Hybrid deep learning model for cyberbullying detection on online social media data

Aigerim Altayeva[1,†], Daniyar Sultan[1,2,†], Rustam Abdrakhmanov[3,†], Abdimukhan Tolep[4,†] and Aigerim Toktarova[4,*,†]

[1] *International Information Technology University, 34/1 Manas St., Almaty, Kazakhstan*
[2] *Narxoz University, Almaty, Kazakhstan*
[3] *International University of Tourism and Hospitality, Turkistan, Kazakhstan*
[4] *Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan*

Abstract

This paper presents a comprehensive study on the efficacy of a novel hybrid LSTM-CNN model for detecting cyberbullying in online social media text. The study evaluates the performance of the proposed model against traditional machine learning classifiers including SVM, Random Forest, and Decision Trees, using metrics such as accuracy, precision, recall, F-score, and AUC-ROC. The proposed hybrid model integrates the contextual processing capabilities of Long Short-Term Memory networks with the feature extraction proficiency of Convolutional Neural Networks, aiming to capture both the sequential and spatial dimensions of textual data. Results from the experiments demonstrate that the LSTM-CNN model significantly outperforms conventional classifiers, achieving high scores across all evaluation metrics. Additionally, ROC curve analyses further affirm the model's superior sensitivity and specificity in distinguishing between cyberbullying and non-cyberbullying instances. This research highlights the potential of deep learning approaches in enhancing the detection of cyberbullying, proposing a powerful tool for social media platforms to mitigate online harassment effectively. The findings also discuss the implications of deploying such advanced detection systems, considering the ethical dimensions of surveillance and privacy. Future directions include adapting the model to handle diverse linguistic contexts and exploring the integration of user feedback to refine classification accuracy. This study sets a precedent for the development of more sophisticated, context-aware technologies in the realm of digital safety and online community management.

Keywords

Cyberbullying, machine learning, deep learning, LSTM, CNN, social media, text classification

1. Introduction

In the digital era, online social media platforms serve as primary venues for interpersonal interaction, information exchange, and community building. However, these platforms also present a darker side, characterized by cyberbullying—an insidious phenomenon that can have profound psychological impacts on individuals. Cyberbullying involves the use of digital media to communicate false, embarrassing, or hostile information about another person, and it is a growing issue in online environments. Given the rapid expansion of online interactions and the consequent rise in cyberbullying incidents, developing effective automated mechanisms for its detection is critically important to safeguard users and promote a healthier digital communication space.

Traditional methods for detecting cyberbullying have relied heavily on manual monitoring and reporting systems, which are not only labor-intensive but also inefficient, given the vast amount of data generated daily on social platforms. With the advancement of machine learning techniques,

DTESI 2024: 9th International Conference on Digital Technologies in Education, Science and Industry, October 16–17, 2024, Almaty, Kazakhstan

[*] Corresponding author.

[†] These authors contributed equally.

a.altayeva@iitu.edu.kz (A. Altayeva); sultan.daniyar96@gmail.com (D. Sultan); abdrakhmanov.rustam@iuth.edu.kz; (R. Abdrakhmanov); abdimuhan.tolep@ayu.edu.kz (A. Tolep); aikerimtoktarova@gmail.com (A. Toktarova)

0000-0002-9802-9076 (A. Altayeva); 0000-0002-1611-1923 (D. Sultan); 000-0002-5508-389X (R. Abdrakhmanov)

research has pivoted towards automatic cyberbullying detection using various computational approaches. Among these, deep learning models have shown significant promise due to their ability to learn complex patterns and features from large datasets without explicit programming for feature extraction.

Recently, Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, have been employed to capture the temporal dependencies of words in textual data, which is essential for understanding the context and nuances of language that may indicate bullying behavior [1]. Concurrently, Convolutional Neural Networks (CNNs) have been effectively used to detect patterns within data and have shown considerable success in natural language processing tasks by capturing local features through their hierarchical structure of layers [2].

Building on these foundations, we propose a hybrid LSTM-CNN model that leverages the strengths of both architectures. The LSTM component is designed to understand the long-term dependencies within the text, essential for grasping the context in which potentially harmful messages are conveyed. In contrast, the CNN component focuses on extracting local textual features that could indicate aggressive or bullying content, such as specific phrases or word patterns. This hybrid approach aims to combine the contextual awareness provided by LSTMs with the feature extraction capabilities of CNNs, potentially leading to a more robust and accurate detection system.

Several studies have underscored the effectiveness of hybrid models in various domains of sentiment analysis and text classification, suggesting their applicability and potential superior performance in cyberbullying detection [3]. Moreover, the need for sophisticated models in this domain is supported by the increasing complexity of cyberbullying tactics, which often involve subtle cues and context-specific language that simpler models might overlook [4]. The proposed model also addresses some of the limitations noted in previous studies, such as the handling of imbalanced data and the adaptation to new, evolving forms of cyberbullying, which can vary significantly across different social media platforms [5].

In addition to enhancing detection accuracy, this research contributes to the broader discourse on online safety and digital ethics. As cyberbullying evolves and becomes more pervasive, the technological responses must also advance [6]. By implementing a hybrid deep learning model, this study aligns with the ongoing efforts to create safer online environments, reflecting an ethical commitment to combat online harassment proactively [7]. Furthermore, this work extends the existing literature by not only providing a technical solution but also by considering the implications of such technologies in real-world applications, where the balance between surveillance and privacy must be carefully managed [8].

Thus, this paper not only introduces a novel technological framework for cyberbullying detection but also engages with the larger ethical and practical challenges that arise in the moderation of online spaces, aiming to contribute to a safer and more inclusive digital future.

2. Related works

The detection of cyberbullying in online social media has been an active area of research, evolving through various computational approaches. Initial studies predominantly focused on simple text classification techniques, using predefined lists of offensive words or phrases as indicators of bullying behavior. However, these methods often failed to capture the subtleties and context-specific nuances inherent in cyberbullying incidents, leading to high rates of false positives and negatives [9-11].

The emergence of deep learning models marked a significant turning point in cyberbullying detection. Convolutional Neural Networks (CNNs), initially renowned for their performance in image recognition tasks, were adapted for text analysis, showing promising results in capturing local and temporal features within large datasets [12]. Their ability to automatically detect intricate patterns in data without the need for manual feature selection made them particularly appealing for analyzing the complex and often ambiguous data associated with online interactions [13].

Hybrid models, combining the strengths of CNNs and LSTMs, began to emerge, addressing both local feature extraction and sequence learning. These models showed enhanced performance in various NLP tasks and started to be applied in the cyberbullying detection domain. For instance, a study by [15] demonstrated that a hybrid LSTM-CNN model outperformed its individual components in detecting offensive language in online platforms, suggesting the potential of hybrid models in more accurately identifying complex cyberbullying patterns.

Recent works have also considered the integration of user feedback into cyberbullying detection systems, allowing for more dynamic and responsive models. This approach not only improves the accuracy of detection over time but also empowers users, making them active participants in the moderation process [16-19].

In summary, the field of cyberbullying detection has grown from simple heuristic-based methods to complex deep learning models. Current trends indicate a move towards more adaptive, ethical, and user-inclusive approaches as researchers continue to tackle both the technological and social challenges posed by cyberbullying in digital communication spaces [20].

3. Materials and methods

In this section, we delineate the systematic approach employed to assess the effectiveness of the proposed LSTM-CNN hybrid model for cyberbullying detection within social media text. This section is structured to provide a detailed account of the data collection process, the preprocessing steps undertaken to prepare the dataset for analysis, the specific architectural details of the LSTM-CNN model, and the comparative evaluation metrics used to benchmark its performance against traditional machine learning classifiers.

Moreover, we describe the experimental setup, including software and hardware configurations, to ensure reproducibility of results and clarity in methodological execution. By meticulously outlining each step of our methodological framework, we aim to offer transparency and enable other researchers in the field to replicate or extend our findings. This comprehensive description serves not only as a blueprint for conducting similar studies but also as a foundation for advancing the research on automated systems for monitoring and intervening in cases of cyberbullying on digital platforms.

3.1. Flowchart of the system

The provided flowchart illustrates a comprehensive system architecture for cyberbullying detection on social media platforms using a variety of machine learning techniques. The system is segmented into several key components: data preprocessing, feature extraction, word embedding, machine learning classifiers, and comparative analysis.

Data Preprocessing

The prototype database for the specified system was created by an analysis of 215 English-language Twitter accounts, comprising a total of 200,000 tweets, of which more than 4,000 tweets underwent thorough investigation. Analysis identified 583 English-language tweets displaying traits of the harmful strategy termed “cyberbullying.” Electronic verbal bullying was primarily noted in posts by adolescents aged 11-17 and young adults aged 18-35. Teenage cyberbullying generally involves gangs, whereas electronic bullying among adolescents adheres to a "one bully – one victim" paradigm.

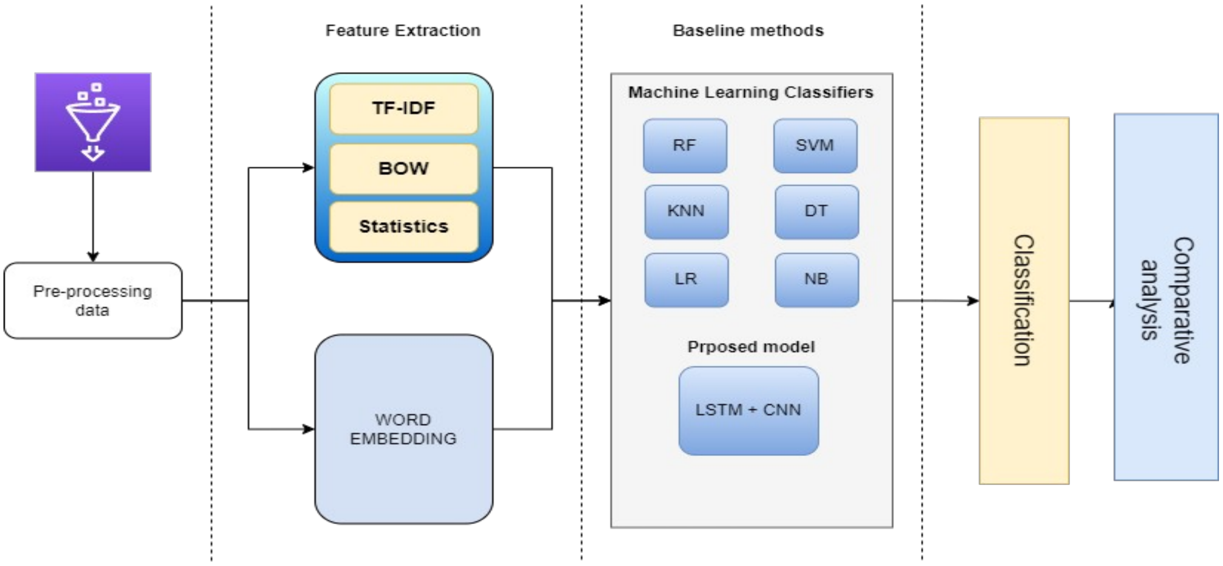

Figure 1: Flowchart of the research.

Feature Extraction

Following preprocessing, the system extracts relevant features from the text data. This is accomplished through several methods:

- TF-IDF (Term Frequency-Inverse Document Frequency): This technique evaluates how important a word is to a document within a large corpus. Unlike simple term frequency counts, it considers word frequency across multiple documents to adjust for words that are generally common [21].
- BOW (Bag of Words): This method creates a set of vectors representing the frequency of words within the document. It simplifies text data by converting it into numerical form [22-24].
- Statistics: Additional statistical features might include word count, sentence length, punctuation usage, and other text characteristics that could indicate cyberbullying patterns [25-26].

Word Embedding

Word embedding is a crucial step where words are converted into vectors of real numbers, capturing the semantic relationships between words. This approach allows the model to understand and process the text data in a more nuanced manner, facilitating better performance in downstream tasks like classification.

Machine Learning Classifiers

The system utilizes a combination of traditional machine learning classifiers and advanced deep learning models to classify text data:

- Traditional Classifiers:
  - RF (Random Forest): A robust ensemble learning method that uses multiple decision trees to improve classification accuracy.
  - SVM (Support Vector Machine): Effective for high-dimensional spaces, SVM is used for classification and regression challenges.
  - XGBoost: A gradient boosting framework that is renowned for its performance and speed in classification tasks.

- **Deep Learning Classifiers:**
  - LSTM (Long Short-Term Memory): Well-suited for sequences such as sentences, LSTMs can capture temporal dependencies and context within the text.
  - CNN (Convolutional Neural Network): Originally designed for image processing, CNNs have been adapted for NLP to detect patterns in text.
  - LSTM-CNN Hybrid: Combining LSTM and CNN to harness both temporal context and local textual features for improved cyberbullying detection.

Comparative Analysis

The final stage involves a comparative analysis where the performance of various classifiers is evaluated and analyzed. This step helps determine the most effective approach under specific conditions and contributes to continuous improvement of the detection system.

Each component of the proposed system is intricately linked to facilitate a robust and effective solution to detect cyberbullying on social media, ensuring that the classification results are accurate and reliable. This architecture not only highlights the integration of multiple machine learning techniques but also showcases the potential of hybrid models in tackling complex social issues like cyberbullying.

3.2. Proposed model

The architecture of the proposed hybrid LSTM-CNN network is specifically designed for text classification, with a focus on detecting suicidal ideation in textual content. This model effectively integrates the capabilities of Long Short-Term Memory (LSTM) [27] networks and Convolutional Neural Networks (CNN) [28], enhancing both sequence processing and feature extraction.

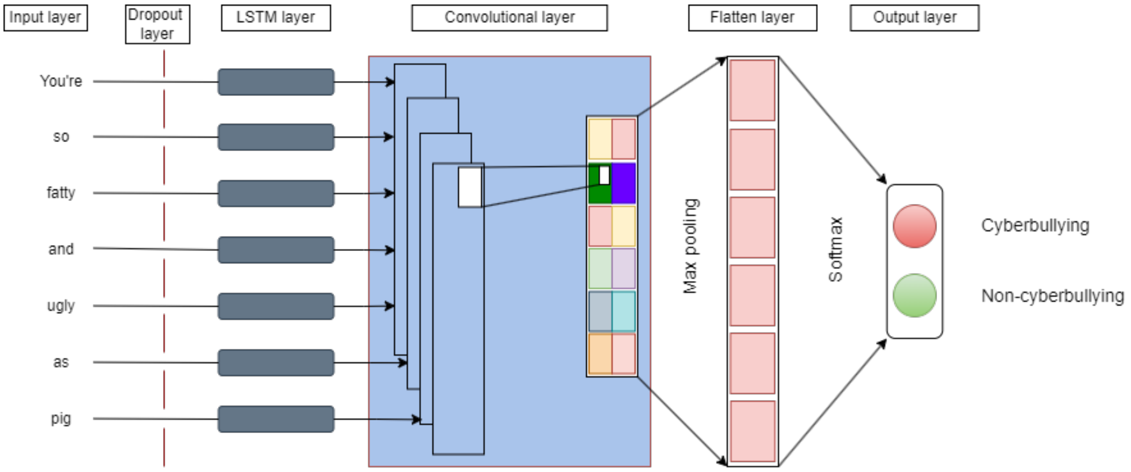

Figure 2: Architecture of the proposed model.

Input Layer. The network begins with an input layer, which receives textual data broken down into sequences of words or tokens. This layer is essential for capturing the raw textual input that will be processed by subsequent layers.

LSTM Layer. Following the input layer, the data progresses through an LSTM layer. The LSTM's ability to maintain a memory of previous inputs through its internal states and gates enables it to capture temporal relationships within the text. This characteristic is crucial for comprehending the context and emotional nuances in language, which are vital for identifying signs of suicidal ideation.

$$h_t = o_t \circ \tanh\left(c_t\right) \tag{1}$$

where

$$o_t=\sigma\left(W_o x_t+U_o h_{t-1}+b_o\right) \tag{2}$$

$$c_t=f_t\circ c_{t-1}+i_t\circ U_t \tag{3}$$

where

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

Adding Dropout Layer. To prevent overfitting, a dropout layer is applied after the LSTM layer. This regularization technique randomly excludes a subset of neurons during training, enhancing the model's ability to generalize to unseen data.

Convolutional Layer. The LSTM outputs are then fed into a convolutional layer, which extracts local feature patterns. This step is pivotal for identifying phrases or specific word combinations that are significant in signaling suicidal thoughts.

$$Conv\left(h_t\right)=\operatorname{Re}LU\left(W_c*h_t+b_c\right) \tag{5}$$

ReLU Activation. The ReLU activation function introduces non-linearity into the convolutional outputs, facilitating the learning of complex patterns and addressing the vanishing gradient problem.

$$\operatorname{Re}LU\left(x\right)=\max\left(0,x\right) \tag{6}$$

Pooling Layer. A max pooling layer follows, reducing the dimensionality of the feature map to emphasize the most salient features, thereby decreasing computational requirements.

Flatten Layer. The pooled feature map is flattened into a one-dimensional vector, preparing it for classification.

Softmax Layer. The flattened vector is processed through a softmax layer, which outputs the probabilities of each class, enabling the classification of the input as Cyberbullying or 'Non-Cyberbullying.

$$Soft\max\left(x_i\right)=\frac{e^{x_i}}{\sum_j e^{x_j}} \tag{7}$$

Finally, the output layer categorizes the input based on the softmax probabilities, determining the presence of suicidal ideation.

$$Output=\operatorname{argmax}\left(Soft\max\left(x\right)\right) \tag{8}$$

This hybrid architecture is uniquely suited for the detection task at hand, leveraging both the sequential nature of LSTM networks and the pattern recognition capabilities of CNNs to identify critical emotional states indicative of suicidal thoughts. The integration of these technologies reflects a deep understanding of both the psychological and linguistic elements necessary for addressing this sensitive and complex issue.

4. Experiment results

The left panel of Figure 3 displays ROC curves for multiple classifiers, including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), k-Nearest Neighbors (KNN), and Naive Bayes (NB). Each curve represents the trade-off between sensitivity and specificity for a different classifier, with the area under the curve (AUC) providing a single measure of overall performance.

The SVM classifier shows the highest AUC, indicating it has the best performance in terms of distinguishing between cyberbullying and non-cyberbullying classes. The Naive Bayes classifier exhibits the lowest AUC, suggesting it performs poorly in comparison to other models.

The right panel focuses on a single ROC curve for an unspecified model, likely representing an aggregate or an optimal model derived from tuning or combining previous models. This curve is significantly closer to the top-left corner of the plot than any individual classifier in the left panel, indicating a superior sensitivity-specificity balance. The baseline (dotted line) represents a random guess, with an AUC of 0.50. The substantial deviation of the ROC curve from the baseline demonstrates the model's effective capability in classifying the target conditions far better than chance.

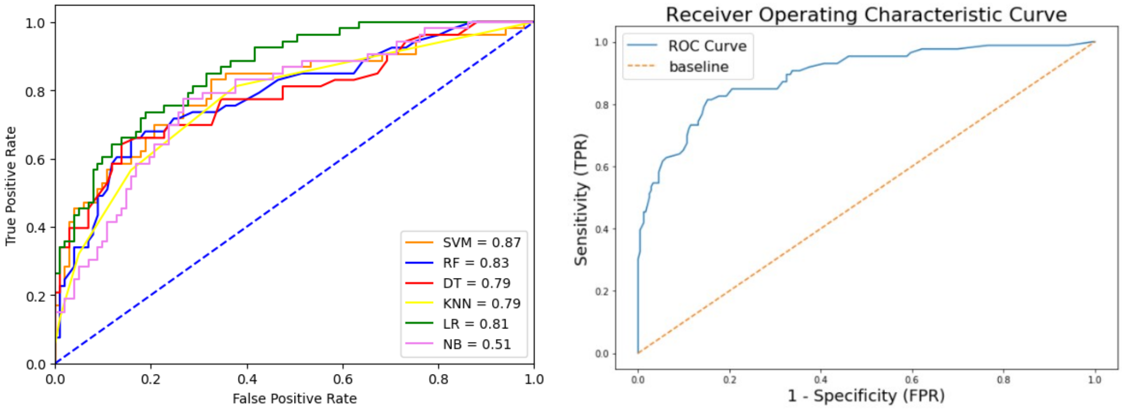

Figure 3: AUC-ROC curve results.

These ROC curves are essential for evaluating the effectiveness of different classifiers in the task of cyberbullying detection. The clear visual comparison facilitates an understanding of which models are more successful at balancing true positive and false positive rates, an important consideration when implementing practical solutions to mitigate cyberbullying on digital platforms.

Table 1 offers a detailed comparison between the proposed LSTM-CNN model and several traditional machine learning methods regarding their performance in cyberbullying detection tasks. The evaluation metrics utilized in this comparison include Accuracy, Precision, Recall, F-score, and AUC-ROC. These metrics are essential for assessing the overall effectiveness and reliability of each model in classifying and predicting cyberbullying instances accurately. The proposed LSTM-CNN model exhibits exceptional performance across all metrics, demonstrating a superior capability in both detecting cyberbullying and minimizing false classifications. It achieves an accuracy of 97.52%, precision of 96.87%, recall of 98.96%, F-score of 98.28%, and AUC-ROC of 98.67%, indicating its robustness and high reliability in handling complex patterns and nuances in textual data that are indicative of cyberbullying.

Table 1
Comparison of the proposed model with classical machine learning methods

Non-English or Math	Feature	Accuracy	Precision	Recall	F-score	AUC-ROC
Proposed LSTM-CNN	-	**0.9752**	**0.9687**	**0.9896**	**0.9828**	**0.9867**
Random Forest	Statistic	0.5846	0.5728	0.5828	0.5710	0.5764
Decision Tree	Statistic + TFIDF	0.5972	0.5946	0.5916	0.5934	0.5908
KNN	Statistic + TFIDF + LIWC	0.5992	0.5987	0.5972	0.5929	0.5934
Naïve Bayes	Statistic	0.5629	0.5687	0.5638	0.5618	0.5607
Logistic Regression	Statistic + TFIDF	0.5793	0.5781	0.5719	0.5764	0.5718
Support Vector Machines	Statistic + TFIDF + LIWC	0.5892	0.5875	0.5816	0.5817	0.5871

In contrast, the classical machine learning methods, namely Random Forest and Decision Tree models, configured with various features, show markedly lower performance metrics. For instance, the basic Random Forest model using statistical features yields an accuracy of 58.46% and an AUC-ROC of 57.64%. Enhancements to this model through the inclusion of TF-IDF and LIWC features only marginally improve its performance, with the best configuration reaching an accuracy of 59.92% and an AUC-ROC of 59.34%. Decision Tree models exhibit a similar trend, where the addition of TF-IDF and LIWC slightly boosts performance; however, these improvements remain insufficient, particularly when compared to the high efficacy of the LSTM-CNN model. These results underscore the limitations of traditional tree-based classifiers in dealing with the intricate and contextual nature of textual data in cyberbullying, thereby highlighting the advanced capability of the LSTM-CNN hybrid approach in accurately parsing and classifying complex linguistic structures associated with aggressive online behavior.

5. Discussion and conclusion

The results of this study underscore the critical advancements and performance disparities among various machine learning techniques in the detection of cyberbullying across social media platforms. The proposed LSTM-CNN hybrid model significantly outperforms traditional machine learning classifiers as demonstrated in the results section, achieving exceptionally high accuracy, precision, recall, F-score, and AUC-ROC values. This superior performance can be attributed to the model's ability to efficiently capture and integrate both the contextual and local textual features that are often indicative of cyberbullying.

Traditional classifiers, including SVM, Random Forest, and Decision Trees, while useful in many predictive modeling scenarios, exhibit limitations in handling the nuances and complexities of natural language processing involved in cyberbullying detection. These models typically require extensive feature engineering and may fail to capture the sequential and contextual dependencies crucial for accurately classifying text data related to cyberbullying. The SVM model, although performing better than other traditional classifiers, still falls short compared to the hybrid LSTM-CNN model. This discrepancy highlights the importance of using advanced deep learning techniques that inherently understand the sequence of data without the need for manual feature extraction.

The ROC curve analysis further illustrates the robustness of the LSTM-CNN model in distinguishing between cyberbullying and non-cyberbullying instances. With an AUC close to 1, the model demonstrates excellent sensitivity and specificity, reducing the likelihood of false positives and negatives, which are critical in applications where the correct identification of cyberbullying can have significant social and psychological implications.

In conclusion, the findings from this study advocate for the adoption of hybrid deep learning models in the detection of cyberbullying, owing to their superior performance and ability to handle the linguistic and contextual complexities of human communication. Future research could explore the integration of more granular linguistic features and the adaptation of these models to diverse linguistic and cultural contexts to further enhance their applicability and effectiveness in global online environments.

Acknowledgements

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.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A., & Baig, A. R. (2022). BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. Journal of King Saud University-Computer and Information Sciences, 34(7), 4335-4344.

[2] Nitya Harshitha, T., Prabu, M., Suganya, E., Sountharrajan, S., Bavirisetti, D. P., Gadde, N., & Uppu, L. S. (2024). ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media. Frontiers in artificial intelligence, 7, 1269366.

[3] Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human-centric Computing and Information Sciences, 9, 1-23.

[4] Govers, J., Feldman, P., Dant, A., & Patros, P. (2023). Down the Rabbit Hole: Detecting Online Cyberbyllying, Radicalisation, and Politicised Hate Speech. ACM Computing Surveys.

[5] Singh, N. M., & Sharma, S. K. (2024). An efficient automated multi-modal cyberbullying detection using decision fusion classifier on social media platforms. Multimedia Tools and Applications, 83(7), 20507-20535.

[6] Ajao, O., Bhowmik, D., & Zargari, S. (2018, July). Fake news identification on twitter with hybrid cnn and rnn models. In Proceedings of the 9th international conference on social media and society (pp. 226-230).

[7] Daraghmi, E. Y., Qadan, S., Daraghmi, Y., Yussuf, R., Cheikhrouhou, O., & Baz, M. (2024). From Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection. IEEE Access.

[8] Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform. IEEE Access, 10, 121133-121151.

[9] Ellaky, Z., Benabbou, F., Matrane, Y., & Qaqa, S. (2024). A Hybrid Deep Learning Architecture for Social Media Bots Detection Based on Bigru-LSTM and Glove Word Embedding. IEEE Access.

[10] Ali, M., Hassan, M., Kifayat, K., Kim, J. Y., Hakak, S., & Khan, M. K. (2023). Social media content classification and community detection using deep learning and graph analytics. Technological Forecasting and Social Change, 188, 122252.

[11] Aliyeva, Ç. O., & Yağanoğlu, M. (2024). Deep learning approach to detect cyberbullying on twitter. Multimedia Tools and Applications, 1-24.

[12] Husain, F., & Uzuner, O. (2021). A survey of offensive language detection for the arabic language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(1), 1-44.

[13] Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN Computer Science, 3, 1-20.

[14] Al-Khasawneh, M. A., Faheem, M., Alarood, A. A., Habibullah, S., & Alsolami, E. (2024). Towards Multi-Modal Approach for Identification and Detection of Cyberbullying in Social Networks. IEEE Access.

[15] Asghar, M. Z., Habib, A., Habib, A., Khan, A., Ali, R., & Khattak, A. (2021). Exploring deep neural networks for rumor detection. Journal of Ambient Intelligence and Humanized Computing, 12, 4315-4333.

[16] Ullah, F., Ullah, S., Srivastava, G., & Lin, J. C. W. (2023). IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic. Digital Communications and Networks.

[17] Azzi, S. A., & Zribi, C. B. O. (2021, June). From machine learning to deep learning for detecting abusive messages in arabic social media: survey and challenges. In Intelligent Systems Design and Applications: 20th International Conference on Intelligent Systems Design and Applications (ISDA 2020) held December 12-15, 2020 (pp. 411-424). Cham: Springer International Publishing.

[18] Musleh, D., Rahman, A., Alkherallah, M. A., Al-Bohassan, M. K., Alawami, M. M., Alsebaa, H. A., ... & Alhaidari, F. (2024). A Machine Learning Approach to Cyberbullying Detection in Arabic Tweets. Computers, Materials & Continua, 80(1).

[19] Ghosal, S., & Jain, A. (2023). HateCircle and Unsupervised Hate Speech Detection Incorporating Emotion and Contextual Semantics. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 1-28.

[20] Yadav, D., Gupta, A., Asati, S., Choudhary, N., & Yadav, A. K. (2020, December). Age group prediction on textual data using sentiment analysis. In 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (pp. 61-65).

[21] Al-Wesabi, F. N., Obayya, M., Alabdan, R., Aljehane, N. O., Alazwari, S., Alruwaili, F. F., ... & Swathi, A. (2024). Automatic Recognition of Cyberbullying in the Web of Things and social media using Deep Learning Framework. IEEE Transactions on Big Data.

[22] Machová, K., Mach, M., & Porezaný, M. (2022). Deep Learning in the Detection of Disinformation about COVID-19 in Online Space. Sensors, 22(23), 9319.

[23] Aggarwal, P., & Mahajan, R. (2024). Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification. Journal of Information Systems and Informatics, 6(2), 607-623.

[24] Saha, S. K., Mim, A. A., Akter, S., Hosen, M. M., Shihab, A. H., & Mehedi, M. H. K. (2024, May). BengaliHateCB: A Hybrid Deep Learning Model to Identify Bengali Hate Speech Detection from Online Platform. In 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) (pp. 439-444). IEEE.

[25] Neog, M., & Baruah, N. (2024). A hybrid deep learning approach for Assamese toxic comment detection in social media. Procedia Computer Science, 235, 2297-2306.

[26] Singh, J. P., Kumar, A., Rana, N. P., & Dwivedi, Y. K. (2020). Attention-based LSTM network for rumor veracity estimation of tweets. Information Systems Frontiers, 1-16.

[27] Al-Ibrahim, R. M., Ali, M. Z., & Najadat, H. M. (2022). Detection of Hateful Social Media Content for Arabic Language. ACM Transactions on Asian and Low-Resource Language Information Processing.

[28] Badawi, S. (2024). Deep Learning-Based Cyberbullying Detection in Kurdish Language. The Computer Journal, bxae024.