Leveraging CNN-Bilstm for Multi-Class Cyber bullying Detection in Hindi Text

Ashu Yadav
Megha Gupta
Bhavesh Shakya
Ambuj Pathak
Department of Computer Science & Engineering, JSSATEN, Noida,Uttar Pradesh, India

Abstract — Cyberbullying has become a tremendous issue in the information age, inflicting notable psychological and emotional harm on its victims. While machine learning (ML) and deep learning (DL) approaches have achieved notable progress in detecting cyberbullying, most studies have concentrated on English-language data, neglecting native languages like Hindi and other regional dialects. This review explores recent progress in automated hurtful sentiment detection, focusing on the integration of traditional (ML) methods and advanced deep learning (DL) frameworks and hybrid models. It places particular attention on challenges unique to the Hindi language, such as the scarcity of annotated datasets, linguistic complexities, and the frequent use of code-switching. The effectiveness of methods like convolutional neural networks (CNN), long short-term memory (LSTM), and bidirectional LSTM (BiLSTM) is examined to understand their ability to identify patterns and circumstantial relationships in text. This research specifies cyberbullying comments using a multi-class classification strategy. Kaggle is used to collect the dataset, which is later used to train the model and evaluate it. The dataset contains

16,923 Hinglish/Hindi comments, consisting of four different classes: Non-Aggressive (NAG), Offensive-Aggressive (OAG), Contextually-Aggressive (CAG), and Gendered (GEN). In this study, the CNN-BiLSTM model is implemented and achieves an accuracy of 96.10%. This paper aims to direct future research toward developing inclusive, precise, and scalable solutions for detecting cyberbullying in Hindi/Hinglish text and cultural landscapes.

Keywords: cyberbullying, machine learning, frameworks, neural networks, bidirectional, adaptability, linguistics

1. Introduction

The exponential growth in social media and other online platforms has changed the way people meet, share information, and form groups. While these platforms have fostered global connectivity and inclusivity, they have also given rise to darker phenomena, such as cyberbullying. Nowadays, cyberbullying has become a pressing societal issue with severe implications for mental health and emotional well-being. Cyberbullying victims generally experience anxiety, downheartedness, and in the worst situations, thoughts of suicide, highlighting the urgent need for effective detection and mitigation strategies.

In recent years, advancements in Artificial Intelligence (AI), particularly Machine Learning (ML) and Deep Learning (DL), have shown promise in tackling cyberbullying. These technologies enable automated systems to analyze text-based interactions and identify harmful behavior patterns. Despite significant progress, much of the research has focused predominantly

DOI: https://doi.org/10.5281/zenodo.15152895

on English-language data, leaving native languages like Hindi and other regional dialects largely unaddressed. This gap poses challenges, as linguistic and cultural nuances can significantly impact the effectiveness of cyberbullying detection models.

In this paper, we implemented a multi-class classification method using a combination of ML and Neural Networks (NN) to detect cyberbullying in the Hindi language and support users from unwanted bullying. Different types of Machine Learning (Linear, SVC, SVM), deep learning (RNN), and fusion models (CNN-BiLSTM) approaches are used in this paper. In the pre-processing stage, we go through the procedure of cleaning, tokenizing, and then extracting features from the data to train the model using some important features. In addition to checking individual words, these models also verify the connections between them, enabling the verification of contextual meanings, words, and subtle clues that are often crucial in the detection of cyberbullying. Hybrid models have advanced the field by fusing the advantages of deep learning and Machine Learning, providing increased accuracy and versatility across a larger range of datasets and situations.

The important outcomes of this paper are as follows:

1. We proposed a Multi Classification approach using Neural Network (NN) models to detect cyberbullying in the Hindi language on social media.
2. The technique includes efficient text-data preprocessing to convert Hindi text-data into a usable text format and feature extraction to obtain important information from the text-data.
3. Finally, NN algorithms are applied to evaluate the performance to find the best model to detect Hindi bullying.

Our paper addresses a significant issue by concentrating on the detection of cyberbullying in Hindi text, a topic that is not widely explored in NLP. A CNN-BiLSTM model is employed in the study to achieve text classification tasks using a deep learning approach that is appropriate and effective. The methodology, which encompasses dataset preprocessing, model training, and evaluation, is well-organized. Though clarity in explaining certain experimental details could be improved, the results are informative and showcase the model's accuracy and effectiveness. The paper makes an important contribution to the field by broadening the research on cyber bullying detection to non-English languages and also making it multi-class detection, thereby bridging a crucial research gap. Enhanced readability can be achieved by enhancing grammatical accuracy and sentence structure. While the research is valuable, it could be improved by better organization and refinement.

2. Literature Survey

The increasing growth in cyber bullying on social applications platforms is becoming a major critical issue, prompting the evolution of automated systems that help us detect and counteract harmful online behavior over the years. A variety of methods have been explored to address this challenge, particularly through Machine Learning and Deep Learning techniques. Despite significant progress in these areas, most research has primarily focused on English language data, leaving many languages, including Hindi and other regional languages, underrepresented. This gap in research has prompted further investigations into how language-specific nuances can be incorporated into detection systems, particularly for languages with distinct linguistic and cultural characteristics.

DOI: https://doi.org/10.5281/zenodo.15152895

A. CNN-BiLSTM-GRU Model Used to Detect Cyberbullying in Arabic Language

Eman-Yaser Daraghmi and his team developed an integrated CNN-BiLSTM-GRU hybrid model to detect Arabic text cyberbullying, which addresses the gap in research. Their study made use of six benchmark datasets based on Facebook, Twitter, and Instagram, as well as an Arabic data cyberbullying lexicon to improve detection accuracy. In their study, they evaluated various machine learning as well as deep learning models, such as Naive Bayes, SVM, Random-Forest, CNN, BiLSTM, and GRU, and concluded that the hybrid approach outperforms single models. Stacked word embeddings were used in the proposed model, which demonstrated superior performance, with an accuracy of 98.83% for Arabic text cyberbullying. This study demonstrated that hybrid deep learning techniques are effective in detecting cyberbullying across diverse datasets, emphasizing the necessity of systems that detect native languages.

B. Detection of Multi-featured Cyberbullying by Deep Learning

Luo et al. (2021) used deep learning and proposed a model to detect cyberbullying, including BiGRU and CNN. An attention mechanism called GCA (BiGRU+CNN+Attention) was employed to improve classification accuracy. Kaggle datasets and social media data were incorporated in their study, which yielded a classification accuracy of 91.07%, surpassing traditional machine learning models. Local text features were extracted by CNN while the BiGRU layer captured contextual relationships. The attention mechanism emphasized the importance of representative words, while CNN extracted local text features. In cyberbullying detection, their research pointed out the significance of integrating sentiment and emoji classification. The study concluded that detection performance is significantly improved by deep learning-based hybrid models in comparison to traditional methods.

C. A Study of Cyberbullying Detection by CNNs

A study conducted by Kargutkar and Chitre (2020) examined how machine learning methods can be used to detect cyberbullying, with an emphasis on content-based analysis for precise calculations. Their investigation used Convolutional Neural Networks to examine textual cyberbullying, showing that deep learning methods are more effective than traditional models for increasing classification accuracy. Their study compared CNN with existing approaches using a publicly available dataset, emphasizing its exceptional accuracy in recognizing cyberbullying incidents. According to their research, CNN-based models offer an efficient and scalable method for identifying cyberbullying.

D. Implementing Methods on Social Networks

A comparative study was conducted on cyberbullying detection by Teng and Varathan (2023) involving two important approaches: conventional machine learning (CML) and transfer learning (TL) using the AMiCS dataset, which includes detailed cyberbullying context. The study incorporated two features, such as toxicity and psycholinguistic from LIWC 2022. The study demonstrated that combining logistic regression with textual data, sentiment features, DistilBERT embeddings, and toxicity metrics resulted in an F-measure of 64.8%, exceeding the performance of Linear SVC. An F-measure of 72.42% was achieved, outperforming CML models. The research found that transfer learning was more effective due to its requirement of less feature engineering and lower computational costs.

DOI: https://doi.org/10.5281/zenodo.15152895

E. Implementation of Deep Learning Techniques on Bangla Facebook Comments

Bhowmik et al. (2023) carried out research on detecting cyberbullying in Bangla Facebook comments by utilizing deep learning techniques, particularly the main subjects of discussion are Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). After preprocessing steps including text cleaning, tokenization, and stop-word removal, 7072 out of 8736 comments collected by the research were utilized. The study demonstrated that GRU was more efficient in handling sequential data for cyberbullying detection than LSTM, with an accuracy of 83.55%. The significance of deep learning models is demonstrated by their findings in detecting online harassment in languages that are under-resourced, such as Bengali.

F. Classification of Cyberbullying by RNNs in Multiple Classes

Sifath et al. (2024) developed a model for text classification in multiple classes based on a Recurrent Neural Network (RNN) to detect cyberbullying in written Bengali, addressing the lack of research on underrepresented languages. Four different classes were categorized, and data was sourced from social media users in Bengal, which included threats, trolling, sexual harassment, and non-bullying comments. Many machine learning models such as SVM, Naïve Bayes, XGBoost, and logistic regression were assessed with RNN, which helped to reach an accuracy of 86%. The study underscored the efficiency of deep learning methods in processing sequential data and further stressed the importance of advanced NLP techniques for successful detection.

Summary of Literature Review

Machine learning, deep learning, and hybrid approaches have been extensively studied to improve classification accuracy in detecting the presence of cyberbullying in text. The performance of DistilBERT embeddings in fine-tuning was demonstrated by Teng and Varathan (2023), showing that they outperformed conventional machine learning models. Transfer learning's efficiency in reducing computational complexity is highlighted by the F-measure of 72.42%. In the same context, a hybrid model used for Arabic cyberbullying detection with stacked word embeddings was developed by Daraghmi et al. (2024). The model achieved an accuracy of 98.83% on six benchmark datasets sourced from social media platforms, significantly outperforming single-model classifiers. Later, multiclass classification was tested using a Recurrent Neural Network (RNN) for Bengali cyberbullying detection, proposed by Sifath et al. (2024). The model, with an accuracy of 86%, categorized comments such as non-bullying, trolling, sexual harassment, and threats, outperforming traditional methods like Logistic Regression and SVM. Additionally, Luo et al. (2021) introduced a multi-featured cyberbullying detection model (GCA: BiGRU+CNN+Attention), achieving a classification accuracy of 91.07%, and included sentiment and emoji classification in the analysis. Bhowmik et al. (2023) focused on Bangla cyberbullying detection using LSTM and GRU techniques, with the GRU model outperforming LSTM at an accuracy of 83.55% using a dataset of 8736 Facebook comments. These studies demonstrate the growing need for deep learning techniques in underrepresented languages, as well as the significance of hybrid and multimodal approaches to optimize cyber bullying detection.

Machine Learning, Deep Learning, and Hybrid approaches have been extensively studied to increase classification accuracy in identifying the presence of cyberbullying in

DOI: https://doi.org/10.5281/zenodo.15152895

text. The performance of DistilBERT embeddings in fine-tuning was demonstrated by Teng and Varathan (2023), who proved that they outperformed conventional machine learning models. Transfer learning's efficiency in reducing computational complexity is highlighted by the F-measure of 72.42%. In the same context, a hybrid model used for Arabic cyberbullying detection using stacked word embeddings was evolved by Daraghmi et al. (2024). The accuracy of 98.83% achieved on six benchmark datasets sourced from platforms of social media was significantly higher than single-model classifiers. Later, multiclass classification was tested using a Recurrent Neural Network (RNN) for Bengali cyberbullying detection proposed by Sifath et al. (2024). The model, with an accuracy of 86%, categorized comments such as non-bullying, trolling, sexual harassment, and threats, outperforming traditional methods like Logistic Regression and SVM. Additionally, Luo and colleagues (2021) introduced a multi-featured cyberbullying detection model (GCA: BiGRU+CNN+Attention) using Kaggle's and social network datasets. The classification accuracy was 91.07%, and it included an analysis of sentiment and emoji classification. Bhowmik et al. (2023) focused on Bangla cyberbullying detection using two architecture techniques: LSTM and GRU. The finding was that the GRU model outperformed LSTM, with an accuracy of 83.55% on a dataset of 8,736 Facebook comments. The study highlights the need for deep learning techniques for underrepresented languages. Kargutkar and Chitre (2020) examined two models of deep learning, comparing them with machine learning models aimed at detecting cyberbullying. Their study reported that CNN-based classifiers boost accuracy by recognizing intricate language patterns. Teng and Varathan (2023) emphasized the value of multimodal detection frameworks by merging psycholinguistic and toxicity features from LIWC 2022, which improved classification performance in social media environments. Their study underscored the increased efficiency of two models for cyberbullying detection: hybrid and deep learning, while stressing the necessity of models that are language-specific and multimodal for optimized precision.

3. Proposed Approach

3.1 Dataset Collection

The quality and variety of datasets used for training and assessment have a significant impact on the performance of cyberbullying detection models. To ensure a comprehensive analysis, we explored open datasets from Kaggle. The selection criteria included datasets containing cyberbullying-related text in English, Hindi, and Romanized Hindi, with proper labeling for offensive, hate speech, and abusive language. Text from social media sites like Twitter and Instagram is included in the **Cyberbullying Classification Dataset** (Kaggle), which is categorized into classes like OAG, CAG, NAG, and GEN (S. Afroze, 2022), as seen in **Fig. 1**.

3.2 Data Augmentation

We implemented random word insertion, deletion, synonym substitution, and other NLP augmentation techniques using **nlpaug**. Augmentation was limited to non-Devanagari texts in order to maintain linguistic integrity. Different augmentation rates were applied to each class for successful balancing of the dataset and to improve model robustness.

Fig.1.Sample of hindi dataset

Data pre-processing

In this research, our Dataset is pre-processed properly to analyze cyber bullying and get the best accuracy for our

Data Cleaning

DOI: https://doi.org/10.5281/zenodo.15152895

Raw text data collected from social media and other platforms contains unnecessary elements that need to be removed or standardized. The key data cleaning steps include:

Removing URLs and User Mentions: Social media posts often contain links and tagged usernames (e.g., @username), which are irrelevant for cyber bullying detection.

- Removing Emojis and Special Characters: Emojis and non-alphanumeric characters can introduce noise and are either removed or replaced with text descriptions.
- Removing Extra White Spaces: Multiple spaces and newlines are normalized.

Handling Code-Mixed Text

Since cyber bullying detection involves analyzing both Hindi (Devanagari script) and Romanized Hindi mixed with English, special preprocessing techniques are required:

- **Script Identification:** A text classification step determines whether the input is in Devanagari or Romanized Hindi.
- **Unicode Normalization:** Standard Unicode forms (NFC/NFKC) are applied to avoid inconsistencies.
- **Standardizing Romanized Variations:** Variants of common Hindi words (e.g., bohot → bahut, kya → kia) are mapped to a standard form.
- **Handling Slang and Abbreviations:** Common slang words such as "LOL", "OMG", or abusive terms are expanded or removed.

Stopword Removal

Stopwords are common words that do not contribute to meaning in a sentence.

- **Hindi stopwords** were removed using NLTK and additional manually curated stopword lists.
- **English stopwords** were filtered using the NLTK stopwords corpus.
- **Romanized Hindi stopwords** were curated separately to ensure better performance in mixed-language settings.

Tokenization

Then the text is broken into mini parts called units (tokens). That process is referred to as Tokenization. Different tokenization techniques were used for different scripts:

or Hindi (Devanagari script): indic_tokenize from the Indic NLP library was used to tokenize sentences efficiently.

- For English and Romanized Hindi: word_tokenize from NLTK was used to break text into words.

Feature Extraction

Feature extraction is a crucial step in cyberbullying detection. Machine learning models can work only on numerical representation, so we need to transform textual data. In this study, deep learning-based embeddings are used for feature representation, specifically leveraging Keras Embedding layers to generate dense vector representations of words.

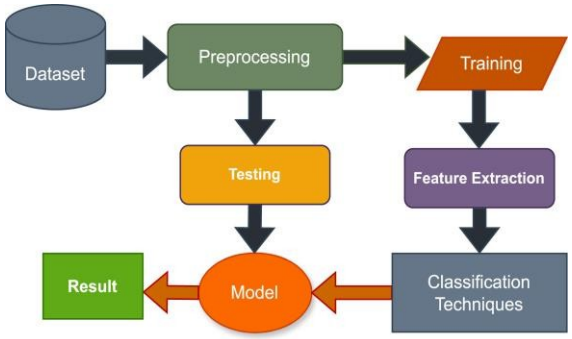
For neural network-based models, semantics and sense of the text data are very important for the betterment of the model and to get better accuracy. Keras embedding converts each word into a vector, where the similar contextual word has a closer vector in the space. The words are represented densely, which helps the model to understand the relationship between words as well as textual data. Text classification improves at a high rate by these steps.

Classification methods & techniques

In this research, we identify and categorize cyberbullying from the social media data. Cyberbullying text data is complex sequential data where the relational structure needs to be considered to get effective results. Considering previous research (R. ALBayari et al., 2022), a CNN-BiLSTM hybrid model is selected to get efficacy. To run the classification algorithm, we separated the dataset into a) training and b) validation sets. Our dataset is now split in 80/20 for the machine learning algorithm. The statistical parameters for both the

DOI: https://doi.org/10.5281/zenodo.15152895

training dataset are 13,538 and the validation dataset is 3,385. Figure 2 depicts the model's



operational process.

Fig. 2.The Architecture of Cyberbullying Classification

NN-BiLSTM Hybrid Model

CNN and Bi-LSTM both have their own unique characteristics and performance attributes when it comes to text cyberbullying classification. Following paragraphs provide an ablation study on each module.

A. CNN:

CNNs, which were originally created for image recognition, have proven to be effective in text classification tasks like cyberbullying detection. Local patterns and features are captured by them in sequences. Convolutional filters are used by 1D CNNs when examining text, and they detect significant word combinations and patterns through their scan. CNNs can identify relevant textual features for cyberbullying using this ability. The CNN architecture consists of convolutional layers followed by pooling layers. The task of extracting spatial patterns and features from the input data is carried out by these layers. Convolutional filters are used by convolutional layers to extract features from the input tokens or characters. Local patterns are captured by these filters and relevant features are detected. By using pooling layers, the dimension of the extracted features can be reduced while keeping the most important information intact. Overfitting and computational complexity can be reduced by this. When it comes to detecting cyberbullying, local patterns are captured by CNNs that identify key features that indicate cyberbullying content, like aggressive language or offensive phrases (R. ALBayari et al., 2022).

B. Bi-LSTM:

Bi-LSTM is a Recurrent Neural Network (RNN) that processes sequences in a bidirectional manner to obtain information from the past and the future simultaneously. The effectiveness of Bi-LSTM in capturing contextual nuances is crucial in understanding cyberbullying instances, which is made possible by its 2D processing capabilities. Bi-LSTM is able to handle dependencies across long distances in text, which makes them ideal for tasks that require understanding the sequential order of words. In Fig. 6, the architecture of Bi-LSTM is depicted. The input tokens are mapped by the embedding layer in Bi-LSTM. During the training process, the semantic meaning of tokens is captured by converting them into dense vectors of fixed size, known as embeddings, which are words or characters. Take note that, as demonstrated in Figure 6, in both Bi-LSTM and CNN models, the embedding layer is the first layer and performs the same function across all architectures (R. ALBayari et al., 2022).

Fig. 3. Bi-LSTM Model Architecture (R. ALBayari et al., 2022).

C. CNN-BiLSTM

In relation to the classification of text cyberbullying, by incorporating the strengths

DOI: https://doi.org/10.5281/zenodo.15152895

of CNN, BiLSTM, and GRU into a hybrid model, performance could be improved by addressing their limitations. By adopting this

hybrid approach, it is possible to emphasize both CNN's pattern recognition capabilities and Bi-LSTM's ability to understand long-range dependencies.

In the Bi-LSTM layer, where both forward and backward LSTM layers are included, long-term dependencies between the input data are captured by processing the sequences of feature vectors generated by the CNN. A Bi-LSTM has the capability to process the input sequence in both forward and backward directions, unlike a regular LSTM, enabling it to gather information from both past and future time steps. The final predictions are generated based on the learned weights after the Bi-LSTM produces hidden states and passes them to a fully connected layer. Comparisons are made between the predictions and the actual target labels. Improving accuracy over time is achieved by propagating the error back to update the weights.

This section explores the architecture of the proposed hybrid model that integrates elements from CNN and Bi-LSTM deep learning models to build an enhanced cyberbullying and the result ant output.

Classification.

The proposed model was implemented and trained using Keras, an API that is built on top of the TensorFlow framework to provide a high-level interface for Neural Network development (R. ALBayari et al., 2022). The implementation process was conducted in PyCharm, with optimal parameter settings.

Table 1 summarizes the key parameters used in training the hybrid CNN-Bi-LSTM model. A dropout rate of 0.5 was applied to prevent overfitting, while the dense layer is set to 1, utilizing the softmax activation function to generate class probabilities. For the CNN component, 128 filters with a kernel size of 5 were specified to extract meaningful features indicative of cyberbullying instances. In the Bi-LSTM layer, 64 hidden nodes captured long-term dependencies bidirectionally. Utilizing the Adam optimizer with a learning rate of 0.001, the model dynamically adapted learning rates for individual parameters during training. Trained for 20 epochs with a batch size of 32, the model leveraged the ReLU activation function in hidden layers to introduce non-linearity and learn complex patterns from input data.

Parameter	**Value**
Dropout rate	0.5
Dense Layer	1
Number of filters	128
Kerner Size	5
Number of Hidden nodes (Bi-LSTM)	64
Optimizer	Adam
ActivationFunction hiddenlayer	ReLU
Activation Function (final output)	Softmax
Number of epochs	20
Batch Size	32
Learning rate	0.001

Table 1. Key parameters used in training the hybrid CNN-Bi-LSTM model.

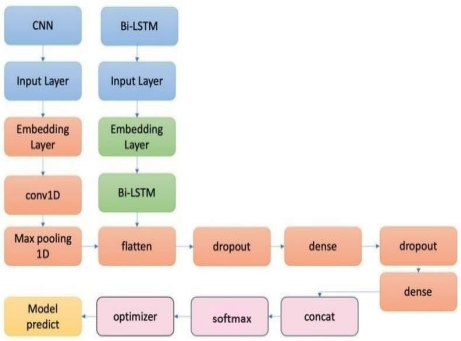
Fig. 4 illustrates the integrated components of our hybrid model:

- **Word Embedding**: Leveraging Keras embeddings, this component captures semantic relationships among words,

DOI: https://doi.org/10.5281/zenodo.15152895

enriching the model's understanding of text.

- **Convolutional Model**: Incorporating a CNN, our model adeptly extracts essential features from text, aiding in pattern recognition (R. ALBayari et al., 2022).
- **Hybrid Architecture**: Combining Bi-LSTM as well as CNN layers, equipped with dropout and batch normalization, comprehends contextual dependencies bidirectionally. A 1D convolutional layer with max-pooling enriches extraction of features across multiple scales, culminating in fully connected dense layers with ReLU activation and a final dense layer with sigmoid activation that will be used for binary classification (R. ALBayari et al., 2022).
- **Fully Connected Model**: Interprets extracted features, establishing a connection between learned features.
- The input layer defines sequence length, while the embedding layer generates 100-dimensional representations aligned with vocabulary size, enhancing semantic understanding. Further architecture includes a Conv1D layer with 128 filters and a kernel size determined by simultaneous word processing, followed by a MaxPooling1D layer to distill essential features. The integrated Flatten layer facilitates concatenation and output transformation, while dropout and weight regularization enhance model robustness.
- Dropout layers play a key role in addressing overfitting and promoting model generalization. The proposed architecture emphasizes the model's capacity to effectively classify input documents as cyberbullying or non-cyberbullying instances, employing robust feature extraction and regularization mechanisms.



- **Fig. 4. The architecture of the proposed hybrid model** (Daraghmi et al., 2024).
- For real-time deployment, we will optimize the model to process data quickly so it can flag harmful content as soon as it's posted. This is crucial for applications such as social media moderation tools where immediate action is necessary to protect users. A mobile devices (Android/Kotlin) approach will be developed so that administrators and users can monitor content and respond in real-time.
- **In conclusion**, the proposed approach is designed to build a powerful, scalable, and accurate system for detecting cyberbullying in Hindi by leveraging the strengths of CNN and BiLSTM in a hybrid model. This approach has the potential to significantly improve cyberbullying detection, making online spaces safer for Hindi-speaking communities and beyond.
- **4. RESULTS**
- **Evaluating the proposed hybrid model** The results of evaluating the proposed Hybrid CNN-BiLSTM-GRU model on the dataset previously illustrated in Section 3 are presented in this section. The strengths of the two models: CNN

DOI: https://doi.org/10.5281/zenodo.15152895

and BiLSTM, are utilized. CNN enables learning local patterns, and Bi-LSTM supports a bidirectional long short-term memory architecture, enabling capturing both forward and backward relationships within textual data.

- **Table 2** summarizes the performance of the proposed model, including Precision, F1-Score, Recall, and Accuracy.
- **hybrid CNN-Bi-LSTM Accuracy**: 96.10%
- **Recall**: 96.10% **Precision**: 96.14% **F1-Score**: 96.10%
- **Table 2. Results of evaluating the proposed hybrid model**
- The above results demonstrate the efficiency of the proposed hybrid CNN-BiLSTM model with an impressive accuracy of 96.10%. The model is capable of distinguishing specific classes/types of cyberbullying. The results show a high recall of 96.10%, which indicates the model's ability to correctly label actual cyberbullying cases. Additionally, the precision of 96.14% indicates the model's ability to correctly label occurrences of cyberbullying, minimizing false positives. The F1-Score of 96.10% provides a balanced justification of the model's precision and recall, proving its overall effectiveness in handling all classes. In general, the above results highlight the potential of the hybrid-proposed model as a valuable tool for cyberbullying detection.
- **Comparison of existing Hindi-English cyberbullying detection approaches**

Reference	**Method**	**Text Representation**	**Accuracy**
Bohra, A., 2018	SVM (Binary)	Character N-grams	71.60%
Mehendale, 2022	Linear SVC (Binary)	TF-IDF, Count vectorization	94%
Our Proposed Model	CNN-BiLSTM (Multi-class)	Keras embeddings, tokenizer, and padding	**96.10%**

- **Table 3. Comparison of existing methods: Cyberbullying detection in Hindi-English content.**
-
- **5. CONCLUSION**
- Cyberbullying detection has come a long way with the help of advancements in Machine Learning and Deep Learning in the past. Researchers mostly relied on traditional Machine Learning techniques like support vector machines and naive Bayes to classify abusive content. While these methods were effective to some extent, the real breakthrough came with Deep Learning models like Convolutional Neural Networks and long short-term memory networks.
- The accuracy and reliability of detecting cyberbullying have increased, moreover combining different types of models like CNN and bidirectional LSTM has helped make detection even more precise by capturing both local features and understanding the broader context in the text. However, there are still challenges, especially when it comes to detecting cyberbullying in non-English languages. For example, Hindi presents its own set of difficulties. The language is rich and complex, with a mix of formal and informal expressions, and often people switch between Hindi and English.

DOI: https://doi.org/10.5281/zenodo.15152895

Additionally, there is a shortage of datasets for training models to detect cyberbullying in Hindi. Therefore, to overcome these challenges, future research needs to focus on creating resources that are specifically tailored to different languages. This includes developing custom lexicons and annotated datasets that reflect the nuances of the Hindi language.

- We also need to improve models that can work across languages, making them more adaptable. Another promising direction is implementing different methods on text and other methods for images and videos and improving the ability to detect cyberbullying in real-time. While a lot of progress has been made, there is still work to be done to truly make a difference, as we need systems that are not only accurate but also scalable and accessible to everyone.
- **6. REFERENCES**

- Daraghmi, E.-Y., Qadan, S., Daraghmi, Y.-A., Yousuf, R., Cheikhrouhou, O., & Baz, M. (2024). From text to insight: An integrated CNN-BiLSTM-GRU model for Arabic cyberbullying detection. *IEEE Access*, 12, 103504-103519, doi: https://doi.org/10.1109/ACCESS.2024.3431939.
- Bohra, A., Puri, A., & Jain, S. (2018). A dataset of Hindi-English code-mixed social media text for hate speech detection. [Online]. Available: https://aclanthology.org/W18-1105.pdf
- Mehendale, Ninad and Shah, Karan and Phadtare, Chaitanya and Rajpara, Keval. (2022). Cyberbullying Detection for Hindi-English Language Using Machine Learning, doi: http://dx.doi.org/10.2139/ssrn.4116143
- S. Afroze. Hindi Offensive Language and Cyber Bullying detect. [Online]. Available: https://www.kaggle.com/code/shadikaafroze/hindi-offensive-language-and-cyberbullying-detect
- Luo, Y., Zhang, X., Hua, J., Shen, W., et al. (2021). Multi-featured cyberbullying detection based on deep learning. *Proceedings of the 16th International Conference on Computer Science & Education (ICCSE)*, 746-751, doi: https://doi.org/10.1109/ICCSE51940.2021.9569270.
- Kargutkar, S. M., Chitre, V., et al. (2020). A study of cyberbullying detection using machine learning techniques. *Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 734-739, doi: https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000137.
- Teng, T.H., Varathan, K.D., et al. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. https://doi.org/10.1109/ACCESS.2023.3275130.
- Shanto, S.B., Islam, M.J., Samad, M.A., et al. (2023). Cyberbullying detection using deep learning techniques on Bangla Facebook comments. *Proceedings of the International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, 1-7, doi: https://doi.org/10.1109/ISACC56298.2023.10083690.
- Sifath, S., Islam, T., Erfan, M., Dey, S. K., Ul Islam, M. M., Samsuddoha, M., Rahman, T., et al. (2024). Recurrent neural network-based multiclass cyberbullying classification. *Natural Language Processing Journal*, 9, Article 100111. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949719124000591

DOI: https://doi.org/10.5281/zenodo.15152895

- R. ALBayari and S. Abdallah. (2022). ‘‘Instagram-based benchmark dataset for cyberbullying detection in Arabic text,’’ *Data*, vol. 7, no. 7, p. 83, doi: 10.3390/data7070083.

DOI: https://doi.org/10.5281/zenodo.15152895