Data in Brief 63 (2025) 112205



Contents lists available at ScienceDirect

Data in Brief

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



Data Article

Bengali cyberbullying detection: A comprehensive dataset for advanced analysis



Gazi Tahsina Sharmin Jahin[a], Firoze Maliha[a], Nuren Nafisa[a], Mohammad Mahadi Hassan[a], M. Jamshed Alam Patwary[b,*]

[a] *Department of Computer Science and Engineering, International Islamic University Chittagong, Kumira, IIUC Avenue, Sonaichhari, Chattogram, 4318, Chattogram, Bangladesh*
[b] *Department of Computer Science and Engineering, Rangamati Science and Technology University, Rangamati Hill District, Rangamati, 4500, Chattogram, Bangladesh*

ARTICLE INFO

Article history:
Received 17 April 2025
Revised 17 July 2025
Accepted 13 October 2025
Available online 21 October 2025

Dataset link: Collected Bangla Comments Dataset

Keywords:
Cyberbullying
Explainable AI (XAI)
BERT embeddings
Topic modeling
Sentiment analysis

ABSTRACT

Cyberbullying has become a major concern in the digital world, with social media platforms facilitating harmful interactions that target individuals. Although there has been significant research, studies specifically focusing on Bengali cyberbullying remain scarce. This paper addresses this gap by analyzing over 70,000 social media comments in the Bengali language. The dataset is preprocessed for sentiment analysis to classify positive and negative comments, followed by topic modeling for negative comments using Latent Dirichlet Allocation (LDA) to extract features from age, gender, ethnicity, religion-based, and miscellaneous comments. Various models are applied, including Support Vector Machine, XGBoost, CNN+BiLSTM+GRU, mBERT, XLM-R, etc. Among the evaluated models, mBERT demonstrated the highest performance, attaining an accuracy of 92%. While the CNN+BiLSTM+GRU hybrid model outperformed with an accuracy of 91%. Further improvements are done by incorporating BERT embeddings into CNN and ANN models, and achieved an accuracy of 93%. To enhance model transparency and trust, Local Interpretable Model-agnostic Explanations is applied to interpret predictions.

* Corresponding author.
E-mail address: jamshed@rmstu.ac.bd (M.J.A. Patwary).

https://doi.org/10.1016/j.dib.2025.112205
2352-3409/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/)

© 2025 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC license
(http://creativecommons.org/licenses/by-nc/4.0/)

Specification Table

Subject Area	Natural Language Processing (NLP).
Specific Subject Area	Cyberbullying Detection in Bengali Language using NLP Models.
Type of Data	Textual Data Bengali social media comments.
Data Collection	The dataset comprises over 70,000 Bengali-language comments. Approximately 50,000 were sourced from a Kaggle dataset focused on social media discourse, while 20,000 were gathered from publicly available Bengali text corpora, including Facebook, Twitter, and forums. These comments reflect diverse communication styles and topics. Data were filtered for relevance and quality, and duplicates were removed. No specialized instruments were used.
Data Source Location	Data were sourced from publicly available platforms relevant to Bangladesh, including Facebook comments and Google-hosted datasets. While user content originated in Bangladesh, no geographic coordinates were recorded due to anonymization. The dataset is stored at the Department of CSE, International Islamic University Chittagong (IIUC).
Data Accessibility	Repository: Zenodo Data Identification Number: 10.5281/zenodo.15230540 Direct URL to data: https://zenodo.org/records/15230540
Related Research Article	None

1. Value of the Data

- The dataset brings together more than 70,000 comments in the Bengali language, making it one of the most significant available public sources. About 50,000 comments have been sourced from Kaggle, while the other 20,000 comments are collected from various online Bengali text sources, providing an extensive mix of content.
- The comments are made from the actual users' interactions on the various online platforms(Facebook, Twitter, and Internet forums), capturing a wide range of natural communication styles and behaviors. It gives a precious foundation to develop and test NLP models based on the Bengali language.
- The dataset can be used for varied research and practical applications. Examples of such include detection of cyberbullying, hate speech analysis, opinion mining, and social behavior understanding in Bengali.

2. Background

Various researchers have applied machine learning (ML) and deep learning (DL) techniques to combat cyberbullying, often relying on small, domain-specific datasets [1]. LIME has been widely used for model interpretability. For instance, [2] used LIME with classifiers like SVM, RF, Bi-LSTM, and BERT, offering interpretability but falling short in model-specific explanations. Similarly, [3] and [4] incorporated LIME in hate speech detection but encountered limitations in capturing individual model behaviors. In sentiment analysis, particularly for Bengali, [5] existing research has shown promising results but remains heavily dependent on limited datasets. [6] used FastText and Skipgram for feature extraction, highlighting the robustness of BanglaBERT. However, the study emphasized the need for domain adaptation and multi-class learning. [7] proposed a hybrid ensemble model achieving 85% accuracy, yet it lacked transformer-based architectures crucial for capturing deeper language features. Similarly, [8] demonstrated the superi-

Table 1
Data Collection Statistics.

Label	Data
Kaggle	50000
Github Repository	20000
Total	70000

ority of transformer-based embeddings such as BERT and GPT-J-6B over traditional methods like TF-IDF and Word2Vec, reaching up to 98% accuracy. Multilingual efforts like those by [9], using WordPiece tokenization with AfriBERTa and RoBERTa, faced challenges of overfitting and data quality, despite achieving 80.88% accuracy. In Bengali-specific settings, [10] applied BERT with 97% accuracy but was limited by a small and imbalanced dataset. Likewise, [11] integrated TF-IDF with BERT, CNN, and MLP, achieving 92.3% accuracy, though the dataset size raised concerns about overfitting. [12] experimented with pre-trained models like BERT, XLM-R, and mBERT, finding BanglaBERT to be the most effective. These studies underline the necessity of fine-tuned transformer models for Bengali, but again highlight the consistent issue of limited, non-diverse datasets. Hybrid and ensemble models are also gaining attention; [13] reached 87.54% using DistilBERT and XLM-R. However, dataset limitations continue to restrict broader applicability and generalization. Despite technological progress, the lack of large, diverse, and representative Bengali datasets remains a significant barrier to the development of reliable and generalizable models. Most existing works depend on narrow-scope datasets, often sourced from single platforms like Facebook or Twitter, which limits cross-platform adaptability and model robustness. Furthermore, the need for domain adaptation, multi-class learning, and model-specific interpretability frameworks remains largely unattended.

To bridge these gaps, we propose an end-to-end system built upon a large-scale, richly diverse Bengali dataset, integrating ML, DL, transformer-based models (TBM), and explainable AI (XAI) techniques. This comprehensive approach aims to enhance model performance, interpretability, and generalizability for effective cyberbullying detection in the Bengali language.

3. Data Description

This section presents a detailed overview of the dataset used in the study, including its collection sources, preprocessing techniques, and preparation steps. We describe the origin and characteristics of the data, followed by the methods applied to clean and tokenize the Bengali textual content to ensure it is suitable for further analysis.

3.1. Data collection

The dataset consists of over 70,000 Bengali comments collected from diverse public domains. This subsection outlines the data sources and collection methods, providing context on how the dataset was compiled to represent real-world social media communication in Bengali.

3.1.1. Data sources

The data set of the present study consists of more than 70,000 comments collected from different public domains, with approximately 50,000 comments sourced from Kaggle and 20,000 comments from other Bengali text resources [Table 1]. Approximately 50,000 comments were obtained from the publicly available Kaggle dataset [14]. These comments were originally scraped from Bangladesh-focused public Facebook posts. An additional 20,000 comments were sourced from a GitHub repository; these were created by other researchers using social media APIs (e.g., Facebook Graph API, Twitter API v2) along with keyword-based filtering and ethical

Before Cleaning	After Cleaning
তুই একটা বাজে ছেলে!/// <> এই ধরনের আচরণ খুবই নিন্দনীয়!	তুই একটা বাজে ছেলে এই ধরনের আচরণ খুবই নিন্দনীয়

(a) Data Cleaning 1

Before Tokenization	After Tokenization
আগে দেশের সমস্যা সমাধান করেন। পাকিস্তান আর ভারত কি করল এটা আমাদের দেখার বিষয় না।	['আগে', 'দেশের', 'সমস্যা', 'সমাধান', 'করেন', '।', 'পাকিস্তান', 'আর', 'ভারত', 'কি', 'করল', 'এটা', 'আমাদের', 'দেখার', 'বিষয়', 'না', '।']

(b) Applying Tokenization

Fig. 1. Comparison of two stages in the data cleaning process.

scraping methods. All data were filtered for Bengali language content and manually or semi-automatically labeled for tasks such as sentiment and toxicity detection. All included data were publicly available and appropriately licensed for academic use. These sources offer the raw data of the social media posts and commentaries that include a broad array of practices and activities typical of Facebook, Twitter, and internet forums [Fig. 3b]. The various types of data achievements with this language will provide a good ground for the analysis of communication patterns within the Bengali language.

3.1.2. Dataset file description

To support reproducibility and facilitate ease of use, we provide a well-organized dataset comprising multiple structured files:

- **Collected Bangla Comments.csv**: A dataset containing raw Bengali comments collected from various online platforms.
- **NotoSansBengali-Regular.ttf**: It is the standard, regular-weight variant of the Noto Sans Bengali font family designed to support the Bengali script with modern typography.
- **SolaimanLipi_22-02-2012.ttf**: It is a widely used Bengali font distributed freely to promote Bengali digital literacy.

3.2. Data preprocessing

This section outlines the essential steps undertaken to prepare the raw Bengali social media comments for analysis. It covers data cleaning, tokenization, and annotation procedures that ensure the dataset's quality and relevance for cyberbullying detection.

3.2.1. Data cleaning

The initial phase of our study involves the careful preparation of Bengali social media textual data. During data pre-processing, the text is cleansed of HTML tags, URLs, emojis, special characters, common stop words, and unnecessary white spaces. Entries without 'text' values are excluded from further analysis [Fig. 1 a]. The 'text' column entries are converted to string format using 'astype(str)' to ensure uniform formatting for tokenization. For word-level parsing in Bengali, language-specific tools are used. However, since transformer-based models inherently handle noise and irrelevant tokens, overly aggressive cleaning is avoided.

3.2.2. Tokenization

For accurate representation of Bengali characters, words, and sentence structures, mBERT is used for tokenization. mBERT is trained on multiple languages, including Bengali, enabling it to handle diverse linguistic patterns and provide effective tokenization for Bengali text. This process involves splitting each input text into tokens, truncating it to a maximum length of 128 tokens, and padding the sequences to maintain uniform input vector lengths [Fig. 1 b].

Table 2
Labeled Dataset.

Label	Data
Age-based cyberbullying	8001
Gender-based cyberbullying	8001
Ethnicity-based cyberbullying	8001
Religion-based cyberbullying	8001
Miscellaneous	8001

Label	Description	Example Comments
Age-based cyberbullying	Targeting someone based on their age, often involving inappropriate content or manipulation	একজন প্রাপ্তবয়স্ক পুরুষের সাথে যৌন দৃশ্য করার জন্য একটি বছর বয়সী শিশুকে কাস্ট করার জন্য তারা সমস্ত বাজে মাদারফাকারদের সম্পর্কে কী করে সাজানো।
Gender-based cyberbullying	Harassment involving sexist or gender-based discrimination	যখন আপনার আশেপাশের লোকেরা সমকামী এবং ধর্ষণকে রসিকতা হিসাবে ব্যবহার করতে শুরু করে
Ethnicity-based cyberbullying	Hateful or racist remarks directed at someone's race or cultural background	আপনার জন্য দুঃখিত একজন রঙ্গিন অন্ধ দেশপ্রেমের বিরুদ্ধে ঘৃণা ও বর্ণবাদে ভরা
Religion-based cyberbullying	Attacking or insulting someone based on their religious beliefs or identity	আমি জানি যে সমস্ত নাইজেরিয়ান এই কমলা বোকাকে সমর্থন করে তারা সবাই উগ্র খ্রিস্টান চরমপন্থী।
Miscellaneous	Threatening or aggressive behavior that doesn't fit specific categories	ইয়াসমিন সাউদি যদি আমি চাই তাহলে আমি একজন ধর্ষক হতে পারি পরের বার দেখব আমি তোমাকে ট্র্যাশে ফেলে দেব হ্যাঁ এটা ঠিক

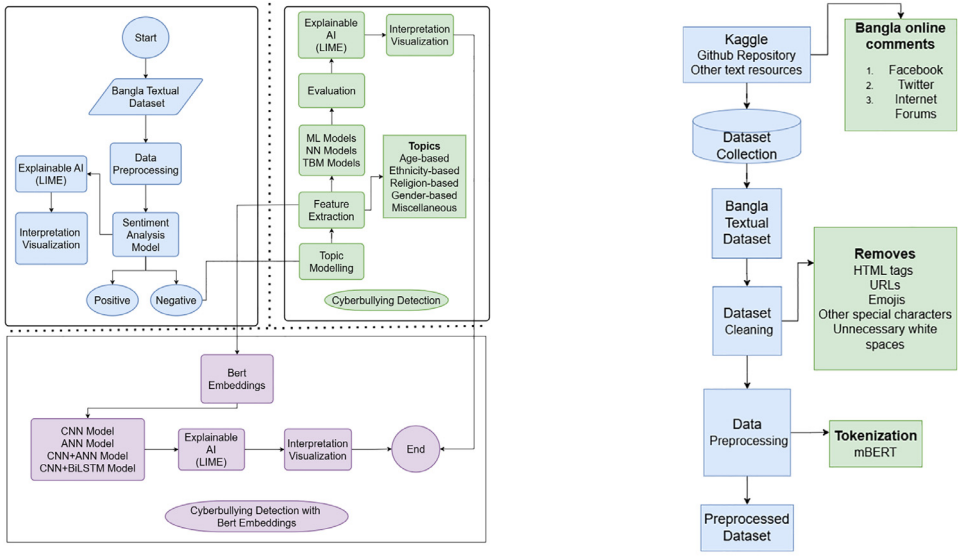
Fig. 2. Description of Dataset Label.

3.2.3. Annotation

The negative dataset is needed to be effective in analyzing various strategies of cyberbullying, which requires manual annotation. The negative sentiments received through the Sentiment Analysis model work as a cyberbullying dataset for the rest of the work of this research. The annotators are native Bengali speakers who also understand online interaction patterns to be noted. Three native Bengali speakers annotated the negative sentiment comments into the defined cyberbullying categories. We calculated pairwise Cohen's kappa on a 20% random sample of the data, obtaining an average κ approximately equal to 0.83 (substantial agreement). Discrepancies were resolved by consensus discussion. Following is a description of the dataset's labels [Table 2, Fig. 2].

4. Experimental Design, Materials and Methods

Cyberbullying is one of the most serious problems in today's highly digital environment, with social media being the prime platform for online abuse [15]. It includes hate speech that targets ethnicity, gender, age, religion, and other such bases [16,17]. Cyberbullying has been studied widely in many languages; however, much work is yet to be done on underrepresented languages like Bengali [18]. Hence, this study proposes to find a way around this gap by studying cyberbullying in Bengali social media, which will provide useful insights for academic research and real-life applications. So, a systematic detection and mitigation of online abuse would be a burning need with more than 250 million speakers and their ever-increasing presence in social media. The work has implemented a systematic approach to the challenges of cyberbullying detection in the Bengali language. A rich dataset of Bengali texts is compiled. The Sentiment Analysis model extracts the positive, negative comments from the dataset [19,20]. [Fig. 7 a]



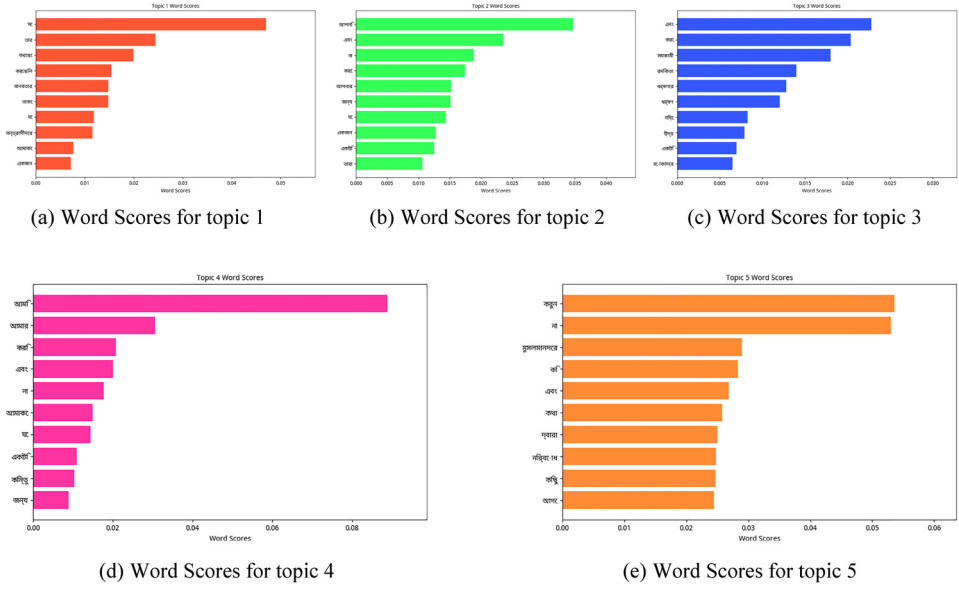
(a) Basic Overview of Framework (b) Data Collection Overview

Fig. 3. Side-by-side comparison of framework and data collection process.

Table 3
Sentiment Label Distribution Statistics.

Label	Data
Positive Sentiment	30000
Negative Sentiment	40000
Total	70000

shows the original sentiment distribution and [Fig. 7 b] the final annotated class distribution The model distinguishes negative sentiments from positive ones, with around 30,000 comments expressing positive emotions and the remaining 40,000 comments reflecting negative feelings on social media platforms [Table 3]. We note this moderate class imbalance (approximately equal to 57% negative vs. 43% positive comments). To mitigate bias, we used stratified train-test splits and applied class weights in the loss function so that misclassifications of the smaller class are penalized more. These steps ensured balanced learning despite the imbalance. The model identifies sentiment as positive or negative using the BERT transformer architecture, which captures efficient context for this text. This is achieved by fine-tuning the model on a labeled dataset, allowing it to adjust its parameters and minimize classification errors, resulting in a more accurate sentiment analysis for Bengali text [Fig. 5]. We have applied LDA methods afterward for latent topic extraction from the negative comments and got useful ideas regarding prevalent themes in cyberbullying speech [21]. [Fig. 4] displays the top words (scores) for each of the five topics. For instance, Topic 1's top terms relate to age, Topic 2 to gender, etc., confirming that LDA captured coherent thematic groups aligned with our label categories. We use Latent Dirichlet Allocation (LDA) as a feature extraction technique in a supervised learning setting for classifying Bengali cyberbullying text. Once we have trained the LDA model with the topics, we extract the topic distribution for each document; the distribution provides the probability of each document belonging to each of the 5 topics. The document-topic probabilities are useful as features for supervised classification. It creates a list of topic probabilities for each document that shows how much that document is affiliated with each topic. For example, if there are 5 topics, then it would have 5 values corresponding to the likelihood of that document to belong to a partic-



(a) Word Scores for topic 1 (b) Word Scores for topic 2 (c) Word Scores for topic 3

(d) Word Scores for topic 4 (e) Word Scores for topic 5

Fig. 4. Overview of Word Scores for Topics.

ular topic. Later, several ML, TBMs, and DL are adopted for analyzing and classifying the data [22–24]. The many-faceted approach, indeed, allows us to evaluate several modeling techniques in terms of their performance with respect to the detection of harmful online behavior [25]. We have introduced a hybrid model named Bengali Cyberbullying Detection(BCD), which integrates CNN, BiLSTM, and GRU architectures to effectively capture both local and sequential features for improved text classification performance. We trained all classification models for 5 epochs using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32. Hyperparameters were chosen based on initial tuning experiments. We used 10-fold cross-validation on the training set to evaluate each model's average performance. Stratified splits ensured proportional class representation in each fold. These settings were chosen because in preliminary tests, accuracy plateaued by epoch 5, and further training yielded diminishing returns. A learning rate of 0.0001 allowed stable convergence without divergence or vanishing gradients, as observed during hyperparameter tuning. To explain better the understanding of how our model has derived at a particular prediction [26], we have integrated LIME into our models [27,28]. The study also uses BERT embeddings, experimenting a bit more with CNN, ANN, and other hybrid models to enhance the detection. This methodology not only tries to increase the detection accuracy of cyberbullying but also ensures that the inferences drawn would be interpretable and actionable [29], helpful towards the needs of the platform moderators and policymakers [Fig. 3 a]. In this way, the critical aspects addressed by our approach have distinguished it from the existing works and have contributed significantly to low-resource language cyberbullying research, such as Bengali (Algorithm 1).

4.1. Model comparison

In the comparative analysis of various model types for cyberbullying detection [Table 4–6], transformer-based models, particularly mBERT, achieved the highest accuracy (92%) and F1-score (91%), showcasing their effectiveness in understanding nuanced Bengali text. We attribute mBERT's superior performance to its extensive pretraining and tokenization strategy. As a multi-

Algorithm 1 Bengali Cyberbullying Detection (BCD)

Initialization
Collect a set of comments from social media
for each comment in the dataset **do**
  Apply sentiment analysis using a sentiment model
  if sentiment is negative **then**
    Extract features using LDA
    Predict class using a classification model
    Generate explanation using LIME
    Store the features and explanation
  end if
end for
return *F* {Set of features and explanations for all negative comments}

Table 4
Performance Results of the ML Models with LIME.

ML Models	Accuracy	Precision	Recall	F1-score
RF	89%	90%	89%	89%
SVM	91%	92%	91%	91%
NB (Multinomial)	81%	82%	81%	80%
NB (Bernoulli)	86%	88%	86%	86%
AdaBoost	85%	86%	85%	85%
XGBoost	91%	91%	91%	91%
LR	90%	90%	90%	90%
DT	86%	86%	86%	86%
GB	88%	88%	88%	89%

Table 5
Performance Results of the DL Models with LIME.

DL Models	Accuracy	Precision	Recall	F1-score	Epoch	Learning Rate
ANN	89%	89%	89%	89%	5	0.0001
LSTM	88%	88%	89%	88%	5	0.0001
GRU	89%	88%	88%	88%	5	0.0001
CNN	90%	89%	89%	89%	5	0.0001
RNN	89%	88%	88%	88%	5	0.0001
CNN+BiLSTM	90%	89%	89%	89%	5	0.0001
CNN+GRU	90%	88%	89%	89%	5	0.0001
CNN+Bi-LSTM+GRU	91%	90%	90%	90%	5	0.0001

Table 6
Performance Results of the TBM Models with LIME.

TBM Models	Accuracy	Precision	Recall	F1-score	Epoch	Learning Rate
DistilBERT	89%	89%	89%	89%	5	0.0001
T5	86%	86%	85%	85%	5	0.0001
mBERT	92%	90%	90%	91%	5	0.0001
XLM-R	91%	90%	90%	91%	5	0.0001

lingual model trained on Bengali text, mBERT effectively captures complex word-level and subword patterns inherent in Bengali. Its WordPiece tokenizer can represent rare words and affixes better than models trained from scratch. Thus, mBERT achieved the highest accuracy (92%) and F1 (91%) in our experiments, outperforming models without such deep pretraining. Following closely, XLM-R also performed well with an accuracy of 91%. Among neural network models, the CNN+BiLSTM+GRU hybrid model outperformed others with 91% accuracy, leveraging the

Negative Sentiment	Positive Sentiment
এখন ইসলামিক সন্ত্রাস শব্দটি যা আপনি আপনার চিন্তার মুখোশ খুলে টেক্সট করেছেন স্যার তাই আমি এই কথোপকথনটি বিক্রি করছি	আল্লাহ নেক হায়াত দান করুক মা আপনাকে
এটি সংখ্যালঘু ধর্মাবলম্বীদের ভর্তুকি মুসলিমদের হজ ভর্তুকি এবং জেরুজালেমে ভ্রমণের জন্য খ্রিস্টানদের ভর্তুকি দেয় বিপরীতে হিন্দুরা সরকারের কাছ থেকে কোনো ভর্তুকি পায় না কোনোটিই নয় কি বোকা তুমি মানুষ	সঠিক ভাবে পদক্ষেপ নেওয়া হয়েছে বলে মনে করি

Fig. 5. Dataset with Sentiment.

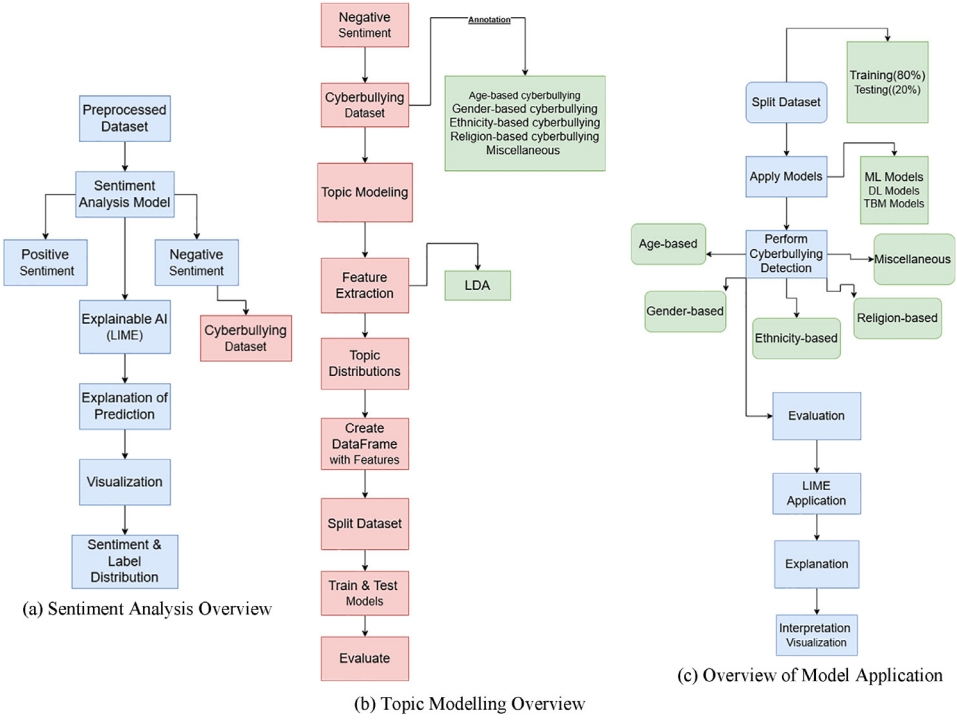
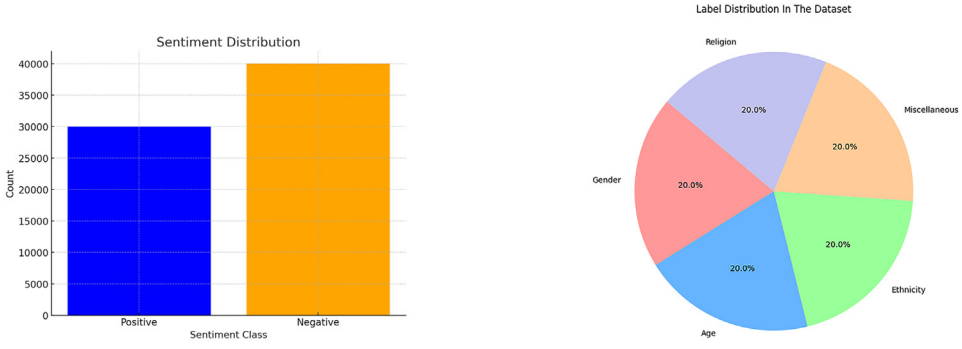


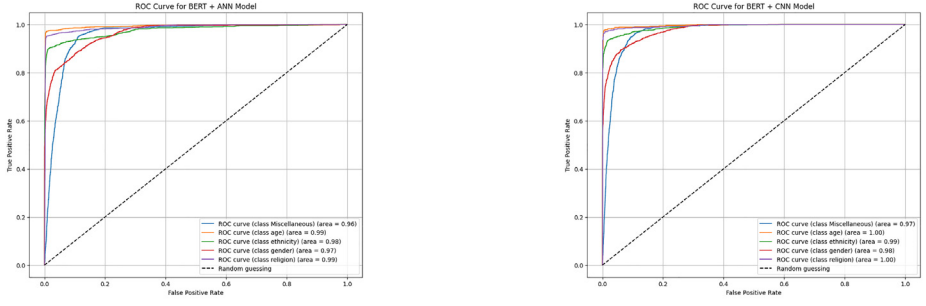
Fig. 6. Overview of the three figures: (a) Sentiment Analysis Overview, (b) Topic Modelling Overview, and (c) Overview of Model Application.

strengths of both convolutional and recurrent layers. In traditional machine learning models, SVM and XGBoost achieved competitive accuracies of 91%, demonstrating strong performance in classification tasks. These results are significantly enabled by the availability of a large and diverse Bengali-language dataset, the volume and linguistic variety in the dataset enhance the model's ability to generalize and handle real-world user comments. Overall, while transformer models excel in capturing linguistic complexities, hybrid neural networks and robust ML models also deliver impressive results, underscoring the critical role of high-quality, representative datasets in achieving reliable model performance (Fig. 6).



(a) Original Sentiment Distribution (b) Annotated Dataset Distribution

Fig. 7. Sentiment and Annotated Dataset Distributions.



(a) BERT+CNN: ROC Curve (b) BERT+ANN: ROC Curve

Fig. 8. ROC Curves for BERT+CNN and BERT+ANN.

Table 7
Performance Results of the Models with LIME.

Models	Accuracy	Precision	Recall	F1-score	Epoch	Learning Rate
BERT+CNN	93%	92%	91%	92%	5	0.0001
BERT+ANN	93%	92%	92%	92%	5	0.0001
BERT+CNN + BiLSTM	92%	92%	92%	92%	5	0.0001
BERT+CNN + ANN	91%	90%	90%	91%	5	0.0001

4.2. LIME Application with BERT Embeddings

According to [Table 7], BERT+CNN and BERT+ANN arrive at the highest accuracy, 93%, along with impressive precision and recall scores, 92% and 91%, respectively. These results point to how effective it is to integrate BERT's contextual embeddings with convolutional and feedforward neural networks to capture the complex patterns in Bengali text [11]. The next best competitor, BERT+CNN+BiLSTM, has an accuracy of 92%, which proves that this model surely can make proper utilization of features from the spatial and temporal dimensions of text. Meanwhile, another model called BERT+CNN+ANN has an accuracy of 91% because while hybrid approaches usually tend to improve a lot, sometimes a certain combination of used architectures does make a difference. BERT-enhanced models (BERT+CNN, BERT+ANN) [Fig. 8a,b], showing the ROC curves

across different identity labels. explain why their performances are much better, with BERT playing the main role. Across all identity classes (age, gender, ethnicity, religion, etc.), the area under the curve (AUC) values are consistently higher. The models, for instance, achieve near-perfect AUC values of 1.00 for age and religion with the BERT+CNN+BiLSTM model, demonstrating outstanding accuracy in classification with almost no false positives. It is the context-sensitive embeddings from BERT that have allowed the models to grasp the subtle nuances of the text and hence differentiate between classes more succinctly than in the no-BERT versions. With high values of AUC, especially for these identity-sensitive categories, the likelihood of misclassifications drastically decreases in BERT; this makes the ROCs steeper and closer to the top-left corner point, meaning near-ideal model behavior.

Limitations

Our data, while large, comes from particular Bengali social media platforms and periods and therefore can be subject to demographic or topical bias. For example, particular offline situations or some dialects are not represented, potentially hindering generalizability. Also, while annotators were experienced, manual detection of cyberbullying is not objective; certain sarcasm or implicit comments might have been interpreted inconsistently. Finally, pretrained models over general corpora (e.g., mBERT) may still miss nuanced domain-specific vocabulary. We understand these shortcomings and propose future research in diversifying data sources and annotators.

Ethics Statement

The data were sourced from publicly available platforms, including Kaggle and other Bengali text corpora. As the data were publicly accessible, no consent was required. The dataset was anonymized, and the study complies with the redistribution policies of the respective platforms, including Kaggle, Facebook, and Twitter.

Data Availability

Collected Bangla Comments Dataset.

CRediT Author Statement

Gazi Tahsina Sharmin Jahin: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Visualization; **Firoze Maliha:** Conceptualization, Software, Data curation; **Nuren Nafisa:** Validation, Supervision; **Mohammad Mahadi Hassan:** Project administration; **M. Jamshed Alam Patwary:** Writing – review & editing, Validation.

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.

References

[1] K. Ashraf, M.H. Hosen, S. Asgar, M.T. Islam, S. Nawar, Analyzing abusive Bangla comments on social media: NLP and explainable AI, in: 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), IEEE, 2024, doi:10.1109/iCACCESS61735.2024.10499547.

[2] J.L. Imbwaga, N.B. Chittaragi, S.G. Koolagudi, Explainable hate speech detection using LIME, Int. J. Speech Technol. 27 (2024) 793–815, doi:10.1007/s10772-024-10135-3.
[3] H. Mehta, K. Passi, Social media hate speech detection using explainable artificial intelligence (XAI), Algorithms 15 (2022) 291, doi:10.3390/a15080291.
[4] M.H. Shakil, M.G.R. Alam, Hate speech classification implementing NLP and CNN with machine learning algorithm through interpretable explainable AI, 2022 IEEE Region 10 Symposium (TENSYMP), IEEE, 2022, doi:10.1109/TENSYMP54529.2022.9864421.
[5] S. Islam, M.J. Islam, M.M. Hasan, S.M.S.M. Ayon, Bengali social media post sentiment analysis using deep learning and BERT model, 2022 IEEE Symposium on Industrial Electronics Applications (ISIEA), IEEE, 2022, doi:10.1109/ISIEA54517.2022.9873680.
[6] M.M. Hossain, I.A.H. Akhi, S.R.S. Raisa, T.S. Onni, Critical analysis of BERT and LSTM model for bengali sentiment analysis across varied datasets, in: 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE, 2023, doi:10.1109/CICN59264.2023.10402353.
[7] M.T. Ahmed, A.S. Urmi, M. Rahman, A.Z.M.T. Islam, D. Das, M.G. Rashed, Cyberbullying detection based on hybrid ensemble method using deep learning technique in bangla dataset, Int. J. Adv. Comput. Sci. Appl. 14 (9) (2023) 545–551, doi:10.14569/IJACSA.2023.0140958.
[8] M. Jobair, D. Das, N.B. Islam, M. Dhar, Bengali hate speech detection with BERT and deep learning models, in: Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning, 2024, doi:10.1007/978-981-99-8937-9_56.
[9] Y. Aliyu, A. Sarlan, K.U. Danyaro, A.S.B.A. Rahman, Comparative analysis of transformer models for sentiment analysis in low-resource languages, Int. J. Adv. Comput. Sci. Appl. 15 (4) (2024), doi:10.14569/IJACSA.2024.0150437.
[10] M. Sen, J. Masih, R. Rajkumar, From tweets to insights: BERT-enhanced models for cyberbullying detection, in: 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), 2024, doi:10.1109/ICETSIS61505.2024.10459672.
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, ArXiv preprint arXiv:1810.04805. 10.48550/arXiv.1810.04805
[12] A. Akhter, U.K. Acharjee, M.A. Talukder, M. Islam, A robust hybrid machine learning model for bengali cyberbullying detection in social media, Nat. Lang. Process. J. 4 (1) (2023) 100027, doi:10.1016/j.nlp.2023.100027.
[13] M.N. Hoque, M.H. Seddiqui, Leveraging transformer models in the cyberbullying text classification system for the low-resource bengali language, in: 2023 26th International Conference on Computer and Information Technology (ICCIT), IEEE, 2023, doi:10.1109/ICCIT60459.2023.10441412.
[14] S.M.N. (Nobel), "Facebook sentiment analysis bangla language,", Kaggle 10 (2021) 2025. [Online]. Available: https://www.kaggle.com/datasets/smnuruzzaman/facebook-sentiment-analysis-bangla-language[Accessed:
[15] V. Bawane, V. Gaupale, A. Wakalkar, K. Rathod, Cyberbullying detection on social media, Int. J. Innov. Eng. Sci. 8 (7) (2023), doi:10.46335/IJIES.2023.8.7.6.
[16] H. Chowdhury, M.A. Supto, Z. Chowdhury, BAAD: A multipurpose dataset for automatic bangla offensive speech recognition, Data Brief 48 (2022) 109067, doi:10.1016/j.dib.2023.109067.
[17] We are social & hootsuite, *Digital 2023: Bangladesh* (2024). [Online]. Available: http://www.datareportal.com/reports/digital-2023-bangladesh[Accessed:
[18] G.I. Hamza, T. Muthaki, S.I. Masuk, F.M. Shah, S.K. Saha Joy, M.M. Rahman, Bengali cyberbullying text detection: a comprehensive study, in: 26th International Conference on Computer and Information technology (ICCIT), IEEE (2023), doi:10.1109/ICCIT60459.2023.10441141.
[19] M.M. Hassan, M.S. Shakil, N. Nessa, M.M. Islam, Sentiment analysis on Bangla conversation using machine learning approach, Int. J. Electr. Comput. Eng. 12 (5) (2022) 5562, doi:10.11591/ijece.v12i5.pp5562-5572.
[20] M.M.H. Manik, F. Haque, M.M.A. Hashem, M.A. Habib, A hybrid framework for sentiment analysis from Bangla texts, in: 2022 25th International Conference on Computer and Information Technology (ICCIT), IEEE, 2022, doi:10.1109/ICCIT57492.2022.10054952.
[21] T. Ahmed, M. Kabir, S. Ivan, H. Mahmud, K. Hasan, Am I being bullied on social media? An ensemble approach to categorize cyberbullying, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, doi:10.1109/BigData52589.2021.9671594.
[22] M.T. Ahmed, M. Rahman, S. Nur, A.A. Islam, 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, 2021, doi:10.1109/ICAECT49130.2021.9392608.
[23] S.S. Nath, R. Karim, M.H. Miraz, Deep learning based cyberbullying detection in Bangla language, Ann. Emerg. Technol. Comput. 8 (1) (2024) 50–65, doi:10.33166/AETiC.2024.01.005.
[24] J. Batani, E. Mbunge, B. Muchemwa, G. Gaobotse, C. Gurajena, S. Fashoto, T.D. Kavu, K. Dandajena, A review of deep learning models for detecting cyberbullying on social media networks, in: R. silhavy (ed.), Cybernetics Perspectives in Systems, Proceedings of the 11th Computer Science On-line Conference 2022 (CSOC 2022), Vol. 3, Lecture Notes in Networks and Systems, vol. 503, Springer, Cham (2022) 528–550, doi:10.1007/978-3-031-09073-8_46.
[25] S. Saha, M.S. Islam, M.M. Alam, M.M. Rahman, M.Z.H. Majumder, M.S. Alam, M.K. Hossain, Bengali cyberbullying detection in social media using machine learning algorithms, in: 2023 5th international conference on sustainable technologies for industry 5.0 (STI), IEEE (2023), doi:10.1109/STI59863.2023.10464740.
[26] F. Ahmmed, A. Jahin, F.T. Zohora, M.J.A. Patwary, An improved hybrid deep learning approach with explainable AI for COVID-19 detection, in: 2025 international conference on electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, IEEE (2025) 1–6, doi:10.1109/ECCE64574.2025.11013503.
[27] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144, doi:10.1145/2939672.2939778.

[28] J. Chen, L. Song, M. Wainwright, M. Jordan, Learning to explain: an information-theoretic perspective on model interpretation, PMLR 80 (2018) 883-892, doi:10.48550/arXiv.1802.07814. in: Proceedings of the 35th International Conference on Machine Learning

[29] V. Pawar, D.V. Jose, A. Patil, Explainable AI method for cyberbullying detection, in: Proceedings of the 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), IEEE, Pune, India, 2022, pp. 119–124, doi:10.1109/ICMNWC56175.2022.10031652.