A Hybrid BERT-CNN with Multihead Self-Attention for Automated Cyberbullying Detection

Meruert Yerekesheva
K. Zhubanov Aktobe Regional University, Kazakhstan
e_meruert@mail.ru

Oxana Akhmetova
Abai Kazakh National Pedagogical University, Kazakhstan
news4oxa@gmail.com

Assel Kaziyeva
Abai Kazakh National Pedagogical University, Kazakhstan
kaziyeva@gmail.com

Daniyar Sultan
Narxoz University, Kazakhstan
daniyar.sultan@narxoz.kz (corresponding author)

Aigerim Toktarova
International University of Tourism and Hospitality, Kazakhstan
toktar.aigerim@list.ru

Rustam Abdrakhmanov
International University of Tourism and Hospitality, Kazakhstan
rustamabdirakhmanov@gmail.com

Tolep Abdimukhan
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan
tolepabdimukhan@gmail.com

Received: 14 June 2025 | Revised: 10 October 2025 | Accepted: 18 October 2025

 DOI: https://doi.org/10.48084/etasr.12741

ABSTRACT

This paper presents a novel hybrid deep learning architecture, the CustomBERTCNNAttentionModel, designed for the automated detection of cyberbullying in social media text. The proposed model integrates the contextual language understanding capabilities of Bidirectional Encoder Representations from Transformers (BERT) with the local feature extraction strengths of Convolutional Neural Networks (CNNs) and the dynamic relevance weighting of multihead self-attention mechanisms. Evaluated on the Kaggle Cyberbullying Dataset, which includes both binary and multiclass labels, the model demonstrates superior performance compared to traditional classifiers and ensemble methods. The architecture effectively handles imbalanced and noisy text data, achieving an accuracy of 0.9853 in binary classification tasks. A comprehensive evaluation using standard metrics and visual analysis through confusion matrices confirms the model's robustness and its capacity to generalize across diverse types of cyberbullying. These results highlight the effectiveness of combining transformer-based embeddings with attention-enhanced convolutional structures for detecting harmful online behavior and contribute to the advancement of intelligent moderation systems.

Keywords-cyberbullying detection; Bidirectional Encoder Representations from Transformers (BERT); Convolutional Neural Network (CNN); multihead self-attention; hybrid deep learning; text classification; social media analysis; Natural Language Processing (NLP)

I. INTRODUCTION

Cyberbullying, characterized by deliberate and repetitive harassment through digital platforms, represents a pressing societal issue amplified by the ubiquitous adoption of social media and instant messaging services [1]. Its adverse consequences extend beyond mere emotional distress, contributing significantly to psychological trauma, depression, and in severe cases, suicidal tendencies among adolescents and young adults [2]. Consequently, detecting and mitigating cyberbullying incidents has become imperative, prompting researchers to develop sophisticated machine learning approaches capable of automating the identification process [3].

The complexity of textual communication, combined with linguistic subtleties, poses a considerable challenge for conventional classification methods, often resulting in suboptimal detection accuracy [4]. Traditional machine learning algorithms, including Support Vector Machines (SVM) and logistic regression, typically rely heavily on handcrafted textual features, which inadequately capture the semantic intricacies inherent in abusive or bullying language [5]. Recent advances in deep learning have paved the way for Natural Language Processing (NLP) models, notably transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT), which demonstrate superior context awareness and generalization capabilities [6].

Nonetheless, transformer-based methods alone may not fully capture local text features critical for precise classification, such as phrase-level nuances that indicate bullying behavior [7]. To address this limitation, Convolutional Neural Networks (CNNs) have shown promise in extracting local semantic features effectively, complementing transformer architectures to enhance detection performance [8]. Moreover, incorporating multihead self-attention mechanisms further enables models to dynamically weigh input features, significantly refining their ability to discern subtle linguistic patterns indicative of cyberbullying [9].

In response to these challenges and opportunities, this study introduces the CustomBERTCNNAttentionModel, an innovative hybrid architecture that synergistically combines BERT, CNN, and multihead self-attention modules. The proposed model aims to achieve improved accuracy and robustness in automated cyberbullying detection.

II. RELATED WORKS

Automated detection of cyberbullying has witnessed substantial growth, driven primarily by advances in NLP and machine learning algorithms [10]. Initial studies often employed traditional machine learning models, such as Naive Bayes, Decision Trees, and SVM, utilizing manually engineered textual features like n-grams, lexical patterns, and sentiment analysis [11]. While these methods established foundational frameworks, they consistently encountered limitations due to their inherent inability to generalize effectively to linguistically nuanced bullying content and context-specific slang [12].

Recognizing these limitations, research progressively shifted towards employing deep learning approaches. CNNs, renowned for their robust local pattern recognition capabilities, have been explored extensively in textual classification tasks, yielding significant improvements in cyberbullying detection accuracy [13]. Similarly, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have been adopted due to their ability to model sequential dependencies and capture contextual relationships within textual data [14]. Notably, CNN–LSTM hybrid models have demonstrated superior classification performance by harnessing both local textual patterns and sequential contextual information [15].

Recently, transformer-based architectures, especially BERT, have revolutionized NLP by delivering substantial enhancements in semantic comprehension through contextual embeddings generated via self-attention mechanisms [16]. BERT-derived models provide fine-grained contextual representations capable of capturing deeper linguistic subtleties critical for effectively distinguishing abusive language [17]. However, exclusive reliance on transformer models occasionally overlooks localized textual patterns indicative of cyberbullying, necessitating combined architectures [18].

Addressing this gap, recent studies have explored integrating transformer architectures with CNNs, creating hybrid frameworks optimized for extracting both global context and localized semantic features [19]. Such hybrid models have proven effective in capturing nuanced bullying behaviors, significantly improving classification metrics such as accuracy and precision compared to standalone models [20]. Additionally, the introduction of attention mechanisms, particularly multihead self-attention, into these hybrid frameworks has further improved their capability by dynamically prioritizing contextually relevant textual features [21].

Parallel to architectural advancements, extensive research has focused on dataset quality and annotation reliability. Annotated benchmark datasets like HateEval, Formspring, and Twitter cyberbullying datasets have served as essential resources for developing and evaluating detection models, highlighting the significance of domain-specific datasets for accurate model benchmarking [22]. Nonetheless, dataset imbalance and class distribution disparities remain critical challenges that significantly affect the robustness of automated detection models [23].

Moreover, research has increasingly emphasized the ethical implications and fairness in automated cyberbullying detection. Studies have underscored potential biases embedded within training data and proposed frameworks for ensuring model fairness and minimizing bias propagation [24]. Furthermore, interpretable AI approaches, including visualization methods

such as Grad-CAM and SHAP, have been integrated to enhance model transparency and facilitate trust in automated classification decisions [25].

Despite these substantial advancements, detecting cyberbullying remains challenging due to continuous linguistic evolution and adaptive behaviors exhibited by perpetrators [26]. Consequently, developing comprehensive, robust, and adaptive models that integrate hybrid architectures with advanced attention mechanisms continues to be a critical research objective [27]. This paper contributes to addressing these open challenges by introducing the CustomBERTCNNAttentionModel, explicitly designed to improve the precision and reliability of automated cyberbullying detection.

III. MATERIALS AND METHODS

The flowchart (Figure 1) illustrates a comprehensive pipeline for automated cyberbullying detection, beginning with the preprocessing of raw text documents. The initial stage involves a series of text normalization operations essential for enhancing the quality and consistency of the input data. These preprocessing tasks include converting text to lowercase to mitigate case-sensitivity issues, removing punctuation and words containing digits to eliminate irrelevant or noisy tokens, and trimming leading and trailing white spaces to ensure uniform formatting. Additionally, URLs are stripped from the text to avoid distractions from content that does not contribute to semantic understanding. These transformations serve to standardize the dataset and reduce variability that could adversely affect model performance.

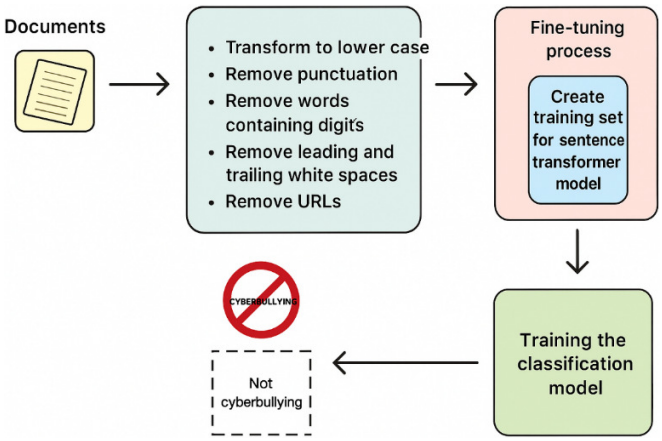

Fig. 1. Data preprocessing and training pipeline for cyberbullying classification.

Following the preprocessing phase, the refined textual data are utilized to generate a training set tailored for a sentence transformer model, which is central to the fine-tuning process. This transformer-based approach facilitates the extraction of high-dimensional contextual embeddings, enabling the model to comprehend the semantic and syntactic intricacies of language associated with cyberbullying. Subsequently, these embeddings are used to train a classification model that can distinguish between cyberbullying and non-cyberbullying content. The model is then evaluated based on its ability to accurately identify and label online texts, contributing to the development of effective automated systems for the early detection and prevention of harmful digital behavior.

A. Proposed Model

The proposed CustomBERTCNNAttentionModel is a hybrid deep learning architecture (Figure 2) that integrates the contextual power of BERT with the local feature extraction capabilities of CNNs and the dynamic relevance weighting provided by multihead self-attention.

The architecture begins by passing input text $X \in R^{1\times128}$ into a pre-trained BERT encoder, which transforms it into a sequence of contextual embeddings:

$$H = \mathrm{BERT}(X), \quad H \in R^{1\times128\times768} \tag{1}$$

To refine semantic understanding, the output from BERT is processed through a multihead self-attention mechanism. This layer computes attention scores between tokens to dynamically capture relationships across the entire sequence. The scaled dot-product attention is defined as:

$$\mathrm{Attention}(Q,K,V) = \mathrm{soft\,max}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{2}$$

Following attention, the tensor is transposed to match the expected input shape for 1D convolutional layers, enabling CNNs to extract local n-gram features along the token dimension. Two stacked Conv1D layers with ReLU activations transform the input, and a max pooling operation reduces the sequence to a condensed representation. After applying dropout for regularization, the features pass through fully connected layers to produce the final logits $\hat{y} \in R^2$, computed as:

$$\hat{y} = \mathrm{soft\,max}\left(W_2 \cdot \mathrm{ReLU}\left(W_1 \cdot x + b_1\right) + b_2\right) \tag{3}$$

This combination of global semantic encoding (BERT), attention-based refinement, and local feature extraction (CNN) allows the model to robustly detect cyberbullying patterns in text with high precision and contextual sensitivity.

B. Dataset

The Kaggle Cyberbullying Dataset [28], compiled by Saurabh Shahane, is a publicly available corpus designed to support the development and evaluation of machine learning models for cyberbullying detection in online textual content. It contains approximately 30,000 labeled comments sourced from social media platforms such as Twitter, categorized into six classes: not cyberbullying, gender-based, religion-based, age-based, ethnicity-based, and other cyberbullying. Each comment is paired with a class label, facilitating supervised learning approaches for distinguishing between neutral and harmful content. While the class distribution is relatively balanced across the cyberbullying subtypes, the overall dataset reveals a predominance of non-cyberbullying examples, introducing challenges for model training and necessitating the application of strategies such as resampling, class weighting, or data augmentation. The dataset is characterized by informal and

noisy language, including slang, abbreviations, and typographical inconsistencies, which accurately represent the linguistic nature of real-world social media discourse. As such, it is highly valuable for training robust models capable of generalizing to unstructured and unpredictable input. In this study, the dataset is employed to fine-tune the CustomBERTCNNAttentionModel for both binary and multiclass classification tasks, leveraging its linguistic complexity and diverse labels to comprehensively evaluate the model's detection performance.

Fig. 2. Architecture of the CustomBERTCNNAttentionModel with tensor dimensions for cyberbullying detection.

Figure 3 illustrates the relationship between sentiment and the prevalence of bullying in tweets, based on percentage distribution. The chart is divided into two sentiment categories: negative and positive. Within each sentiment group, the proportion of bullying and non-bullying tweets is shown. Notably, bullying content is more frequently associated with negative sentiment, comprising a larger share within that group. Conversely, positive sentiment tweets are predominantly non-bullying, with bullying tweets forming a smaller percentage. This trend indicates a clear correlation between negative sentiment and the likelihood of cyberbullying behavior, reinforcing the role of sentiment analysis as a valuable feature for predictive modeling in cyberbullying detection tasks. By highlighting the emotional tone of messages, this insight can be used to enhance feature engineering and guide the integration of sentiment-aware modules into classification models.

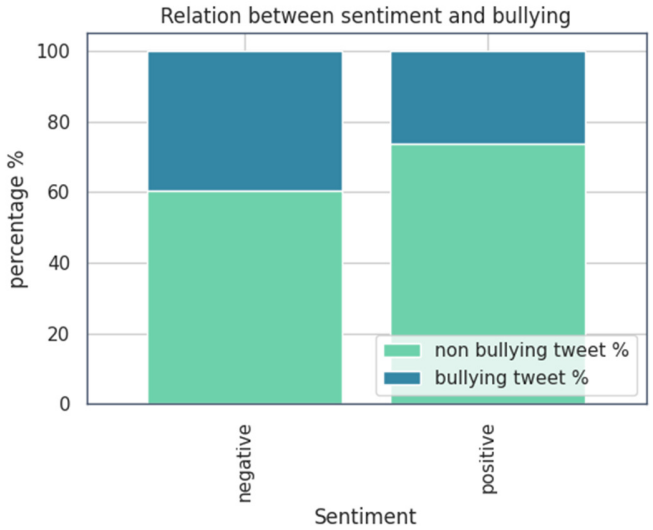

Fig. 3. Percentage distribution of bullying and non-bullying tweets by sentiment polarity.

C. Evaluation Metrics

To evaluate the CustomBERTCNNAttentionModel in detecting cyberbullying, several standard evaluation metrics are employed, including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model and is defined as the ratio of correctly predicted instances to the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision quantifies the model's ability to correctly identify positive instances and is given by:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

Recall, also known as sensitivity, reflects the model's ability to detect all relevant positive cases:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

The F1-score, which represents the harmonic mean of precision and recall, provides a balanced evaluation metric especially in the presence of class imbalance:

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

These metrics collectively offer a comprehensive evaluation of the model's effectiveness in identifying cyberbullying content, particularly in imbalanced datasets where accuracy alone may be misleading.

IV. RESULTS

The results section presents a comprehensive evaluation of the proposed CustomBERTCNNAttentionModel for automated cyberbullying detection, employing both binary and multiclass classification tasks. The model's performance is assessed using standard metrics such as accuracy, precision, recall, and F1-score, alongside visual tools including confusion matrices and class distribution analyses. Comparative experiments were conducted against several baseline classifiers, including traditional machine learning algorithms, ensemble methods, and optimized variants, to contextualize the efficacy of the proposed architecture. Additionally, the dataset's structural characteristics, such as class balance and diversity, are examined to ensure the reliability of performance outcomes.

Figure 4 illustrates the class distribution of the binary-labeled dataset, training set, and testing set used for cyberbullying detection. Each bar chart represents the count of two classes: class 0, denoting non-cyberbullying content, and class 1, representing cyberbullying instances. The first subplot, depicting the full dataset, reveals a significant class imbalance, with class 0 comprising the vast majority of samples. This skewed distribution persists in both the training and testing sets, as shown in the second and third subplots, respectively. While class 0 dominates with over 120,000 samples in the training set and approximately 30,000 in the testing set, class 1 samples remain relatively scarce, numbering only in the tens of thousands and a few thousand, respectively. Such imbalance poses a challenge for model training, as classifiers may become biased toward the majority class, leading to poor generalization on minority class predictions. This imbalance underscores the necessity of employing appropriate strategies such as weighted loss functions, oversampling, or data augmentation to mitigate bias. Despite the imbalance, the consistent proportion of classes across all subsets indicates that the data splitting process was stratified, preserving representativeness and enabling a fair and valid evaluation of the model's classification capabilities.

Figure 5 illustrates the distribution of six cyberbullying categories: religion, age, gender, ethnicity, not_cyberbullying, and other_cyberbullying. Each category contains around 7,900 to 8,000 samples, reflecting a nearly balanced dataset. This balance is important for multiclass classification, as it reduces the risk of bias toward any single class. Although not_cyberbullying and other_cyberbullying have slightly fewer instances, the difference is minimal. Such uniformity supports

robust model training and improves generalization in detecting various types of cyberbullying in online text.

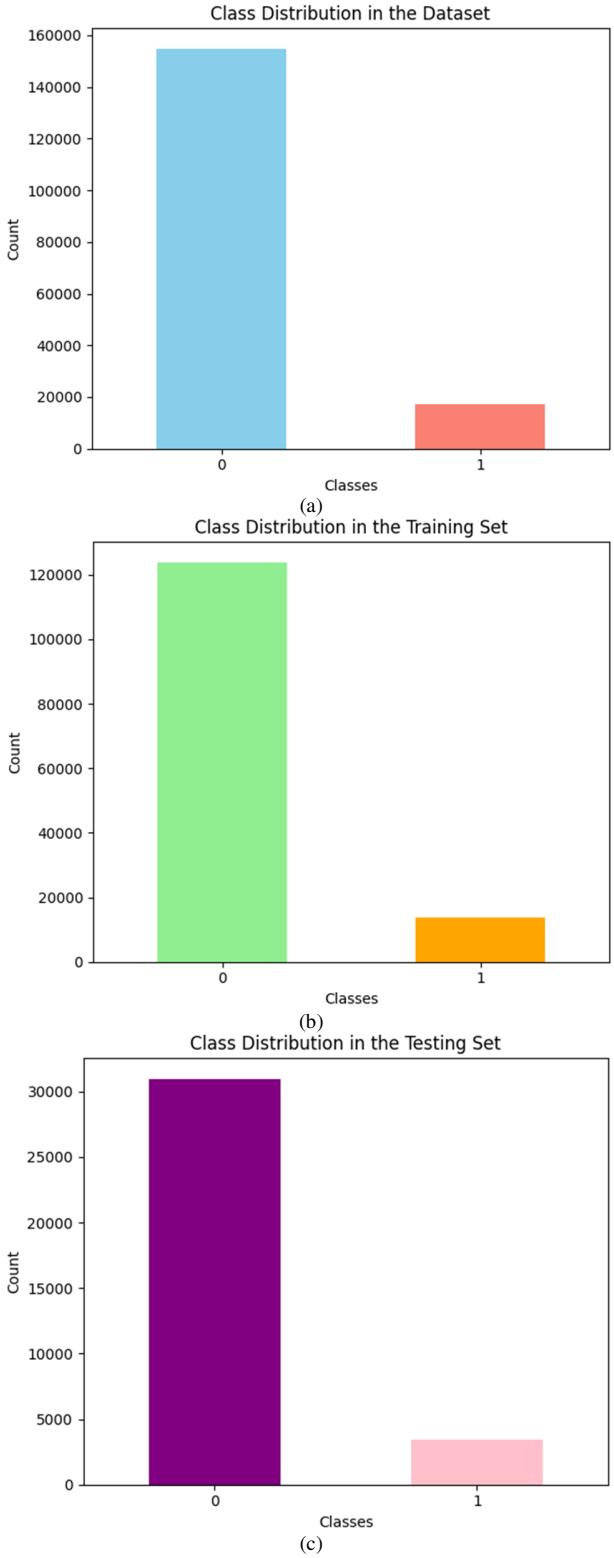

Fig. 4. Class distribution of cyberbullying and non-cyberbullying instances in the: (a) full dataset, (b) training set, and (c) testing set.

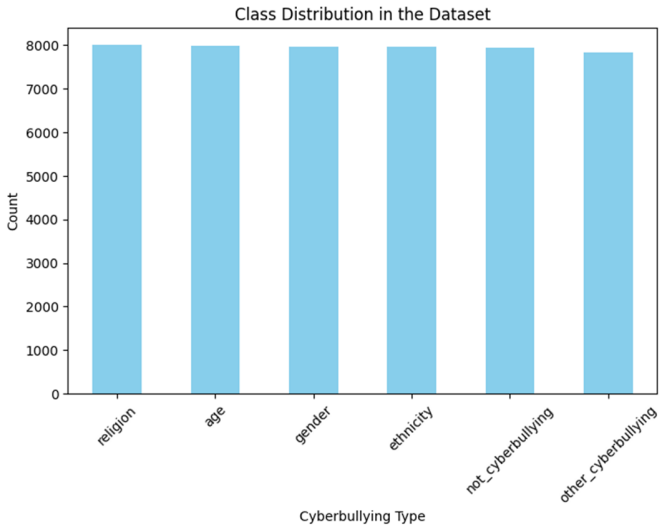

Fig. 5. Distribution of cyberbullying types in the multiclass dataset.

Figure 6 presents the confusion matrix for multiclass cyberbullying detection, assessing the model's classification across five categories: religion, age, gender, ethnicity, and not bullying. The model achieves high accuracy in age and ethnicity, with 398 and 408 correct predictions. However, misclassifications between gender and not bullying suggest semantic overlap. Minor confusion also appears in the not bullying category, reflecting difficulties with subtle or ambiguous language. Overall, the matrix highlights strong performance and reveals opportunities for improving nuanced content detection.

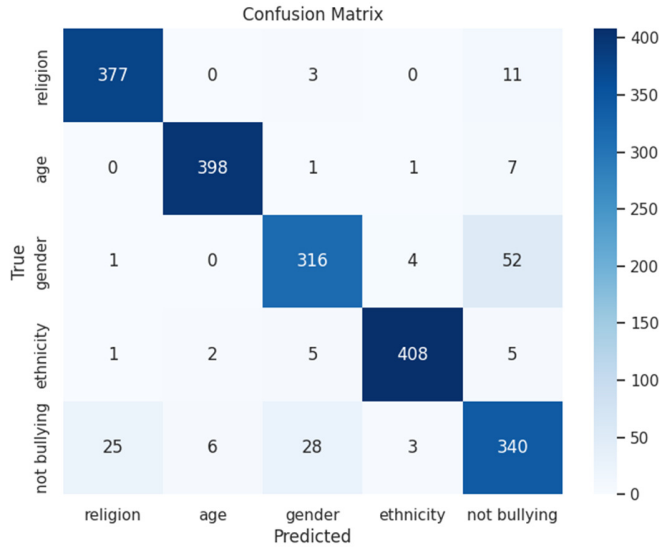

Fig. 6. Confusion matrix of the proposed model for multiclass cyberbullying classification.

Figure 7 presents a comprehensive comparative analysis of classification accuracy across a diverse range of baseline and advanced models applied to the cyberbullying detection task. The evaluated models include conventional machine learning

algorithms such as logistic regression, random forest, gradient boosting, and linear SVC, alongside more sophisticated methods like tuned variants, ensemble models (voting and stacking), and the proposed CustomBERTCNNAttentionModel. Among the baseline models, linear SVC achieved the highest performance with an accuracy of 0.8033, closely followed by the voting ensemble at 0.8011, indicating the effectiveness of combining multiple classifiers. However, these models still struggle to capture the complex contextual and linguistic patterns often found in online abuse, limiting their ability to generalize in nuanced classification settings.

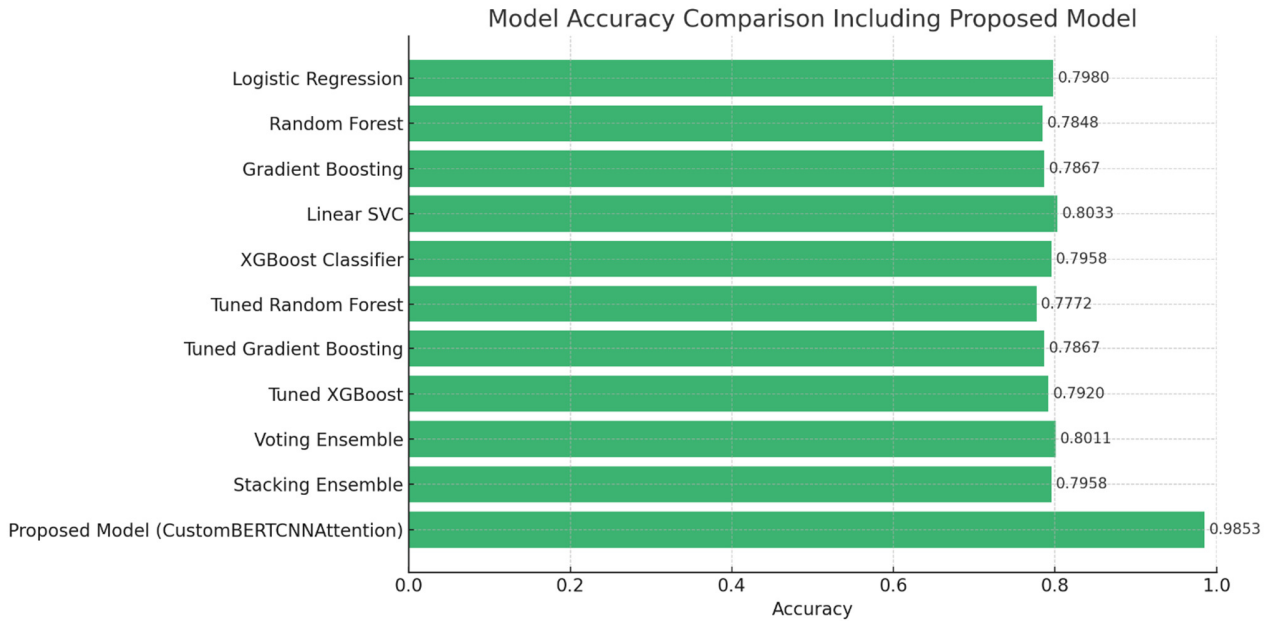

Fig. 7. Comparison of classification accuracy across traditional, ensemble, and proposed models for cyberbullying detection.

In contrast, the CustomBERTCNNAttentionModel demonstrates a substantial improvement, achieving a remarkably high accuracy of 0.9853. This performance leap underscores the model's enhanced ability to extract deep semantic features using BERT embeddings, refine them through convolutional layers, and dynamically prioritize relevant information via multihead self-attention mechanisms. Such a hybrid architecture is particularly well-suited for cyberbullying detection, where subtle cues, sarcasm, and shifting discourse require both local and global text understanding. The significant margin by which the proposed model outperforms all others in the comparison highlights its robustness, scalability, and potential as a state-of-the-art solution for automated content moderation in social media environments.

The marked performance gain underscores the effectiveness of integrating contextualized BERT embeddings with convolutional layers and multihead self-attention mechanisms. This comparison confirms that the proposed architecture not only captures both global semantics and local linguistic patterns more effectively but also provides a robust alternative to conventional classifiers and ensemble approaches in detecting nuanced cyberbullying behaviors in text.

Table I provides a comparative evaluation of various cyberbullying detection models reported in recent literature, highlighting their performance across key metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Among the models listed, traditional deep learning approaches such as RNN-LSTM and BiLSTM-GRU show strong performance, with accuracies ranging from 91.82% to 97%, but often lack full metric reporting. The BERT-based model by authors in [29] demonstrates high consistency across all metrics, achieving 98% in precision, recall, and F1-score. Notably, the proposed CustomBERTCNNAttentionModel surpasses all baseline and previously reported methods with an accuracy of 98.53%, precision of 98.70%, recall of 98.72%, F1-score of 98.70%, and AUC-ROC of 98.30%. These results suggest that the integration of BERT with CNN and multihead self-attention not only enhances contextual understanding but also strengthens local feature extraction and relevance weighting, leading to more robust and generalizable classification performance in cyberbullying detection tasks.

TABLE I. COMPARATIVE PERFORMANCE ANALYSIS OF EXISTING CYBERBULLYING DETECTION MODELS AND THE PROPOSED CUSTOMBERTCNNATTENTIONMODEL

Ref.	Method	Acc (%)	Prec (%)	Rec (%)	F1-score (%)	AUC-ROC (%)
[30]	LSTM + Fuzzy logic	93.67	-	93.62	93.64	96.41
[31]	RNN-LSTM	91.82	-	-	-	-
[32]	1D VGG16 + SVM	97.86	-	-	-	-
[29]	BERT-based cyberbullying detection	96	98	98	98	-
[13]	BiLSTM-GRU	97	98	97	-	-
[33]	BiLSTM-GRU	91	89	81	85	-
Proposed model	CustomBERTCNNAttentionModel	98.53	98.7	98.72	98.7	98.3

V. CONCLUSION

In conclusion, this study proposed a novel hybrid deep learning architecture, the CustomBERTCNNAttentionModel, for the automated detection of cyberbullying in online textual content. By integrating contextual embeddings from Bidirectional Encoder Representations from Transformers (BERT) with Convolutional Neural Network (CNN) layers and multihead self-attention mechanisms, the model effectively captures both global semantics and localized discriminatory patterns indicative of abusive language. Experimental results on the Kaggle Cyberbullying Dataset demonstrated that the proposed model outperforms a broad range of traditional machine learning classifiers and ensemble techniques, achieving a significantly higher accuracy of 0.9853. Moreover, both binary and multiclass classification evaluations confirmed the model's robustness and generalization capabilities, even in the presence of data imbalance and informal linguistic structures. Visual tools, such as confusion matrices and comparative performance graphs, further validated its superior discriminative power. These findings underscore the potential of hybrid transformer-CNN frameworks in advancing the state of automated cyberbullying detection and promoting safer online environments through intelligent content moderation.

ACKNOWLEDGMENT

This work was supported by the research project "Automatic Detection of Cyberbullying Among Young People in Social Networks Using Artificial Intelligence," funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. IRN AP23488900).

REFERENCES

[1] L. M. Al-Harigy, H. A. Al-Nuaim, N. Moradpoor, and Z. Tan, "Building towards Automated Cyberbullying Detection: A Comparative Analysis," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, June 2022, Art. no. 4794227, https://doi.org/10.1155/2022/4794227.

[2] N. M. Singh and S. K. Sharma, "An efficient automated multi-modal cyberbullying detection using decision fusion classifier on social media platforms," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 20507–20535, Feb. 2024, https://doi.org/10.1007/s11042-023-16402-w.

[3] C. Amol, L. Wanzare, and J. Obuhuma, "Modelling Misinformation in Swahili-English Code-switched Texts," *International Journal of Information Technology and Computer Science*, vol. 17, no. 1, pp. 67–80, Feb. 2025, https://doi.org/10.5815/ijitcs.2025.01.05.

[4] A. Kumar, R. Sharma, and P. Bedi, "Towards Optimal NLP Solutions: Analyzing GPT and LLaMA-2 Models Across Model Scale, Dataset Size, and Task Diversity," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14219–14224, June 2024, https://doi.org/10.48084/etasr.7200.

[5] R. Narayan and P. Samanta, "A Machine Learning Approach for Sentiment Analysis Using Social Media Posts," *International Journal of Information Technology and Computer Science*, vol. 16, no. 5, pp. 23–35, Oct. 2024, https://doi.org/10.5815/ijitcs.2024.05.02.

[6] R. Shamim and M. Lahby, "Automated Detection and Analysis of Cyberbullying Behavior Using Machine Learning," in *Combatting Cyberbullying in Digital Media with Artificial Intelligence*, 1st ed., M. Lahby, A.-S. K. Pathan, and Y. Maleh, Eds. Boca Raton, FL, USA: Chapman and Hall/CRC, 2023, pp. 116–136, https://doi.org/10.1201/9781003393061-9.

[7] M. Al-Hashedi, L.-K. Soon, H.-N. Goh, A. H. L. Lim, and E.-G. Siew, "Cyberbullying Detection Based on Emotion," *IEEE Access*, vol. 11, pp. 53907–53918, 2023, https://doi.org/10.1109/ACCESS.2023.3280556.

[8] A. A. Olagunju and I. O. Awoyelu, "Performance Evaluation of Fake News Detection Models," *International Journal of Information Technology and Computer Science*, vol. 16, no. 6, pp. 89–100, Dec. 2024, https://doi.org/10.5815/ijitcs.2024.06.07.

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

[10] S. S. Alzahrani, "Data Mining Regarding Cyberbullying in the Arabic Language on Instagram Using KNIME and Orange Tools," *Engineering, Technology & Applied Science Research*, vol. 12, no. 5, pp. 9364–9371, Oct. 2022, https://doi.org/10.48084/etasr.5184.

[11] T. H. Teng and K. D. Varathan, "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches," *IEEE Access*, vol. 11, pp. 55533–55560, 2023, https://doi.org/10.1109/ACCESS.2023.3275130.

[12] K. S. Ganguly, K. S. Ganguly, and A. Dutta, "A Comparative Study of Statistical (SARIMA) Vis-À-Vis Some Traditional Machine-Learning and Deep-Learning Techniques to Forecast Malaria Incidences in Kolkata of India," *International Journal of Information Technology and Computer Science*, vol. 17, no. 5, pp. 68–83, Oct. 2025, https://doi.org/10.5815/ijitcs.2025.05.06.

[13] G. Jaradat, M. Shehab, D. Ibrahim, S. Najdawi, and R. Sihwail, "Deep Learning Approaches for Detecting Cyberbullying on Social Media," *Journal of Computational and Cognitive Engineering*, Mar. 2025, https://doi.org/10.47852/bonviewJCCE52024162.

[14] M. K. Mali *et al.*, "Automatic detection of cyberbullying behaviour on social media using Stacked Bi-Gru attention with BERT model," *Expert Systems with Applications*, vol. 262, Mar. 2025, Art. no. 125641, https://doi.org/10.1016/j.eswa.2024.125641.

[15] Y. Tashtoush, A. Banysalim, M. Maabreh, S. Al-Eidi, O. Karajeh, and P. Zahariev, "A Deep Learning Framework for Arabic Cyberbullying Detection in Social Networks," *Computers, Materials & Continua*, vol. 83, no. 2, pp. 3113–3134, Apr. 2025, https://doi.org/10.32604/cmc.2025.062724.

[16] K. I. Arce-Ruelas, O. Alvarez-Xochihua, L. Pellegrin, L. Cardoza-Avendaño, and J. Á. González-Fraga, "Automatic Cyberbullying Detection: a Mexican case in High School and Higher Education students," *IEEE Latin America Transactions*, vol. 20, no. 5, pp. 770–779, May 2022, https://doi.org/10.1109/TLA.2022.9693561.

[17] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839–1852, June 2023, https://doi.org/10.1007/s00530-020-00701-5.

[18] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," *Information*, vol. 14, no. 8, Aug. 2023, Art. no. 467, https://doi.org/10.3390/info14080467.

[19] M. Fahaad Almufareh, N. Zaman Jhanjhi, M. Humayun, G. Naif Alwakid, D. Javed, and S. Naif Almuayqil, "Integrating Sentiment Analysis With Machine Learning for Cyberbullying Detection on Social Media," *IEEE Access*, vol. 13, pp. 78348–78359, 2025, https://doi.org/10.1109/ACCESS.2025.3558843.

[20] V. Balakrisnan and M. Kaity, "Cyberbullying detection and machine learning: a systematic literature review," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 1375–1416, Oct. 2023, https://doi.org/10.1007/s10462-023-10553-w.

[21] S. Pericherla and E. Ilavarasan, "Cyberbullying detection and classification on social media images using Convolution Neural Networks and CB-YOLO model," *Evolving Systems*, vol. 16, no. 2, Feb. 2025, Art. no. 43, https://doi.org/10.1007/s12530-025-09656-2.

[22] A. Al-Marghilani, "Artificial Intelligence-Enabled Cyberbullying-Free Online Social Networks in Smart Cities," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, Jan. 2022, Art. no. 9, https://doi.org/10.1007/s44196-022-00063-y.

[23] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan, "Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying," *Social Network Analysis and Mining*, vol. 12,

no. 1, Aug. 2022, Art. no. 99, https://doi.org/10.1007/s13278-022-00934-4.

[24] K. Subhashree and S. M. Kumar, "Enhanced quantum long short-term memory neural network based multi-task learning for sentimental analysis and cyberbullying detection," *Expert Systems with Applications*, vol. 282, July 2025, Art. no. 127555, https://doi.org/10.1016/j.eswa.2025.127555.

[25] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," *Information Processing & Management*, vol. 60, no. 5, Sept. 2023, Art. no. 103454, https://doi.org/10.1016/j.ipm.2023.103454.

[26] D. L. Hall, Y. N. Silva, B. Wheeler, L. Cheng, and K. Baumel, "Harnessing the Power of Interdisciplinary Research with Psychology-Informed Cyberbullying Detection Models," *International Journal of Bullying Prevention*, vol. 4, no. 1, pp. 47–54, Mar. 2022, https://doi.org/10.1007/s42380-021-00107-5.

[27] M. Al-Ajlan and M. Ykhlef, "Firefly-CDDL: A Firefly-Based Algorithm for Cyberbullying Detection Based on Deep Learning," *Computers, Materials & Continua*, vol. 75, no. 1, pp. 19–34, Feb. 2023, https://doi.org/10.32604/cmc.2023.033753.

[28] "Cyberbullying Dataset." Kaggle. [Online]. Available: https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset.

[29] B. Ogunleye and B. Dharmaraj, "The Use of a Large Language Model for Cyberbullying Detection," *Analytics*, vol. 2, no. 3, pp. 694–707, Sept. 2023, https://doi.org/10.3390/analytics2030038.

[30] M. H. Obaid, S. K. Guirguis, and S. M. Elkaffas, "Cyberbullying Detection and Severity Determination Model," *IEEE Access*, vol. 11, pp. 97391–97399, 2023, https://doi.org/10.1109/ACCESS.2023.3313113.

[31] 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, July 2022, Art. no. 401, https://doi.org/10.1007/s42979-022-01308-5.

[32] S. Viswanath and A. K. K M, "Hybrid 1D VGG16 and SVM Framework for Early Detection and Intensity Classification of Cyberbullying," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 3, pp. 543–555, Apr. 2025, https://doi.org/10.22266/ijies2025.0430.37.

[33] A. O. Akinwumi, A. O. Ige, J. R. Obafemi, O. D. Akinrolabu, and B. O. Akingbesote, "CBDS-ConvNet: A Cyber-Bullying Detection Model using Convolutional Neural Network," *Communications on Applied Electronics*, vol. 7, no. 40, pp. 11–21, Jan. 2025, https://doi.org/10.5120/cae2025652905.