Toward Safer Digital Communication: A Deep Hybrid Model for Detecting Abusive Language on Social Networks

Akbayan Aliyeva
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
akbayan.aliyeva@ayu.edu.kz

Balnur Kenjayeva
International University of Tourism and Hospitality, Turkistan, Kazakhstan
balnur.kendjaeva@iuth.edu.kz

Moldir Kizdarbekova
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
moldir.kizdarbekova@ayu.edu.kz

Bolganay Kaldarova
Zhanibekov University, Shymkent, Kazakhstan
bolganaykaldarova@gmail.com

Satmyrza Mamikov
University of Friendship of People's Academician A. Kuatbekov, Shymkent, Kazakhstan
Satmyrza85@mail.ru

Bauyrzhan Omarov
Al-Farabi Kazakh National University, Almaty, Kazakhstan
Bauyrzhanomarov01@gmail.com

Nurlan Omarov
International University of Tourism and Hospitality, Turkistan, Kazakhstan
nurlan.omarov@iuth.edu.kz

Aigerim Toktarova
International University of Tourism and Hospitality, Turkistan, Kazakhstan
aikerimtoktarova@gmail.com (corresponding author)

Eshref Adaly
Istanbul Technical University, Istanbul, Turkiye
adali@itu.edu.tr

Received: 13 June 2025 | Revised: 19 July 2025 and 27 July 2025 | Accepted: 1 August 2025

 DOI: https://doi.org/10.48084/etasr.12721

ABSTRACT

The prevalence of abusive language and cyberbullying on social media platforms presents a growing challenge to the user safety and digital well-being, necessitating the development of effective automated content moderation systems. This study proposes a hybrid deep learning model that combines Long Short-

Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs) to classify the abusive text with enhanced accuracy and contextual awareness. The LSTM component captures long-range dependencies and semantic context, while the CNN module extracts discriminative local n-gram features. The model was trained and evaluated on three benchmark datasets: HatebaseTwitter, HatEval, and TRAC. The experimental results demonstrated that the proposed architecture outperforms traditional classifiers, such as SVM, Random Forest, and Logistic Regression, as well as standalone CNN and LSTM models, achieving superior performance across all standard evaluation metrics. Notably, the model attained AUC scores of up to 0.97, indicating robust discriminatory power. These findings underscore the effectiveness of the hybrid LSTM–CNN model for abusive language detection and highlight its potential for deployment in real-time content moderation tools aimed at fostering safer online communication environments.

Keywords-abusive language detection; cyberbullying classification; deep learning; LSTM-CNN hybrid model; text classification; social media analysis; natural language processing; sentiment analysis

I. INTRODUCTION

The proliferation of social media platforms has reshaped the digital communication landscape, offering users access to share information and engage in discourse globally [1]. However, this digital expansion has also facilitated the widespread dissemination of offensive language, hate speech, and cyberbullying, posing serious risks to the societal harmony and user well-being [2]. Given the massive volume and velocity of user-generated content, traditional manual moderation approaches have become increasingly impractical and resource-intensive [3]. As a result, the development of automated abusive language detection systems has emerged as a pressing need within Natural Language Processing (NLP) and Artificial Intelligence (AI) research [4].

Initial solutions predominantly relied on classical machine learning algorithms, such as Support Vector Machines, Naïve Bayes classifiers, and Logistic Regression models [5]. These approaches typically used manually engineered features including n-grams, TF-IDF vectors, and sentiment lexicons to classify text [6]. Although effective in certain domains, these models often struggled with capturing deeper semantic structures, contextual meaning, and the evolving linguistic patterns inherent in online discourse [7]. Moreover, challenges, such as class imbalance, noise in data, and contextual ambiguity limited their effectiveness in real-world applications [8].

The advent of deep learning methodologies marked a significant advancement in text classification tasks. CNNs became prominent for their ability to extract local textual features and recognize specific patterns within data [9]. CNNs demonstrated strong performance in sentiment analysis and initial hate speech detection efforts [10]. Nevertheless, CNN architectures inherently focus on local dependencies and are limited in their capacity to capture long-term sequential information [11]. This limitation makes them insufficient when abusive language detection requires understanding broader discourse contexts or semantic nuances spread across longer text spans [12].

To overcome these limitations, Recurrent Neural Networks (RNNs), especially LSTM networks, have been adopted for their strength in modeling sequential dependencies and retaining context over time [13]. LSTMs have shown promise in various text classification tasks, including offensive language detection [14]. However, they may underperform in isolating specific local features critical for identifying subtle abusive cues within the text [15]. Consequently, hybrid architectures combining CNNs for localized pattern detection with LSTMs for sequence modeling have emerged as effective solutions [16]. These hybrid models capture both immediate lexical patterns and extended contextual information, improving the detection accuracy for explicit and implicit abusive language forms [17].

Further advancement introduced attention mechanisms and transformer-based models, allowing a focus on the key text elements and enhancing the model interpretability [18]. While these architectures offer state-of-the-art results, they often require significant computational resources and remain sensitive to the dataset quality and linguistic variability [19]. Publicly available annotated datasets, such as HatebaseTwitter [20], HatEval [21], and TRAC [22], have become essential for benchmarking the model performance, but even these datasets often suffer from class imbalance and annotation inconsistencies, underscoring the need for robust preprocessing and augmentation techniques [23]. Ethical considerations including fairness, bias mitigation, and transparency are equally critical to ensure the responsible deployment of automated moderation systems [24]. Against this backdrop, this study proposes a hybrid LSTM-CNN model designed to effectively integrate convolutional and sequential learning for robust abusive language detection across diverse social media contexts.

II. MATERIALS AND METHODS

The methodological framework adopted in the current study focuses on developing a robust deep learning pipeline for detecting offensive language in user-generated social media content. The process begins with the collection of textual data, which includes annotated posts labeled as abusive or non-abusive based on predefined criteria. To prepare the data for modeling, a preprocessing phase is implemented, which involves standard NLP techniques, such as lowercasing, removal of punctuation and special characters, stopword elimination, and tokenization. Additionally, lemmatization and noise reduction are applied to enhance the semantic quality of the input. Once preprocessed, the data are split into training and testing subsets to enable both the model optimization and performance evaluation.

Rather than relying on generic neural architectures, a hybrid LSTM–CNN model specifically designed to leverage both the

local and sequential features in text classification was applied. The LSTM component captures long-term contextual dependencies, while the CNN component identifies local patterns, such as abusive n-grams. The comparative evaluation of this hybrid model against baseline machine learning algorithms, like SVM, Random Forest, and Naïve Bayes as well as standalone deep learning models, such as CNN, LSTM and Bi-LSTM is central. Performance metrics including accuracy, precision, recall, F1-score, and AUC-ROC are used to quantitatively assess the advantages of the proposed approach. The results demonstrate the hybrid model's consistent superiority across multiple benchmark datasets, further validating the methodological contributions of this work.

A. *Dataset Collection*

The dataset used was collected by automatically extracting user-generated content from the VKontakte (VK) social network through the VK API. Public profile details, including user ID, name, surname, date of birth, city, and post history, were gathered and systematically grouped by user identifiers for analysis (Figure 1). Each post was labeled based on predefined sentiment categories: neutral, depressive, or aggressive, creating a labeled dataset for model training and evaluation. A Python-based interface in PyCharm initiated API queries, retrieving user data via HTTPS requests in JSON format for efficient processing (Figure 2).

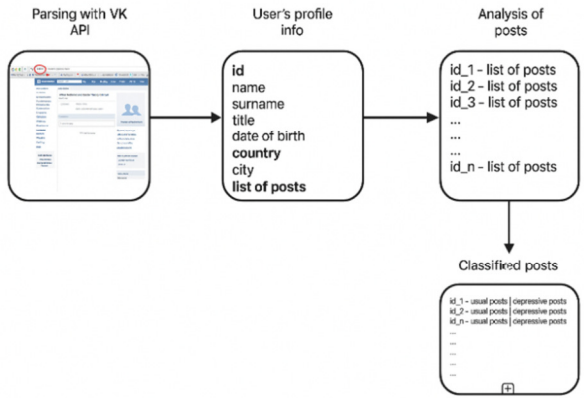

Fig. 1. Dataset collection and classification pipeline using VK API.

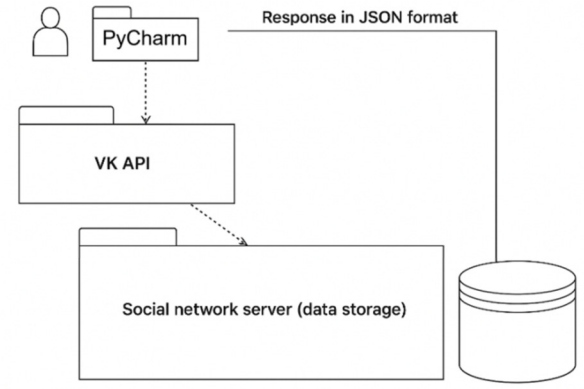

Fig. 2. Architecture of data acquisition via VK API.

The VK API allows access to both user profile metadata and their associated posts, which are stored on the VK server and returned in a structured format. Upon receiving the response, the client-side script parses the JSON data and stores them locally for subsequent preprocessing and analysis [25]. This architecture ensures scalable and repeatable data extraction, enabling large-scale data collection with minimal manual intervention. By leveraging the API's structured endpoints and response formats, the system guarantees consistency in data retrieval while adhering to the platform usage guidelines and ethical standards.

B. *Proposed Model*

The proposed model (Figure 3) integrates convolutional feature extraction with recurrent sequence modeling to detect abusive language in social media text.

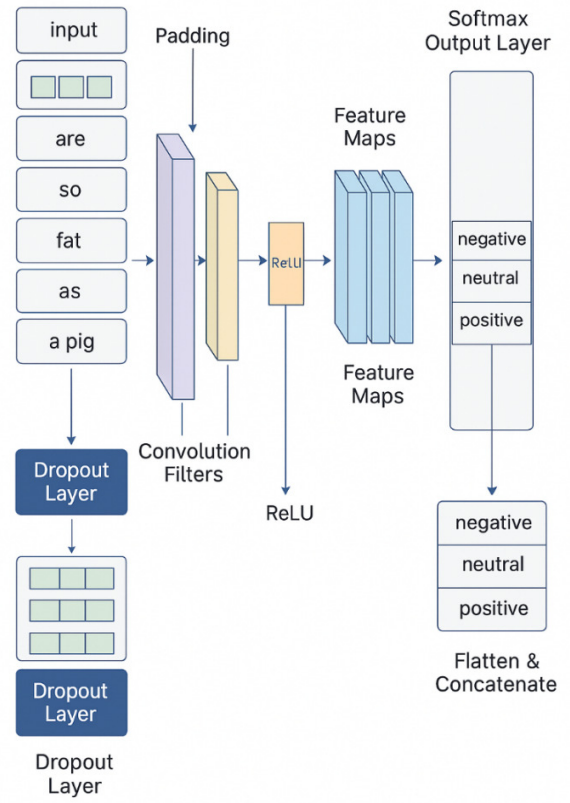

Fig. 3. Proposed hybrid LSTM–CNN architecture for automated abusive language detection.

The input sequence of token embeddings $x = [x_1, x_2, ..., x_n]$ is first passed through a dropout layer to mitigate overfitting. Convolutional filters of varying window sizes h then slide over the embeddings, producing feature maps $c^{(h)}$ via:

$$c_i^{(h)} = \mathrm{Re}\,LU\left(W^h \cdot x_{i:i+h-1} + b^{(h)}\right) \quad (1)$$

where W^h and b^h denote the filter weights and bias, respectively, and $\mathrm{Re}LU(z)$ is the $\max(0, z)$.

A max-over-time pooling operation then condenses each feature map into a scalar:

$$\hat{c}^{(h)} = \max_i c_i^{(h)} \quad (2)$$

The gathered convolutional features $\left\{\hat{c}^{(h)}\right\}$ are concatenated with the final hidden state of a bidirectional LSTM, which processes the original embeddings to capture the long-range dependencies. The LSTM component updates the gating equations:

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \quad (3)$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (4)$$

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \quad (5)$$

$$c_t = f_t \oplus c_{t-1} + i_t \tilde{c}_t \quad (6)$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \quad (7)$$

$$h_t = o_t \oplus \tanh\left(c_t\right) \quad (8)$$

where *i, f, c*, and *o* stand for input, forget, cell, and output, respectively. Also *W* are the weight matrices, *U* are the weight matrices for hidden state h_{t-1} in respective gates, σ is the sigmoid activation function, and $\oplus$ is the element-wise addition.

The combined feature vector $z = \left[\left\{\hat{c}^{(h)}\right\} \| h_n\right]$ is then fed into a fully connected layer with softmax activation function to produce class probabilities $\hat{y}$ [26]:

$$\hat{y} = \mathrm{softmax}\left(W_y z + b_y\right) \quad (9)$$

The model training minimizes the categorical cross-entropy loss *L*, over *K* classes:

$$L = -\sum_{k=1}^{K} y_k \log\left(\hat{y}_k\right) \quad (10)$$

where y_k is the true label for class *k,* represented as a one hot vector and $\hat{y}_k$ is the model's predicted probability for class *k.*

C. Evaluation Metrics

To quantitatively assess the performance of the proposed model, a set of standard evaluation metrics was used tailored to each of the three core tasks: classification, localization, and segmentation. For the classification task, accuracy, precision, recall, and F1-score were employed, to evaluate the model's ability to distinguish abusive from non-abusive cases [27-30]. These metrics are defined by:

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (11)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (12)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (13)$$

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision+recall}} \quad (14)$$

where TP, TN, FP, and FN represent True Positive (the model correctly predicts an abusive case when the case is abusive), True Negative (the model correctly predicts a non-abusive case when the case is not abusive), False Positive (the model incorrectly predicts an abusive case, but the case is not abusive), and False Negative (the model incorrectly predicts a non-abusive case, but the case is abusive), respectively.

III. RESULTS

The experimental evaluation conducted highlights the superior performance of the proposed hybrid LSTM–CNN model for detecting offensive language across three benchmark datasets: HatebaseTwitter, HatEval, and TRAC. The model consistently outperforms both traditional machine learning classifiers such as Naïve Bayes, Support Vector Machine, and Logistic Regression, as well as deep learning baselines including CNN, LSTM, and Bi-LSTM. Comparing the average results for accuracy, precision, recall, and F1-score across the datasets for each model (Table I), it is concluded that the hybrid model achieves the highest overall performance, with accuracy exceeding 97%, precision and recall above 98%, and the highest F1-score across all models evaluated. These improvements reflect not only statistical superiority, but also meaningful gains in the practical classification reliability.

The superior performance of the proposed hybrid model arises from its architectural integration of CNN and RNN components. The convolutional layers effectively extract local n-gram features, capturing specific abusive expressions, while the LSTM layers model long-range dependencies to preserve the contextual meaning. This complementary design allows for the accurate identification of both obvious and subtle offensive language, enabling the model to offer a robust, generalizable, and scalable solution for offensive language detection in real world content moderation across heterogeneous social media platforms.

The proposed hybrid LSTM–CNN model consistently surpasses various baseline classifiers. It achieves higher accuracy, precision, recall, and F1-score, with improvements of approximately 2-3 percentage points over Bi-LSTM. By combining convolutional feature extraction with sequential modeling, the hybrid architecture effectively captures both the local patterns and long-term dependencies, enhancing the detection of explicit and implicit abusive content. This synergy provides a reliable tool for automated content moderation.

TABLE I. PERFORMANCE COMPARISON OF THE PROPOSED MODEL WITH MACHINE LEARNING AND DEEP LEARNING MODELS

Dataset	Type	Method	Accuracy	Precision	Recall	F1-score	AUC-ROC
Hatebase Twitter dataset	Proposed method	Hybrid LSTM–CNN architecture for automated abusive language detection	97.2%	97.1%	97.2%	97.1%	97.1%
	Machine learning methods	Random Forest	70.6%	70.3%	70.4%	70.3%	70.2%
		Decistion Tree	74.5%	74.2%	74.1%	74.3%	74.1%
		Logistic Regression	79.2%	78.7%	78.6%	78.5%	78.3%
		KNN	77.7%	77.4%	77.3%	76.4%	76.2%
		Naïve Bayes	68.8%	68.5%	68.4%	68.4%	68.4%
		SVM	79.8%	79.6%	79.3%	79.4%	79.4%
	Deep learning	CNN	83.4%	82.9%	82.0%	82.4%	82.4%
		LSTM	87.4%	87.9%	88.0%	87.4%	87.4%
		BiLSTM	90.4%	89.9%	90.0%	89.4%	89.4%
		CNN-LSTM	92.2%	91.7%	91.8%	91.6%	91.5%
		CNN-BiLSTM	92.4%	92.4%	92.3%	92.2%	92.3%
HatEval dataset	Proposed method	Hybrid LSTM–CNN architecture for automated abusive language detection	97.8%	97.4%	97.3%	97.4%	97.4%
	Machine learning methods	Random Forest	68.6%	68.3%	68.4%	68.3%	68.2%
		Decistion Tree	72.9%	72.2%	72.8%	72.6%	72.4%
		Logistic Regression	78.2%	77.9%	77.6%	77.4%	77.3%
		KNN	74.7%	74.4%	74.3%	73.4%	73.2%
		Naïve Bayes	66.8%	66.5%	66.4%	66.4%	66.4%
		SVM	78.8%	78.6%	78.3%	78.4%	78.4%
	Deep learning	CNN	82.4%	81.9%	82.0%	81.4%	81.4%
		LSTM	86.4%	85.9%	85.0%	85.4%	85.5%
		BiLSTM	89.4%	88.9%	89.0%	88.4%	88.4%
		CNN-LSTM	91.7%	91.3%	91.2%	91.2%	91.2%
		CNN-BiLSTM	91.8%	91.7%	91.7%	91.6%	91.4%
TRAC dataset	Proposed method	Hybrid LSTM–CNN architecture for automated abusive language detection	96.5%	96.3%	96.3%	96.0%	95.3%
	Machine learning methods	Random Forest	71.6%	71.3%	71.4%	71.3%	71.2%
		Decistion Tree	76.9%	76.2%	76.8%	76.6%	76.4%
		Logistic Regression	77.2%	76.9%	76.6%	76.4%	76.3%
		KNN	76.7%	76.4%	76.3%	76.4%	76.2%
		Naïve Bayes	69.8%	69.5%	69.4%	69.4%	68.4%
		SVM	80.8%	80.6%	80.3%	80.4%	80.4%
	Deep learning	CNN	85.4%	84.9%	85.0%	84.4%	84.5%
		LSTM	90.4%	89.9%	90.0%	89.4%	89.4%
		BiLSTM	92.4%	91.9%	92.0%	91.4%	91.4%
		CNN-LSTM	92.8%	92.3%	92.3%	92.4%	92.1%
		CNN-BiLSTM	93.0%	92.9%	93.8%	92.8%	92.7%

Figure 4 shows the ROC curves for all evaluated models on the HatebaseTwitter dataset, where the diagonal line represents random chance (AUC = 0.5). The proposed hybrid LSTM-CNN model demonstrates the best performance, achieving an AUC of approximately 0.96, reflecting its strong ability to differentiate between abusive and non-abusive content. The Bi-LSTM and CNN baselines follow with AUCs around 0.92, confirming their effectiveness in sequential and local feature extraction. Traditional models, such as SVM and Random Forest, score lower with AUCs between 0.82 and 0.86. These results validate the hybrid model's superior classification capability.

Figure 5 displays the ROC curve results for all classifiers evaluated on the HatEval dataset, a key benchmark for hate speech detection. The diagonal line represents random performance (AUC = 0.50) for baseline comparison. The proposed hybrid LSTM–CNN model achieves the highest AUC of approximately 0.95, highlighting its superior ability to distinguish abusive from non-abusive text. This performance indicates the model's strength in capturing both explicit and subtle forms of hate speech. The Bi-LSTM and CNN baselines follow with AUCs of around 0.92 and 0.90, respectively. In contrast, traditional classifiers show lower AUCs between 0.82 and 0.87.

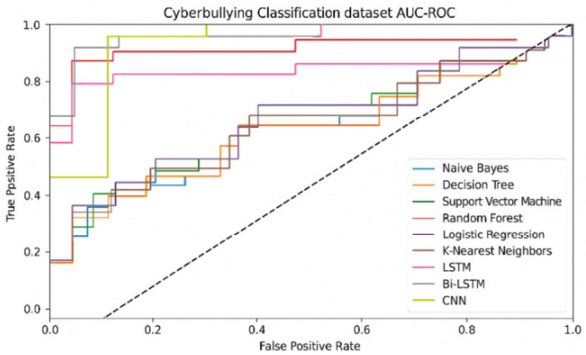

Fig. 4. The ROC curve results for the HatebaseTwitter dataset.

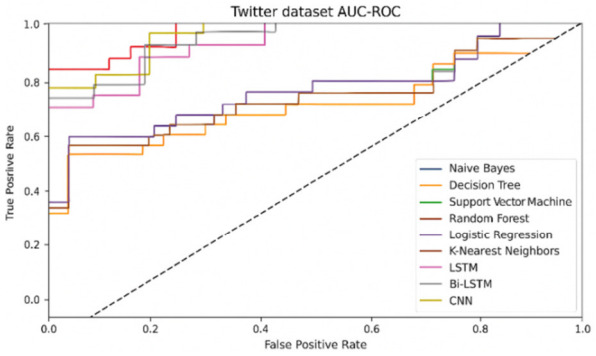

Fig. 5. The ROC curve results for the HatEval dataset.

Figure 6 illustrates the ROC curves for all evaluated classifiers on the TRAC dataset, where the diagonal line represents random performance (AUC = 0.50) as in the other cases. The proposed hybrid LSTM–CNN model achieves the highest AUC of approximately 0.97, confirming its strong capability to differentiate between abusive and non-abusive posts. The Bi-LSTM and CNN baselines follow with AUCs of about 0.93 and 0.91, respectively, highlighting their respective strengths in sequence modeling and local pattern extraction. Traditional machine learning models, including SVM, Random Forest, and Logistic Regression, yield lower AUCs between 0.81 and 0.87. These results validate the hybrid model's superior performance.

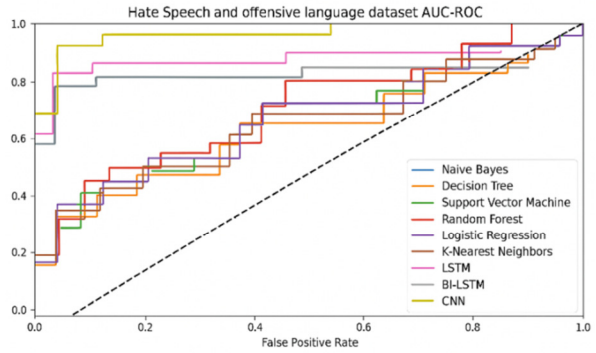

Fig. 6. The ROC curve results for the TRAC dataset.

IV. CONCLUSIONS

This study introduced a hybrid Convolutional Neural Network (CNN)-Long Short Term Memory (LSTM) model for automated abusive language detection on social media, designed to leverage both convolutional and sequential learning mechanisms. By integrating CNN's capacity to extract local n-gram features with LSTM's strength in modeling long-term contextual dependencies, the proposed model achieved consistent performance improvements over traditional machine learning classifiers, such as SVM, Random Forest, and Logistic Regression, as well as over standalone deep learning models, like CNN, LSTM, and Bi-LSTM. Across three benchmark datasets (AbuseTwitter, HatEval, and TRAC), the hybrid model recorded higher accuracy, precision, recall, and F_1-score, with relative gains of 2-3 percentage points compared to the strongest baselines.

Compared to previous studies relying solely on CNN or LSTM architectures, which often struggled to balance the local feature extraction with the global context understanding, the proposed hybrid framework consistently delivered superior classification outcomes. This finding aligns with recent research advocating for model architectures that combine multiple learning paradigms to address the nuanced nature of abusive language detection. Notably, the high AUC values, ranging from 0.95 to 0.97, underscore the model's enhanced discriminative capacity, surpassing the performance typically reported for single-architecture models on similar datasets.

In addition to addressing class imbalance and linguistic variability, the model demonstrates practical applicability for real-time content moderation tasks. Future research will focus on integrating attention mechanisms, transformer-based encodings, and multilingual data to further refine the detection capabilities and broaden the cross-platform adaptability. This work confirms that the hybrid architectures hold significant potential for advancing automated moderation tools, contributing to safer and more responsible digital communication environments.

REFERENCES

[1] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, and T. Zia, "Abusive language detection from social media comments using conventional machine learning and deep learning approaches," *Multimedia Systems*, vol. 28, no. 6, pp. 1925–1940, Dec. 2022, https://doi.org/10.1007/s00530-021-00784-8.

[2] P. K. Roy and A. Kumar, "Ensuring safety in digital spaces: Detecting code-mixed hate speech in social media posts," *Data & Knowledge Engineering*, vol. 156, Mar. 2025, Art. no. 102409, https://doi.org/10.1016/j.datak.2025.102409.

[3] R. Prabhu and V. Seethalakshmi, "A comprehensive framework for multi-modal hate speech detection in social media using deep learning," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, Art. no. 13020, https://doi.org/10.1038/s41598-025-94069-z.

[4] H. Han, M. Asif, E. M. Awwad, N. Sarhan, Y. Y. Ghadi, and B. Xu, "Innovative deep learning techniques for monitoring aggressive behavior in social media posts," *Journal of Cloud Computing*, vol. 13, no. 1, p. 19, Jan. 2024, https://doi.org/10.1186/s13677-023-00577-6.

[5] R. Kumar and A. Bhat, "A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media," *International Journal of Information Security*, vol. 21, no. 6, pp. 1409–1431, Dec. 2022, https://doi.org/10.1007/s10207-022-00600-y.

[6] A. T. Azar, H. M. Noori, A. R. Mahlous, A. Al-Khayyat, and I. K. Ibraheem, "Quasi-Reflection Learning Arithmetic Firefly Search Optimization with Deep Learning-based Cyberbullying Detection on Social Networking," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17162–17169, Oct. 2024, https://doi.org/10.48084/etasr.8314.

[7] S. Narynov, D. Mukhtarkhanuly, and B. Omarov, "Dataset of depressive posts in Russian language collected from social media," *Data in Brief*, vol. 29, Apr. 2020, Art. no. 105195, https://doi.org/10.1016/j.dib.2020.105195.

[8] M. Neog and N. Baruah, "A hybrid deep learning approach for Assamese toxic comment detection in social media," *Procedia Computer Science*, vol. 235, pp. 2297–2306, Jan. 2024, https://doi.org/10.1016/j.procs.2024.04.218.

[9] N. A. Samee, U. Khan, S. Khan, M. M. Jamjoom, M. Sharif, and D. H. Kim, "Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and Emotional Features for Cyberbullying Detection," *IEEE Access*, vol. 11, pp. 124524–124541, 2023, https://doi.org/10.1109/ACCESS.2023.3329347.

[10] S. Abarna, J. I. Sheeba, and S. Pradeep Devaneyan, "A novel ensemble model for identification and classification of cyber harassment on social media platform," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 1, pp. 13–36, Jul. 2023, https://doi.org/10.3233/JIFS-230346.

[11] B. Omarov, Z. Zhumanov, A. Gumar, and L. Kuntunova, "Artificial Intelligence Enabled Mobile Chatbot Psychologist using AIML and Cognitive Behavioral Therapy," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 6, 2023, https://doi.org/10.14569/IJACSA.2023.0140616.

[12] A. Rawat, S. Kumar, and S. S. Samant, "Hate speech detection in social media: Techniques, recent trends, and future challenges," *WIREs Computational Statistics*, vol. 16, no. 2, 2024, Art. no. e1648, https://doi.org/10.1002/wics.1648.

[13] D. Sultan *et al.*, "A Review of Machine Learning Techniques in Cyberbullying Detection," *Computers, Materials & Continua*, vol. 74, no. 3, pp. 5625–5640, 2022, https://doi.org/10.32604/cmc.2023.033682.

[14] Md. A. Rahman, S. M. N. Sadat, A. T. Asyhari, N. Refat, M. N. Kabir, and R. A. Arshah, "A Secure and Sustainable Framework to Mitigate Hazardous Activities in Online Social Networks," *IEEE Transactions on Sustainable Computing*, vol. 6, no. 1, pp. 30–42, Jan. 2021, https://doi.org/10.1109/TSUSC.2019.2911188.

[15] S. Unnava and S. R. Parasana, "A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15607–15613, Aug. 2024, https://doi.org/10.48084/etasr.7621.

[16] R. Singh *et al.*, "Deep Learning for Multi-Class Antisocial Behavior Identification From Twitter," *IEEE Access*, vol. 8, pp. 194027–194044, 2020, https://doi.org/10.1109/ACCESS.2020.3030621.

[17] F. N. Al-Wesabi *et al.*, "Automatic Recognition of Cyberbullying in the Web of Things and social media using Deep Learning Framework," *IEEE Transactions on Big Data*, vol. 11, no. 1, pp. 259–270, Oct. 2025, https://doi.org/10.1109/TBDATA.2024.3409939.

[18] V. U. Gongane, M. V. Munot, and A. D. Anuse, "Detection and moderation of detrimental content on social media platforms: current status and future directions," *Social Network Analysis and Mining*, vol. 12, no. 1, Sep. 2022, Art. no. 129, https://doi.org/10.1007/s13278-022-00951-3.

[19] K. K. Mohbey, B. Agarwal, N. Kesswani, M. Sterjanov, Y. Nikol, and V. Margarita, "Hate Speech Identification and Categorization on Social Media Using Bi-LSTM: An Information Science Perspective," *Journal of Information Science Theory and Practice*, vol. 13, no. 1, pp. 51–69, 2025, https://doi.org/10.1633/JISTaP.2025.13.1.4.

[20] "Hatebase," *HATEBASE*. https://hatebase.org/.

[21] V. Basile *et al.*, "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA, Mar. 2019, pp. 54–63, https://doi.org/10.18653/v1/S19-2007.

[22] "TRAC-2024," *kaggle*. https://www.kaggle.com/datasets/rajsingh16/trac-2024.

[23] R. T. Potla, "AI-Powered Threat Detection in Online Communities: A Multi-Modal Deep Learning Approach," *Journal of Computer and Communications*, vol. 13, no. 2, pp. 155–171, Feb. 2025, https://doi.org/10.4236/jcc.2025.132010.

[24] G. Ramos *et al.*, "A comprehensive review on automatic hate speech detection in the age of the transformer," *Social Network Analysis and Mining*, vol. 14, no. 1, Oct. 2024, Art. no. 204, https://doi.org/10.1007/s13278-024-01361-3.

[25] A. Khan, A. Ahmed, S. Jan, M. Bilal, and M. F. Zuhairi, "Abusive Language Detection in Urdu Text: Leveraging Deep Learning and Attention Mechanism," *IEEE Access*, vol. 12, pp. 37418–37431, 2024, https://doi.org/10.1109/ACCESS.2024.3370232.

[26] K. Mnassri *et al.*, "A survey on multi-lingual offensive language detection," *PeerJ Computer Science*, vol. 10, Mar. 2024, Art. no. e1934, https://doi.org/10.7717/peerj-cs.1934.

[27] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, "An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques," *SN Computer Science*, vol. 3, no. 5, Jul. 2022, Art. no. 401, https://doi.org/10.1007/s42979-022-01308-5.

[28] N. Khanduja, N. Kumar, and A. Chauhan, "Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation," *Systems and Soft Computing*, vol. 6, Dec. 2024, Art. no. 200112, https://doi.org/10.1016/j.sasc.2024.200112.

[29] G. Ramos *et al.*, "Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts," *IEEE Access*, vol. 12, pp. 101374–101389, 2024, https://doi.org/10.1109/ACCESS.2024.3430848.

[30] E. N. Crothers, N. Japkowicz, and H. L. Viktor, "Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods," *IEEE Access*, vol. 11, pp. 70977–71002, 2023, https://doi.org/10.1109/ACCESS.2023.3294090.