# AI-Powered Cyberbullying Detection in Indian Languages: An Ensemble Approach With XAI

***Manav Shah, **Dr. Ranjithkumar S**
*Vellore Institute of Technology, Vellore
**Professor, Vellore Institute of Technology, Vellore

*DOI:10.37648/ijrst.v16i01.007*

[1]*Received: 21 December 2025; Accepted: 20 January 2026; Published: 08 February 2026*

**Abstract**

In recent times, all of us have been a clear witness to how social media has become so inevitable for us and how we are surrounded with data all around us. With the extensive use of social media, the major downside has been cyberbullying. Therefore, detecting cyberbullying and prevention of cyberbullying is of vital importance. Majority of the work that has been done in this field includes only English and Arabic. In this study, a cyberbullying detection model based on DL techniques has been used. An Ensemble model consisting of ERNIE and RNN has been proposed that detects if the data present is cyberbully content or not and provides a good accuracy. The model is evaluated using a dataset sourced from Kaggle and social media, featuring content in three languages: English, Hindi, and Bengali. Also, the model is tested on English as well as some regional languages spoken widely in a multilingual country like India. The dataset consists of comments and posts in English as well as other languages like Hindi, and Bengali. The model provides an accuracy of 98.4% for English language, 86.6% for Hindi and 83.75% for Bengali.

**Index Terms:** *Artificial Neural Networks (ANN); Explainable Artificial Intelligence (XAI); Bidirectional Long-Short Term Memory (Bi-LSTM); Convolutional Neural Networks (CNN); Cyberbullying; Deep learning (DL); Enhanced representation through Knowledge Integration (ERNIE); Convolutional Neural Networks (CNN); Long Short- Term Memory (LSTM); Machine learning; Multilingual; Natural Language Processing (NLP); Bidirectional Encoder Representations from Transformers (BERT); Recurrent Neural Network (RNN)*

## 1. Introduction

Cyberbullying has become a major issue in the modern digital era, characterized by the act of harassing, intimidating, or harming others using digital communication tools. While content-sharing forums have facilitated communication, especially during the pandemic, they have also become breeding grounds for cyberbullying. The lack of cyberbullying detection tools on these platforms often leads to hateful comments and posts, which affect the victim's psychological and emotional health. In countries like India, where multiple languages are spoken, detecting cyberbullying only in English is insufficient, as most users communicate in their regional languages.

Existing models often fall short in addressing the multilingual aspect and fail to provide explanations for their decisions, limiting their effectiveness and trustworthiness [5], [24], [29]. This study proposes an approach that can detect cyberbullying in multiple languages.

This study is driven by the aim to build a reliable system capable of accurately identifying instances of cyberbullying in English as well as other Indian regional languages like Hindi and Bengali. By incorporating Explainable AI techniques, such as Attention Visualization, the model not only aims to detect whether any comment or text contains

[1] How to cite the article: Shah M., Ranjithkumar S..; February 2026; SDG 7: AI-Powered Cyberbullying Detection in Indian Languages: An Ensemble Approach With XAI; *International Journal of Research in Science and Technology*, Vol 16, Issue 1, 57-75, DOI: http://doi.org/10.37648/ijrst.v16i01.007

cyberbullying content but also provides an understanding of why certain content is identified as cyberbullying. Such transparency is essential for establishing trust in the system and verifying the accuracy of the detected content.

The proposed approach involves using an ensemble model, specifically ERNIE-RNN, for the task across multiple languages. The ERNIE model is well-suited for handling multilingual text, while the RNN model can capture the sequential nature of language, making them complementary. Additionally, the model incorporates XAI techniques, such as Attention Visualization, to identify the most impactful elements of the input data that drive the system's decision on whether the content constitutes cyberbullying.

Our proposed method seeks to create a safer and more supportive online space for all users by developing a system that effectively and transparently identifies abusive content in user interactions. By developing a system that can effectively detect multilingual cyberbullying and provide explanations for its decisions, this study seeks to address the growing concern of online abuse on communication platforms and promote a more positive online experience for users.

## 2. Aim and Motivation

Due to the rise of technology and the abundance of data available all around us, traditional human oversight becomes useless because of the absence of scalable and efficient methods to identify cyberbullies and address the issue. Therefore, it is imperative to tackle the concern through automated processes that are rapid, effective, and precise, ultimately promoting social welfare.

The paper focuses on cyberbullying detection for various Indian regional languages using DL, NLP, and XAI. The existing focus on languages like English or Arabic/Urdu, is not as effective for the Indian context. By targeting Indian regional languages, the paper addresses a significant gap in the literature and potentially provides more relevant and accurate results for users in India.

The paper incorporates an ensemble model which is ERNIE+RNN. ERNIE is designed to understand the meaning of words in context, which is important for languages with rich contextual nuances like those in India. Combining ERNIE with RNN enables the ensemble model to grasp both the contextual embeddings from ERNIE as well as the sequential aspects of the language, resulting in a more robust and effective model for your task compared to models like RNN or CNN-BiLSTM. The ensemble model proposed in this study delivers superior performance on the evaluated dataset compared to previously used models in the field. In addition, the research includes the implementation of Explainable AI (XAI) methods. It not only helps to explain the predictions of the model but also provides insight into how the model is making decisions. The generation of attention visualization highlights the impactful areas of the input data that contribute to the decision-making activity of the model, making the outcome more interpretable and actionable. Overall, the paper’s approach and methodology show promise for substantial avenue for tackling the online bullying epidemic, particularly for Indian languages, and contribute valuable insights to the research community.

## 3. Related Work

The purpose of this study is to evaluate previous research, draw conclusions from it. To better understand the drawbacks in the prevailing work and for the efficiency and accuracy of the work proposed a thorough review of the body of literature was done.

Muneer et al. [1] proposed to use Stacking Ensemble Learning with an Enhanced version of BERT. Multiple models like SVM, Gradient Boosting and Random Forest are also combined to make it perform better.

López-Vizcaino et al. [2] employed ML and DL techniques like SVM and CNN. The metrics like recall, precision, F1-score, and accuracy are used in evaluating the proposed model. Jain et al. [3] employs Random Forest and SVM which is then evaluated using various performance metrics. In Sultan et al. [4], CNN used for image classification and SVM used for classification based on features extracted by CNN are employed. This provides a scope to enhance the capabilities of traditional text-based cyberbullying detection methods. Alduailaj et al. [5] focuses on word-

based data in the English and Arabic language majorly. The proposed mechanism employs SVM to identify cyberbullying utilizing data scraped from popular platforms like YouTube and Twitter. In Obaid et al. [6], LSTM is used to classify the data from Twitter and Kaggle. Subsequently, fuzzy logic was utilized to assess the intensity of the comments. The findings showed accuracy, F1-score, and recall values of 93.67%, 93.64%, and 93.62%, respectively.

Xingyi et al. [7] proposes the BERT architecture to get the hidden information about the semantics along with the BiSRU++ model. The model's performance is enhanced by using this combination. The results indicate that the combination outperforms existing models to flag cyberbully comments.

Fati et al. [8] uses CBOW's attention mechanisms to identify cyberbullying behaviour. CBOW builds semantic relations between different words and enhances the architecture's ability to grasp the meaning of the input text.

Ahmed et al. [9] proposes a VGG16-BiLSTM architecture to classify cyberbully content in memes. The visual bullying content is predicted by the VGG16 CNN while the textual one using the BiLSTM model. It also has the capacity to interpret Bengali memes to detect cyberbullying.

Rani et al. [10] implemented Natural Language Processing (NLP) and machine learning techniques to identify cyberbullying and hate speech in user-generated content, focusing on personal attacks within Twitter and Wikipedia discussions. Their model achieved over 90% accuracy on Twitter data and surpassed 80% accuracy on Wikipedia data. Similarly, Awate et al. [11] extracted textual, behavioural, and demographic features from the dataset, where textual indicators such as threatening language contributed significantly to accurate detection outcomes. Evaluation ofsystem's performance is based on the SVM and the Bernoulli NB. The SVM classifier demonstrated superior performance, achieving an overall accuracy of 87.14%, surpassing the accuracy of the Bernoulli Naive Bayes classifier, which was comparatively lower.

Muhariya et al. [12] proposed a K-means clustering-based method to detect cyberbullying content on Instagram, achieving 64.25% accuracy using 10-fold cross-validation. Huang et al. [13] applied logistic regression, decision tree, random forest, and SVM, with SVM delivering the highest true positive accuracy of 87%.

Balakeishnan et al. [14] explored whether personality traits and emotions gleaned from YouTube comments could improve cyberbullying detection. Using a combination of classifiers, they achieved over 95% accuracy, suggesting these psychological factors significantly aid in identifying cyberbullying.

Nahar et al. [15] focuses on cyberbullying detection, recognition, and determination of type using machine learning. Its lays emphasis on determining the type of cyberbullying, which adds complexity as it requires the model to classify incidents into specific categories such as verbal abuse or cyberstalking. However, potential drawbacks include the need of a large and diverse dataset to train the model effectively, and the challenge of classifying the model's results into specific cyberbullying types.

Dewani A et al. [16], their approach lies in addressing the challenges of limited labelled data and the informal nature of Roman Urdu. They propose a multi-stage methodology that includes data preprocessing, feature extraction, and ensemble classification. Their work contributes to the field by achieving competitive performance on cyberbullying detection tasks in low-resource, colloquial language settings. Key milestones include the development of an effective preprocessing pipeline for Roman Urdu text, the exploration of various ML algorithms for classification, and the design of an ensemble model that combines multiple classifiers for improved performance.

Mehta H. et al. [17] focused on interpreting the decision- making processes of complex AI models. They utilized the Google Jigsaw and HateXplain datasets, comprising content from Wikipedia, Twitter, and Gab. For the Jigsaw dataset, models such as decision trees, KNN, multinomial Naive Bayes, random forest, logistic regression, and LSTM were employed. Considering the second dataset, XAI methods such as LIME were applied. Variants of the BERT model were also created for effective performance in terms of explainability using the ERASER benchmark.

Neelakandan S et al. [18] explores CNNs, RNNs, and possibly Transformer-based architectures to classify and

categorise social media posts into cyberbullying or non-cyberbullying labels. They discuss dataset selection, preprocessing steps, model training, and evaluation, presenting results that show the effectiveness of DL models in identifying online abuse content.

Naveed Ejaz et al. [19] introduces a new dataset for detecting cyberbullying that covers a wide range of behaviours, such as peeress, aggressive texts, repetition, and intent to harm. This dataset is valuable because it includes multiple examples of online abuse, making it more effective for training cyberbullying detection algorithms. It helps researchers and developers improve their understanding and detection of cyberbullying online.

Sultan et al. [20] covers techniques such as SVM, Random Forest, Neural Networks, and others used in different studies for detecting cyberbullying behaviour in texts. The novelty of this review lies in its comprehensive analysis and analogy of different ML approaches, providing insights into their performance and suggesting potential directions for future research.

Roy et al. [21] explores the application of deep transfer learning (DTL) for online abuse detection. They use a pretrained DL model, such as a CNN or a Transformer-based model, and fine-tune it on a cyberbullying detection task. The discovery made in this study involve showcasing the impact of DTL in improving the performance of detection of cyberbullying models compared to traditional approaches. The novelty of this work lies in its use of transfer learning, which allows the model to leverage knowledge from a pretrained model and adapt it to the specific characteristics of cyberbullying detection tasks.

Luo et al. [22] introduced a deep learning-based approach for cyberbullying detection, likely utilizing architectures such as CNNs or RNNs to extract features from textual data. Their study highlights the effectiveness of combining multiple features—linguistic, semantic, and contextual—to achieve higher detection accuracy than single-feature models, showcasing the novelty of their integrated feature approach. Alam et al. [23] presented an ensemble machine learning method for cyberbullying detection, combining models like Decision Trees, Random Forest, and Gradient Boosting to build a robust classifier for identifying abusive behavior in text. Dewani A. et al. [24], on the other hand, introduced a novel technique focused on detecting cyberbullying in Roman Urdu content.

Additionally, a DL architecture, such as a CNN or a LSTM network, to classify the text data into cyberbullying or non-cyberbullying. The findings of this study contribute to the development of cyberbullying detection systems for languages with limited resources, such as Roman Urdu. Mahat et al. [25] presents an approach for detecting online bullying that is applicable to various communication platforms. The approach likely involves using DL techniques, such as CNNs or RNNs, to analyse social media posts and identify instances of cyberbullying.

Jain et al. [26] proposes unconventional method for detection of cyberbullying and hate speech content. The approach likely involves a combination of conventional machine learning techniques and heuristic-based methods to identify patterns and characteristics of cyberbullying and hate speech in text data. The novelty lies in the integration of conventional machine learning techniques with heuristic- based methods, which allows for a more comprehensive and accurate detection of cyberbullying and hate speech in text data.

Yadav et al. [27] conducted a study in comparing various DL techniques for detecting hate speech in the input text. They compare models such as RNNs, and CNNs and possibly Transformer-based architectures like BERT or GPT. The study evaluates the performance of these models using different performance metrics.

Lee E et al. [28] proposes a novel approach for detecting racism in tweets. They likely use sentiment analysis techniques to analyse the differential opinions expressed in tweets and identify instances of racist content. The model used is likely a Graph Convolutional Recurrent Neural Network (GCRNN) trained on a stacked ensemble of models to enhance the accuracy of racism detection. The discovery made in this study demonstrates the potent of their approach in accurately detecting instances of racism in tweets. The originality of their work lies in the use of sentiment analysis and the GCRNN model for racism detection, which allows for a more nuanced understanding of the context and intent behind racist tweets.

Berrimi et al. [29] proposes the use of attention-based neural networks for analysing unsuitable speech in Arabic. They likely use models such as Attention Mechanisms in RNNs or Transformer-based architectures to capture the context and semantic meaning of Arabic text. The discovery made in this study involves demonstrating the efficacy of attention mechanisms in improving the accuracy of detecting unsuitable speech compared to traditional approaches. The originality of their work lies in the application of attention-based networks specifically to Arabic text, which presents unique linguistic challenges compared to other languages.

Dubey et al. [30] presents a method for detecting toxic comments using LSTM networks. LSTMs are RNNs that are well-suited for processing sequential data like text. The discovery made in this study involves demonstrating the effectiveness of LSTM networks in accurately detecting toxic comments contrast to other machine learning approaches. d'Sa et al. [31] proposes a system to automatically detect toxic speech in online text messages. It leverages two word embedding techniques, BERT and fastText, to capture word meaning and relationships. These are then fed into DL classifiers to categorise text as toxic or non-toxic. This approach has the potential to improve online safety by filtering harmful content, but considerations like dataset bias and the challenge of interpreting context in language remain.

Yuvaraj N et al. [32] introduced a method for detecting online harassment by leveraging various features from textual data, including linguistic cues, sentiment insights, and contextual details. Their approach utilizes a deep decision tree classifier, with its novelty rooted in the integration of these diverse features to enable more precise and nuanced identification of cyberbullying behaviour.

## 4. Proposed Methodology

This study utilizes a deep learning-based model for the purpose of flagging abusive content. The problem is to detect if the data present is cyberbullying content or not which falls under text classification. A multilingual dataset consisting of English, Hindi, Bengali has been used. The model uses ERNIE and RNN models.

ML or DL algorithms alone can also perform this task of prediction but the above combination of algorithms provides a better accuracy and has its own advantages which makes it superior as compared to the existing algorithms.

### A. *ERNIE*

ERNIE was developed by Baidu, a Chinese Technology Company. It was specifically designed for NLP tasks itself and it integrates the knowledge learned from pre-trained language models and factual knowledge from external knowledge bases. It is capable to merge external knowledge from encyclopaedias, dictionaries, or other information repositories with the model. Its design aims to enhance and optimize accuracy and efficiency by incorporating knowledge and information from outside sources, thereby adding value to the model. It performs exceptionally well in NLP tasks like question answering, text classification, machine translation and language generation.

Figure 1 represents the working of the ERNIE Architecture.


FIGURE 1: ERNIE Model

The mathematical equations employed for the ERNIE Model are as follows:

1) Text Vectorization

In equation 1 first, every comment $t0$ in the dataset T is transformed into a sequence of fixed length consisting of characters $T0$ with a maximum length of $Lmax$.

$$T_0 = [t_{0,1}, t_{0,2}, \ldots, t_{0,c}, t_{0,len(T_0)}] \quad (1)$$

2) Word Embedding

In equation 2, each token $Wd$ in the comment $t0$ is transformed into a word vector H($We$). This is done using the word embedding layer of the ERNIE model. A sequential combination of word vectors $Si$ is obtained by summing all the vectors.

$$Si = [H(W1), H(W2), \ldots, H(We), \ldots, H(WLmax)] \quad (2)$$

3) Final Output

In equation 3, the word vectors $Si$ for all the comments in the dataset are combined into a sequence S which is the final output.

$$S = [S1, S2, \ldots, Si, \ldots, Slen(T0)] \quad (3)$$

### B. *RNN*

One type of ANNs is RNN's where the concept of directed cycles is followed and the output of one cycle is fed to the next cycle as an input. They also remember the data in the previous steps using the concept of Hidden State or the Memory State. RNNs provide exceptional performance for analysing sequences of data like the NLP tasks. RNNs are used in popular NLP applications like Siri, Google Translate, Voice Search etc. Figure 2 represents the working of the RNN Model.

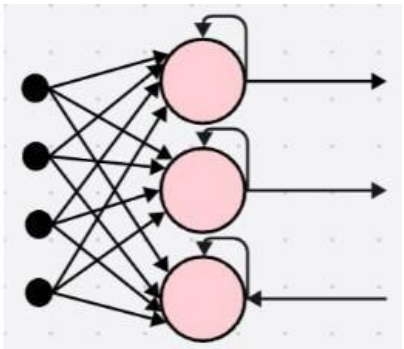

FIGURE 2: Recurrent Neural Networks

The mathematical equation for the RNN model is as follows: In equation 4, x is an input sequential combination of length T, the hidden states $ht$ for t = 1, 2, ..., T are computed as:

$$ht = \text{RNN}(xt, ht-1) \quad (4)$$

Where:

- at any time $t$, the hidden state is $ht$.

- at any time *t*, the input is *xt*.
- RNN represents the recurrent function.

The ultimate hidden state *ht* serves as the representation of the complete sequence. In equation 5, this hidden state undergoes processing through a linear layer, succeeded by a SoftMax function, to derive the output probabilities for each class:

$yt$ = SoftMax (Linear($ht$)) (5)

Where:

- at any time *t*, the hidden state is *ht*.
- at any time *t*, the output is *yt*.
- SoftMax represents the SoftMax function.

The equation 6 is a representation of the loss function used is the model which is the cross-entropy loss defined as:

$$Loss = -\sum_{i=1}^{N} (yi \log(\hat{y}i)) \quad (6)$$

Where:

- there are N classes.
- *yi* is the true label (0 or 1 in this case)
- at any time *t*, the output is *yt*.
- at any time *t*, the predicted probability for class i is *y^i*.

## 5. System Architecture

A hybrid architecture consisting of a pre-trained ERNIE combined with RNN is utilised in this study.

Both the models are combined so that the benefits of both of them can be combined and an efficient cyberbullying detection system can be developed. The proposed model consists of two components:

1) ERNIE pretrained model - Here, the tokenizer and model are loaded from the ERNIE pretrained functions Auto Tokenizer and Auto Model respectively.
2) RNN model - The RNN model architecture consists of RNN layer and the linear layer which does the classification task. A many-to-one RNN model is used here. Many words are given as input and the output is only one i.e. the label.

Both these models are given as an input and combined into the ERNIE_RNN model for joint training as visually depicted in Figure 3.

Figure 4 is a representation of the block diagram of the system architecture of our proposed model. As shown in the

figure, the steps followed are:

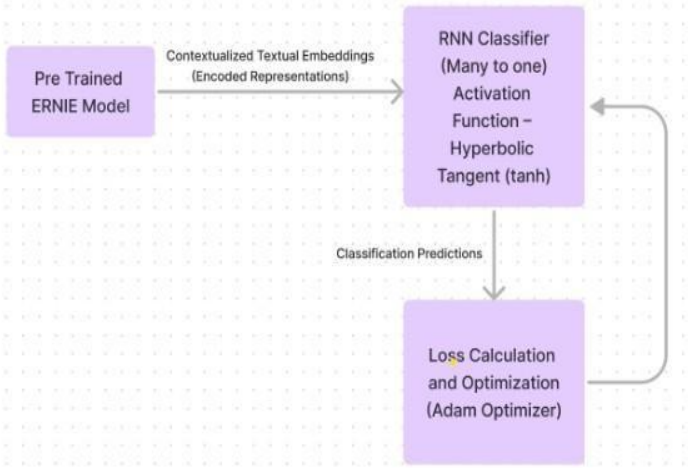


FIGURE 3: Crisp Diagram of the Proposed Model

TABLE 1: Model Parameters

| Parameter | Value |
|---|---|
| Batch Size | 8 |
| Optimizer | Adam |
| Learning Rate | 0.001 |
| Loss Function | Cross Entropy Loss |
| Activation Function | Hyperbolic Tangent (tanh) Function |

### C. *DATA ACQUISITION*

The data is gathered from various sources which include the comments and texts posted on Twitter, Instagram, YouTube and Kaggle. All this gathered data is then curated into a single dataset for the languages English, Hindi, and Bengali.

### D. *DATA PREPROCESSING AND STANDARDIZATION*

The dataset is then sent for cleaning and preprocessing. This process includes actions like removal of stop words, URLs, mentions, special characters and numerical data from the textual data. To ensure consistency, all text data is converted to lowercase. It is then tokenized using the Hugging Face Transformers library. The dataset is labelled using a 0-1 classifier where 0 means that the text does not contain cyberbullying content while the text labelled as 1 does.

### E. *MODEL TRAINING*

After the preprocessing and cleaning stage is completed, the data is fed to the model for training. The ensemble model works by extracting features from the pre-trained ERNIE model and then passing them through the RNN classifier for classification.

1) The input data is shuffled and divided into batches using a Data Loader.

2) The gradients and the model parameters are set to zero after every batch for calculating the same for the next batch.

3) The values are assigned to input, attention mask, and label for each batch. Here,

- Inputs refer to the textual data i.e. comments that are fed to the model for prediction.

- Attention Masks are applied to each input and these exist to help the model decide on the tokens that should be given importance while predicting and those which can be potentially ignored.
- Labels refer to the output that we get from the model. The label is either 0 which indicates no cyberbullying content or 1 which indicates cyberbullying content.

4) The output is then obtained by passing the data through the model.

5) The loss is calculated by computing the difference between the predicted outputs and the actual outputs. The loss function which is used in the proposed model is the cross-entropy loss function.

6) The model's parameters are fine-tuned using the Adam optimizer, with a learning rate of 0.001, and the batch's accuracy is calculated.

7) The average loss and the average accuracy of the entire epoch is then calculated by taking in the values calculated for each batch.

8) A heat map is generated for every epoch where the attention weights are visualised. This visualisation depicts the amount of importance given to different tokens while prediction.

### F. *EXPLAINABLE AI*

Explainable AI (XAI) encompasses techniques aimed at making the decisions of AI and machine learning models transparent and understandable to humans. Integrating explainability into these models helps build trust and ensures clarity, enabling developers, regulators, and users to more effectively interpret and engage with the systems. The benefits of XAI are shown in the Figure 5.

The graph in Figure 6 shows relation between interpretability and accuracy.

So, to better understand models with high accuracy like our model Ernie-RNN, XAI is used which explains the users and the model developers with the black box area of the code making it more interpretable. Figure 7 shows how to select the eXplainable AI techniques that can be used for specific models. Using this, Attention Visualisation XAI technique was used for our proposed model.

The code utilises model-specific XAI techniques, specifically:

1) Attention Visualisation: While the code uses an ERNIE model with an attention mechanism, attention visualisation itself is not model-specific. It's a general technique applicable to various deep learning models, particularly those involving sequence data like text, to understand how the model prioritizes the specific areas of the input text.

2) Performance Metrics: The code calculates and reports accuracy, precision, and F1-score. These are common performance metrics used to evaluate the overall performance of the model, offering insights into its ability to correctly classify sentiment. These metrics are not specific to XAI but are crucial for understanding model performance and potential biases.

## 6. Dataset Statistics

Table 2 shows dataset statistics in English, Hindi, and Bengali, with varying sizes and attributes which is scraped from popular social media forums and Kaggle. The English dataset contains 887 records, the Hindi dataset includes 4579 records, and the Bengali dataset comprises of 2928 records. Each dataset consists of 2 attributes: 'Text' and 'Label'.

TABLE 2: Dataset Statistics for the various languages employed

| Language | No. Of Records | No. Of Attributes | Training Ratio | Training Records | Testing Records |
|---|---|---|---|---|---|
| English | 887 | 2 | 0.8 | 709 | 178 |
| Hindi | 4579 | 2 | 0.8 | 3663 | 916 |
| Bengali | 2928 | 2 | 0.8 | 2342 | 586 |

## 7. Results and Discussion

Figure 8 is a bar chart with X-axis as the length of comments and Y-axis as the number of comments, the given bar chart in Figure 8 shows we have the maximum number of comments whose length lies between the range of 0-200.

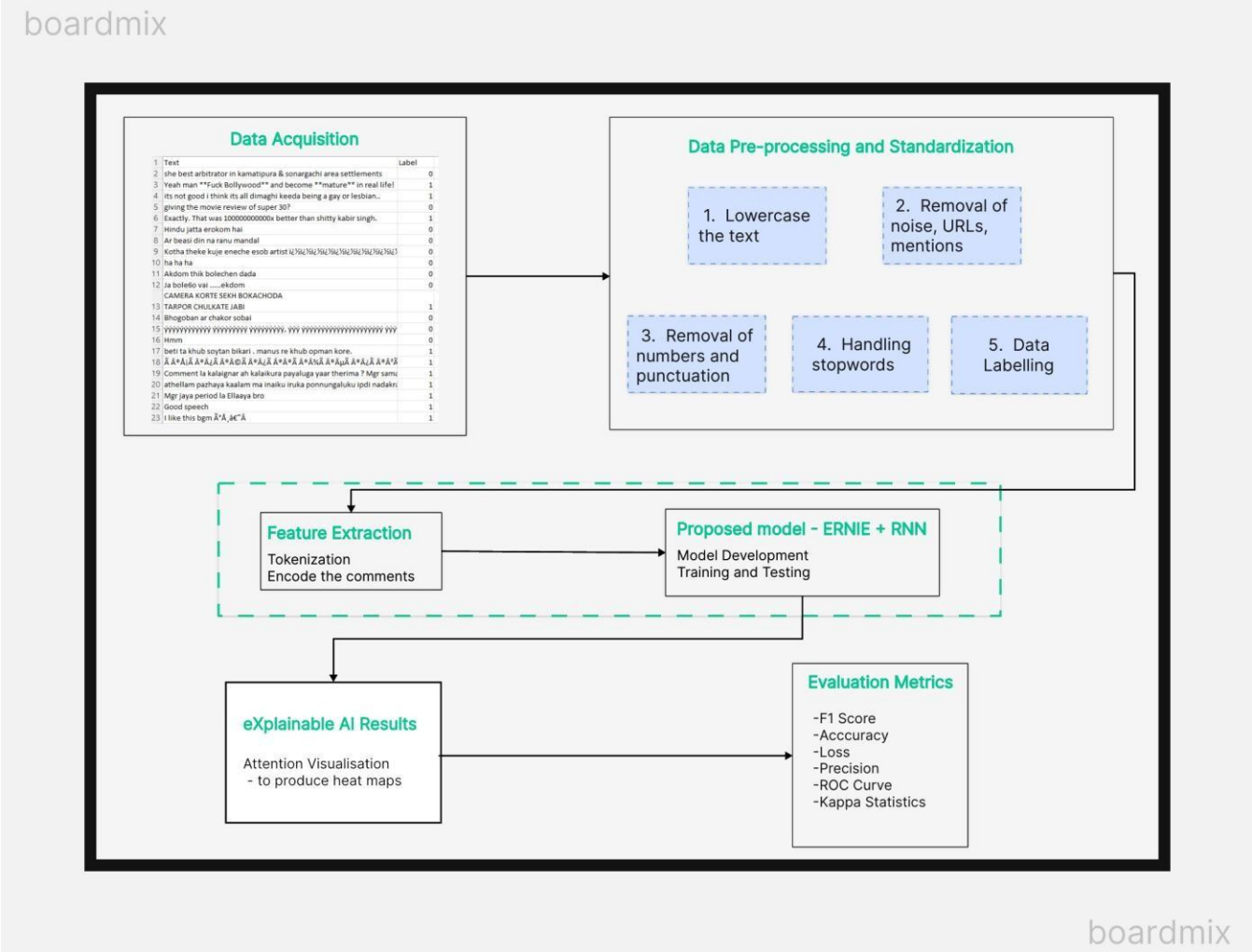


FIGURE 4: System architecture

Figure 9 is a pie chart showing that the multilingual dataset (Hindi, English, Bengali) contains 52.6% of data which is not cyberbully content and is labelled as 0 and 47.4% data which contains cyberbully content and is labelled as 1.

Figure 11 depicts a word cloud, visually representing text data—specifically, cyberbully content labelled as 1 in the dataset. In this visualization, the size of each word corresponds to its frequency or significance within the text. Larger, more prominent words occur more frequently. Word clouds serve as a helpful tool for swiftly pinpointing the most prevalent words in a dataset and gaining insight into the overarching themes or subjects within the text.

Figure 10 is a bar chart with X-axis representing the different architecture model tested against the multilingual dataset (Hindi, English, Bengali) and Y-axis representing the accuracy of each model. There are 9 bar plots as the languages English, Hindi and Bengali have been tested against RNN, CNN-BiLSTM and Ernie+RNN(proposed model). The visualisation shows that the proposed model gives better accuracy for all the 3 languages compared to the other architecture models.

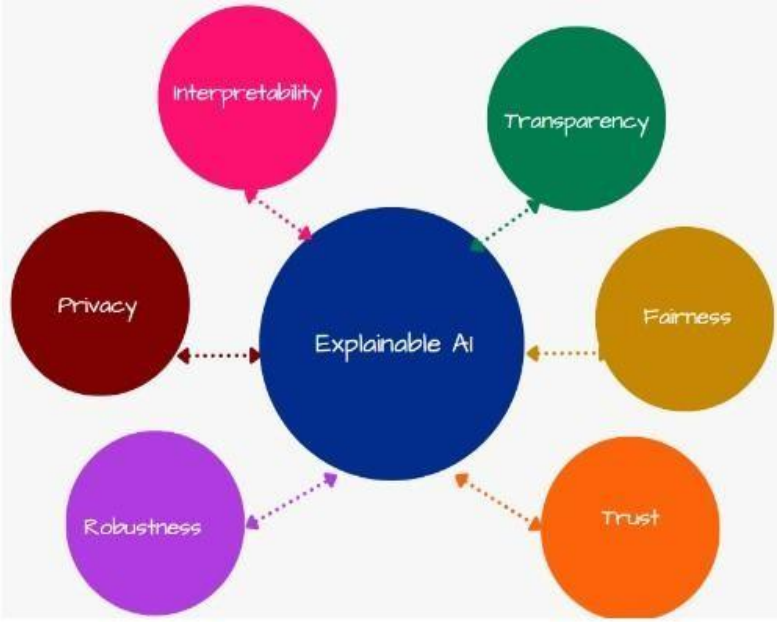


FIGURE 5: Benefits of Explainable AI

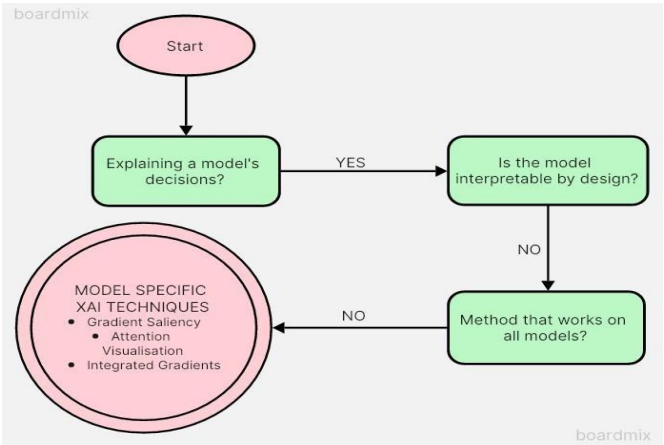


FIGURE 6: Graph between interpretability and accuracy

Table 3 shows a comparison table showing the different parameters and their results- for three different architecture models and their loss, prediction, f1-score, accuracy, and ROC-curve area for the multilingual dataset (Hindi, English,

Bengali).

Figure 12 illustrates an attention visualization in the form of a heat map, which was generated during the 13th training epoch of the model while it was being trained on English- language data. This heat map serves as a graphical representation of the attention mechanism employed by the model, highlighting the relative importance assigned to various tokens (words) within a sentence during the learning process. Both the x-axis and y-axis display the tokens, allowing for a clear interpretation of how each word relates to and influences the others in the context of the input sentence. The attention values, which are visualized through varying colour intensities, reflect how much focus the model places on specific tokens while processing the text. These values play a crucial role in guiding the model's decision- making process, particularly in determining whether the sentence contains any form of cyberbullying. The deeper or more intense the colour, the higher the level of importance assigned to that token, indicating its influence on the model's final classification output. the amount of importance given to that token. The darker the colour, more is the importance given.

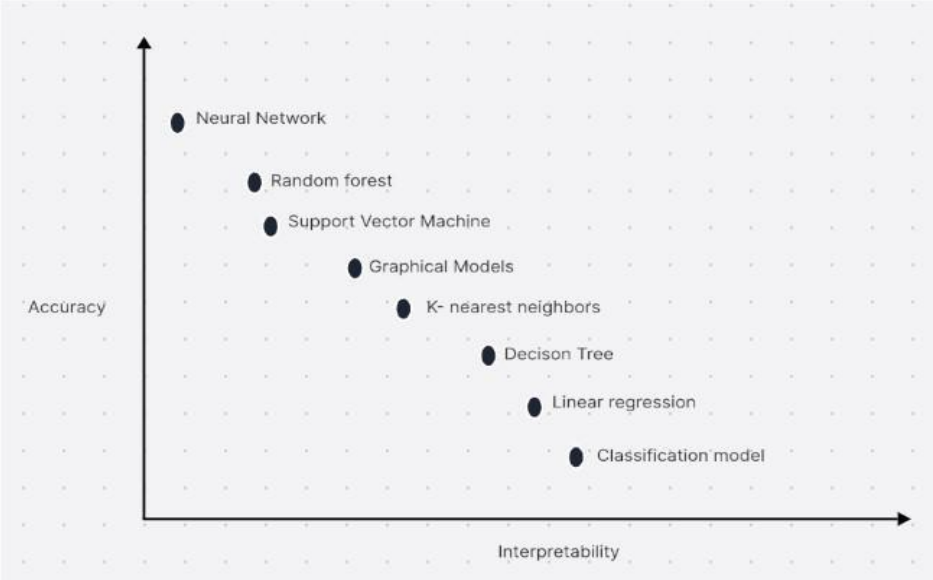


FIGURE 7: Explainable AI

Here are the ROC Curves (Figures 13, 14, and 15) depicting the performance of our proposed Ernie-RNN models across English, Hindi, and Bengali languages. The ROC Curve illustrates the True Positive Rate (TPR) plotted against the False Positive Rate (FPR), with a higher area under the curve indicating better model performance. Specifically, Figure 13 represents the ROC Curve for English, Figure 14 for Hindi, and Figure 15 for the Bengali language.

## 8. Conclusion

In conclusion, the paper aims to develop a cyberbullying detection system for various Indian regional languages using an ensemble model (ERNIE-RNN) and eXplainable Artificial Intelligence (XAI). By combining ERNIE and RNN models, the system leverages the strengths of both to improve accuracy and robustness. The integration of XAI techniques enhances transparency and trust in the model's decisions. Overall, this paper has the potential to significantly contribute to combating cyberbullying in Indian regional languages, ultimately fostering a safer online environment.

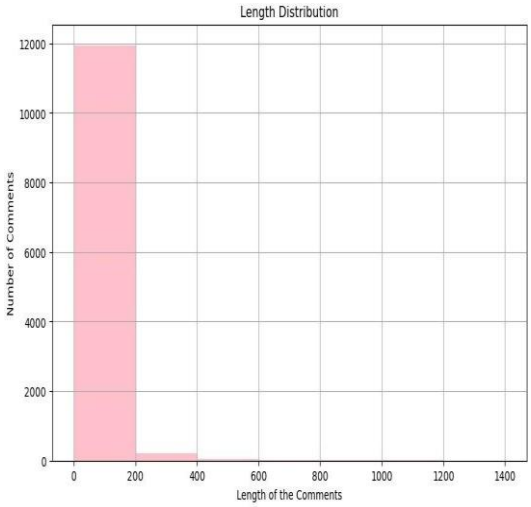


FIGURE 8: Length of Comments

**TABLE 3: Comparison of Performance metrics for various models**

| Architecture | Language | Loss | Precision | F1 Score | Accuracy | Area Under ROC Curve | Kappa Statistic Score |
|---|---|---|---|---|---|---|---|
| RNN | English | 0.47 | 76% | 76% | 76% | 0.86 | 0.457 |
| | Hindi | 0.56 | 78% | 78% | 79% | 0.86 | 0.561 |
| | Bengali | 0.43 | 80% | 79% | 81% | 0.86 | 0.517 |
| CNN-BiLSTM | English | 0.66 | 77% | 76% | 79% | 0.86 | 0.537 |
| | Hindi | 0.77 | 79% | 79% | 79% | 0.86 | 0.572 |
| | Bengali | 0.61 | 77% | 76% | 78% | 0.79 | 0.379 |
| RNN-ERNIE | English | 0.035 | 97% | 98% | 98% | 1.00 | 0.939 |
| | Hindi | 0.265 | 84% | 87% | 87% | 0.98 | 0.828 |
| | Bengali | 0.337 | 84% | 84% | 84% | 0.96 | 0.767 |

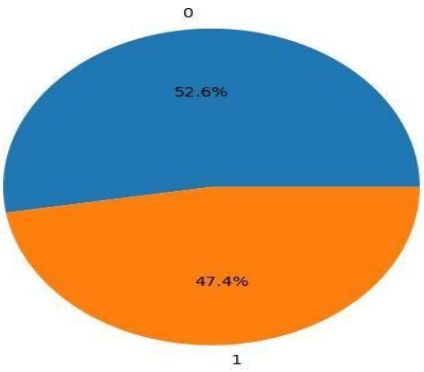


FIGURE 9: Distribution of Cyberbully Labels Languages

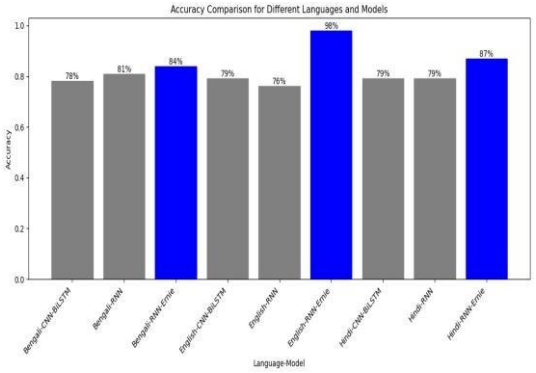


FIGURE 10: Accuracy Comparison for different and different Models

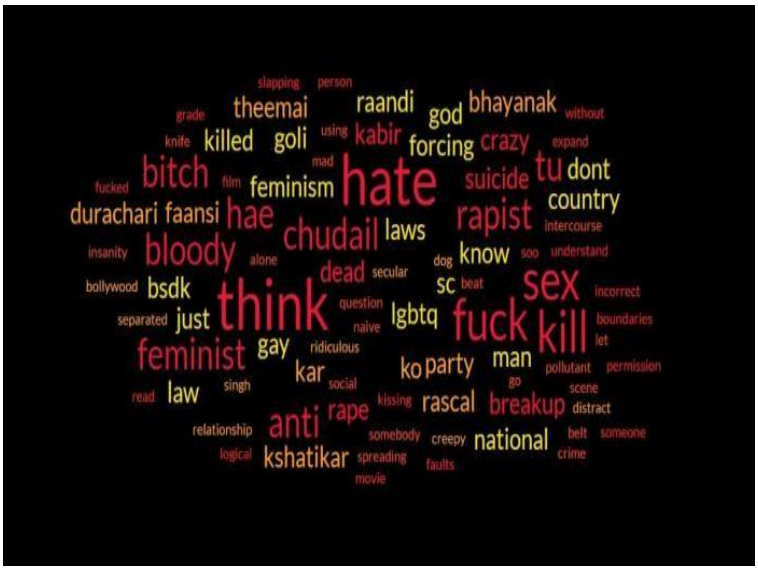


FIGURE 11: Word Cloud Visualisation

Moreover, the future scope of this paper on cyberbullying detection for various Indian regional languages using deep learning, Natural Language Processing (NLP), and eXplainable Artificial Intelligence (XAI) is highly promising. One avenue for future development is to enhance the performance of the ERNIE-RNN model by fine-tuning the hyper parameters and consolidating additional linguistic features specific to Indian languages. Additionally, expanding the dataset and model training to include more Indian regional languages can increase the reach and effectiveness of this paper's cyberbullying detection system. Integrating the paper's model with social media platforms or online forums for real time cyberbullying detection and intervention is another valuable application. Furthermore, incorporating other modalities such as images, videos, and audio in addition to text could provide a more comprehensive approach to cyberbullying detection. Developing a user-friendly interface that allows users to interact with the model, view explanations for predictions, and provide feedback could enhance the usability and adoption of your system. Collaborating with experts in psychology and education to understand the psychological impact of cyberbullying and to develop prevention and intervention strategies based on the insights from the paper's model is also a promising direction.

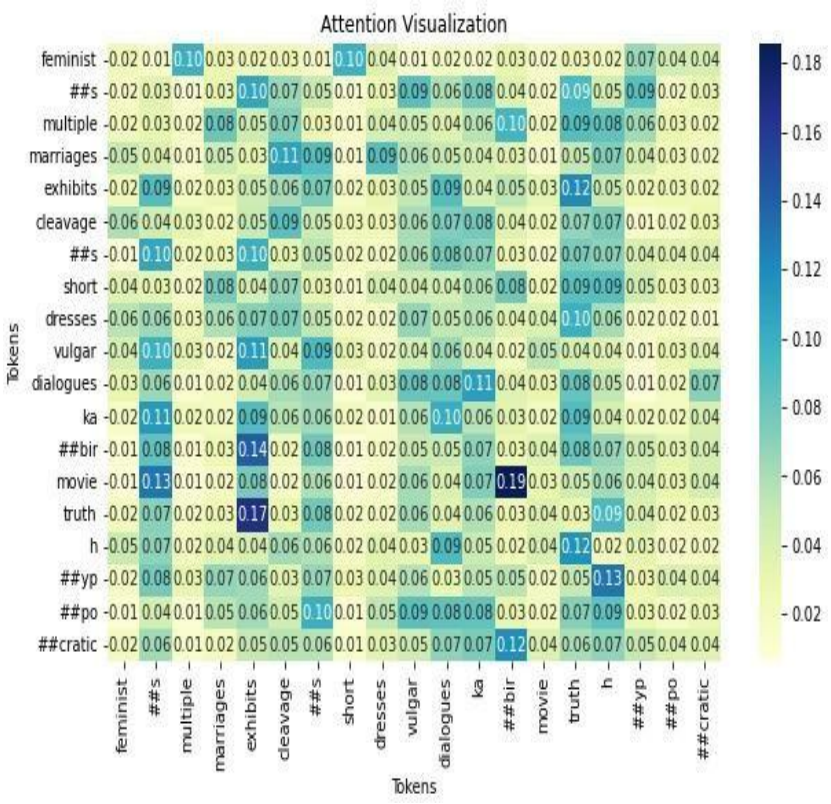


FIGURE 12: Attention Visualisation

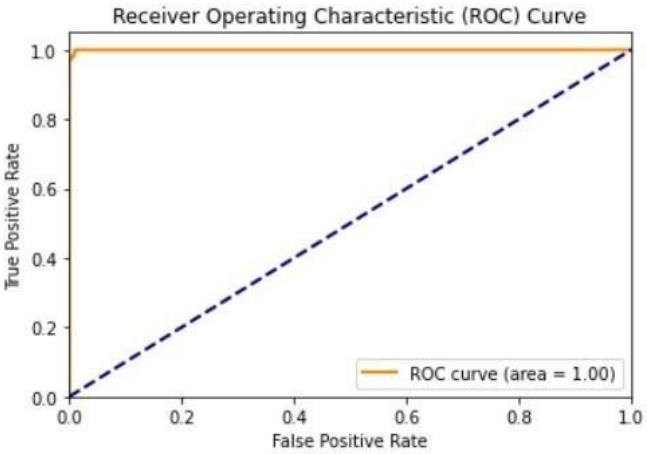


FIGURE 13: ROC Curve - English

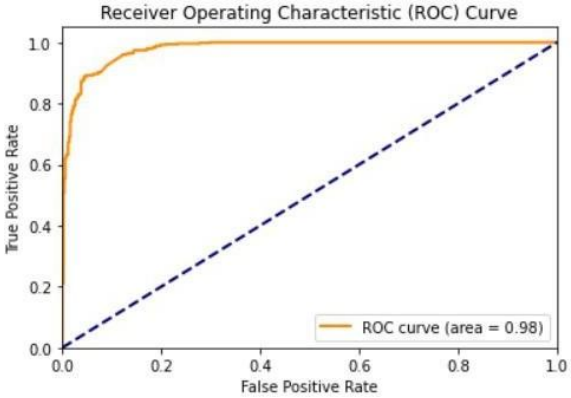


FIGURE 14: ROC Curve - Hindi

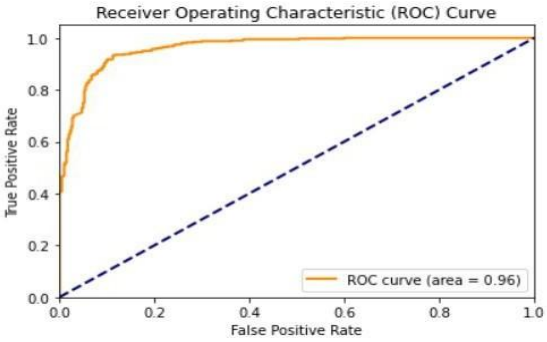


FIGURE 15: ROC Curve - Bengali

**References**

1. Muneer, A., Alwadain, A., Ragab, M. G., & Alqushaibi, A. (2023). Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. *Information*, *14*(8), 467. https://doi.org/10.3390/info14080467

2. López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., & Cacheda, F. (2021). Early detection of cyberbullying on social media networks. *Future Generation Computer Systems*, *118*, 219–229. https://doi.org/10.1016/j.future.2021.01.006

3. Jain, V., Kumar, V., Pal, V., & Vishwakarma, D. K. (2021, April). Detection of cyberbullying on social media using machine learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1091–1096). IEEE. https://doi.org/10.1109/ICCMC51019.2021.9417777

4. Sultan, T., Jahan, N., Basak, R., Jony, M. S. A., & Nabil, R. H. (2023). Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition. *International Journal of Intelligent Systems and Applications (IJISA)*, *15*(2), 1–13. https://doi.org/10.5815/ijisa.2023.02.01

5. Alduailaj, A. M., & Belghith, A. (2023). Detecting Arabic cyberbullying tweets using machine learning. *Machine Learning and Knowledge Extraction*, *5*(1), 29–42. https://doi.org/10.3390/make5010018

6. Obaid, M. H., Guirguis, S. K., & Elkaffas, S. M. (2023). Cyberbullying detection and severity determination model. *IEEE Access*, *11*, 110723–110735. https://doi.org/10.1109/ACCESS.2023.3321237

7. Xingyi, G., & Adnan, H. (2024). Potential cyberbullying detection in social media platforms based on a multi-task learning framework. *International Journal of Data and Network Science*, *8*(1), 25–34. https://doi.org/10.5267/j.ijdns.2023.10.021

8. Fati, S. M., Muneer, A., Alwadain, A., & Balogun, A. O. (2023). Cyberbullying detection on Twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction. *Mathematics*, *11*(16), 3567. https://doi.org/10.3390/math11163567

9. Ahmed, M. T., Akter, N., Rahman, M., Das, D., Azm, T., & Rashed, G. (2023). Multimodal cyberbullying meme detection from social media using deep learning approach. *International Journal of Computer Science and Information Technology (IJCSIT)*, *15*(4), 27–37. https://doi.org/10.5121/ijcsit.2023.15403

10. Rani, M. U., Ramesh, M. A., Srinivas, M. G., Ganesh, M. S., & Lakshmi, M. D. V. (2023). Detection of cyberbullying on social media. *Journal of Engineering Sciences*, *14*(4), 1–15. https://doi.org/10.36897/jes.2023.14.04.01

11. Awate, V., Bagad, V., Jadhav, S., & Jadhao, B. (2023). Detection of cyberbullying on social media using machine learning. *Advancement in Image Processing and Pattern Recognition*, *6*(2), 6–12.

12. Muhariya, A., Riadi, I., Prayudi, Y., & Saputro, I. A. (2023). Utilizing K-means clustering for the detection of cyberbullying within Instagram comments. *Ingénierie des Systèmes d'Information*, *28*(4).

13. Huang, H., & Qi, D. (2023). Cyberbullying detection on social media. *Higher Education and Oriental Studies*, *3*(1).

14. Balakrishnan, V., & Ng, S. K. (2023). Personality and emotion based cyberbullying detection on YouTube using ensemble classifiers. *Behaviour and Information Technology*, *42*(13), 2296–2307. https://doi.org/10.1080/0144929X.2022.2079756

15. Nahar, K. M., Alauthman, M., Yonbawi, S., & Almomani, A. (2023). Cyberbullying detection and recognition with type determination based on machine learning. *Computers, Materials & Continua*, *75*(3).

16. Dewani, A., Memon, M. A., Bhatti, S., Sulaiman, A., Hamdi, M., Alshahrani, H., Alghamdi, A., & Shaikh, A. (2023). Detection of cyberbullying patterns in low resource colloquial Roman Urdu microtext using natural language processing, machine learning, and ensemble techniques. *Applied Sciences*, *13*(4), 2062. https://doi.org/10.3390/app13042062

17. Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms*, *15*(8), 291. https://doi.org/10.3390/a15080291

18. Neelakandan, S., Sridevi, M., Chandrasekaran, S., Murugeswari, K., Pundir, A. K. S., Sridevi, R., & Bheema, T. (2022). Deep learning approaches for cyberbullying detection and classification on social media.

19. Ejaz, N., Razi, F., & Choudhury, S. (2024). Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm. *Computers in Human Behavior*, *153*, 108123. https://doi.org/10.1016/j.chb.2023.108123

20. Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A., Zholdassov, Y., & Abdrakhmanov, R. (2023). Review of machine learning techniques in cyberbullying detection.

21. Roy, P. K., & Mali, F. U. (2022). Cyberbullying detection using deep transfer learning. *Complex & Intelligent Systems, 8*, 5449–5467. https://doi.org/10.1007/s40747-022-00756-6

22. Luo, Y., Zhang, X., Hua, J., & Shen, W. (2021). Multi-featured cyberbullying detection based on deep learning. In *2021 16th International Conference on Computer Science and Education (ICCSE)* (pp. 746–751). IEEE. https://doi.org/10.1109/ICCSE52927.2021.9548577

23. Alam, K. S., Bhowmik, S., & Prosun, P. R. K. (2021). Cyberbullying detection: An ensemble-based machine learning approach. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 710–715). IEEE. https://doi.org/10.1109/ICICV52345.2021.9515567

24. Dewani, A., Memon, M. A., & Bhatti, S. (2021). Cyberbullying detection: Advanced preprocessing techniques and deep learning architecture for Roman Urdu data. *Journal of Big Data, 8*(1), 1–20. https://doi.org/10.1186/s40537-021-00504-3

25. Mahat, M. (2021). Detecting cyberbullying across multiple social media platforms using deep learning. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 299–301). IEEE. https://doi.org/10.1109/ICACITE51222.2021.9538727

26. Jain, N., Hegde, A., Jain, A., Joshi, A., & Madake, J. (2021). Pseudo-conventional approach for cyberbullying and hate-speech detection. In *2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)* (pp. 1–8). IEEE. https://doi.org/10.1109/ICAC353735.2021.9591367

27. Yadav, Y., Bajaj, P., Gupta, R. K., & Sinha, R. (2021). A comparative study of deep learning methods for hate speech and offensive language detection in textual data. In *2021 IEEE 18th India Council International Conference (INDICON)* (pp. 1–6). IEEE. https://doi.org/10.1109/INDICON52577.2021.9701347

28. Lee, E., Rustam, F., Washington, P. B., El Barakaz, F., Aljedaani, W., & Ashraf, I. (2022). Racism detection by analysing differential opinions through sentiment analysis of tweets using stacked ensemble GCRNN model. *IEEE Access, 10*, 9717–9728. https://doi.org/10.1109/ACCESS.2022.3143125

29. Berrimi, M., Moussaoui, A., Oussalah, M., & Saidi, M. (2020). Attention based networks for analysing inappropriate speech in Arabic text. In *2020 4th International Symposium on Informatics and Its Applications (ISIA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISIA50717.2020.9311357

30. Dubey, K., Nair, R., Khan, M. U., & Shaikh, S. (2020). Toxic comment detection using LSTM. In *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)* (pp. 1–8). IEEE. https://doi.org/10.1109/ICAECC48674.2020.9312322

31. d'Sa, A. G., Illina, I., & Fohr, D. (2020). BERT and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)* (pp. 1–5). IEEE. https://doi.org/10.1109/OCTA51217.2020.9312707

32. Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G., & Rajan, A. R. (2021). Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. *Computers & Electrical Engineering, 92*, 107186. https://doi.org/10.1016/j.compeleceng.2021.107186