Hate speech detection on social media using machine learning

Aigerim Toktarova[1,2,*,†], Aigerim Altayeva[3,†], Rustam Abdrakhmanov[4,†], Danyar Sultan[3,†] and Baktykul Jakhanova[5,†]

[1] *M. Auezov South Kazakhstan University, Shymkent, 160000, Kazakhstan*
[2] *Khoja Akhmet Yassawi International Kazakh – Turkish University, Turkistan, 161200, Kazakhstan*
[3] *International Information Technology University, 34/1 Manas St., Almaty, 050000, Kazakhstan*
[4] *International University of Tourism and Hospitality, Turkistan, 161200, Kazakhstan*
[5] *Asfendiyarov Kazakh National Medical University, 050000, Almaty, Kazakhstan*

Abstract

This research study uses a thorough analysis of numerous machine learning and deep learning techniques to address the crucial problem of cyberbullying detection in the social media arena. By carefully assessing these methods using industry-standard measures including F-measure, AUC-ROC, precision, accuracy, and recall, the study investigates how effective these approaches are. The outcomes show how well deep learning models—specifically, the bidirectional long-short-term memory (BiLSTM) architecture—perform, consistently surpassing other techniques on a range of categorization tasks. Confusion matrices and graphical depictions provide more insight into the model's functionality, showcasing the extraordinary capacity of the BiLSTM-based model to correctly identify and categorize instances of cyberbullying. The significance of sophisticated neural network architectures in identifying the intricacy of hateful and objectionable content on the internet underscores by these findings. This study offers insightful information for encouraging early detection and mitigation of cyberbullying, which in turn promotes secure and welcoming online communities. Future studies could look into real-time detection systems, hybrid techniques, or the integration of complementing elements to further develop and enhance cutting-edge technology in tackling this significant social issue.

Keywords

Machine learning, deep learning, hate speech, CNN, RNN, LSTM

1. Introduction

Cyberbullying is a contemporary epidemic rapidly infiltrating the online realm. The issues of psychological and physical violence, once confined to the social sphere, have now transitioned to the virtual realm. Initially, this type of persecution appears innocuous [1]. The distinctions between cyberbullying and conventional bullying stem from the characteristics of the Internet: anonymity, a vast audience, continuous accessibility for attacks, and the potential for deception. The issue of online violence is more significant, as it jeopardizes the psychological well-being of adolescents.

Hate speech transcends mere words. It can occur both in-person and online, manifesting through various mediums such as photos, cartoons, games, movies, objects, gestures, and symbols.

Cyberbullying is the transmission, publication, or spread of negative, harmful, false, or malicious content about another individual. It may entail revealing personal or secret information about another individual, leading to embarrassment or disgrace. Hate speech is distinguished from cyberbullying as it constitutes abuse aimed especially at an inherent, immutable characteristic of a certain group of individuals.

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.
toktar.aigerim@list.ru (A. Toktarova); rustam.abdyrakhmanov@gmail.com (R. Abdrakhmanov); aigerim.altayeva@gmail.com (A. Altayeva); daniyarsultan916@gmail.com (D. Sultan)
0000-0002-6265-9236 (A. Toktarova); 0000-0002-9802-9076 (A. Altayeva); 0000-0002-1611-1923 (D. Sultan)

Hate speech has spread as a result of this accessibility and openness, though. The identification and mitigation of hate speech has gained popularity, leading scholars to investigate several methodologies like deep learning (DL) and machine learning (ML).

Hate can lead to violence, social discord, and psychological injury to the targeted individuals or communities. It typified by offensive, damaging, or discriminating content [2]. Its existence on social media not only jeopardizes user safety but also damages the platforms' credibility and reputation.

Furthermore, these techniques demonstrate exceptional proficiency in managing the rapid and intricate nature of social media writing, which distinguished by its brevity, casual tone, and frequent spelling errors. They have the ability to efficiently handle and examine textual information from diverse origins, such as tweets, comments, and forum postings [6].

Recent research has demonstrated encouraging outcomes in the identification of hate speech through the utilization of machine learning (ML) and deep learning (DL) methodologies. These methods have successfully attained high levels of accuracy, precision, and recall rates, presenting a promising answer to the persistent problem of controlling hate speech on the internet [7-9].

2. Related works

Cyberbullying can transpire continuously, providing no opportunity for individuals to feel secure; messages and comments may arrive unexpectedly at any moment, resulting in significant psychological effects on adolescents. Furthermore, the anonymity of the Internet may prevent the teenager from identifying the one perpetrating the bullying, potentially exacerbating their dread. In contrast to physical violence, the repercussions of emotional abuse ultimately impact psychological well-being. Identifying a victim of emotional abuse is challenging. The automatic detection of cyberbullying can prevent it promptly [7-12].

As a result, social media corporations and policymakers have taken proactive steps to combat the dissemination of hate speech [2].

Machine learning techniques utilize natural language processing (NLP) technologies to automatically detect and categorize hate speech material [3]. Significantly, they do not depend exclusively on conventional methods that use keywords, which frequently fail to identify nuanced manifestations of hate speech [4].

An internet benefit of utilizing machine learning (ML) and deep learning (DL) techniques in hate speech identification is their capacity to adapt. Hate speech undergoes a process of development over time, assimilating novel derogatory terms, symbols, and phrases that may not be effectively identified by rigid rule-based systems. Machine learning (ML) and deep learning (DL) models have the ability to learn and adjust to new patterns in a continuous manner [5].

This work provides a comprehensive overview of several approaches, methodologies, and datasets employed in hate speech research, emphasizing their respective advantages and disadvantages [10]. Through a thorough examination of the complexities involved in identifying hate speech, our objective is to make a valuable contribution to the ongoing discussion on how to tackle this crucial problem. Additionally, we seek to offer valuable insights that can guide future research and advancements in this particular domain [11]. In addition, we explore the approaches and empirical findings, providing a comprehensive examination of the efficacy of machine learning (ML) and deep learning (DL) methods in identifying hate speech on social media platforms.

3. Problem statement

The issue of early detection of cyberbullying within the realm of social networking platforms may inherently differ from the challenge associated with classifying distinct manifestations of cyberbullying [12]. In the context delineated herein, we identify a cohort of social media interactions collectively denoted as "S." Consequently, it becomes plausible that a subset of these interactions may indeed represent instances of cyberbullying. The progression of such interactions on a given social network can be succinctly characterized using the following equation (1):

$$S=\left\{s_1,s_2,\ldots,s_{|S|}\right\} \tag{1}$$

Within the scope of this investigation, the variable "S" denotes the aggregate count of sessions, while the variable "i" signifies the present session under consideration. It is noteworthy that the order in which submissions occur during a given session can undergo modifications at distinct temporal junctures, influenced by an array of multifaceted determinants.

$$P_s=\left(\left\langle P_1^S,t_1^S\right\rangle,\left\langle P_2^S,t_2^S\right\rangle,\ldots,\left\langle P_n^S,t_n^S\right\rangle\right) \tag{2}$$

In the context of this study, the tuple denoted as "P" symbolizes the kth post within the context of the social network session, while "s" corresponds to the timestamp indicating the precise moment at which post P disseminated.

Simultaneously, a distinctive vector of attributes harnessed for the unequivocal identification of each individual post.

$$P_k^S=\left[f_{k_1}^S,f_{k_2}^S,\ldots,f_{k_n}^S\right],\ k\in[1,n] \tag{3}$$

Hence, the primary aim of this endeavor is to amass the requisite insights, enabling the formulation of a function denoted as "f," which possesses the capability to discern the association between a given text and the presence of hate speech.

4. Materials and methods

The prototype database for the aforementioned system was established by an examination of 215 English-language Twitter accounts, encompassing a total of 200,000 tweets, of which over 4,000 tweets were subjected to detailed analysis. Analysis revealed 583 English-language tweets exhibiting characteristics of the detrimental tactic known as “cyberbullying”. Electronic verbal bullying was predominantly observed in posts by adolescents aged 11-17 and young adults aged 18-35. Teenage cyberbullying typically involved groups, while electronic bullying among adolescents followed a "one bully – one victim" model.

Illustration of the developed model designed for the classification of hate speech instances is visually depicted in Figure 1. The model comprises distinct stages, which include preprocessing, feature extraction, classification, and evaluation. This section entails a comprehensive exploration of each of these stages, with a deliberate emphasis on the intricacies involved.

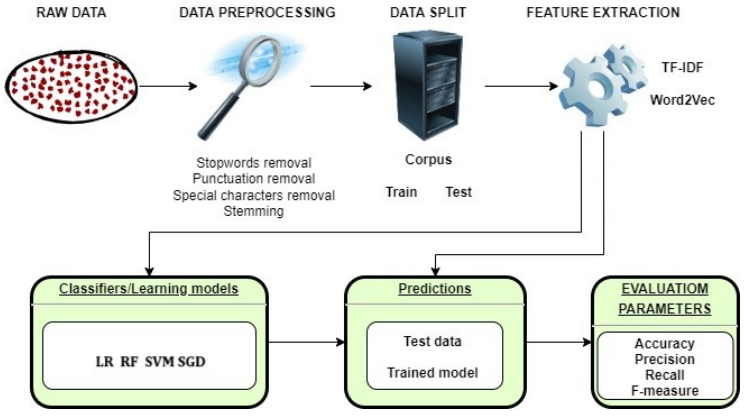


Figure 1: Proposed framework.

Word2Vec is a widely used feature representation technique in NLP [13]. It belongs to the family of word embedding methods that transform words into continuous vector representations in a high-dimensional space. Word2Vec captures semantic and contextual relationships between words by learning from large text corpora [14].

This technique assigns each word a vector in such a way that words with similar meanings are closer to each other in the vector space [15]. Word2Vec enhances NLP tasks by enabling models to understand the context and semantics of words, which is particularly valuable for applications like sentiment analysis, document clustering, and information retrieval [16]. By converting words into vectors, Word2Vec contributes to more effective and accurate text analysis and natural language understanding.

Bag of Words (BoW) The Bag of Words (BoW) model stands as a foundational technique in the field of natural language processing (NLP) and text mining, facilitating the transformation of textual information into numerical data, thereby enabling computational algorithms to process language. This model operates by constructing a vocabulary of unique words from a corpus and then converting text documents into vectors, where each vector element represents the frequency of a particular word in the document [17]. Despite its simplicity, the BoW model has been instrumental in numerous NLP applications, including document classification, sentiment analysis, and topic modeling [18]. However, it is not without limitations; notably, the model's disregard for word order and context can lead to a loss of semantic meaning [19]. Furthermore, the high dimensionality of the resulting vectors, especially with large vocabularies, poses challenges for computational efficiency [20]. Nonetheless, the BoW model's ease of implementation and interpretability continues to make it a valuable tool in the initial stages of text analysis projects.

4.1. Machine learning for hate speech detection

In the realm of hate speech detection within social networks, various machine learning models have been employed to address the complex task of distinguishing between offensive language and benign content. Each of these models offers distinct advantages and trade-offs, making them suitable for different aspects of the problem [21].

Decision Trees: Decision tree models provide a structured representation of decision-making processes. They are interpretable and can be valuable for identifying explicit patterns and features indicative of hate speech [22]. However, they may struggle to capture more subtle contextual cues.

Logistic Regression allows for the estimation of probabilities and predictions in situations where the outcome is categorical, such as spam email detection or medical diagnosis. Logistic Regression's simplicity and interpretability make it a valuable tool in various fields, including data analysis, healthcare, and marketing.

Naive Bayes: Naive Bayes models are based on probabilistic principles. They are especially adept at handling text data due to their independence assumptions. Naive Bayes models can efficiently process large volumes of text and can adapt well to the high perplexity of social media content.

K-Nearest Neighbors [24] can be useful for identifying similar posts with similar hate speech content, yet it may struggle with high-dimensional data.

Support Vector Machines (SVM) is robust against overfitting and can handle high-dimensional feature spaces [25]. SVMs can be effective in capturing complex decision boundaries in hate speech detection.

The choice of machine learning model should consider the specific characteristics of the hate speech detection problem, such as the prevalence of subtle hate speech, the dimensionality of the text data, and the need for interpretability. Often, a combination of these models in ensemble techniques or hybrid approaches is employed to harness their individual strengths and mitigate their limitations, ultimately improving the overall performance of hate speech detection systems.

4.2. Deep learning for hate speech detection

In the domain of hate speech detection in social networks, deep learning models have emerged as potent tools due to their capacity to capture intricate linguistic nuances and contextual dependencies within textual data. Three prominent deep learning architectures, Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Bidirectional LSTMs (BiLSTMs), have been widely employed to address the complexities inherent in this task [26].

Convolutional Neural Networks (CNNs): CNNs, initially designed for image processing, have been adapted for text analysis (Figure 2). They employ convolutional layers to detect local patterns and hierarchies of features within text. In hate speech detection, CNNs can effectively identify significant textual structures and are particularly adept at capturing short-range dependencies such as n-grams and patterns indicative of hate speech expressions.

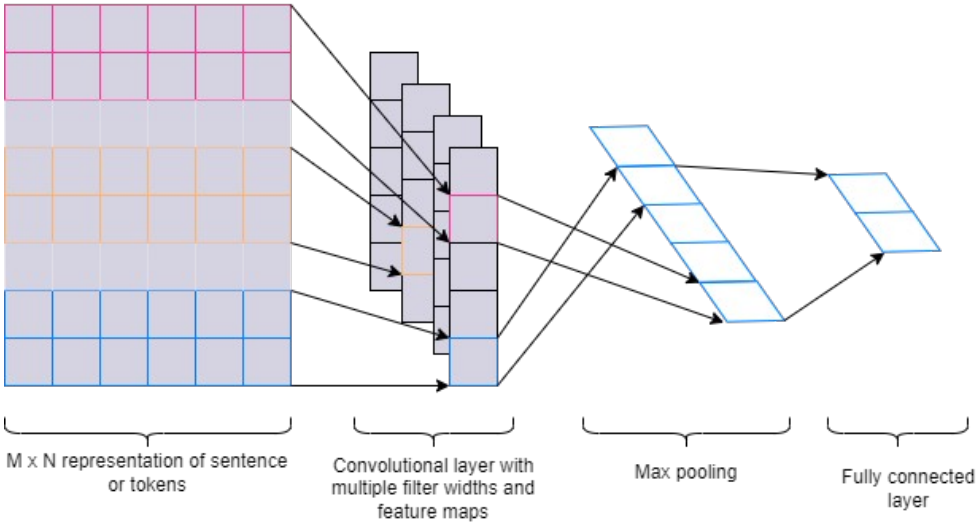


Figure 2: CNN for Hate Speech Detection.

Long Short-Term Memory networks (LSTMs): LSTMs are recurrent neural networks (RNNs) designed to capture sequential information over longer distances (Figure 3). They excel in modeling dependencies over time and have proven valuable in understanding the temporal aspects of hate speech evolution. LSTMs can detect contextually relevant information and provide a dynamic understanding of text.

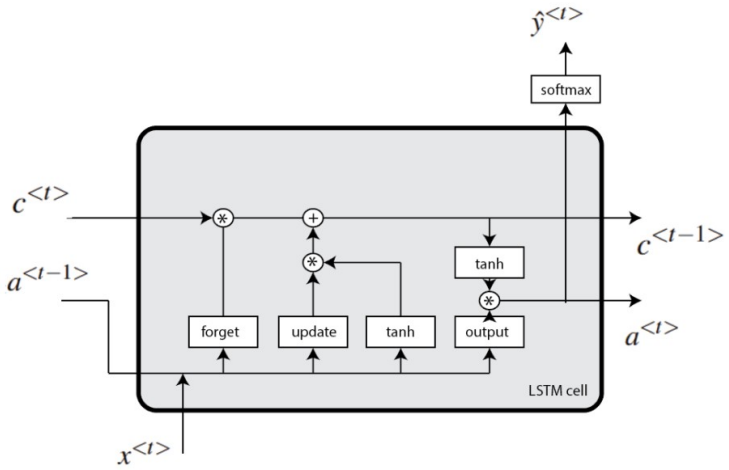


Figure 3: LSTM for Hate Speech Detection.

Bidirectional LSTMs (BiLSTMs): BiLSTMs extend the LSTM architecture by processing sequences in both forward and backward directions, allowing them to capture bidirectional dependencies (Figure

4). In hate speech detection, BiLSTMs are particularly effective in understanding contextual nuances and capturing relationships between words in both preceding and succeeding contexts.

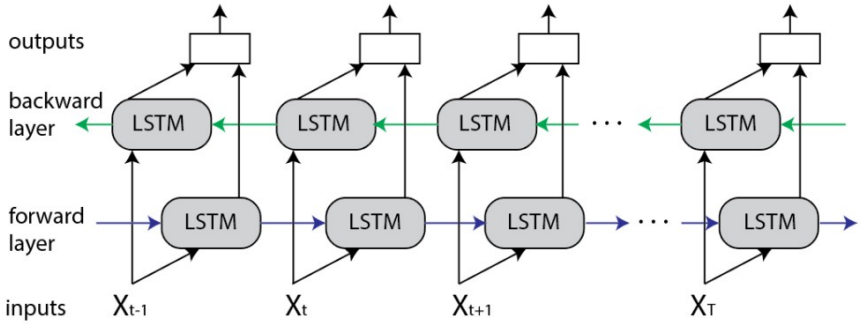


Figure 4: BiLSTM for Hate Speech Detection.

5. Experiment results

5.1. Evaluation parameters

In the context of hate speech detection within social networks, evaluating the performance of machine learning and deep learning models is crucial for assessing their effectiveness in mitigating the spread of offensive content. Several evaluation parameters commonly employed to gauge the performance of such models comprehensively.

$$accuracy = \frac{TP + TN}{P + N} \tag{6}$$

$$preision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

In the context of hate speech detection, a balance between precision and recall is often sought, as falsely classifying non-hate speech as hate speech (false positives) or failing to detect hate speech (false negatives) can have significant real-world consequences. Researchers and practitioners may also consider domain-specific evaluation metrics and adjust the thresholds based on the desired trade-offs between precision and recall. Robust evaluation methodologies are essential to developing and deploying effective hate speech detection systems that contribute to fostering safer and more inclusive online communities.

6. Results

Evaluation metrics are essential for quantifying the effectiveness of algorithms in classifying instances within the cyberbullying classification dataset.

Confusion matrices, as depicted in Figure 5, play a pivotal role in visualizing the outcomes of these classification techniques. They provide a clear representation of the actual distribution of classification results across different classes.

By utilizing confusion matrices, researchers can discern the true positive, true negative, false positive, and false negative predictions, enabling a comprehensive understanding of the model's performance in distinguishing between cyberbullying and non-cyberbullying instances. These evaluations are essential for refining and optimizing cyberbullying detection algorithms to enhance their accuracy and reliability in addressing the critical issue of online harassment and bullying.

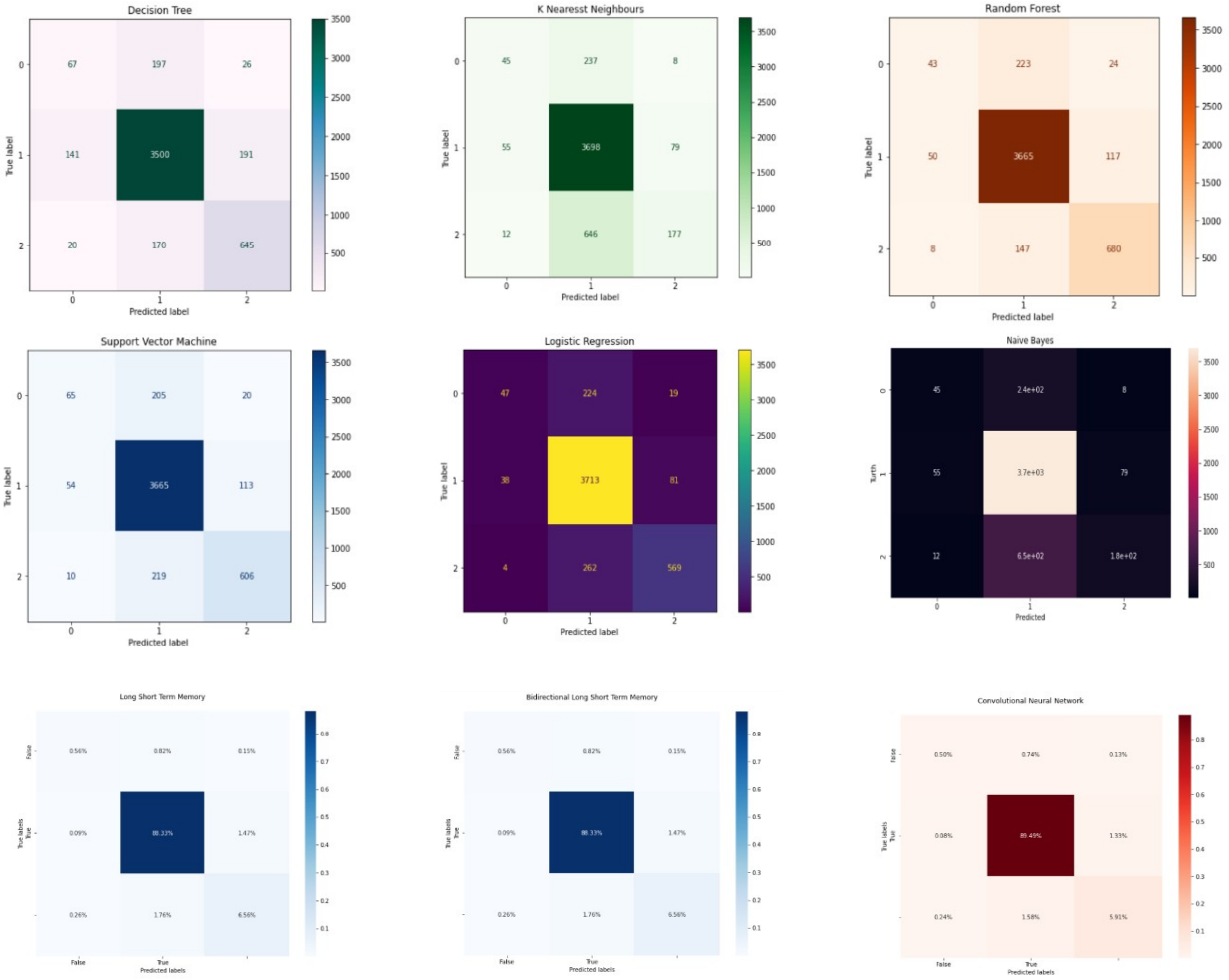
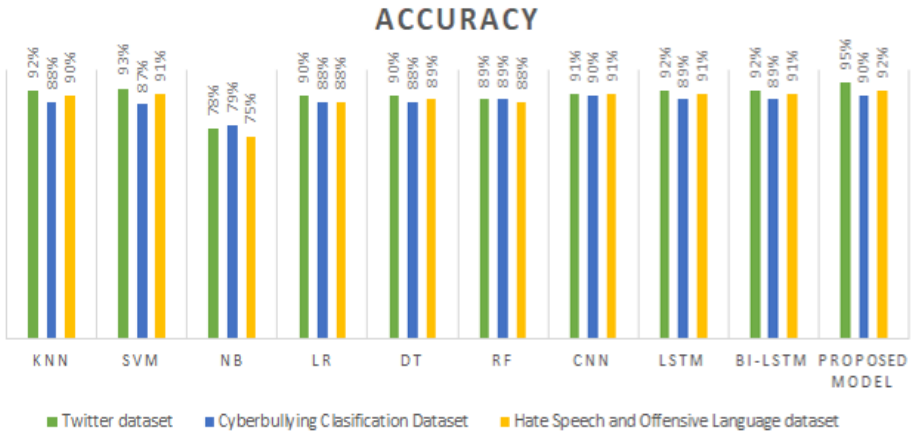
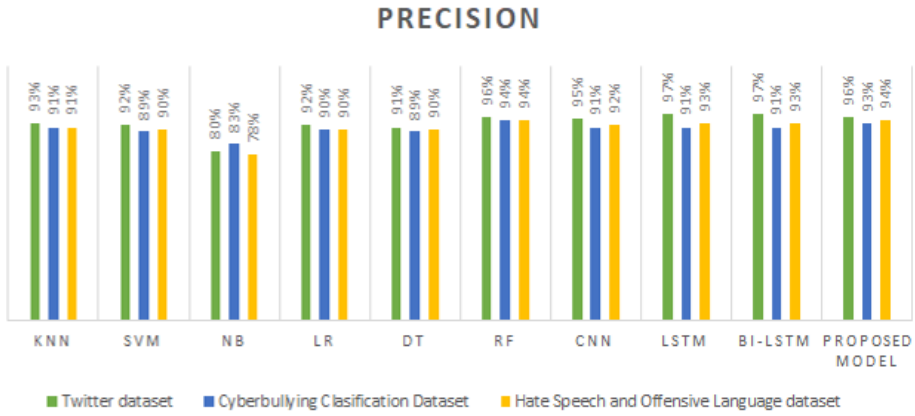
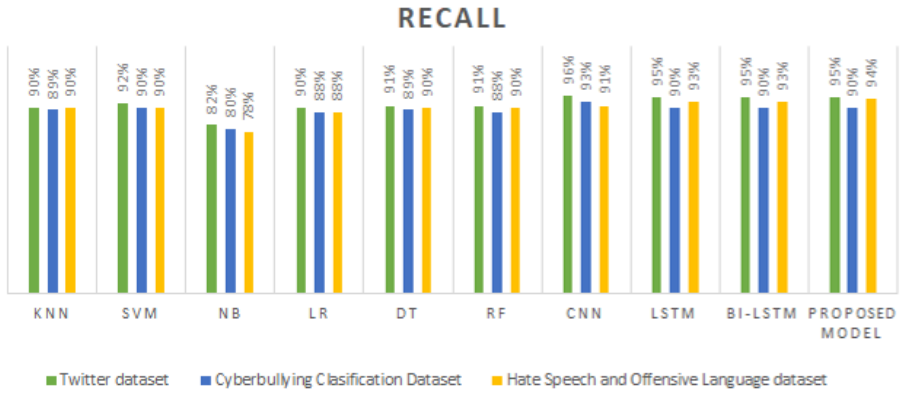


Figure 5: Confusion matrices results in hate speech detection.

Figure 6 presents a comparative analysis between the proposed model and a range of other machine learning and deep learning models employed in this study. The performance evaluation in each classification scenario is conducted by computing the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), encompassing all extracted features. This approach allows for a comprehensive assessment of the discriminatory power and effectiveness of the suggested model in comparison to alternative methodologies, thereby providing valuable insights into its performance across different classification tasks.







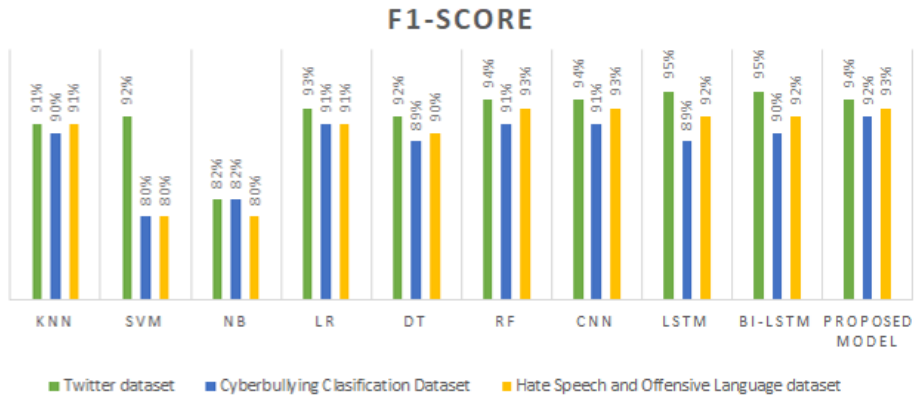


Figure 6: Results in Hate Speech Detection

These findings underscore the efficacy and robustness of the BiLSTM-based model in effectively discriminating and classifying the target classes, further substantiating the merit of deep learning paradigms in the context of the study.

7. Conclusion

In conclusion, this research paper has delved into the critical realm of cyberbullying detection within the context of social networks. Through a comprehensive exploration of various machine learning and deep learning methodologies, coupled with meticulous evaluation using metrics such as Accuracy, Precision, Recall, F-measure, and AUC-ROC, we have endeavored to shed light on the effectiveness of these techniques in addressing the m ultifaceted challenge of identifying instances of cyberbullying.

Our findings underscore the pivotal role that deep learning models, particularly the Bidirectional Long Short-Term Memory (BiLSTM) architecture, play in enhancing the discriminatory power and accuracy of cyberbullying detection systems. The consistent superiority of the BiLSTM-based model across various classification tasks reaffirms the potential of advanced neural network structures in capturing the intricacies of online hate speech and offensive content. Moreover, the utilization of confusion matrices and visualizations has allowed for a nuanced understanding of model performance. This research contributes valuable insights into the ongoing efforts to create safer and more inclusive online spaces, where the early identification and mitigation of cyberbullying are paramount. Future research endeavors may explore hybrid approaches, leverage additional features, or delve into real-time cyberbullying detection systems to further refine and enhance the state-of-the-art in this vital domain

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] T. Alsubait and D. Alfageh, “Comparison of machine learning techniques for cyberbullying detection on youtube arabic comments,” International Journal of Computer Science and Network Security, vol. 21, no. 1, pp. 1–5, 2021.

[2] A. Dewani, M. Memon and S Bhatti, “Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for roman urdu data,” Journal of Big Data, vol. 8, no. 1, pp. 1–20, 2021.

[3] D. Hall, Y. Silva, Y. 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, 2021.

[4] K. Arce-Ruelas, “Automatic cyberbullying detection: A Mexican case in high school and Higher Education Students,” IEEE Latin America Transactions, vol. 20, no. 5, pp. 770–779, 2022.

[5] T. Ahmed, M. Rahman, S. Nur, A. Islam and D. Das, “Natural language processing and machine learning based cyberbullying detection for Bangla and romanized bangla texts,” TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 20, no. 1 pp. 89–97, 2021.

[6] Toktarova, A., Sultan, D., & Azhibekova, Z. (2024, May). Review of Machine Learning Models in Cyberbullying Detection Problem. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 233-238). IEEE.

[7] A. Al-Marghilani, "Artificial intelligence-enabled cyberbullying-free online social networks in smart cities," International Journal of Computational Intelligence Systems, vol. 15, no. 1, pp. 1–13, 2022.
[8] C. Theng, N. Othman, R. Abdullah, S. Anawar, Z. Ayop et al., "Cyberbullying detection in twitter using sentiment analysis," International Journal of Computer Science & Network Security, vol. 21, no. 11, pp. 1-10, 2021.
[9] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. Choi et al., "Aggression detection through deep neural model on twitter," Future Generation Computer Systems, vol. 114, no. 1, pp. 120–129, 2021.
[10] E. Sarac Essiz and M. Oturakci, "Artificial bee colony–based feature selection algorithm for cyberbullying," The Computer Journal, vol. 64, no. 3, pp. 305–313, 2021.
[11] C. E. Gomez, M. O. Sztainberg and R. E. Trana, "Curating cyberbullying datasets: a human-AI collaborative approach," International journal of bullying prevention, vol. 4, no. 1, pp. 35-46, 2022.
[12] S. Salawu, J. Lumsden and Y. He, "A mobile-based system for preventing online abuse and cyberbullying," International Journal of Bullying Prevention, vol. 4, no. 1, pp. 66–88, 2022.
[13] M. Mladenović, V. Ošmjanski and S. V. Stanković, "Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges," ACM Computing Surveys (CSUR), vol. 54, no.1, pp. 1–42, 2021.
[14] S. R. Sangwan and M. P. S. Bhatia, "Denigrate comment detection in low-resource Hindi language using attention-based residual networks," Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 1, pp. 1–14, 2021.
[15] T. T. Aurpa, R. Sadik and M. S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models," Social Network Analysis and Mining, vol. 12, no.1, pp. 1–14, 2022.
[16] R. Yan, Y. Li, D. Li, Y. Wang, Y. Zhu et al., "A Stochastic Algorithm Based on Reverse Sampling Technique to Fight Against the Cyberbullying," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 15, no. 4, pp. 1–22, 2021.
[17] C. J. Yin, Z. Ayop, S. Anawar, N. F. Othman and N. M. Zainudin, "Slangs and Short forms of Malay Twitter Sentiment Analysis using Supervised Machine Learning," International Journal of Computer Science & Network Security, vol. 21, no. 11, pp. 294–300, 2021.
[18] G. Jacobs, C. Van Hee and V. Hoste, "Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?," Natural Language Engineering, vol. 28, no. 2, pp. 141–166, 2022.
[19] A. Jevremovic, M. Veinovic, M. Cabarkapa, M. Krstic, I. Chorbev et al., "Keeping Children Safe Online With Limited Resources: Analyzing What is Seen and Heard," IEEE Access, vol. 9, no. 1, pp. 132723–132732, 2021.
[20] K. Kumari, J. P. Singh, Y. K. Dwivedi and N. P. Rana, "Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization," Future Generation Computer Systems, vol. 118, no. 1, pp. 187–197, 2021.
[21] A. M. Abbas, "Social network analysis using deep learning: applications and schemes," Social Network Analysis and Mining, vol.11, no. 1, pp. 1–21, 2021.
[22] Toktarova, A., et al. "Automatic offensive language detection in online user generated contents." Journal of Theoretical and Applied Information Technology 99.9 (2021): 2054-2067.
[23] S. Mohammed, W. C. Fang, A. E. Hassanien and T. H. Kim, "Advanced Data Mining Tools and Methods for Social Computing," The Computer Journal, vol. 64, no. 3, pp. 281–285, 2021.
[24] Baimakhanova, Aigerim, et al. "Automatic Classification of Scanned Electronic University Documents using Deep Neural Networks with Conv2D Layers." International Journal of Advanced Computer Science and Applications 14.5 (2023).
[25] Toktarova, A., et al. "Automated Hate Speech Classification using Emotion Analysis in Social Media User Generated Texts." J. Theor. Appl. Inf. Technol 100 (2022): 6621-6634.

[26] Makhanova, Zlikha, et al. "A Deep Residual Network Designed for Detecting Cracks in Buildings of Historical Significance." International Journal of Advanced Computer Science & Applications 15.5 (2024).