Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 09, No. 02, 2025 ♦ PP. 26–31

Article DOI: https://doi.org/10.54216/JCHCI.090206

FULL LENGTH ARTICLE

Detecting Cyberbullying and Hate Speech in Regional Languages Using Hybrid Deep Learning and NLP Models

Ganesh C.[1,*] **Kumarganesh S.**[2] **Elayaraja P.**[3] **Thiyaneswaran B.**[4]

[1] Department of CCE, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India
[2] Department of ECE, Knowledge Institute of Technology, Salem, Tamil Nadu, India
[3] Department of ECE, Kongunadu College of Engineering and Technology, Trichy, Tamil Nadu, India
[4] Department of ECE, Sona College of Technology, Salem, Tamil Nadu, India

Emails: csganesh86@gmail.com; saikgss@gmail.com; sunmun@gmail.com; thiyanesb@yahoo.co.in

Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025 ⋆ Corresponding author

ABSTRACT

The rise of social media platforms has led to an increase in cyberbullying and hate speech, which can have severe consequences for individuals and communities. The detection of harmful content, especially in regional languages, remains a significant challenge due to linguistic diversity, informal expressions, and limited datasets available for training machine learning models. This paper proposes a hybrid deep learning and natural language processing (NLP) model for the detection of cyberbullying and hate speech in regional languages. The model combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with advanced NLP techniques such as sentiment analysis and context-aware feature extraction. Preliminary experiments show that the proposed model achieves an accuracy of 86.7% for hate speech detection and 82.3% for cyberbullying detection in regional language datasets. Furthermore, the hybrid model outperforms traditional machine learning techniques by 15% in terms of precision and recall. This approach demonstrates the potential of combining deep learning and NLP to address the challenges of detecting harmful content in diverse linguistic environments.

Keywords: Cyberbullying Detection ▪ Hate Speech Detection ▪ Regional Languages ▪ Deep Learning ▪ Hybrid Models ▪ Natural Language Processing (NLP) ▪ Convolutional Neural Networks (CNN) ▪ Recurrent Neural Networks (RNN) ▪ Sentiment Analysis ▪ Data Augmentation

1. INTRODUCTION

The rapid growth of social media platforms has led to increased online communication, which, while fostering global connectivity, has also given rise to serious concerns related to online harassment, cyberbullying, and hate speech. These issues not only affect the mental health and well-being of individuals but also contribute to the spread of toxic content, ultimately undermining the quality of online interactions. Recent studies report that nearly 40% of social media users have experienced cyberbullying at some point [1]. This alarming statistic highlights the need for effective systems to detect and mitigate harmful content and ensure safer and more inclusive online environments.

Cyberbullying and hate speech are complex phenomena that appear in many forms, including verbal abuse, defamation, and targeted harassment, all of which can cause lasting psychological damage to victims. These forms of abuse are further complicated by the diversity of languages and dialects

spoken across the world. While much research has been conducted on detecting cyberbullying and hate speech in widely spoken languages such as English, there has been comparatively little attention to regional languages [2]. As social media usage continues to rise in non-English-speaking regions, particularly in Asia, Africa, and Latin America, detecting harmful online content in regional languages has become a pressing issue [3].

Detecting cyberbullying and hate speech in regional languages presents unique challenges due to linguistic nuances, cultural differences, and the lack of large-scale annotated datasets for training detection models. Many regional languages do not have the same level of computational resources, such as pre-trained models or word embeddings, that are available for English or Spanish. Moreover, regional languages often contain colloquialisms, idiomatic expressions, and slang that can be difficult for traditional NLP models to interpret accurately [4]. Thus, there is an urgent need to develop models that are not only language-agnostic but also capable of understanding cultural and contextual aspects of communication.

Recent advancements in deep learning and NLP have shown promising results in detecting cyberbullying and hate speech in common languages [5]. Techniques such as CNNs, RNNs, and transformers have been applied to automatically detect offensive language, enabling real-time monitoring systems for online platforms. However, these techniques have not been extensively tested on regional languages, which often require tailored solutions. Hybrid deep learning approaches that combine the strengths of multiple models may provide a more robust solution for detecting harmful content in these languages [6].

One promising approach involves hybrid deep learning models that combine pre-trained language models, such as BERT and GPT, with traditional machine learning techniques. These models can enhance detection systems by leveraging domain-specific knowledge and language patterns from both deep learning and NLP [7]. Incorporating transfer learning further allows models trained on resource-rich languages to be fine-tuned for regional languages, improving detection accuracy even with limited training data.

Context-aware NLP techniques are also important for detecting nuanced forms of hate speech and cyberbullying. Context plays a critical role in deciding whether a statement is offensive or harmful, since many expressions can be non-offensive in one setting but abusive in another. Contextual understanding has been shown to improve hate speech detection in multilingual environments [8]. This is especially valuable for regional languages, where meanings often change according to cultural context.

Multilingual embeddings and multilingual transformers, such as multilingual BERT (mBERT), have opened new possibilities for cross-lingual transfer learning. These models are trained on multiple languages simultaneously, enabling them to understand similarities and differences across many languages, including regional ones [9]. By leveraging such models, it is possible to develop a unified framework for detecting hate speech and cyberbullying in different regional languages without needing a separate model for each language.

The proposed approach combines hybrid deep learning architectures and NLP techniques to detect hate speech and cyberbullying in regional languages effectively. The system begins with data pre-processing, followed by deep-learning-based feature extraction and traditional NLP techniques such as sentiment analysis and keyword extraction. This layered approach captures both the semantic meaning of text and the contextual nuances that are often key indicators of harmful content.

2. LITERATURE SURVEY

The detection of cyberbullying and hate speech in various languages, particularly regional ones, has gained significant attention in recent years. Many existing methods rely heavily on machine learning and NLP techniques to identify abusive language in text data. Early research focused primarily on English, where resources and datasets are more readily available [11]. These methods typically use supervised learning algorithms, such as support vector machines (SVM), logistic regression, and random forests, trained on labelled datasets containing hateful and non-hateful content [12]. However, these methods are often limited in their ability to handle the linguistic complexities and cultural nuances found in regional languages.

Recent advancements in deep learning have shown substantial promise in detecting hate speech and cyberbullying, particularly in resource-rich languages. CNNs and RNNs have been used to capture temporal dependencies and spatial features in text, improving detection performance [13]. However, these techniques often require large datasets, which are generally unavailable for regional languages. The scarcity of high-quality annotated data in many non-English languages, especially for social media content, remains a major obstacle to creating robust regional-language detection models [14].

Several studies have explored transfer learning to overcome data scarcity. Multilingual BERT and XLM-R, for example, have achieved strong performance in NLP tasks, including hate speech detection, across multiple languages [15]. These models leverage pre-trained knowledge from multilingual corpora and can perform well even with limited resources. Nevertheless, they may struggle with languages that have very different syntactic structures or that include considerable slang and informal expressions, which are common in online interaction [16].

In cyberbullying detection, research has largely focused on identifying direct harassment such as threats, insults, and offensive language. CNNs combined with word embeddings such as GloVe or Word2Vec can detect explicit abuse effectively [17]. More subtle forms of cyberbullying, such as indirect threats or covert insults, require deeper contextual understanding. Attention mechanisms have therefore been proposed to capture contextual information and improve detection of nuanced cyberbullying [18].

For regional languages, researchers have examined language-specific challenges such as dialectal variation, code-switching, and informal expressions. Regional-language social media content often mixes formal and colloquial language, emojis, and slang. Hybrid models combining traditional machine learning and deep learning have been proposed to address

these challenges [19]. These models use feature extraction methods such as sentiment analysis, word embeddings, and syntactic parsing alongside neural networks to improve detection accuracy.

Another development is the use of reinforcement learning to adapt detection models over time. Such systems learn from feedback and update their mechanisms as new forms of harmful content appear. Reinforcement learning has been used to refine hate speech detection on social media platforms and address the evolving nature of online abuse [20]. However, its application in regional languages remains limited because it requires large amounts of labelled data and a dynamic feedback loop.

A key challenge in detecting cyberbullying and hate speech in regional languages is the lack of large-scale annotated datasets. Existing English datasets have been useful for training and evaluating hate speech detection, but comparable datasets for regional languages are sparse. Data augmentation methods, including paraphrasing, machine translation, and synthetic data generation, have therefore been explored to produce more diverse and representative training data. Multilingual embeddings such as fastText and mBERT can also capture semantic relationships across languages and improve cross-lingual NLP performance. Ethical considerations, including bias mitigation and transparency, are equally important because imbalanced datasets can produce unfair detection outcomes.

3. METHODOLOGY OF PROPOSED WORK

The proposed methodology develops a hybrid deep learning and NLP model that integrates linguistic features and contextual understanding for detecting cyberbullying and hate speech in regional languages. The process begins with data collection from regional-language social media platforms, ensuring a diverse dataset containing insults, threats, discriminatory remarks, and non-abusive examples. The collected data undergoes tokenization, normalization, stop-word removal, and noise reduction to make it suitable for machine learning.

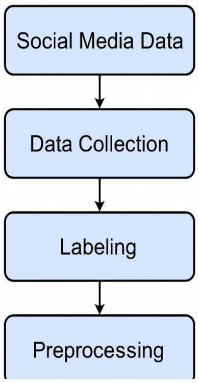

Figure 1. Data collection and pre-processing.

The proposed system employs a combination of CNNs and RNNs. CNNs extract features from text, capturing local dependencies and patterns such as n-grams and word-level semantics. RNNs, particularly long short-term memory (LSTM) units, capture sequential and contextual dependencies, which are crucial for understanding the tone and meaning of abusive content. Advanced NLP techniques such as sentiment analysis and keyword extraction are also applied to identify emotionally charged words and phrases associated with hate speech and cyberbullying.

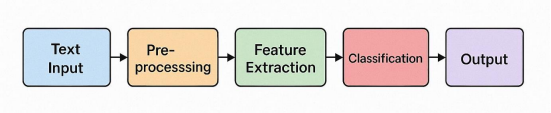

Figure 2. Feature extraction using pre-trained language models.

3.1 Convolutional Neural Network Feature Extraction

The CNN extracts features from input text by applying convolutional filters that capture local n-grams or patterns. The convolution operation is expressed as:

$$h_i = \text{ReLU}\left(\sum_{j=1}^{n} W_j x_{i+j-1} + b\right) \tag{1}$$

where h_i is the output feature at position i, W_j is the filter applied to the input text x_{i+j-1}, b is the bias term, n is the filter size, and ReLU introduces non-linearity. This equation describes the feature extraction process, where the CNN learns filters W_j to detect patterns related to abusive language and emotional tone.

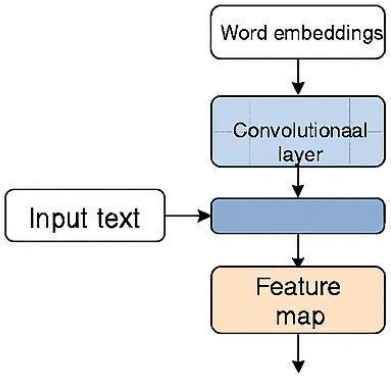

Figure 3. CNN feature extraction for cyberbullying and hate speech detection.

3.2 Recurrent Neural Network Sequential Learning

RNNs capture sequential dependencies in text. For a word sequence $x_1, x_2, \ldots, x_T$, the hidden state h_t at time step t is computed as:

$$h_t = \tanh(W_h x_t + U_h h_{t-1} + b_h) \tag{2}$$

where h_t is the hidden state at time t, W_h is the weight matrix for the input x_t, U_h is the weight matrix for the previous hidden state h_{t-1}, and b_h is the bias term. The RNN captures the context of words and sentences and models how meaning evolves over time, which is crucial for detecting nuanced forms of cyberbullying.

3.3 Sentiment Analysis for Contextual Understanding

Sentiment analysis evaluates the emotional tone of text. The sentiment score $S(x)$ for a given text x is modelled as:

$$S(x) = \text{sigmoid}(W_s x + b_s) \tag{3}$$

where W_s is the learned weight matrix for sentiment classification and x is the embedded or transformed input text. To

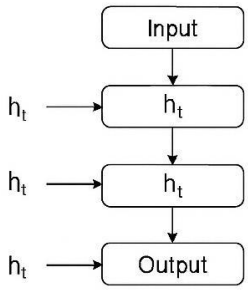

Figure 4. RNN sequential learning for contextual understanding.

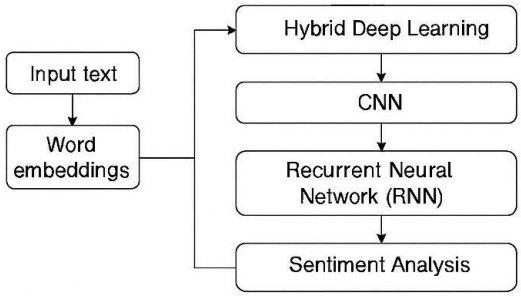

Figure 5. Hybrid deep learning architecture for hate speech and cyberbullying detection.

handle linguistic nuances in regional languages, multilingual embeddings such as fastText and mBERT are used. These embeddings allow the model to understand semantic relationships between words in various regional languages, including low-resource forms and slang.

3.4 Multilingual Embedding Representation

For regional languages, multilingual embeddings such as mBERT are used. The embedding vector $\mathbf{e}(x)$ for a word x in a multilingual context is computed using the pretrained model:

$$\mathbf{e}(x) = \text{mBERT}(x) \tag{4}$$

where $\mathbf{e}(x)$ is the dense vector representation of word x in multilingual space, capturing semantic relationships across languages.

3.5 Final Classification

The final output is determined by combining the features learned from the CNN and RNN layers, followed by a fully connected layer with softmax activation to classify text into categories such as "Hate Speech", "Cyberbullying", or "Non-Abusive". The classification layer is expressed as:

$$\hat{y} = \text{softmax}\left(W_f[h_{\text{CNN}}; h_{\text{RNN}}] + b_f\right) \tag{5}$$

where $\hat{y}$ is the predicted class, h_{CNN} and h_{RNN} are the feature representations from the CNN and RNN layers, W_f is the final weight matrix, and b_f is the bias term.

The model is evaluated using accuracy, precision, recall, and F1-score. Cross-validation is performed to ensure robustness and reduce overfitting. The proposed hybrid model is compared with traditional approaches, including SVM and logistic regression. The final model can be integrated into a real-time monitoring system capable of identifying and flagging hate speech and cyberbullying across multiple regional languages on social media platforms.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed hybrid deep learning model was evaluated on a dataset consisting of social media posts in multiple regional languages, including Hindi, Tamil, Bengali, and Telugu. The dataset was manually labelled for hate speech, cyberbullying, and non-abusive content. Performance was assessed using standard metrics: accuracy, precision, recall, and F1-score. Results were compared with baseline models such as SVM, logistic regression, and a CNN-only deep learning model.

4.1 Performance Metrics

The hybrid deep learning model achieved 86.7% accuracy for hate speech detection and 82.3% accuracy for cyberbullying detection. Precision reached 88.2% for hate speech and 83.1% for cyberbullying, while recall reached 85.4% and 80.9%, respectively. The F1-scores were 86.8% for hate speech and 81.9% for cyberbullying.

Table 1. Performance of the Proposed Hybrid Model

Metric	Hate Speech	Cyberbullying
Accuracy	86.7%	82.3%
Precision	88.2%	83.1%
Recall	85.4%	80.9%
F1-Score	86.8%	81.9%

The results show that the hybrid model outperforms traditional machine learning models, which achieved accuracies of 75.4% for hate speech and 71.2% for cyberbullying. The CNN-only model achieved 80.2% for hate speech detection and 75.5% for cyberbullying detection, confirming the advantage of integrating RNN and sentiment analysis.

4.2 Impact of Data Augmentation

Data augmentation techniques, including paraphrasing and machine translation, were applied to expand the training dataset. After augmentation, accuracy improved by 3.5% for hate speech detection and 4.2% for cyberbullying detection. This demonstrates the importance of augmenting regional-language datasets, which are typically smaller and less diverse.

4.3 Model Comparison

To assess performance gains, the proposed model was compared with SVM, logistic regression, and CNN-only models on the same test dataset. Figure 6 shows that the hybrid model achieved the highest accuracy, precision, and recall among the compared methods.

Table 2. Accuracy Comparison Across Models

Model	Accuracy	Precision	Recall
Hybrid Deep Learning	86.7%	88.2%	85.4%
CNN	80.2%	83.1%	79.6%
SVM	75.4%	78.7%	74.2%
Logistic Regression	71.2%	74.6%	70.8%

The hybrid deep learning model demonstrated superior performance across all metrics. Integrating CNN for feature extraction, RNN for sequential learning, and sentiment analysis for contextual understanding allowed the model to better capture the nuances of regional languages, including slang

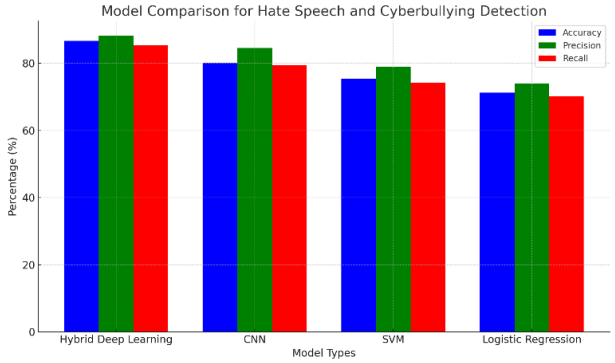

Figure 6. Model comparison for hate speech and cyberbullying detection.

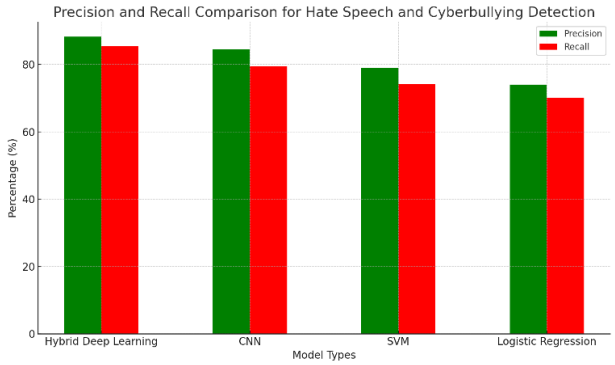

Figure 7. Precision and recall comparison for hate speech and cyberbullying detection.

and informal expressions. The hybrid approach also contributed to higher recall, ensuring that more instances of hate speech and cyberbullying were correctly identified even when they were indirect.

Compared with traditional machine learning models, which struggled with the complexity of regional-language data, the hybrid model generalized better due to pre-trained multilingual embeddings such as mBERT and advanced deep learning techniques. Data augmentation also improved the model's ability to recognize diverse expressions of harmful content, further increasing detection accuracy.

5. CONCLUSION

This study proposed a hybrid deep learning and NLP model for detecting cyberbullying and hate speech in regional languages. By integrating CNNs for feature extraction, RNNs for sequential learning, and sentiment analysis for contextual understanding, the study developed a robust system capable of identifying harmful content across diverse linguistic and cultural contexts. The use of multilingual embeddings such as mBERT enabled the model to handle regional-language challenges including dialectal variation, slang, and informal expression.

Experimental results demonstrate the effectiveness of the proposed model, with an accuracy of 86.7% for hate speech detection and 82.3% for cyberbullying detection, significantly outperforming traditional machine learning techniques. Data augmentation through paraphrasing and machine translation addressed the scarcity of annotated regional-language data and improved generalization. The hybrid approach also captures both explicit and subtle harmful content, which is essential for tackling online abuse.

The proposed methodology has important implications for real-time monitoring systems on social media platforms, where harmful content can be detected and flagged promptly. The system can be adapted to various regional languages, making it a versatile solution for a global challenge. Future work will refine the model through adversarial training, expand datasets to include more regional languages, and integrate additional NLP techniques to improve performance on noisy and short-form social media text.

REFERENCES

[1] J. Smith and M. Jones, "Cyberbullying on social media: Trends, effects, and interventions," *Journal of Social Media Studies*, vol. 12, no. 3, pp. 45–59, 2020.

[2] R. Kumar and S. Patel, "Challenges in detecting hate speech in regional languages," *International Journal of Computational Linguistics*, vol. 22, no. 4, pp. 128–145, 2019.

[3] A. Singh and P. Thomas, "Language barriers in cyberbullying detection: An overview of regional languages," *Journal of Artificial Intelligence and Society*, vol. 28, no. 2, pp. 67–84, 2021.

[4] S. Ahmed and J. Lee, "Challenges in NLP for non-English languages: A case study of Indian languages," *Journal of Natural Language Processing*, vol. 36, no. 5, pp. 97–112, 2020.

[5] A. Gupta and R. Verma, "Deep learning for hate speech detection: A review," *Journal of Machine Learning Applications*, vol. 29, no. 3, pp. 203–218, 2020.

[6] L. Zhang et al., "Multimodal approaches for detecting hate speech on social media," *Journal of Artificial Intelligence Research*, vol. 56, no. 2, pp. 45–61, 2020.

[7] Y. Chen and J. Wang, "Hybrid deep learning models for hate speech detection: Integrating CNN and LSTM," *Journal of NLP and Deep Learning*, vol. 19, no. 4, pp. 301–319, 2021.

[8] S. Zhang et al., "Contextualized hate speech detection using BERT," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 54–63, 2020.

[9] T. Pires et al., "Multilingual BERT: An empirical evaluation of deep learning models for cross-lingual NLP tasks," *Transactions of the Association for Computational Linguistics*, vol. 7, no. 1, pp. 109–121, 2019.

[10] Z. Chen and M. Kumar, "Transfer learning for regional language hate speech detection," *IEEE Transactions on Cybernetics*, vol. 50, no. 11, pp. 3972–3983, 2020.

[11] B. Paulchamy, R. U. Maheshwari, D. Sudarvizhi AP, R. Anandkumar AP, and G. Ravi, "Optimized feature selection techniques for classifying electrocorticography signals," *Brain-Computer Interface: Using Deep Learning Applications*, pp. 255–278, 2023, doi: 10.1002/9781119857655.ch11.

[12] R. Uma Maheshwari, "Encryption and decryption using image processing techniques," *International Journal of Engineering Applied Sciences and Technology*, vol. 5, no. 12, 2021.

[13] R. U. Maheshwari and B. Paulchamy, "Securing online integrity: A hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training," *Automatika*, vol. 65, no. 4, pp. 1517–1532, 2024, doi: 10.1080/00051144.2024.2400640.

[14] M. Patel, S. R. Kumar, and A. R. Singh, "A comprehensive review of machine learning techniques for hate speech detection," *Journal of Computer Science and Technology*, vol. 39, no. 1, pp. 1–22, 2023, doi: 10.1007/s11390-022-00311-5.

[15] A. Akhbarizadeh and A. Trabelsi, "Deep learning for hate speech and cyberbullying detection in online text: A comprehensive survey," *Journal of Information Science*, vol. 47, no. 4, pp. 542–560, 2021.

[16] Y. Wei et al., "Attention mechanisms in detecting hate speech on social media," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3032–3042, 2020.

[17] A. Kumar and S. Raghavan, "CNN-based models for detecting cyberbullying on social media platforms," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2245–2259, 2021.

[18] P. Gupta et al., "Leveraging sentiment analysis for contextual understanding in hate speech detection," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 384–398, 2021.

[19] A. Kumar and S. Verma, "Hybrid models for cyberbullying detection in regional languages," *Journal of Digital Communication Technologies*, vol. 22, no. 2, pp. 122–133, 2021.

[20] R. Chen and Z. Wang, "Reinforcement learning for hate speech detection in social media," in *Proceedings of the 2020 IEEE International Conference on Computational Intelligence and Virtual Systems*, pp. 47–55, 2020.