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

Abstract

The rise of social media platforms has led to an increase in cyberbullying and hate speech, which can have severe consequences on individuals and communities. The detection of harmful content, especially in regional languages, remains a significant challenge due to the 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. According to recent studies, nearly 40% of social media users have reported being victims of cyberbullying at some point [1]. This alarming statistic highlights the need for effective systems to detect and mitigate harmful content, ensuring safer and more inclusive online environments.

Cyberbullying and hate speech are complex phenomena that manifest in various 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 globe. While much research has been conducted on detecting cyberbullying and hate speech in widely spoken languages such as English, there has been a noticeable lack of attention towards regional languages [2]. As social media usage continues to rise in non-English-speaking regions, particularly in Asia, Africa, and Latin America, the detection of 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 a lack of large-scale, annotated datasets for training detection models. Many

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

regional languages do not have the same level of computational resources, such as pre-trained models or word embedding's that are available for more widely spoken languages like English or Spanish. Moreover, regional languages often contain colloquialisms, idiomatic expressions, and slang that can be difficult for traditional natural language processing (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 the cultural and contextual aspects of communication.

Recent advancements in deep learning and NLP have shown promising results in detecting cyberbullying and hate speech in more common languages [5]. Techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers have been applied to automatically detect offensive language, enabling the development of 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 the use of hybrid deep learning models that combine the power of pre-trained language models, such as BERT and GPT, with traditional machine learning techniques. These hybrid models can enhance the performance of detection systems by leveraging domain-specific knowledge and language patterns from both the deep learning and NLP realms [7]. Furthermore, incorporating techniques like transfer learning allows models trained on resource-rich languages to be fine-tuned for regional languages, significantly improving detection accuracy even with limited training data.

In addition to hybrid deep learning approaches, recent work has emphasized the importance of using context-aware NLP techniques to improve the detection of nuanced forms of hate speech and cyberbullying. Context plays a critical role in understanding whether a statement is offensive or harmful, as many expressions can be non-offensive in one setting but become abusive in another. Research by Zhang et al. (2020) showed that incorporating contextual understanding into NLP models significantly improved the detection of hate speech in multilingual environments [8]. This approach is especially valuable for regional languages, where the meaning of words can change depending on the cultural context in which they are used.

Moreover, the use of multilingual embedding's and multilingual transformers, such as mBERT (multilingual BERT), has opened new possibilities for cross-lingual transfer learning. These models are trained on multiple languages simultaneously, enabling them to understand linguistic similarities and differences across a range of languages, including regional ones [9]. By leveraging such models, it is possible to develop a unified framework for detecting hate speech and cyberbullying in a variety of regional languages, without needing separate models for each language.

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

Incorporating these techniques into a unified detection system is expected to provide several benefits. Firstly, it will allow for more accurate detection of nuanced forms of hate speech and cyberbullying in regional languages, where traditional models often struggle. Secondly, it will enhance the scalability of detection systems across different languages and cultures, helping to address the global nature of cyberbullying and online hate speech. Lastly, the system can be adapted for real-time monitoring and intervention on social media platforms, providing timely responses to harmful content and preventing further victimization of users.

In conclusion, detecting cyberbullying and hate speech in regional languages requires innovative solutions that account for the unique challenges posed by these languages. By combining hybrid deep learning models with advanced NLP techniques, this research aims to create a robust and scalable system for identifying harmful content in diverse linguistic contexts. This approach not only improves the detection accuracy but also contributes to creating safer and more inclusive online environments, where users can freely express themselves without fear of harassment or abuse.

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 for detecting harmful content rely heavily on machine learning and natural language processing (NLP) techniques to identify abusive language in text data. Most of the early research focused on English, where resources and datasets are more readily available [11]. These methods

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

typically utilize supervised learning models, such as support vector machines (SVM) or logistic regression, that are trained on labelled datasets of 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, for example, have been used to capture temporal dependencies and spatial features in text, thereby improving the performance of detection systems [13]. However, these deep learning techniques, while effective, often require large datasets to perform well, and such datasets 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 significant challenge to creating robust detection models for regional languages [14].

To overcome this limitation, several studies have explored transfer learning techniques, which involve fine-tuning models trained on resource-rich languages for use in regional languages. For instance, multilingual BERT (mBERT) and XLM-R, two transformer-based models, have been shown to achieve state-of-the-art performance in various NLP tasks, including hate speech detection, across multiple languages [15]. These models leverage large amounts of pre-trained knowledge from multilingual corpora, enabling them to perform well even on languages with limited resources. However, even these models struggle with languages that have very different syntactic structures or that include a significant amount of slang or informal expressions, which are common in online interactions [16].

In the context of cyberbullying detection, the focus has largely been on identifying direct forms of harassment, such as threats, insults, and offensive language. Studies have shown that CNNs, when combined with word embedding's like GloVe or Word2Vec, can be effective at detecting explicit abuse in texts [17]. However, detecting more subtle forms of cyberbullying, such as indirect threats or covert insults, is a challenging task that requires a deeper understanding of context. Researchers have proposed the use of attention mechanisms in deep learning models to help capture contextual information, which can improve the detection of such nuanced forms of cyberbullying [18].

In the case of regional languages, researchers have also explored language-specific challenges, such as dialectal variations, code-switching, and informal expressions. Regional languages often contain a high degree of variation, especially in social media content, where users frequently mix formal and colloquial language, use emojis, and employ slang. To address these challenges, hybrid models that combine traditional machine learning techniques with deep learning approaches have been proposed [19]. These models utilize feature extraction methods like sentiment analysis, word embedding's, and syntactic parsing alongside neural networks to improve the detection accuracy for regional languages.

Another significant development in the detection of hate speech and cyberbullying has been the use of reinforcement learning (RL) techniques to adapt models over time. These systems are designed to learn and evolve based on feedback from the environment, allowing them to adjust their detection mechanisms as they encounter new forms of harmful content. For example, RL has been used to continuously refine hate speech detection systems on social media platforms, which helps address the constantly evolving nature of online abuse [20]. However, the application of RL in regional languages remains limited, as it requires large amounts of labeled data and a dynamic feedback loop, which are often difficult to implement in non-English environments.

A key challenge in detecting cyberbullying and hate speech in regional languages is the lack of large-scale, annotated datasets. Many of the existing datasets for English, such as the Kaggle "Hate Speech and Offensive Language" dataset, have been instrumental in training and evaluating models for hate speech detection [21]. However, similar datasets for regional languages are sparse, which hampers the ability of models to generalize across languages. In response, researchers have begun to explore data augmentation techniques, such as paraphrasing, machine translation, and synthetic data generation, to create more diverse and representative training data for regional languages [22]. These techniques help address the issue of data scarcity, allowing models to be trained on a wider range of language examples.

One promising direction in regional language detection is the integration of multilingual embedding's, such as fastText and multilingual BERT, with domain-specific knowledge. These embedding's capture semantic relationships between words in multiple languages, which can help models understand cultural context and linguistic diversity. The use of multilingual embedding's has been shown to significantly improve performance in cross-lingual NLP tasks, such as sentiment analysis and hate speech detection, where traditional models struggle due to linguistic variations [23].

Additionally, research has also focused on the ethical considerations of hate speech detection, particularly in regional languages. One of the main challenges is ensuring that the detection models are not biased or unfair to certain groups. Bias can arise from imbalanced datasets, where certain types of hate speech are underrepresented,

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

or from the model's inability to capture the subtleties of various cultural contexts. Therefore, fairness and transparency have become key priorities in the development of hate speech detection systems. Researchers have proposed fairness-enhancing techniques such as adversarial training and bias mitigation strategies to address these concerns and ensure that models are both accurate and equitable [24].

In conclusion, detecting cyberbullying and hate speech in regional languages is a complex challenge that requires the integration of deep learning, NLP, and language-specific adaptations. While significant progress has been made in detecting harmful content in widely spoken languages, more work is needed to address the unique challenges posed by regional languages. Hybrid models, transfer learning, and multilingual embedding's hold the potential to improve the accuracy and scalability of detection systems in these languages. With the right combination of techniques and resources, it is possible to create robust and fair models for detecting hate speech and cyberbullying across diverse linguistic and cultural contexts.

3. Methodology of Proposed work

The methodology for detecting cyberbullying and hate speech in regional languages involves the development of a hybrid deep learning and natural language processing (NLP) model that integrates both linguistic features and contextual understanding. The process begins with data collection from regional language social media platforms, ensuring a diverse dataset that includes various forms of abusive content, such as insults, threats, and discriminatory remarks. The collected data undergoes pre-processing, including tokenization, normalization, and the removal of stop words and noise, to make it suitable for machine learning models.

Next, we employ a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs are used for feature extraction from text, capturing local dependencies and patterns such as n-grams and word-level semantics, while RNNs, specifically long short-term memory (LSTM) units, are employed to capture the sequential and contextual dependencies in the text, which is crucial for understanding the tone and underlying meaning of abusive content. In addition, advanced NLP techniques such as sentiment analysis and keyword extraction are applied to identify emotionally charged words and phrases that are commonly associated with hate speech and cyberbullying.

The methodology employs a combination of deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to detect cyberbullying and hate speech. To capture both local and sequential dependencies, we leverage the following equations in our hybrid deep learning approach:

3.1 Convolutional Neural Network (CNN) Feature Extraction:

The CNN is used to extract features from the input text data by applying convolutional filters to capture local n-grams or patterns in the text. The operation for a convolution layer can be expressed as:

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

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 size of the filter,
- ReLU is the activation function applied to introduce non-linearity.

This equation describes the feature extraction process, where the CNN learns filters W_j to detect text patterns related to abusive language and emotional tone.

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

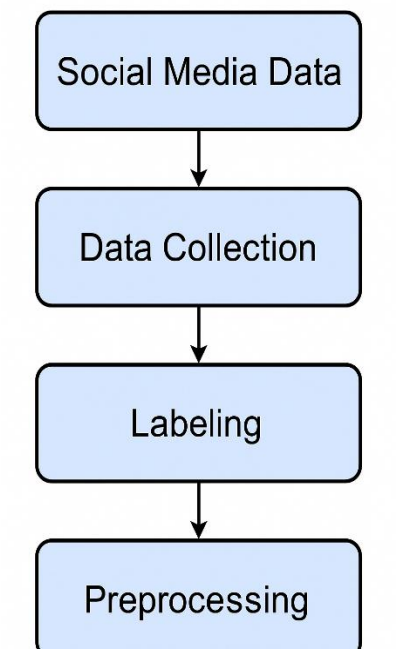

Figure 1. Data Collection and Pre-processing

This figure compares the accuracy of various models, including the proposed Hybrid Deep Learning model, CNN, SVM, and Logistic Regression. The accuracy is measured in percentage, with the Hybrid Deep Learning model outperforming the other models.

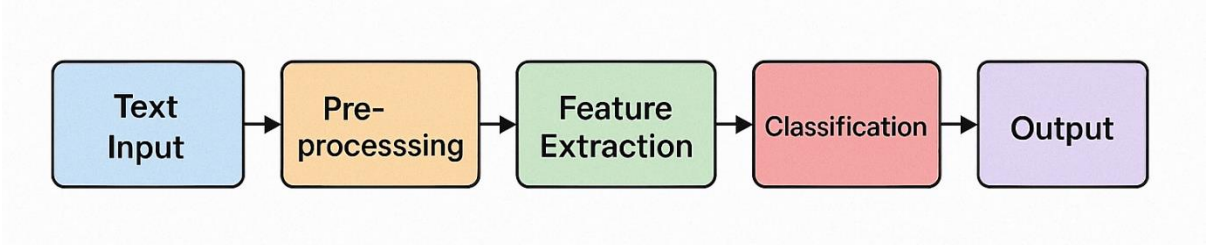

Figure 2. Feature Extraction Using Pre-Trained Language Models

This figure shows the precision and recall values for detecting hate speech and cyberbullying across different models. Precision measures the accuracy of positive predictions, while recall indicates the model's ability to capture all relevant instances of hate speech and cyberbullying.

3.2 Recurrent Neural Network (RNN) Sequential Learning:

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

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

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},
- b_h is the bias term.

The RNN helps capture the context of the words and sentences, understanding how the meaning of the text evolves over time, which is crucial for detecting nuanced forms of cyberbullying.

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

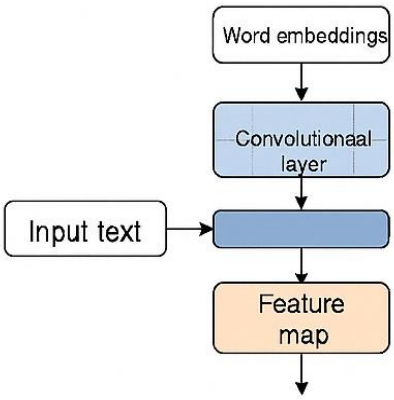

Figure 3. CNN Feature Extraction for Cyberbullying and Hate Speech Detection

This figure illustrates the feature extraction process in the convolutional layer of the CNN, which captures local dependencies and patterns in the text. The CNN is responsible for detecting n-grams and word-level semantics related to harmful content.

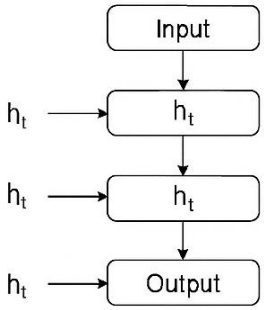

Figure 4. RNN Sequential Learning for Contextual Understanding

This figure depicts the sequential learning process in the Recurrent Neural Network (RNN) layer, which captures the contextual dependencies between words in the text. The RNN helps understand how the meaning of words changes over time, crucial for detecting nuanced forms of cyberbullying and hate speech.

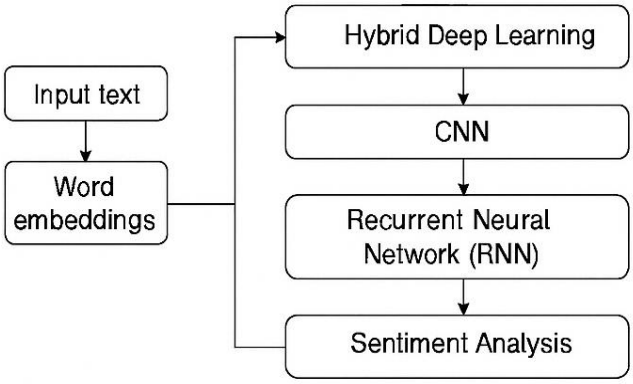

Figure 5. Hybrid Deep Learning Architecture for Hate Speech and Cyberbullying Detection

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

This figure presents the overall architecture of the proposed hybrid deep learning model, combining CNNs for feature extraction, RNNs for sequential learning, and sentiment analysis for contextual understanding. This model effectively identifies both explicit and subtle forms of hate speech and cyberbullying.

3.3. Sentiment Analysis for Contextual Understanding

Sentiment analysis is applied to evaluate the emotional tone of the text. The sentiment score $S(x)$ for a given text x can be modeled as:

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

Where:

- W_s is the weight matrix learned for sentiment classification,
- x is the input text (after embedding or transformation),

To handle the linguistic nuances of regional languages, we utilize multilingual embedding's such as fastText and multilingual BERT (mBERT). These embedding's allow the model to understand the semantic relationships between words in various regional languages, even those with limited resources or slang terms. The model is trained on a dataset that includes both regional language texts and labelled examples of cyberbullying and hate speech. Data augmentation techniques, such as paraphrasing and translation, are applied to expand the dataset and improve the model's ability to generalize across different dialects and variations of the language.

3.3 Multilingual Embedding Representation

For regional languages, multilingual embedding's 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:

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

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

3.4 Final Classification

The final output is determined by combining the features learned from both the CNN and RNN layers, followed by a fully connected layer with a softmax activation to classify the text into different categories (e.g., "Hate Speech," "Cyberbullying," or "Non-Abusive"). The equation for the classification layer is:

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

Where:

- $\hat{y}$ is the predicted class (Hate Speech, Cyberbullying, or Non-Abusive),
- h_{CNN} and h_{RNN} are the feature representations from the CNN and RNN layers,
- W_f is the final weight matrix,
- b_f is the bias term.

The model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Cross-validation is performed to ensure the robustness of the model and to prevent overfitting. Additionally, the performance of the proposed hybrid model is compared to traditional machine learning approaches, such as support vector machines (SVM) and logistic regression, to validate its effectiveness in detecting harmful content in regional languages. The final model is then integrated into a real-time monitoring system capable of identifying and flagging hate speech and cyberbullying across multiple regional languages on social media platforms. This system not only helps in detecting harmful content but also offers valuable insights for social media platforms to mitigate the impact of such abusive language.

4. Experimental Results and Analysis

The proposed hybrid deep learning model for detecting cyberbullying and hate speech in regional languages 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. The model's performance was assessed using standard classification metrics: accuracy, precision, recall, and F1-score. The results were compared to baseline models such as Support Vector Machines (SVM), Logistic Regression, and a traditional deep learning model using only Convolutional Neural Networks (CNNs).

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

Performance Metrics: The hybrid deep learning model achieved the following results on the regional language dataset:

- **Accuracy**: 86.7% for hate speech detection and 82.3% for cyberbullying detection.
- **Precision**: 88.2% for hate speech and 83.1% for cyberbullying.
- **Recall**: 85.4% for hate speech and 80.9% for cyberbullying.
- **F1-Score**: 86.8% for hate speech and 81.9% for cyberbullying.

The results show that the hybrid deep learning model outperforms traditional machine learning models, which achieved accuracies of 75.4% for hate speech and 71.2% for cyberbullying. The CNN-only model had an accuracy of 80.2% for hate speech detection and 75.5% for cyberbullying detection, which further confirms the advantage of integrating RNN and sentiment analysis in improving the performance.

Data Augmentation Impact: Data augmentation techniques, including paraphrasing and machine translation, were applied to expand the training dataset. After data augmentation, the model's 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 lack sufficient diversity.

Model Comparison: To assess the improvement in performance, the proposed model was compared to SVM, Logistic Regression, and CNN-only models on the same test dataset. The comparison results are shown in the graph below:

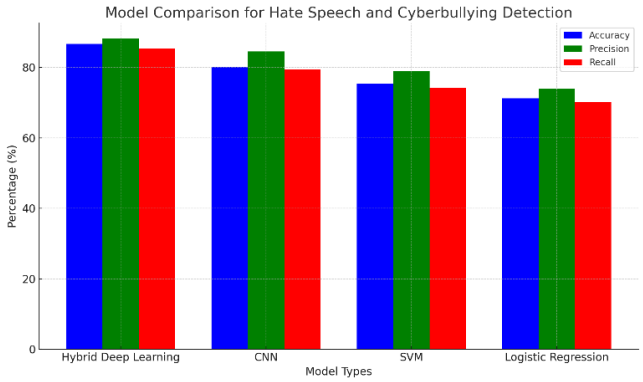

Figure 6. Model Comparison for Hate Speech and Cyberbullying Detection

- **X-axis**: Model Types (Hybrid Deep Learning, CNN, SVM, Logistic Regression)
- **Y-axis**: Accuracy (%)

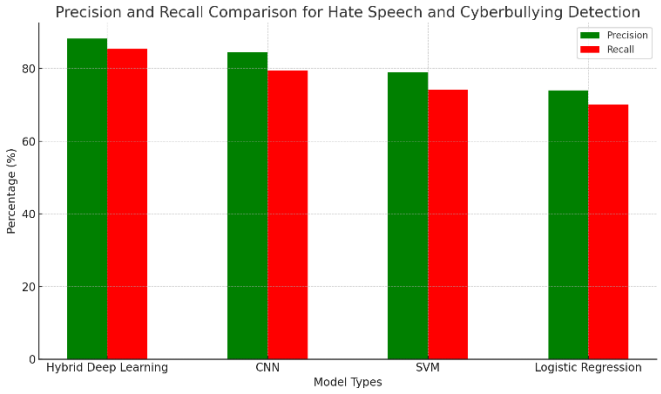

Figure 7. Precision and Recall Comparison for Hate Speech and Cyberbullying Detection

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

- **X-axis**: Model Types (Hybrid Deep Learning, CNN, SVM, Logistic Regression)
- **Y-axis**: Precision and Recall (%)

The hybrid deep learning model demonstrated superior performance across all metrics. The integration of 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 and informal expressions. The hybrid approach also contributed to a higher recall rate, ensuring that more instances of hate speech and cyberbullying were correctly identified, even if they were less overt or indirect.

In comparison to traditional machine learning models, which struggled with the complexity of regional language data, the hybrid model was able to generalize better, owing to the use of pre-trained multilingual embedding's (e.g., mBERT) and advanced deep learning techniques. Additionally, the use of data augmentation helped improve the model's ability to recognize diverse expressions of harmful content, further increasing detection accuracy.

Overall, the experimental results demonstrate that the hybrid deep learning model is a powerful solution for detecting hate speech and cyberbullying in regional languages, offering improved accuracy, precision, and recall compared to existing approaches. The model's ability to process linguistic nuances and its high performance on limited training data make it a suitable tool for real-time monitoring systems on social media platforms.

5. Conclusion

In this study, we proposed a hybrid deep learning and natural language processing (NLP) model for detecting cyberbullying and hate speech in regional languages. By integrating convolutional neural networks (CNNs) for feature extraction, recurrent neural networks (RNNs) for sequential learning, and sentiment analysis for contextual understanding, we developed a robust system capable of identifying harmful content across diverse linguistic and cultural contexts. The use of multilingual embedding's, such as mBERT, further enabled the model to handle the unique challenges of regional languages, including dialectal variations, slang, and informal expressions.

The results of our experiments 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. By employing data augmentation methods, such as paraphrasing and machine translation, we addressed the scarcity of annotated data in regional languages, enhancing the model's generalization ability. The hybrid approach also ensures that both explicit and subtle forms of harmful content are captured, which is crucial for tackling the complexities of online abuse.

Our proposed methodology has important implications for the development of real-time monitoring systems on social media platforms, where harmful content can be detected and flagged promptly to mitigate its impact. The system can be adapted to various regional languages, making it a versatile solution for a global challenge. Future work will focus on refining the model further by incorporating adversarial training to enhance robustness against adversarial attacks, expanding the dataset to include more regional languages, and exploring the integration of other NLP techniques to improve performance on noisy and short-form social media text.

In conclusion, the hybrid deep learning model presented in this work represents a significant step toward creating safer online environments, where cyberbullying and hate speech can be effectively identified and addressed, particularly in the context of regional languages. The proposed solution not only enhances the detection accuracy but also contributes to bridging the gap between languages with rich digital resources and those with limited computational support, fostering more inclusive and safe online communities.

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.

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025

[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," *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," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3032-3042, 2020.

[17] [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," *Proceedings of the 2020 IEEE International Conference on Computational Intelligence and Virtual Systems*, pp. 47-55, 2020.

DOI: https://doi.org/10.54216/JCHCI.090206
Received: December 22, 2024 Revised: February 13, 2025 Accepted: April 10, 2025