

Automatic Detection of Cyberbullying in Multilingual Texts Using XLM-RoBERTa

¹Vignesh Pinni, ²Dr. S. Sri Harsha

^{1, 2} Department of CSE, Koneru Lakshmaiah Foundation, Vaddeswaram, Andhra Pradesh, India

Abstract: - Detection of cyberbullying in multilingual and code-mixed social media content is difficult as it is not written in formal English, translated through transliteration and has ambiguity in context especially in the Hinglish text. The aim of this research is to suggest an architecture of transformer, fine-tuned with XLM-RoBERTa to classify the data into two classes: cyberbullying and non-cyberbullying for the identification of cyberbullying in Hinglish social media posts. The proposed model is tested against the conventional baselines of ML and DL with the use of classical classification measures. The experimental results demonstrate that the XLM-RoBERTa model demonstrates an accuracy of 94.05%, a balanced precision-recall trade-off of 0.97 and an AUC of 0.97, which are all higher than the base approaches. The results show that multilingual transformer models can learn the pattern of code-mixed multilingual text, both with respect to the context and semantics, which makes them suitable to be used in cyberbullying detection in real-world social media setting.

Keywords: Cyberbullying Detection, Hinglish, Code-Mixed Text, Multilingual NLP, XLM-RoBERTa, Transformer Models, Social Media, Natural Language Processing (NLP), Deep Learning, Text Classification.

1. Introduction

With the rise of social media platforms, interactions between linguistic groups are more common in the online space, and there is an increase in cyberbullying. So the detection of cyberbullying has been turned to be automatic, as manual moderation can no longer cope with large, real-time content created by users. Though great progress has been made in the domain of English language literature, the identification of cyberbullying in multilingual and code-mixed setting is a challenging research subject.

Depending on the context (e.g. India), users may also use the so called Hinglish, which is a combination of English and Romanized Hindi, informal syntax, transliteration, slang and words and phrases which are context dependent. These language characteristics cause confusion, and restrict the performance of the standard ML methods built upon surface characteristics like n-grams or TF-IDF vectors. While conventional DL models like CNN and LSTM models have achieved success in sequential modelling, they are not able to capture long range contextual dependency and cross-lingual semantics in code-mixed text.

Recently, transformer-based language models have been shown to be effective for collecting contextual and semantic representations of over one language. When dealing with code-mixed and low-resource languages, large-scale corpora-based multilingual transformers are a promising base. However, its application in the field of Cyberbullying in Hinglish is somewhat restricted and detailed comparisons with the classical and DL baselines are not presented elaborately.

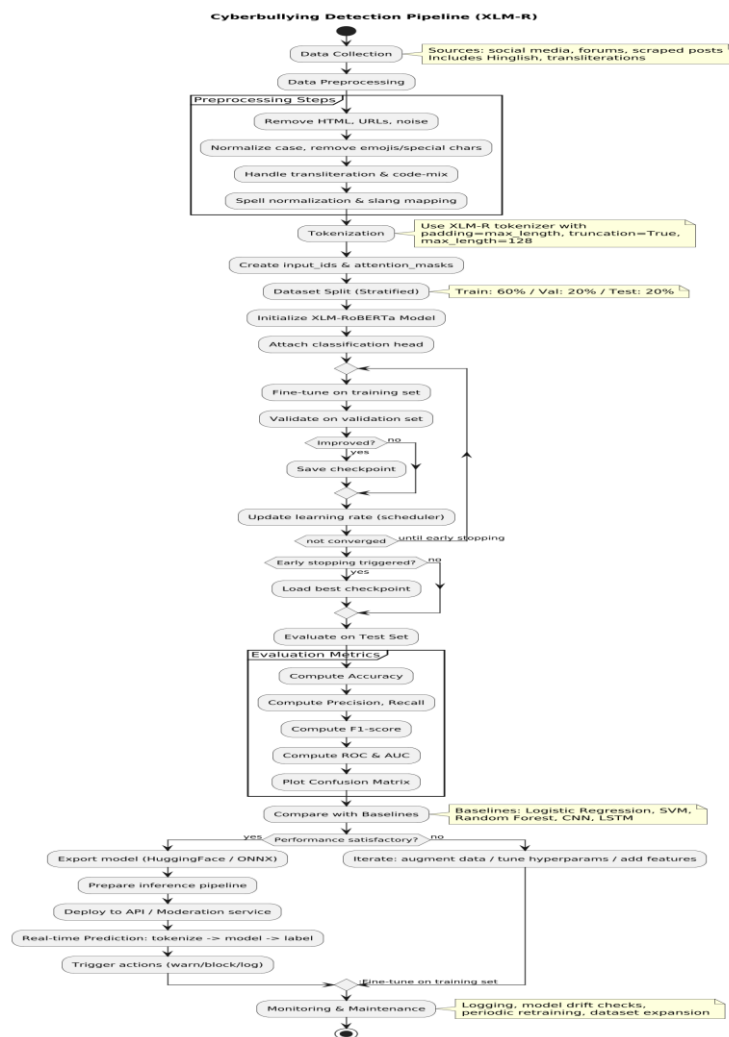
To address these limitations, the current study suggests a transformer-based cyberbullying detection model using XLM-RoBERTa, with the aim of training a binary classification model for Hinglish social media content. The proposed approach is tested on a big real-world dataset and benchmarked with the classic models of ML and DL through the commonly used performance metrics. The results demonstrate the improvement of the performance of multilingual contextual representations to detect cyberbullying in linguistically complicated social media situations, and the transformer-based models are suitable to apply to the real content moderation tasks.

2. Objectives

The objective of this research is to investigate the challenges of detecting cyberbullying in code-mixed and multilingual Hinglish social media posts. In this study, a dataset is curated and preprocessed with informal Hinglish text to identify cyberbullying and non-cyberbullying posts. It aims to address transliteration, slang, abbreviations, spelling variations and contextual ambiguity common in social media content. The research also aims to analyze the contextual and semantic characteristics of Hinglish text and to develop a transformer-based approach for cyberbullying detection. XLM-RoBERTa is fine-tuned to classify social media posts into cyberbullying and non-cyberbullying classes. The accuracy, precision, recall, F1-score, and AUC of the proposed model are compared to traditional machine learning and deep learning techniques. The study also investigates the extent to which XLM-RoBERTa can understand patterns and meaning in code-mixed multilingual text. Finally, the research aims toward determining the effectiveness and practical feasibility of the proposed approach for automated cyberbullying detection in real-world social media environments.

3. Methods

In this section, the proposed cyberbullying detection framework is introduced, which is based on XLM-RoBERTa for multi-lingual and code-mixed Hinglish text. The methodology is structured upon three phases: pre-processing of data, pre-processing of contextual embedding generation, supervised fine-tuning and quantitative evaluation. The whole workflow is properly summarised in Algorithm 1.



3.1 Problem Definition

Suppose the data are given in the following form:

$$D = \{(x_i, y_i)\}_{i=1}^N$$

x_i represents a sample of 'Hinglish' social media text.

y_i is the label associated with the label x_i .

The numbers 0 and 1 indicate respectively mean no cyberbullying and cyberbullying.

The aim is to learn a classification function:

$$f_{\theta}: X \rightarrow Y$$

The model is trained on the parameters θ and the output is the predicted labels Y_{pred} , which are generated by the XLM-RoBERTa-based model.

3.2 Dataset Acquisition and Preprocessing

The Hinglish corpus D adopted for this study includes 17069 multilingual and code-mixed social media texts of mixed language constructs which are common in the social media platforms of India and english, romanised hindi and mixed language constructs. The number of cyberbullying cases is 10,588 while the number of non-cyberbullying cases is 6,481, which forms a naturally unbalanced representation of classes in real life online interactions.

All text samples are passed through a typical pre-processing pipeline which incorporates lower-casing, punctuations marks and other special characters removal and noise removal to reduce the lexical variability. Special attention is paid to code-mixed tokens and words in transliteration to minimize the number of errors in informal grammar and spelling that would result in change in semantic meaning. This pre-processing method would be consistent with transformer-based tokenisation, which would not lose any contextual information, essential for the identification of implicit and explicit meanings.

The data is pretty wide in terms of the length of text as well as structure of its texts with an average text length of 279 and the most length text of 6249, which is basically brief conversational and long context-based posts. Moreover, the dataset reveals that the mean Flesch readability score is poor (35.8), which means that the material is informal, complex, and non-standard language also contributing to the deployment of contextual transformer models.

The data is pre-processed and then divided into training, validation and test sets in the ratio of 60:20:20 with the stratification to maintain the class proportions within each set. This stratification allows training and testing of models in a reliable manner which closely mimics real life classes.

3.3 Tokenization using XLM-RoBERTa

All the texts are preprocessed and tokenized into x_i using the XLM-RoBERTa tokeniser, which works with multilingual and code-mixed text. Tokenisation: transforms raw text into input token ids and attention masks:

$$\mathcal{T}(x_i) = \{\text{input_ids}_i, \text{attention_mask}_i\}$$

All sequences are padded or shortened to a fixed length $L=128$ to achieve a balance between the contextual representation and computation performance.

3.4 Contextual Embedding Generation

We feed the tokenised inputs to the pre-trained XLM-RoBERTa encoder to get contextual embeddings. Let:

$$H_i \in \mathbb{R}^{L \times d}$$

where d is the hidden dimension of the encoder, and represent the output embedding matrix. The context of the special categorisation token [CLS] is taken as:

$$h_i = H_i^{[CLS]} \in \mathbb{R}^d$$

This model models the semantic, syntactic and contextual relationships of code-mixed Hinglish text.

3.5 Classification and Model Fine-Tuning

Then the recovered embedding h_i s are fed into a fully connected classification layer and then passed through a Softmax layer:

$$\hat{y}_i = \text{Softmax}(Wh_i + b)$$

W is an $2 \times d$ dimensional trainable parameter, and b is a 2-dimensional trainable parameter.

The model is trained with the cross-entropy loss function:

$$\mathcal{L} = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

The model is optimized with the AdamW optimiser, 8 samples per batch and 3 epochs. The validation performance of the model is checked at every epoch and early halting with checkpointing is done to prevent overfitting and to guarantee stable convergence.

3.6 Algorithmic Workflow

The entire cyberbullying detection process is summarized in Algorithm 1, which shows the steps of loading data sets, training, testing and evaluation.

Algorithm 1: Cyberbullying Detection using XLM-RoBERTa

Input:

Hinglish dataset

$$D = \{X, Y\}$$

X : Text samples and Y : class label $\{0, 1\}$.

Output:

Predicted labels Y_{pred}

Step 1: Data Collection

Load the Hinglish dataset D .

Step 2: Data Preprocessing

Normalise text by removing punctuation and converting to lower case.

Code-mixed token processing.

Use XLM-RoBERTa tokeniser to tokenise text.

Step 3: Embedding Generation

Get contextual embeddings from the pre-trained XLM-RoBERTa encoder.

Step 4: Model Fine-Tuning

Split the dataset into training, validation and test sets:

Fine-tune the classification head with cross-entropy loss.

Step 5: Prediction

Y_{pred} = Predicted labels for test set.

Step 6: Evaluation

Compute accuracy, precision and recall.

Return:

The scores and y_{pred}.

3.7 Model Evaluation Metrics

The trained model is tested on the test set using the traditional classification metrics:

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

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

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

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

These metrics collectively assess the accuracy and sensitivity to the cyberbullying cases as well as the robustness in the context of class imbalance.

4. Results

4.1. CONFUSION MATRIX:

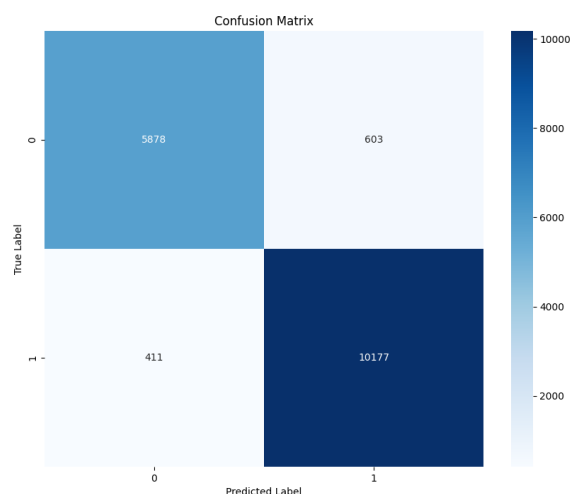


Fig :1 confusion matrix This section gives extensive evaluation of the model's performance using confusion matrix. The matrix shows the cyberbullying and non-cyberbullying content classification accuracy of the cyberbullying detection model.

From this matrix are computed the following standard evaluation metrics:

4.1.1. Accuracy

Accuracy is a measure of the overall correctness of the model.

$$\text{Accuracy} = \frac{10177 + 5878}{10177 + 5878 + 603 + 411} = \frac{16055}{17069} \approx 94.05\%$$

4.1.2. Precision

Precision: Percentage of the cyberbullying comments that were predicted and right:

$$\frac{10177}{10177 + 603} = \frac{10177}{10780} \approx 94.41\%$$

4.1.3. Recall (Sensitivity)

The ability of the model to detect true positive is measured by recall:

$$\frac{10177}{10177 + 411} = \frac{10177}{10588} \approx 96.12\%$$

4.1.4. F1 Score

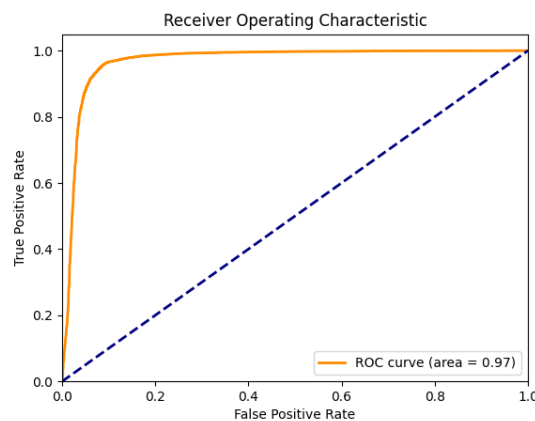
The F1 Score is the harmonic mean of precision and recall:

$$2 \cdot \frac{0.9441 \cdot 0.9612}{0.9441 + 0.9612} \approx 95.21\%$$

The accuracy of the model is very high, being 94.05% which demonstrates the general performance of the model. A recall of 96.12% shows a good ability to retrieve the real cases of cyberbullying, minimising the false negatives. But the 94.41% rate of accuracy assures little false alarm in the delicate content filtering environment, essential for sensitive content moderation.

The high and balanced F1 Score of 95.21% reflects the potential of the system in real applications, where both false charges and missed detection can have an impact. The performance is useful for deployment of the model to a real time system to assist in reducing online harm and keeping users safe.

4.2 ROC CURVE



The ROC curve is used to test the developed cyberbullying detection model as shown in Figure 2. ROC curve is a widely used technique for the assessment of the performance of binary classification systems. The ROC curve is a graph of the TPR as a function of the FPR, over a range of threshold values.

Key Metrics:

True Positive Rate (TPR) or Sensitivity / Recall:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}$$

TP, FP, FN, TN in the above equations are the True Positives, False Positives, True Negatives, and False Negatives, respectively. True Positives are the cases of cyberbullying that are correctly classified, False Positives are the instances that are not cyberbullying but are classified as such, True Negatives

are the instances that are not cyberbullying but are correctly classified, and False Negatives are the instances that are cyberbullying but are not detected by the model.

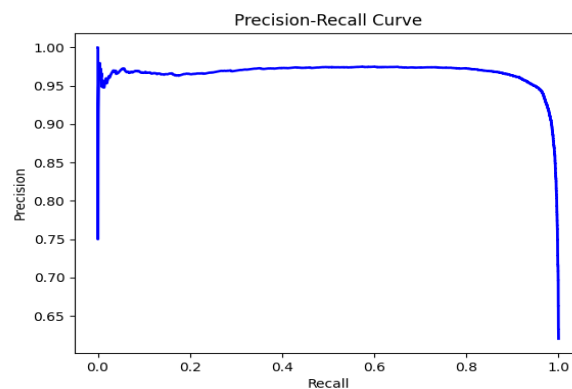
4.3 ROC Curve Observation:

The ROC curve of our model has a decent predictability as shown in Figure 2. As threshold value changes, the orange line plots the relationship between TPR and FPR, the baseline is the diagonal blue dash line, baseline of non-discrimination-power classifier (AUC=0.5). Our model has a significant curvature towards the top-left of the ROC curve, which means it has high sensitivity and specificity.

The model has a strong performance in terms of AUC score with the value of 0.97.

These AUC measures are usually considered poor = 0.5 - 0.6, fair = 0.6 - 0.7. The AUC metric is 0.97 indicating that the model is able to accurately identify text as bullying or not bullying 97% of the time. These findings support the suitability of the DL approach used in this study especially the usage of XLM-R to interpret rich linguistic inputs.

4.4 PRECISION-RECALL CURVE:

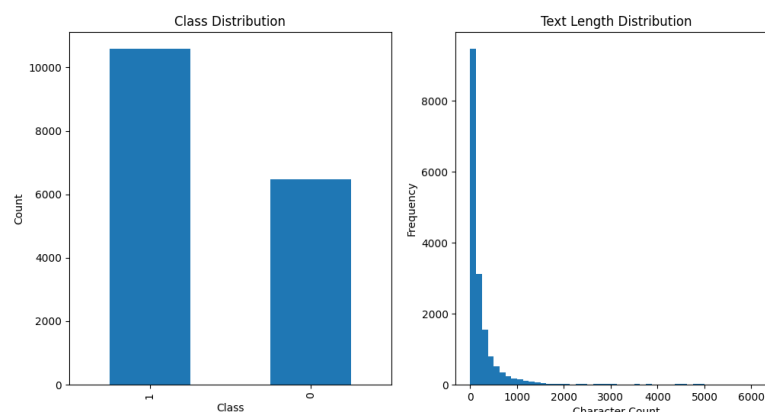


The PR curve in Figure 3 has recall on the x-axis and precision on the y-axis for different threshold settings. PR curve is extremely useful in the assessment of a model for imbalanced data. The greater the area of the PR curve, the greater the performance, that is, the model can achieve high recall and high precision. Given the following PR curve:

- High precision (around 95-98%) for majority of recall values.

There is a trade-off between precision and recall as close to 1 and a little less precision.

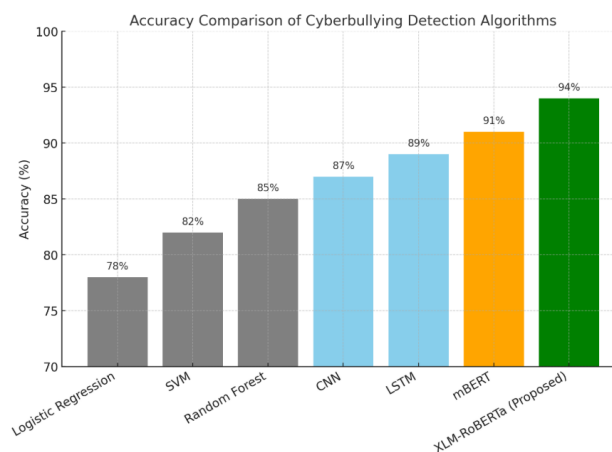
4.5 Exploratory Data Analysis – Distribution



The distribution plot of the class (Fig:6) indicates that data is highly skewed with the class of cyberbullying (Class 1) having approximately 10,500 instances, while the non-cyberbullying class (Class 0) has about 6,500 instances. It suggests that there is an increase in the amount of cyberbullying content in the dataset and thus must be considered when training the algorithm to account for the bias with appropriate class weighting techniques.

Distribution of text length (Figure Y). The majority of the samples are short. Most of the entries are less than 500 characters long. A long tail of relatively short text (up to 6000 characters). This is a common way of online communication as short conversation is commonplace, but in some instances long debates do occur. The properties suggest the need for a model that is capable of effectively handling short and long textual data.

4.6 Accuracy Analysis in Cyberbullying Detection



In Fig:7, it is demonstrated that several algorithms have a high precision in detecting cyberbullying. The LR (78%) and SVM (82%) are basic ML algorithms that achieve a moderate accuracy. RF (85%): Many decision tree models and vote to improve precision. Efficiently processing sequential text data does help to achieve a better performance in algorithms for DL such as CNN (87%) and LSTM (89%). These transformer models include mBERT (91%) and XLM-RoBERTa (94%) which are based on large-scale pre-training to learn the contextual and semantic information from the text. The suggested model XLM-RoBERTa is the most accurate (94%) and it has shown that transformer models are the most effective in identifying complex and subtle cyberbullying language patterns. This is in line with the observation that detection performance is improved by more complex models.

5. Discussion

Findings show that the multilingual transformer models such as XLM-RoBERTa perform better than classical ML and standard DL models in detecting cyberbullying in language mixed Hinglish text. This is because the model has a special sensitivity to the contextual and semantic relations of mixed languages and transliterated words and phrases that they might not capture for other classifiers like n-gram based classifiers and sequential models like CNN and LSTM. The accuracy of the observed balanced behaviour of precision and recall in the presence of class imbalance constitutes a reliable differentiation between the bullying and non-bullying information. Transformer models are more costly to compute, but its higher contextual understanding could be beneficial for cyberbullying detection in the real world social media environment in which the language is more complex.

References

- [1] Akhter, A., Acharjee, U. K., Talukder, M. A., Islam, M. M., & Uddin, M. A. (2023). A robust hybrid machine learning model for Bengali cyberbullying detection in social media. *Natural Language Processing Journal*, 4, 100027. <https://doi.org/10.1016/j.nlp.2023.100027>

-
- [2] Alabdulwahab, A., Haq, M. A., & Alshehri, M. (2023). Cyberbullying detection using machine learning and deep learning. *International Journal of Advanced Computer Science and Applications*, 14(10). Science and Information (SAI) Organization Limited. <https://doi.org/10.14569/IJACSA.2023.0141045>
- [3] Azumah, S., Elsayed, N., Elsayed, Z., Ozer, M., & Guardia, A. (2024, May). *Deep learning approaches for detecting adversarial cyberbullying and hate speech in social networks*. arXiv. <https://doi.org/10.48550/arXiv.2406.17793>
- [4] Behzadi, M., Harris, I. G., & Derakhshan, A. (2021). Rapid cyber-bullying detection method using compact BERT models. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)* (pp. 199–202). IEEE. <https://doi.org/10.1109/ICSC50631.2021.00043>
- [5] Ghosh, S., Priyankar, A., Ekbal, A., & Bhattacharyya, P. (2023). Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data. *Knowledge-Based Systems*, 260, 110182. <https://doi.org/10.1016/j.knosys.2022.110182>
- [6] Hoque, M. N., & Seddiqui, M. H. (2023). 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)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICCIT60459.2023.10441412>
- [7] Jha, V. K., P, H., P N, V., Vijayan, V., & Prabakaran, P. (2020). DHOT-Repository and classification of offensive tweets in the Hindi language. *Procedia Computer Science*, 171, 2324–2333. <https://doi.org/10.1016/j.procs.2020.04.252>
- [8] Lin, H., Siarry, P., Gururaj, H. L., Rodrigues, J., & Jain, D. K. (2022). Special issue on deep learning methods for cyberbullying detection in multimodal social data. *Multimedia Systems*, 28(6), 1873–1875. <https://doi.org/10.1007/s00530-022-01000-x>
- [9] Mahmud, T., Ptaszynski, M., Eronen, J., & Masui, F. (2023). Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Information Processing & Management*, 60(5), 103454. <https://doi.org/10.1016/j.ipm.2023.103454>
- [10] Mehendale, N., Shah, K., Phadtare, C., & Rajpara, K. (2022, May 21). Cyber bullying detection for Hindi-English language using machine learning. *SSRN*. <https://doi.org/10.2139/ssrn.4116143>
- [11] Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access*, 10, 25857–25871. <https://doi.org/10.1109/ACCESS.2022.3153675>
- [12] Obaida, M. H., Elkaffas, S. M., & Guirguis, S. K. (2024). Deep learning algorithms for cyber-bullying detection in social media platforms. *IEEE Access*, 12, 76901–76908. <https://doi.org/10.1109/ACCESS.2024.3406595>
- [13] Paul, S., Saha, S., & Hasanuzzaman, M. (2022). Identification of cyberbullying: A deep learning based multimodal approach. *Multimedia Tools and Applications*, 81, 26989–27008. <https://doi.org/10.1007/s11042-020-09631-w>
- [14] Perera, A., & Fernando, P. (2024). Cyberbullying detection system on social media using supervised machine learning. *Procedia Computer Science*, 239, 506–516. <https://doi.org/10.1016/j.procs.2024.06.200>
- [15] Raj, M., Singh, S., Solanki, K., & others. (2022). An application to detect cyberbullying using machine learning and deep learning techniques. *SN Computer Science*, 3(401). <https://doi.org/10.1007/s42979-022-01308-5>
- [16] Rehman, M. Z. U., Raghuvanshi, D., Pachar, H., Raghaw, C. S., & Kumar, N. (2025). Hierarchical attention-enhanced contextual capsuleNet for multilingual hate speech detection. *Expert Systems with Applications*, 268, 126285. <https://doi.org/10.1016/j.eswa.2024.126285>
- [17] Roy, P. K., & Mali, F. U. (2022). Cyberbullying detection using deep transfer learning. *Complex & Intelligent Systems*, 8(6), 5449–5467. <https://doi.org/10.1007/s40747-022-00772-z>

- [17] Sihab-Us-Sakib, S., Rahman, M. R., Forhad, M. S. A., & Aziz, M. A. (2024). Cyberbullying detection of resource-constrained language from social media using transformer-based approach. *Natural Language Processing Journal*, 9, 100104. <https://doi.org/10.1016/j.nlp.2024.100104>
- [18] Srivastava, A., Hasan, M., Bhargva, Y., Walambe, R., & Kotecha, K. (2021, May). *Role of artificial intelligence in detection of hateful speech for Hinglish data on social media*. arXiv. <https://doi.org/10.48550/arXiv.2105.04913>
- [19] Tabassum, I., & Nunavath, V. (2024). A hybrid deep learning approach for multi-class cyberbullying classification using multi-modal social media data. *Applied Sciences*, 14(24), 12007. <https://doi.org/10.3390/app142412007>