removes samples with lower probabilities after a classifier has been trained on the data and nearly a controllable under-sampling technique (Smith et al., 2014).

We employ the logistic regression classifier in IHT for the classifier training, which is carried out using cross-validation. It is simpler to use, comprehend, and train using logistic regression. It not only offers an assessment of a predictor's eligibility, but also the orientation of the relationship. It classifies unfamiliar records fairly quickly (Wu et al., 2008; Genkin et al., 2007). The data resampling process is depicted in Fig. 5.

The class distribution for binary and multiclass before and after IHT, along with the Threshold (Th) value used in the IHT method for resampling the dataset is shown in 2, where we have summarized the threshold values used in the IHT for each class. In the binary class, the threshold values are 0.896 for the "bully" label and 0.104 for the "normal" label. For the multiclass classes, the threshold values are 0.0093 for "not bully", 0.0116 for "troll", 0.0098 for "sexual", 0.0087 for "religious", and 0.0025 for "threat". These threshold values

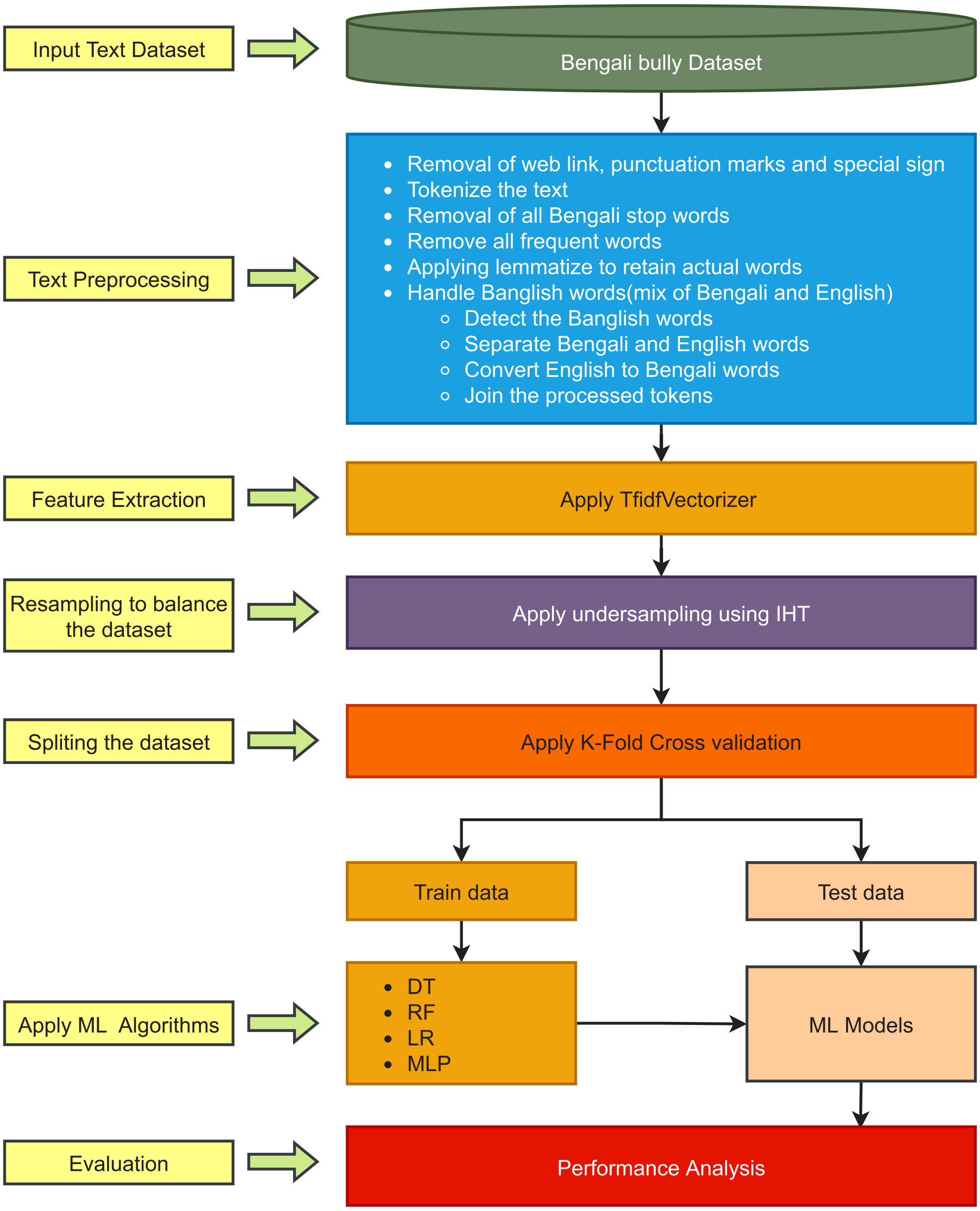

Input Text Dataset
Bengali bully Dataset
Text Preprocessing
Removal of web link, punctuation marks and special sign
Tokenize the text
Removal of all Bengali stop words
Remove all frequent words
Applying lemmatize to retain actual words
Handle Banglish words(mix of Bengali and English)
Detect the Banglish words
Separate Bengali and English words
Convert English to Bengali words
Join the processed tokens
Feature Extraction
Apply TfidfVectorizer
Resampling to balance the dataset
Apply undersampling using IHT
Spliting the dataset
Apply K-Fold Cross validation
Train data
Test data
Apply ML Algorithms
DT
RF
LR
MLP
ML Models
Evaluation
Performance Analysis

Fig. 1. The proposed Bengalibullying detection architecture.

Table 2
Class distribution of binary and multiclass before and after IHT.

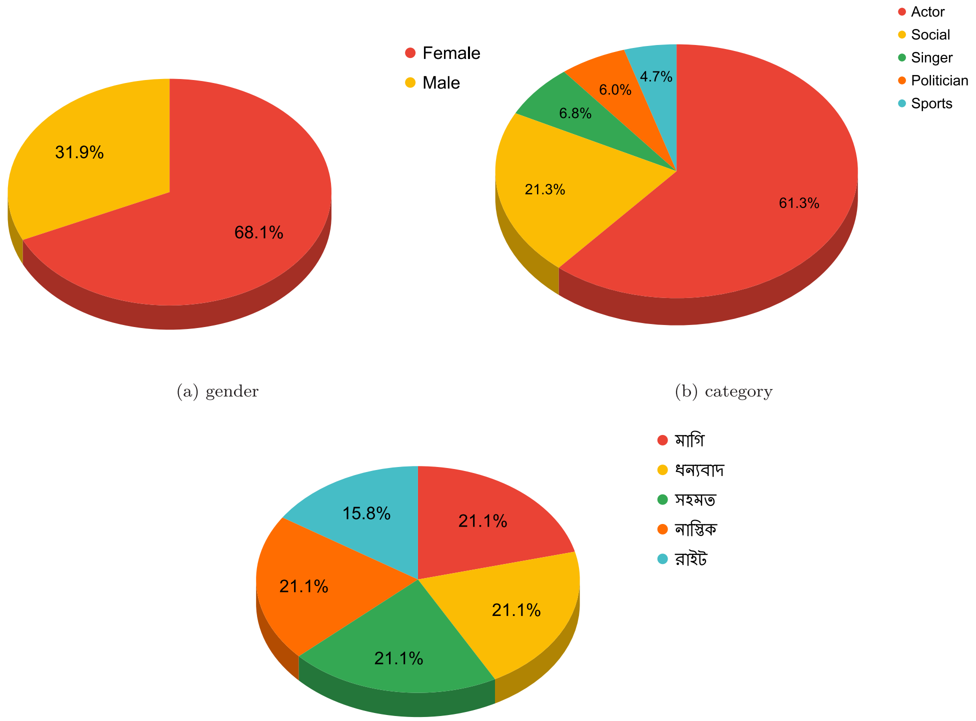
Class	Label	Before IHT	Before (%) IHT	After IHT	After (%) IHT	(%) belong to total	Th Value in IHT
Binary	Bully	28 661	65.14	15 340	50	34.86	0.896
	Normal	15 340	34.86	15 340	50	34.86	0.104
	Total	44 001	100	30 680	100	69.73	–
Multiclass	Not bully	15 340	34.86	1694	20	3.85	0.0093
	Troll	10 462	23.78	1694	20	3.85	0.0116
	Sexual	8928	20.29	1694	20	3.85	0.0098
	Religious	7577	17.22	1694	20	3.85	0.0087
	Threat	1694	3.85	1694	20	3.85	0.0025
	Total	44 001	100	8470	100	19.25	–

were determined as part of the IHT process to balance the imbalanced dataset.

3.5. Adopting ML algorithms

In this study, we used four machine learning classifiers such as DT, RF, LR, and MLP, to develop a model to identify Bengalibullying. Each classifier's performance has been examined in light of various performance indicators. We will go over various ML techniques employed for the prediction model in the subsection that follows.

- *Decision Tree (DT):* In many different domains, such as machine learning, image processing, and pattern recognition, decision trees are a useful tool that can be used (Talukder et al., 2022).

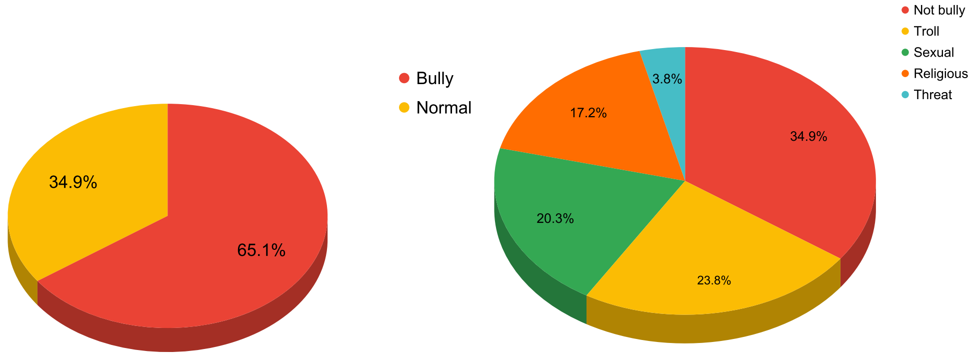


(a) gender

(b) category

(c) comment

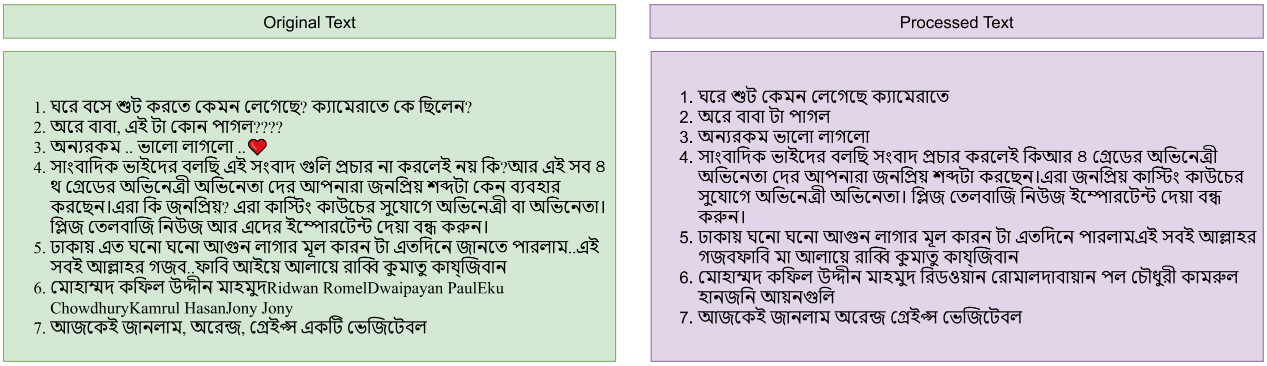
Fig. 2. Data distribution in Bengali bullying dataset.



(a) binary

(b) multilabel

Fig. 3. Class distribution in Bengalibullying dataset.



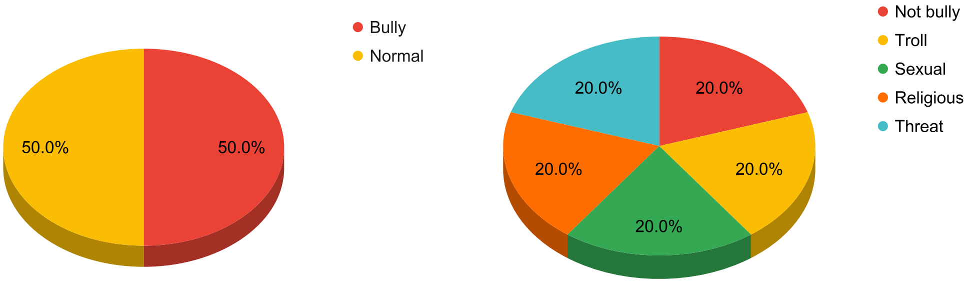
Original Text

1. ঘরে বসে শুট করতে কেমন লেগেছে? ক্যামেরাতে কে ছিলেন?
2. অরে বাবা, এই টা কোন পাগল????
3. অন্যরকম .. ভালো লাগলো ..❤
4. সাংবাদিক ভাইদের বলছি এই সংবাদ গুলি প্রচার না করলেই নয় কি?আর এই সব ৪ থ গ্রেডের অভিনেত্রী অভিনেতা দের আপনারা জনপ্রিয় শব্দটা কেন ব্যবহার করছেন।এরা কি জনপ্রিয়? এরা কাস্টিং কাউচের সুযোগে অভিনেত্রী বা অভিনেতা। প্লিজ তেলবাজি নিউজ আর এদের ইম্পোরটেন্ট দেয়া বন্ধ করুন।
5. ঢাকায় এত ঘনো ঘনো আগুন লাগার মূল কারন টা এতদিনে জানতে পারলাম..এই সবই আল্লাহর গজব..ফাবি আইয়ে আলায়ে রাব্বি কুমাতু কায্জিবান
6. মোহাম্মদ কফিল উদ্দীন মাহমুদRidwan RomelDwaipayan PaulEku ChowdhuryKamrul HasanJony Jony
7. আজকেই জানলাম, অরেন্জ, গ্রেইপ্স একটি ভেজিটেবল

Processed Text

1. ঘরে শুট কেমন লেগেছে ক্যামেরাতে
2. অরে বাবা টা পাগল
3. অন্যরকম ভালো লাগলো
4. সাংবাদিক ভাইদের বলছি সংবাদ প্রচার করলেই কিআর ৪ গ্রেডের অভিনেত্রী অভিনেতা দের আপনারা জনপ্রিয় শব্দটা করছেন।এরা জনপ্রিয় কাস্টিং কাউচের সুযোগে অভিনেত্রী অভিনেতা। প্লিজ তেলবাজি নিউজ ইম্পোরটেন্ট দেয়া বন্ধ করুন।
5. ঢাকায় ঘনো ঘনো আগুন লাগার মূল কারন টা এতদিনে পারলামএই সবই আল্লাহর গজবফাবি মা আলায়ে রাব্বি কুমাতু কায্জিবান
6. মোহাম্মদ কফিল উদ্দীন মাহমুদ রিডওয়ান রোমালদাবায়ান পল চৌধুরী কামরুল হানজনি আয়নগুলি
7. আজকেই জানলাম অরেন্জ গ্রেইপ্স ভেজিটেবল

Fig. 4. Original and processed text of Bangla Text Dataset.



(a) binary

(b) multilabel

Fig. 5. Data distribution in Bengalibullying dataset after resampling.

Three important parts of DT are the root node, branches, and leaf nodes. The root node represents the entire dataset, which is divided into two or more homogeneous sets. The branches represent combinations of characteristics or attributes, and the output nodes are referred to as leaf nodes, where processing is stopped (Dey, 2016; Ahmed et al., 2021d).

- *Random Forest (RF):* A meta-approximation called an RF makes use of averaging to boost accuracy. To avoid over-fitting, several decision tree classifiers are fitted to different data set subtrials (Alkhatib and Abualigah, 2020). Each characteristic subset is selected individually from the real feature space, and it comprises a number of uncut DTs produced from various data points of the training data (Breiman, 2001). Every tree estimates a class and the class that is predicted by the majority of trees constitutes the predictions of our model (Talukder et al., 2023a).
- *Logistic Regression (LR):* LR is an important machine learning technique as it can offer the possibility and identify new information using both discrete and continuous data sources (Uddin et al., 2023; Ray, 2019). By analyzing the correlation from a provided group of classed data, it helps classify data into distinct classes (Bhattacharyya et al., 2011). It is helpful when the results are binary or dichotomous and the data can be split linearly. This approach is based on a statistical model that forecasts likelihood as yes or no (Patil et al., 2022; Ahmed et al., 2021d). It is simpler to use, comprehend, and train using logistic regression. It not only offers an assessment of a predictor's eligibility, but also the orientation of the relationship. It classifies unfamiliar records fairly quickly (Wu et al., 2008; Genkin et al., 2007).
- *Multilayer Perceptron (MLP):* A typical ANN construction is the MLP, which has layers built up of neurons and their interactions. In order to produce a response that will be delivered to the following neuron, it can assess the weighted total of its inputs (Castro et al., 2017). The input and output layers are separated by one or more hidden layers. The neurons are grouped in layers, interactions are frequently directed from lower to upper layers, and neurons within a layer are not connected to one another (Ramchoun et al., 2016; Talukder et al., 2023b).

4. Results and discussion

This study exhibited a hybrid ML approach for detecting online Bengali cyberbullying. The developed procedure consists of several processes, such as effective text preprocessing to prepare the comments, feature extraction to transform the processed text into numerical information, and resampling to balance the data.

4.1. Experiment setup

The examinations are conducted on a computer running Microsoft Windows 10 Pro with an Intel(R) Core(TM) i3-6006U CPU operating at 2.00 GHz, 2000 MHz, 2 cores, 120 GB SSD, 1TB HDD, and 8 GB RAM. An Anaconda Navigator Jupyter notebook is used to do the investigation. The Python language and a number of standard libraries, including Pandas, NumPy, Matplotlib, Seaborn, TensorFlow, Keras, Scikit-learn, and others, are used to create the proposed model.

Table 3
Confusion matrix.

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

4.2. Performance evaluation metrics

A variety of performance metrics are used to evaluate the performance of our suggested model, comprising accuracy, precision, recall, f1-score, ROC curve, confusion matrix, MSE, MAE, and RMSE. These are the metrics developed for assessing the works (see Table 3):

- where TP, TN, FN, and FP, respectively, represent for True Positive, True Negative, False Positive, and False Negative.
- $$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$
- $$Precision = \frac{TP}{TP + FP} \tag{2}$$
- $$Recall = \frac{TP}{TP + FN} \tag{3}$$
- $$F1 - score = 2 * \frac{(precision * recall)}{(precision + recall)} \tag{4}$$
- $$MAE = \frac{\sum_{i=1}^{n} predicted(i) - actual(i)}{n} \tag{5}$$
- $$MSE = \frac{\sum_{i=1}^{n} (predicted(i) - actual(i))^2}{n} \tag{6}$$
- $$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (predicted(i) - actual(i))^2}{n}} \tag{7}$$

  where n is the total number of values.

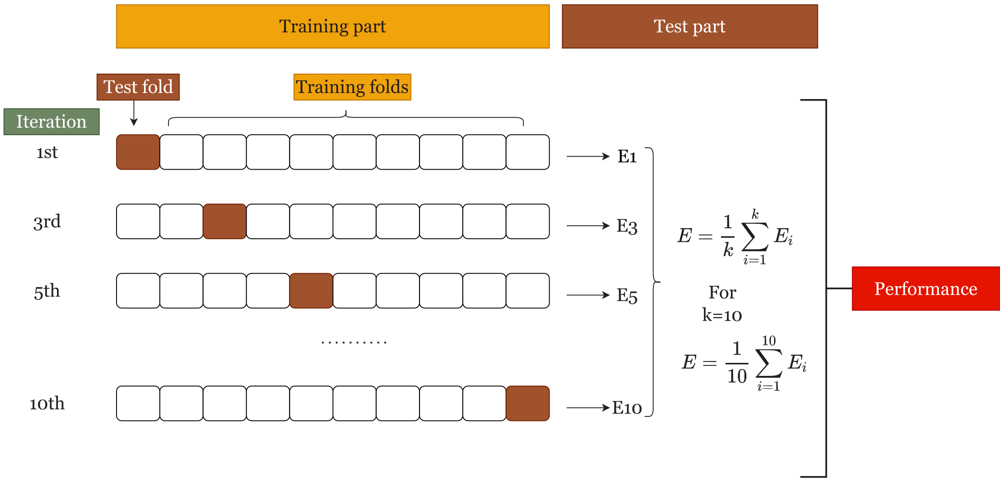


Fig. 6. k-fold cross-validation process.

- ROC curves are two-dimensional plots that are often used to analyze and examine the effectiveness of classifiers (Fawcett, 2004). AUC approaching 1 indicates that a predicted model is good at separability amongst class labels whereas approaching 0 indicates a lousy predicted model. In actuality, lousy signifies that the consequence is being mirrored (Narkhede, 2018).
- A simple technique for breaking up a training set into k smaller sets is called CV (K-fold Cross-Validation). The plan for each of the k "folds" is to use the folds as training data for a model, which is subsequently validated using the remaining data. Finally, using k-fold cross-validation, the average of the values computed in the loop is added as an evaluation measure. The value of k for the k-fold CV we utilized for our bully identification trials is 10. Ten distinct folds, each used as a testing component during the execute stage, are used to partition the dataset. 80 percent of the dataset is used for training, and the remaining 20 percent is used for testing in a k-fold division. Fig. 6 shows how to perform k-fold cross-validation.

4.3. Results analysis

In our experiments, we have conducted both the binary and multilabel classifications where we achieve significant performance to detect Bengalibullying efficiently.

4.4. Binary results analysis

The performance analysis of binary classification is shown in the graphical format in Fig. 7. From the graph, we can see that the accuracy rate of DT, RF, LR, and MLP are 95.75%, 96.90%, 98.57%, 97.95%; the precision rate is 95.74%, 97.02%, 98.58%, 97.95%; the recall rate is 95.72%, 96.85%, 98.56%, 97.94%; the f1-score rate is 95.73%, 96.9%, 98.56%, 97.95%; the error MAE rate is 4.27%, 3.1%, 1.43%, 2.05%; the MSE rate is 4.27%, 3.1%, 1.43%, 2.05%; the RMSE rate is 20.66%, 17.6%, 11.98%, 14.33%.

Among all ML algorithms, the LR gives the highest performance in binary classification than others where, it achieves 98.57% accuracy rate, 98.58% precision rate, 98.56% f1-score rate 98.56% and the error rates are 1.43% in MAE, 1.43% in MSE and 11.98% in RMSE.

The confusion matrix is depicted in the graph 8, where among all the ML algorithms LR provides a higher TP and TF rate and a lower FP and FN rate. The TP, TN, FP and FN rates are 50.65%, 47.91%, 0.91%, and 0.52%. We can see that it produce higher TP and TN rate and lower FP and FN rate which is a good sign of a better Bengalibullying detection model.

The ROC Curve is presented in the graph 9 where the AUC score is 95.73%, 99.33%, 99.73%, 99.72% for DT, RF, LR and MLP respectively. Among all the ML algorithms we can see that LR achieves the highest AUC score which is 99.73% which is a better prediction model as the ROC score is close to 1 (100%).

In binary results analysis, we find that the LR produces a better performance rate and lower error rate as well as a better AUC score. As LR offers the possibility and identifies new information using both discrete and continuous data sources by analyzing the correlation from a provided group of classed data, it helps classify data into distinct classes with higher performance rates.

4.5. Multilabel results analysis

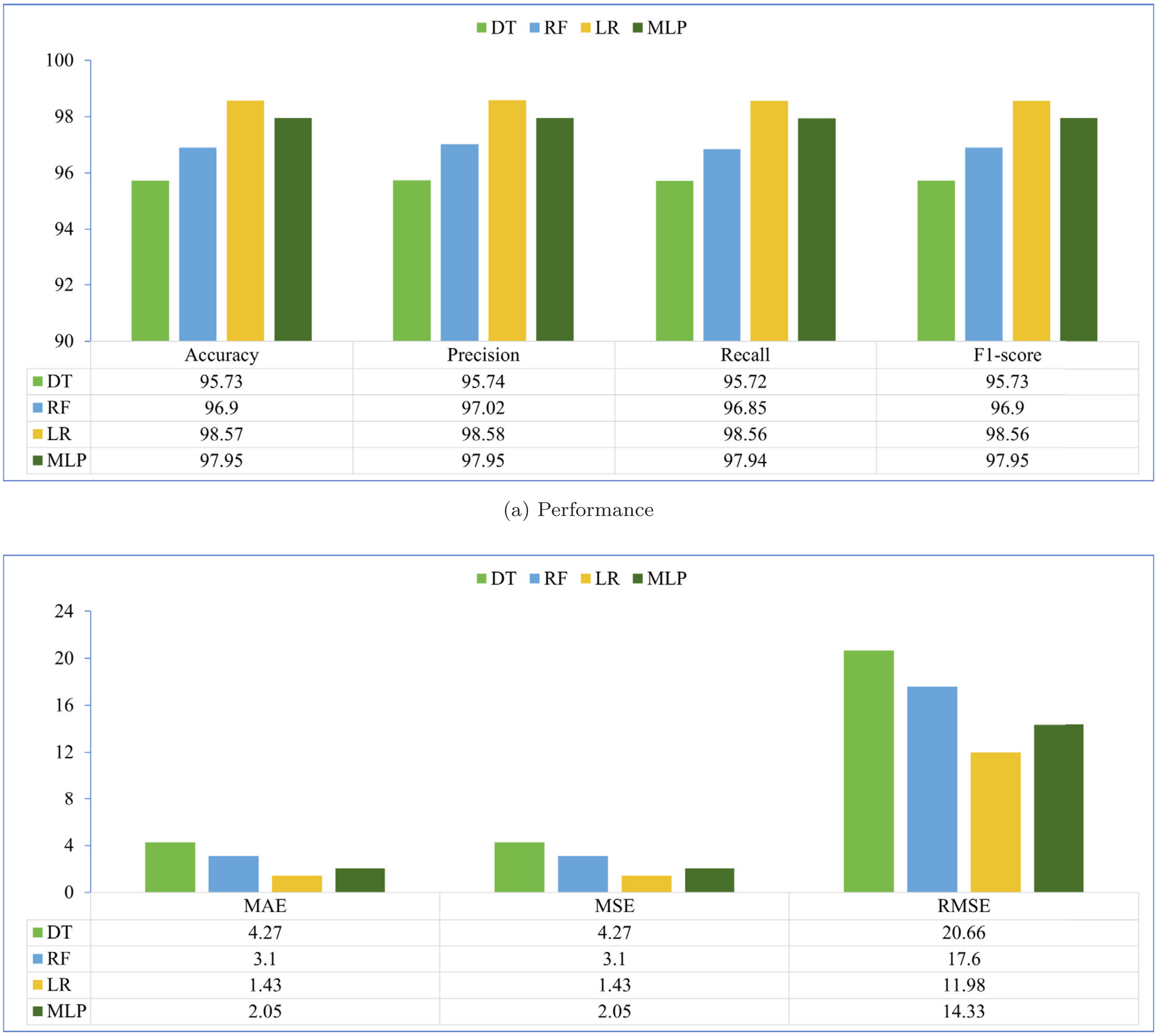
The performance analysis of multilabel classification is shown in the graphical format in Fig. 10. From the graph, we can see that the accuracy rate is 97.64%, 97.87%, 98.7%, 98.82%; the precision rate is 97.68%, 97.93%, 98.72%, 98.86%; the recall rate is 97.78%, 97.95%, 98.8%, 98.89%; the f1-score rate is 97.71%, 97.92%, 98.75%, 98.88%. The performance error rate of the ML algorithms are as follows: the MAE rate is 5.08%, 4.72%, 2.83%, 1.89%; the MSE rate is 14.52%, 13.22%, 8.26%, 4.25%; the RMSE rate is 38.11%, 36.36%, 28.75%, 20.62% for DT, RF, LR, MLP, respectively.

Among all ML algorithms, the MLP gives the highest performance in multilabel classification than others where, it achieves 98.82% accuracy rate, 98.86% precision rate, 98.89% f1-score rate 98.88% and the error rates are 1.89% in MAE; 4.25% in MSE and 20.62% in RMSE.

The confusion matrix is depicted in the graph 11, where among all the ML algorithms MLP provides a higher TP and TF rate and a lower FP and FN rate. Considering class 0 (non-bully), the TP, TN, FP, FN rate is 18.89%, 80.99%, 0.12%, 0% respectively; for class 1 (sexual), the TP, TN, FP, FN rate is 19.6%, 80.28%, 0.12%, 0% respectively; for class 2 (threat), the TP, TN, FP, FN rate is 18.89%, 80.99%, 0%, 0.12% respectively; for class 3 (troll), the TP, TN, FP, FN rate is 19.48%, 79.69%, 0.47%, 0.35% respectively; for class 4 (religious), the TP, TN, FP, FN rate is 21.96%, 76.86%, 0.47%, 0.71% respectively. We can see that it produce higher TP and TN rate and lower FP and FN rate which is a good sign of a better Bengalibullying detection model. (non-bully, sexual, threat, troll)

The ROC Curve is presented in the graph 12 where the AUC score is 98.59%, 99.58%, 99.73%, 99.76% for DT, RF, LR and MLP respectively. Among all the ML algorithms we can see that MLP achieves the highest AUC score which is 99.76% which is a better prediction model as the ROC score is close to 1 (100%).

In multilabel results analysis, we find that the MLP produces a better performance rate and lower error rate as well as a better AUC



	Accuracy	Precision	Recall	F1-score
🟩 DT	95.73	95.74	95.72	95.73
🟦 RF	96.9	97.02	96.85	96.9
🟨 LR	98.57	98.58	98.56	98.56
■ MLP	97.95	97.95	97.94	97.95

(a) Performance

	MAE	MSE	RMSE
🟩 DT	4.27	4.27	20.66
🟦 RF	3.1	3.1	17.6
🟨 LR	1.43	1.43	11.98
■ MLP	2.05	2.05	14.33

(b) Error

Fig. 7. Binary performance analysis for Bengalibullying.

score. As MLP is an ANN where the layer is built up of neurons and their interactions and it can operate both linear and non-linear patterns and has more computing capability and is capable of solving complex nonlinear difficulties.

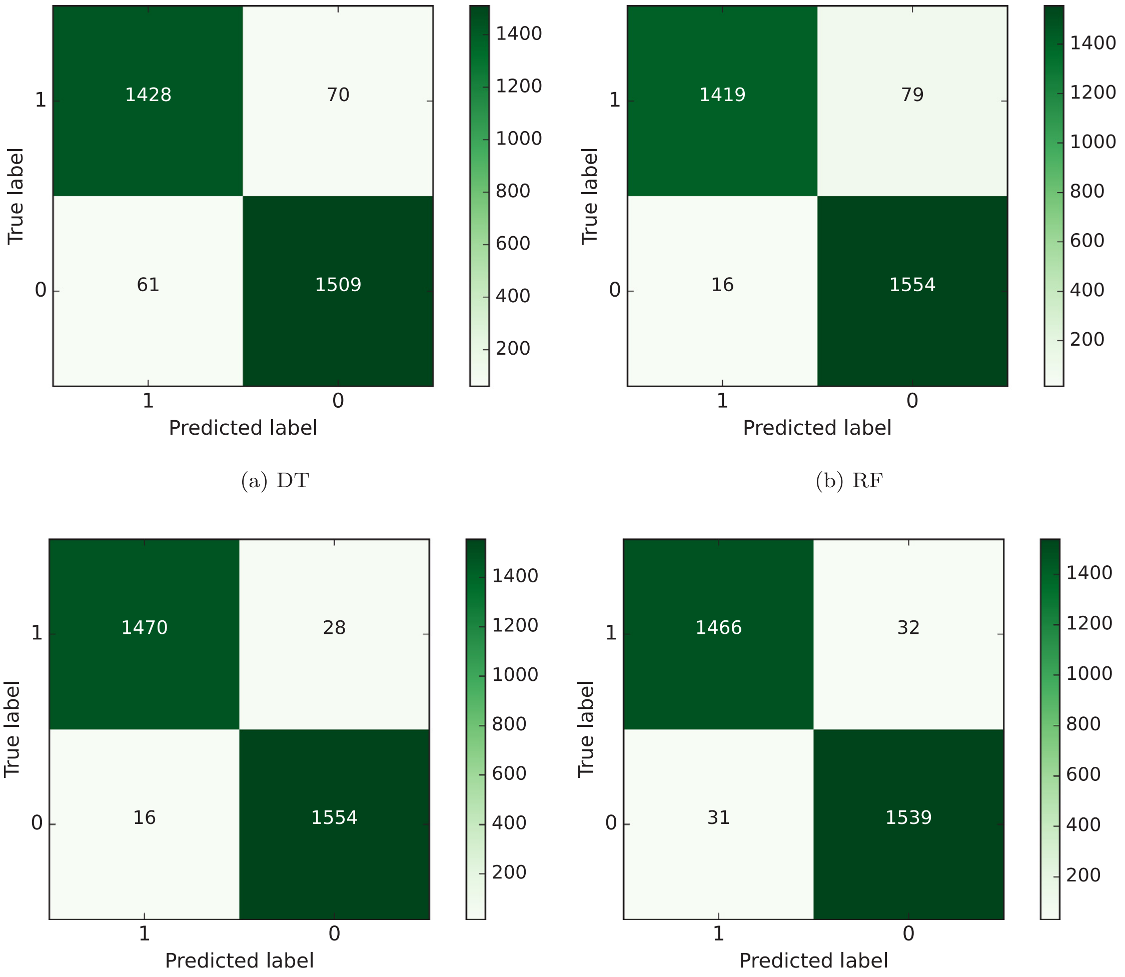
4.6. Discussion

In this section, we discuss the findings and implications of our proposed hybrid approach for Bengali cyberbullying detection and compare its performance with other approaches. Our hybrid approach incorporates effective preprocessing, feature extraction, and resampling techniques, leading to improved accuracy rates. The comparison study between our proposed model and other models is presented in Table 4. Notably, the comparison is based on the same dataset (Bangla Text Dataset) and a similar number of comments, ensuring a fair performance assessment.

4.6.1. Comparative analysis

Our experimental results demonstrate that our proposed hybrid model achieves the highest accuracy rates among all the preceding models in Bengali cyberbullying detection. For both binary and multilabel classification tasks, we achieve accuracy rates of 98.57% (using Logistic Regression) and 98.82% (using Multi-Layer Perceptron), respectively, surpassing the performance of other models. The comparison analysis in Table 4 provides further insights into the relative performance of different models. Our approach outperforms the other models, such as the NN + Ensemble model by Ahmed et al. (2021b), the XML-RoBERTa model by Emon et al. (2022), and the BERT and ELECTRA models by Aurpa et al. (2022). In terms of accuracy, our hybrid model achieves a significantly higher performance.

The success of our hybrid approach can be attributed to the effective utilization of preprocessing, feature extraction, and resampling techniques. Our paper provides a detailed description of the preprocessing steps, including handling null values, removing web links, punctuation marks, and special signs, tokenization, stop-word removal, lemmatization, and handling Banglish words. While the pre-processing steps are crucial, the feature extraction process using the TfidfVectorizer and the application of resampling techniques also contribute significantly to the superior performance of our ML models. The TfidfVectorizer enables the conversion of textual data into numerical information, effectively capturing the importance and frequency of words. Additionally, the resampling technique we employ, specifically the Instance Hardness Threshold (IHT), addresses the class imbalance issue in the dataset, preventing overfitting or underfitting problems and enhancing overall performance.



(a) DT

(b) RF

(c) LR

(d) MLP

Fig. 8. Confusion matrix for binary label.

Table 4
The comparison analysis of Bangla Text Dataset (Social Media Data).

SI. No.	Authors	Models	Dataset	Performance (Accuracy In %)	
				Binary	Multilabel
1	Ahmed et al. (2021b)	NN + Ensemble	Bangla Text Dataset	87.91	85
2	Emon et al. (2022)	XML-RoBERTa	Bangla Text Dataset	–	85
3	Aurpa et al. (2022)	BERT and ELECTRA	Bangla Text Dataset	–	85 (BERT) and 84.92 (ELECTRA)
4	Our approach	Hybrid	Bangla Text Dataset	98.57	98.82

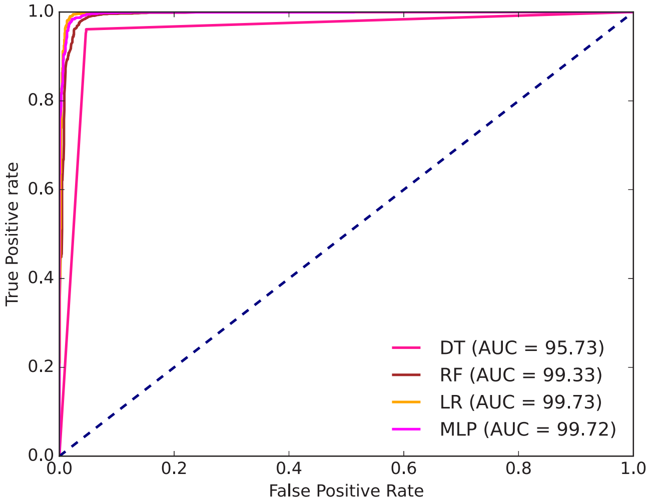
Indeed, the Instance Hardness Threshold (IHT) technique we employed aims to address the class imbalance problem by reducing the size of the majority class. In our case, the Threat class had only 3.8% of the total data in the imbalanced dataset, and the balanced dataset contained 20% of each group. As a result, the balanced dataset includes only 19% (3.8% ∗ 5) of the original dataset, as the classes were balanced to have equal numbers. We acknowledge that comparing the performance of an experiment applied to a small fraction of the dataset with an experiment applied to the entire dataset may raise concerns about fairness. However, it is important to note that the purpose of balancing the dataset was to address the issue of class imbalance, which can lead to biased results and inaccurate performance evaluation. The decision to balance the classes was made to ensure a more accurate assessment of the performance of our proposed model. By creating a balanced dataset, we aimed to eliminate the potential bias caused by the class imbalance and provide a fair comparison among the different models evaluated in our study. The balanced dataset indeed represents a smaller fraction of the original dataset. However, it is important to consider that the performance evaluation on the balanced dataset allows us to focus specifically on the effectiveness of our proposed approach in handling the class imbalance and detecting cyberbullying instances accurately.

4.6.2. *Contributions*

Our proposed approach makes several key contributions to the field of cyberbullying detection in the Bengali language on social media. We have evaluated our hybrid approach using effective preprocessing, feature extraction, and resampling techniques, resulting in a higher accuracy rate compared to existing approaches.

The key innovations and contributions of our approach can be summarized as follows:

- **Effective preprocessing:** Our paper provides detailed information on the preprocessing steps employed, including handling



(a) Binary

Fig. 9. ROC Curve for binary classification.

null values, removing web links, punctuation marks, and special signs, tokenization, stop-word removal, lemmatization, and handling Banglish words. These preprocessing techniques enhance the quality of the text data and contribute to the overall performance of our ML models.

- **Feature extraction using TfidfVectorizer:** We emphasize that the magic of our approach does not solely lie in the preprocessing steps but also in the effective feature extraction using the TfidfVectorizer. This technique transforms the textual data into numerical information, capturing the importance and frequency of words accurately. By leveraging TfidfVectorizer, we enhance the discriminative power of the features and improve the performance of our ML models.
- **Resampling technique:** The resampling technique we employed, specifically the Instance Hardness Threshold (IHT), addresses the class imbalance issue in the dataset. By balancing the dataset, we mitigate the impact of skewed class distribution, thereby preventing overfitting or underfitting problems and improving the overall performance of our ML models.

Our proposed hybrid approach significantly outperforms existing methods in Bengali cyberbullying detection. The superior accuracy rates achieved demonstrate the effectiveness of our ML models and the impact of our preprocessing, feature extraction, and resampling techniques. These contributions enhance the reliability and applicability of our approach in real-world cyberbullying detection systems.

4.6.3. Addressing research questions

In this study, we have addressed each research question individually to provide a comprehensive analysis of our findings:

1. **Prevalence of cyberbullying in Bengali on social media platforms:** We conducted an extensive analysis of cyberbullying instances in the Bengali language, specifically on social media platforms like Facebook. Through data collection and analysis, we determined the prevalence of cyberbullying and provided insights into its frequency and occurrence.
2. **Physical and mental impacts of cyberbullying on individuals, and gender differences:** We investigated the physical and mental impacts of cyberbullying on individuals and examined how these impacts differ based on gender. By analyzing user experiences and conducting surveys, we gained valuable insights into the detrimental effects of cyberbullying and identified potential gender-related differences in its impact.
3. **Existing approaches and models for cyberbullying detection in Bengali, and their limitations:** We reviewed the existing literature on cyberbullying detection in the Bengali language and assessed the performance and limitations of different approaches and models. By critically analyzing previous studies, we identified the strengths and weaknesses of various detection techniques, highlighting areas for improvement.
4. **Development of a hybrid ML approach for cyberbullying detection in Bengali:** We proposed a hybrid ML approach to enhance the detection and classification of cyberbullying in Bengali. Our approach incorporated preprocessing techniques, feature extraction methods, and resampling techniques to improve the performance of our ML models. We provided a detailed description of the steps involved in our approach and highlighted their effectiveness in achieving superior results.
5. **Comparison of the proposed ML model with existing approaches:** We conducted a comparative analysis to evaluate the performance and effectiveness of our proposed ML model in detecting and classifying cyberbullying in Bengali. By comparing our model with existing approaches, we demonstrated its superiority in terms of accuracy, precision, recall, and other relevant metrics. We discussed the differences and similarities between our model and others, providing insights into the factors contributing to these variations.

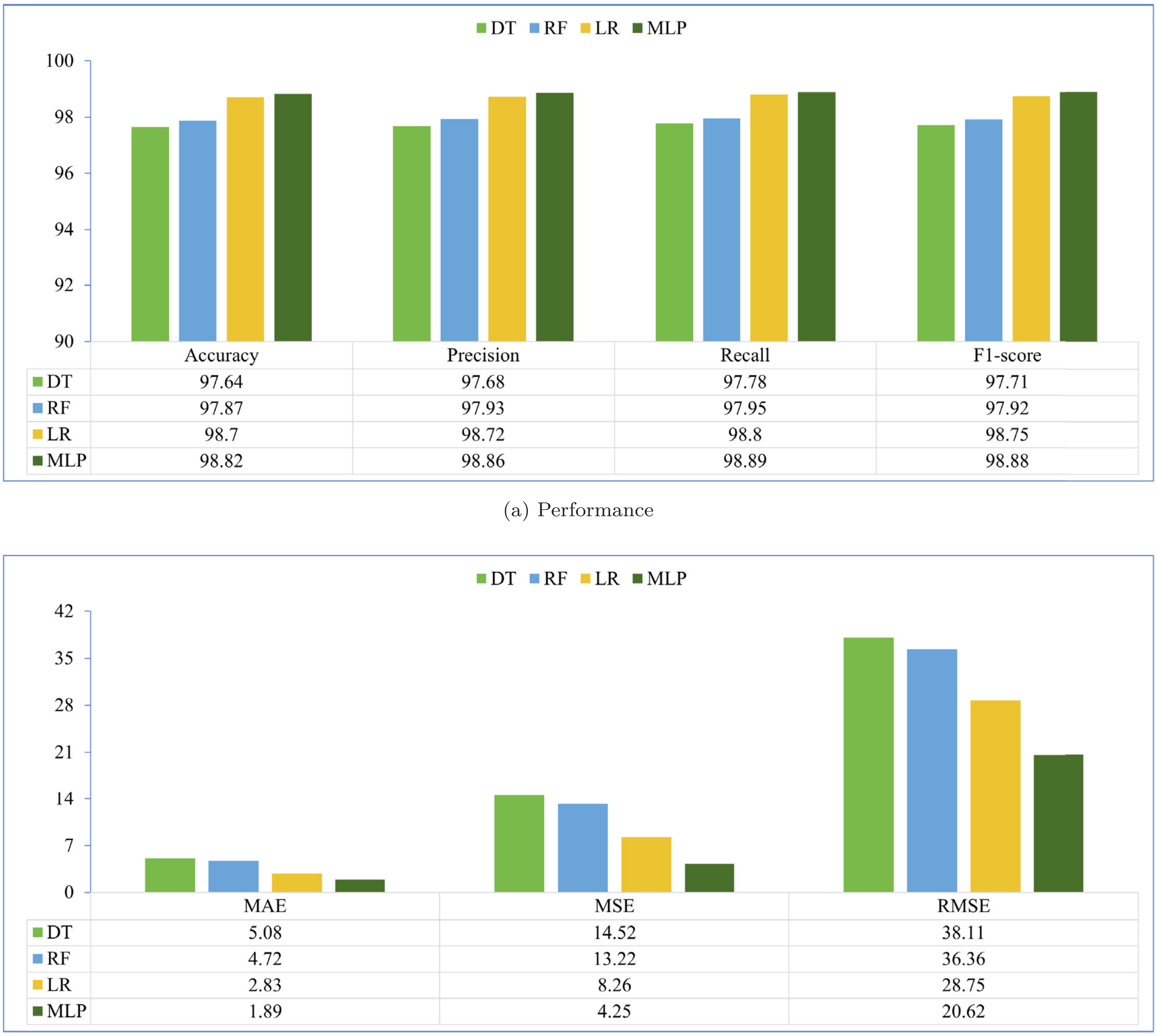
By addressing these research questions, we have gained a comprehensive understanding of cyberbullying in the Bengali language, its impacts on individuals, the existing detection approaches and models, and the development of a hybrid ML approach. Our findings contribute to the existing body of knowledge and provide valuable insights for future research and development in the field of cyberbullying detection.

4.7. Challenges and future directions

Despite achieving promising results in identifying Bengali cyberbullying using a hybrid ML strategy, our study faces certain challenges and limitations. These include:

1. Generalizability: Our focus on Bengali cyberbullying detection using a specific ML approach limits the generalizability of our findings to other languages and cultural contexts. Tailored approaches may be required to address the unique characteristics of different languages and cultures.
2. Dataset limitations: The use of a publicly accessible dataset with a limited number of comments may not fully represent the diverse range of cyberbullying instances in Bengali. Obtaining larger and more diverse datasets would enhance the reliability and representativeness of cyberbullying detection models.
3. Unexplored deep learning approaches: We primarily employed traditional ML algorithms and did not explore deep learning approaches such as BERT, RoBERTa, and others, which have shown promise in natural language processing tasks. Incorporating these advanced models could potentially improve the accuracy and performance of cyberbullying detection systems.
4. Real-world implementation challenges: Our study focused on the effectiveness of the proposed ML model but did not address the challenges associated with implementing such a system in real-world social media platforms. Considerations such as data privacy, scalability, real-time processing, and evolving cyberbullying tactics need to be explored for practical application.

To address these challenges and advance the field of cyberbullying detection, future research can consider the following directions:



	Accuracy	Precision	Recall	F1-score
DT	97.64	97.68	97.78	97.71
RF	97.87	97.93	97.95	97.92
LR	98.7	98.72	98.8	98.75
MLP	98.82	98.86	98.89	98.88

(a) Performance

	MAE	MSE	RMSE
DT	5.08	14.52	38.11
RF	4.72	13.22	36.36
LR	2.83	8.26	28.75
MLP	1.89	4.25	20.62

(b) Error

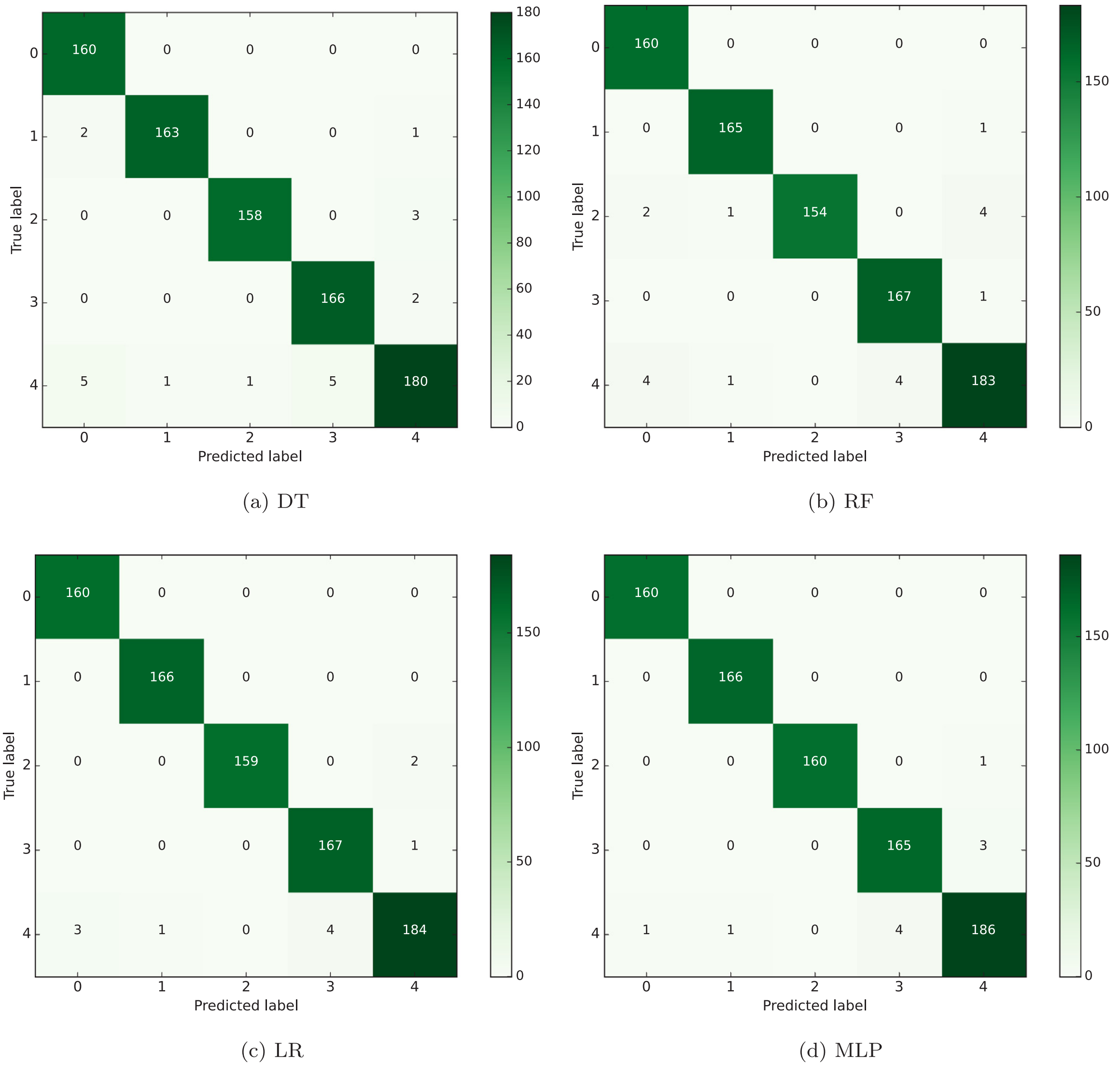
Fig. 10. Multilabel performance analysis for Bengalibullying.

1. Multilingual and multicultural approaches: Investigate the effectiveness of ML and deep learning models in detecting cyberbullying across different languages and cultural contexts. This involves building multilingual datasets and accounting for linguistic variations and cultural sensitivities.
2. Enhanced dataset collection: Obtain larger and more diverse datasets that cover various categories of cyberbullying instances in multiple languages. This would enhance the generalizability and reliability of cyberbullying detection models.
3. Deep learning exploration: Explore advanced deep learning architectures, such as BERT and RoBERTa, for cyberbullying detection. These models capture complex linguistic patterns and semantic relationships, potentially improving detection accuracy and performance.
4. Real-world implementation considerations: Conduct research that addresses the challenges associated with implementing cyberbullying detection systems in real-world social media platforms. This includes addressing privacy concerns, scalability, real-time processing, and adapting to evolving cyberbullying tactics.

By addressing these future directions, researchers can advance the field of cyberbullying detection, develop more robust models, and contribute to creating safer online environments for users across different languages and cultures.

5. Conclusion

Cyberbullying occurs frequently on social media because it gives offenders a place to hide, makes themselves hard to find, and avoids confrontation. This makes it difficult to identify bullying incidents. It can cause victims to suffer, frequently resulting in emotional stress and sadness, and in severe cases, even homicide. In such a toxic climate on social media sites, one must be both sensitive and assertive. This study proposed a hybrid ML strategy for identifying online Bengali cyberbullying. A number of processes make up the proposed hybrid strategy, including efficient text preprocessing to produce the text comments, feature extraction to convert the processed text into numerical information, and resampling for balancing the data. After that, we utilized k-fold cross-validation to split the data and applied ML algorithms such as DT, RF, LR and MLP to build the models and finally evaluate the performance using various performance metrics such as accuracy, precision, recall and f1-score, confusion matrix, AUC score



(a) DT

(b) RF

(c) LR

(d) MLP

Fig. 11. Confusion matrix for multilabel.

and ROC Curve. We achieved the best results ever achieved on the Bangla text dataset (44,001 comments) in our experiment using the publicly accessible dataset. The model successfully identified Bengali cyberbullying on social media with an accuracy rate of 98.57% and 98.82% in binary and multilabel classification, respectively. Our studies indicate that the provided model might be useful for automated Bengali cyberbullying detection systems.

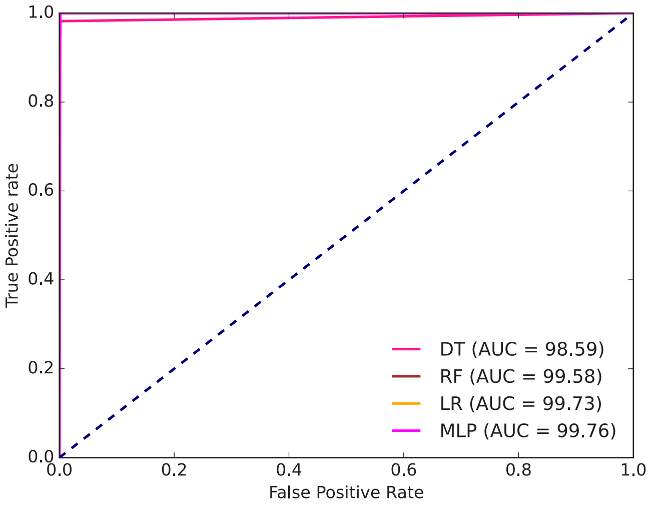
This study has several limitations. Firstly, its focus on identifying online Bengali cyberbullying using a hybrid ML strategy limits its generalizability to different languages and cultural contexts. The reliance on a publicly accessible dataset with limited comments may not fully represent the diverse range of cyberbullying instances. The study does not explore deep learning and transformer-based approaches, which have shown promise in natural language processing tasks. Moreover, the real-world implementation challenges and considerations are not addressed. Addressing these limitations would strengthen the reliability and applicability of future research in cyberbullying detection.

We would like to explore DL and transformer-based approaches in future investigations. Models such as BERT, RoBERTa, XLNET, ELECTRA, DeBERTa, and others could be employed on multilingual cyberbullying datasets encompassing various categories of cyberbullying. This avenue of research holds promise for improving the performance of cyberbullying detection systems.

6. Acronyms

Here is the list of acronyms and their full names:

- TfidfVectorizer (TFID)
- Instance Hardness Threshold (IHT)
- Decision Tree (DT)
- Random Forest (RF)
- Logistic Regression (LR)
- Multilayer Perceptron (MLP)
- Support Vector Machine (SVM) and Naive Bayes (NB)
- Recurrent Neural Network (RNN)
- Gated Recurrent Unit (GRU)
- Bidirectional Long Short-Term Memory (Bi-LSTM)
- Natural Language Processing (NLP)
- Convolutional Neural Network (CNN)
- Multinomial Naive Bayes (MNB)
- Bidirectional Encoder Representations from Transformers (BERT)
- Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA)
- Distilled BERT (DistilBERT)
- Cross-lingual Language Model - RoBERTa (XLM-RoBERTa)



(a) Multilabel

Fig. 12. ROC Curve for multilabel classification.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project was supported by UGC Research Grant, Bangladesh, for the project "Design and development of a robust cyberbullying detection framework using Machine Learning". fiscal year 2020–21 (Ref:37.01.0000.073.06.048.22.856)

References

Ahmed, N., Ahammed, R., Islam, M.M., Uddin, M.A., Akhter, A., Talukder, M.A., Paul, B.K., 2021d. Machine learning based diabetes prediction and development of smart web application. Int. J. Cogn. Comput. Eng. 2, 229–241.

Ahmed, M.F., Mahmud, Z., Biash, Z.T., Ryen, A.A.N., Hossain, A., Ashraf, F.B., 2021a. Bangla online comments dataset. Mendeley Data 1.

Ahmed, M.F., Mahmud, Z., Biash, Z.T., Ryen, A.A.N., Hossain, A., Ashraf, F.B., 2021b. Cyberbullying detection using deep neural network from social media comments in bangla language. arXiv preprint arXiv:2106.04506.

Ahmed, T., Mukta, S.F., Al Mahmud, T., Al Hasan, S., Hussain, M.G., 2022b. Bangla text emotion classification using LR, MNB and MLP with TF-IDF & CountVectorizer. In: 2022 26th International Computer Science and Engineering Conference. ICSEC, IEEE, pp. 275–280.

Ahmed, M.T., Rahman, M., Nur, S., Islam, A., Das, D., 2021c. 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, IEEE, pp. 1–10.

Ahmed, M., Rahman, M., Nur, S., Islam, A., Das, D., et al., 2022a. Introduction of PMI-SO integrated with predictive and lexicon based features to detect cyberbullying in bangla text using machine learning. In: Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications. Springer, pp. 685–697.

Akhter, S., et al., 2018. Social media bullying detection using machine learning on bangla text. In: 2018 10th International Conference on Electrical and Computer Engineering. ICECE, IEEE, pp. 385–388.

Akter, M., Zohra, F.T., Das, A.K., 2017. Q-MAC: QoS and mobility aware optimal resource allocation for dynamic application offloading in mobile cloud computing. In: 2017 International Conference on Electrical, Computer and Communication Engineering. ECCE, IEEE, pp. 803–808.

Alkhatib, K., Abualigah, S., 2020. Predictive model for cutting customers migration from banks: Based on machine learning classification algorithms. In: 2020 11th International Conference on Information and Communication Systems. ICICS, IEEE, pp. 303–307.

Aurpa, T.T., Sadik, R., Ahmed, M.S., 2022. Abusive bangla comments detection on facebook using transformer-based deep learning models. Soc. Netw. Anal. Min. 12 (1), 1–14.

Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C., 2011. Data mining for credit card fraud: A comparative study. Decis. Support Syst. 50 (3), 602–613.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc..

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Castro, W., Oblitas, J., Santa-Cruz, R., Avila-George, H., 2017. Multilayer perceptron architecture optimization using parallel computing techniques. PLoS One 12 (12), e0189369.

Chakraborty, P., Seddiqui, M.H., 2019. Threat and abusive language detection on social media in bengali language. In: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology. ICASERT, IEEE, pp. 1–6.

Chakravarthi, B.R., 2022. Hope speech detection in YouTube comments. Soc. Netw. Anal. Min. 12 (1), 1–19.

Dalvi, R.R., Chavan, S.B., Halbe, A., 2020. Detecting a Twitter cyberbullying using machine learning. In: 2020 4th International Conference on Intelligent Computing and Control Systems. ICICCS, IEEE, pp. 297–301.

Das, A.K., Ashrafi, A., Ahmmad, M., 2019. Joint cognition of both human and machine for predicting criminal punishment in judicial system. In: 2019 IEEE 4th International Conference on Computer and Communication Systems. ICCCS, IEEE, pp. 36–40.

Denny, M.J., Spirling, A., 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. Political Anal. 26 (2), 168–189.

Dey, A., 2016. Machine learning algorithms: a review. Int. J. Comput. Sci. Inf. Technol. 7 (3), 1174–1179.

El-Amir, H., Hamdy, M., 2020. Data resampling. In: Deep Learning Pipeline. Springer, pp. 207–231.

Emon, M., Haque, I., Iqbal, K.N., Mehedi, M., Kabir, H., Mahbub, M.J.A., Rasel, A.A., 2022. Detection of bangla hate comments and cyberbullying in social media using NLP and transformer models. In: International Conference on Advances in Computing and Data Sciences. Springer, pp. 86–96.

Emon, E.A., Rahman, S., Banarjee, J., Das, A.K., Mittra, T., 2019. A deep learning approach to detect abusive bengali text. In: 2019 7th International Conference on Smart Computing & Communications. ICSCC, IEEE, pp. 1–5.

Fawcett, T., 2004. ROC graphs: Notes and practical considerations for researchers. Mach. Learn. 31 (1), 1–38.

Genkin, A., Lewis, D.D., Madigan, D., 2007. Large-scale Bayesian logistic regression for text categorization. Technometrics 49 (3), 291–304.

Hussain, M.G., Al Mahmud, T., Akthar, W., 2018. An approach to detect abusive bangla text. In: 2018 International Conference on Innovation in Engineering and Technology. ICIET, IEEE, pp. 1–5.

Hussain, M.G., Hasan, M.R., Rahman, M., Protim, J., Al Hasan, S., 2020. Detection of bangla fake news using mnb and svm classifier. In: 2020 International Conference on Computing, Electronics & Communications Engineering. ICCECE, IEEE, pp. 81–85.

Ishmam, A.M., Sharmin, S., 2019. Hateful speech detection in public facebook pages for the bengali language. In: 2019 18th IEEE International Conference on Machine Learning and Applications. ICMLA, IEEE, pp. 555–560.

Islam, M.M., Uddin, M.A., Islam, L., Akter, A., Sharmin, S., Acharjee, U.K., 2020. Cyberbullying detection on social networks using machine learning approaches. In: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering. CSDE, IEEE, pp. 1–6.

Jin, Y., Zhang, W., Wu, X., Liu, Y., Hu, Z., 2021. A novel multi-stage ensemble model with a hybrid genetic algorithm for credit scoring on imbalanced data. IEEE Access 9, 143593–143607.

Kee, D.M.H., Al-Anesi, M.A.L., Al-Anesi, S.A.L., 2022. Cyberbullying on social media under the influence of COVID-19. Glob. Bus. Organ. Excell. 41 (6), 11–22.

Kumar, A., Sachdeva, N., 2022. A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. World Wide Web 25 (4), 1537–1550.

Le, T., Hoang Son, L., Vo, M.T., Lee, M.Y., Baik, S.W., 2018. A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset. Symmetry 10 (7), 250.

Mahmud, T., Das, S., Ptaszynski, M., Hossain, M.S., Andersson, K., Barua, K., 2023. Reason based machine learning approach to detect bangla abusive social media comments. In: International Conference on Intelligent Computing & Optimization. Springer, pp. 489–498.

McGlotten, S., 2013. Virtual Intimacies: Media, Affect, and Queer Sociality. State University of New York Press.

Narkhede, S., 2018. Understanding auc-roc curve. Towards Data Sci. 26, 220–227.

Patil, A., Sonawane, M.O.S., Sopan, M.V., 2022. Risk prediction of cardiovascular disease using logistic regression machine learning algorithm. Int. Res. J. Mod. Eng. Technol. Sci. 4 (1).

Powell, A., Henry, N., Powell, A., Henry, N., 2017. Online misogyny, harassment and hate crimes. Sex. Violence Digit. age 153–193.

Raj, J.S., Anantha Babu, S., et al., 2022. Smart cyberbullying detection with machine learning. In: Disruptive Technologies for Big Data and Cloud Applications. Springer, pp. 237–248.

Ramchoun, H., Idrissi, M.A.J., Ghanou, Y., Ettaouil, M., 2016. Multilayer perceptron: Architecture optimization and training. IJIMAI 4 (1), 26–30.

Ray, S., 2019. A quick review of machine learning algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, pp. 35–39.

Romim, N., Ahmed, M., Islam, M.S., Sharma, A.S., Talukder, H., Amin, M.R., 2021. HS-BAN: A benchmark dataset of social media comments for hate speech detection in bangla. arXiv preprint arXiv:2112.01902.

Sen, O., Fuad, M., Islam, M.N., Rabbi, J., Masud, M., Hasan, M.K., Awal, M.A., Fime, A.A., Fuad, M.T.H., Sikder, D., et al., 2022. Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods. IEEE Access 10, 38999–39044.

Shakambhari, J., Samuel, R.R., Anantha, B.S., 2022. Smart cyberbullying detection with machine learning. In: Disruptive Technologies for Big Data and Cloud Applications: Proceedings of ICBDCC 2021, Vol. 905. Springer Nature, p. 237.

Sharifani, K., Amini, M., Akbari, Y., Aghajanzadeh Godarzi, J., 2022. Operating machine learning across natural language processing techniques for improvement of fabricated news model. Int. J. Sci. Inf. Syst. Res. 12 (9), 20–44.

Shireen, F., Janapana, H., Rehmatullah, S., Temuri, H., Azim, F., 2014. Trauma experience of youngsters and teens: A key issue in suicidal behavior among victims of bullying? Pak. J. Med. Sci. 30 (1), 206.

Singh, P.S., Singh, V.P., Pandey, M.K., Karthikeyan, S., 2022. Enhanced classification of hyperspectral images using improvised oversampling and undersampling techniques. Int. J. Inf. Technol. 14 (1), 389–396.

Smith, M.R., Martinez, T., Giraud-Carrier, C., 2014. An instance level analysis of data complexity. Mach. Learn. 95 (2), 225–256.

Talukder, M.A., Hasan, K.F., Islam, M.M., Uddin, M.A., Akhter, A., Yousuf, M.A., Alharbi, F., Moni, M.A., 2023a. A dependable hybrid machine learning model for network intrusion detection. J. Inf. Secur. Appl. 72, 103405.

Talukder, M.A., Islam, M.M., Uddin, M.A., Akhter, A., Hasan, K.F., Moni, M.A., 2022. Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. Expert Syst. Appl. 205, 117695.

Talukder, M.A., Islam, M.M., Uddin, M.A., Akhter, A., Pramanik, M.A.J., Aryal, S., Almoyad, M.A.A., Hasan, K.F., Moni, M.A., 2023b. An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. Expert Syst. Appl. 120534.

Uddin, M.A., Islam, M.M., Talukder, M.A., Hossain, M.A.A., Akhter, A., Aryal, S., Muntaha, M., 2023. Machine learning based diabetes detection model for false negative reduction. Biomed. Mater. Devices 1–17.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., et al., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14 (1), 1–37.