Offensive Language Detection on Social Media using Machine Learning

Rustam Abdrakhmanov[1], Serik Muktarovich Kenesbayev[2], Kamalbek Berkimbayev[3],
Gumyrbek Toikenov[4], Elmira Abdrashova[5], Oichagul Alchinbayeva[6], Aizhan Ydyrys[7]
International University of Tourism and Hospitality, Turkistan, Kazakhstan[1]
Kazakh National Women’s Teacher Training University, Almaty, Kazakhstan[2, 4]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[3]
M. Auezov South Kazakhstan University, Shymkent, Kazakhstan[5, 6]
International Information Technology University, Almaty, Kazakhstan[7]

***Abstract*—This research paper addresses the critical issue of cyberbullying detection within the realm of social networks, employing a comprehensive examination of various machine learning and deep learning techniques. The study investigates the performance of these methodologies through rigorous evaluation using standard metrics, including Accuracy, Precision, Recall, F-measure, and AUC-ROC. The findings highlight the notable efficacy of deep learning models, particularly the Bidirectional Long Short-Term Memory (BiLSTM) architecture, in consistently outperforming alternative methods across diverse classification tasks. Confusion matrices and graphical representations further elucidate model performance, emphasizing the BiLSTM-based model's remarkable capacity to discern and classify cyberbullying instances accurately. These results underscore the significance of advanced neural network structures in capturing the complexities of online hate speech and offensive content. This research contributes valuable insights toward fostering safer and more inclusive online communities by facilitating early identification and mitigation of cyberbullying. Future investigations may explore hybrid approaches, additional feature integration, or real-time detection systems to further refine and advance the state-of-the-art in addressing this critical societal concern.**

***Keywords*—Machine learning; deep learning; hate speech; CNN; RNN; LSTM**

I. Introduction

The advent of social media has revolutionized the way individuals communicate, providing platforms that facilitate rapid information dissemination and interaction across global communities. While these platforms have empowered users to share information and foster connections, they have also become breeding grounds for various forms of online abuse, including hate speech. Hate speech encompasses any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. It poses severe risks to community harmony, individual safety, and democratic discourse [1]. Consequently, the detection and mitigation of hate speech on social media is of paramount importance for maintaining social cohesion and protecting vulnerable groups.

The challenge of combating hate speech on social media is amplified by the vast amount of data generated daily and the fluid nature of online communication. Traditional content moderation methods, which rely heavily on human moderators to review content, are not scalable to the volumes of data produced on platforms such as Facebook, Twitter, and Instagram. Furthermore, manual moderation is prone to inconsistencies and errors, making it an inefficient solution in the dynamic and diverse environment of social media [2]. As a result, there has been a significant shift toward automated systems, particularly those utilizing machine learning (ML) and deep learning (DL), to address the complexities associated with identifying and managing hate speech [3].

Machine learning offers a promising approach to automate the detection of hate speech by learning from large datasets of labeled examples. It uses natural language processing (NLP) to parse and understand the textual content of social media posts, learning to differentiate between harmful and harmless expressions based on training data [4]. Unlike rule-based systems, which fail to adapt to the evolving language of online communities, ML algorithms can update their knowledge as new data becomes available, thereby adapting to changes in the lexicon used in hate speech [5].

Deep learning, a subset of ML characterized by models that learn through layers of neural networks, has shown exceptional capability in handling the intricacies and subtleties of human language. DL models, particularly those based on recent advancements such as transformer architectures, have demonstrated high accuracy in contextual understanding and sentiment analysis [6]. These models are particularly adept at capturing the contextual nuances that differentiate hostile or derogatory speech from benign usage of potentially sensitive words [7].

The application of ML and DL in detecting hate speech is not without challenges. One significant issue is the balance between accuracy and the rate of false positives—where benign content is incorrectly flagged as hate speech. High rates of false positives can lead to unnecessary censorship and could impact user engagement and trust in social media platforms [8]. Another challenge is the development of models that can operate across different languages and cultural contexts, as hate speech often involves cultural references and idioms that are not universally recognized [9].

Recent studies have applied various ML and DL models to address these challenges, employing sophisticated algorithms and a range of feature extraction techniques to improve detection accuracy [10]. Furthermore, researchers have explored the use of ensemble methods, where multiple models are used in conjunction to make final predictions, thereby reducing the likelihood of errors that might occur when relying on a single model [11].

The continuous evolution of social media necessitates ongoing research and development to refine these technological approaches. By enhancing the accuracy and adaptability of ML and DL models, researchers aim to contribute effectively to the global effort to mitigate hate speech on social media. This will not only protect individuals from the harms associated with such speech but also preserve the integrity of digital platforms as spaces for free but respectful expression.

In this paper, we delve into the methodologies, experimental results, and implications of using ML and DL for hate speech detection, providing a comprehensive overview of the current landscape and future directions in this critical area of research. Through detailed analysis and discussion, we aim to further the understanding of technological capabilities and limitations in combating hate speech and to explore potential pathways for innovative solutions.

II. 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 Eq. (1):

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

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 was disseminated.

Simultaneously, a distinctive vector of attributes is 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.

III. Materials and Methods

Illustration of the developed model designed for the classification of hate speech instances is visually depicted in Fig. 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.

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].

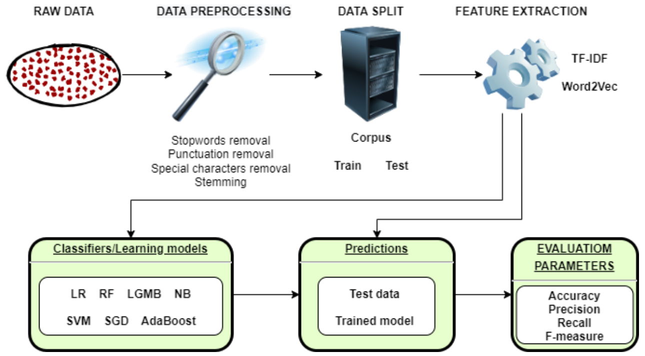

Fig. 1. Proposed framework.

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.

$$w_{i,j} = TF_{i,j} \times \log\left(\frac{N}{DF_i}\right) \tag{4}$$

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. The overarching objective is to enhance the likelihood of success under the prevailing circumstances:

$$\arg\max_{\theta} \prod_{w \in T} \left[\prod_{c \in C} p(c \mid w; \theta) \right] \tag{5}$$

A. 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 [23].

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.

B. 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 [26-27].

Convolutional Neural Networks (CNNs): CNNs, initially designed for image processing, have been adapted for text analysis (see Fig. 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.

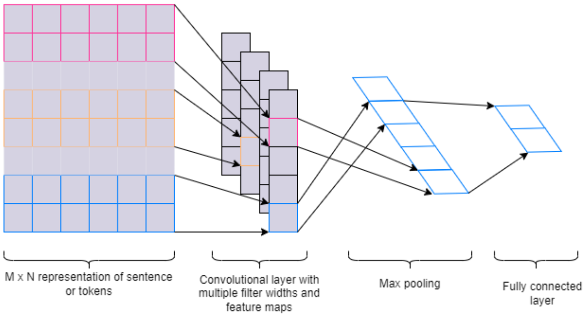
M x N representation of sentence or tokens
Convolutional layer with multiple filter widths and feature maps
Max pooling
Fully connected layer

Fig. 2. CNN for hate speech detection.

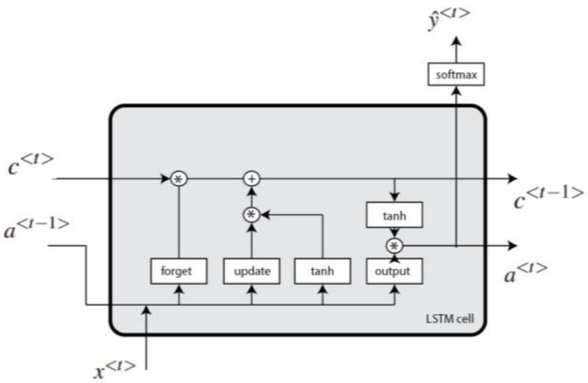

Fig. 3. LSTM for hate speech detection.

Long Short-Term Memory networks (LSTMs) represent a specialized category of recurrent neural networks (RNNs) engineered to process and retain information across extended temporal intervals (see Fig. 3). These networks are particularly adept at modeling long-distance dependencies within sequential data, making them highly effective for tasks that require an understanding of temporal dynamics, such as the evolution of hate speech. LSTMs maintain a structured memory cell that captures relevant context over time, enabling them to discern and retain contextually significant information amidst a flow of input data. This capability allows LSTMs to offer a nuanced and dynamic understanding of text, which is crucial for effectively detecting and interpreting the progressive nature of communicative patterns, including the subtleties and shifts in hate speech across social media platforms.

Bidirectional LSTMs (BiLSTMs): Bidirectional Long Short-Term Memory networks (BiLSTMs) are an advanced variant of the traditional Long Short-Term Memory (LSTM) networks, designed to enhance the model's context capturing capabilities by processing data in both forward and backward directions. Unlike standard LSTMs that propagate information through time in a single direction, BiLSTMs consist of two separate layers that operate synchronously: one processes the input sequence from start to end, while the other processes it from end to start. This dual-pathway architecture allows BiLSTMs to gather contextual information from both past and future states, providing a comprehensive understanding of the sequence at any given point. This feature is particularly beneficial for complex sequence prediction tasks where context from both directions is crucial for accurate interpretation. Applications in natural language processing, such as sentiment analysis or text classification, have demonstrated the effectiveness of BiLSTMs in capturing nuanced linguistic patterns that a unidirectional approach might miss.

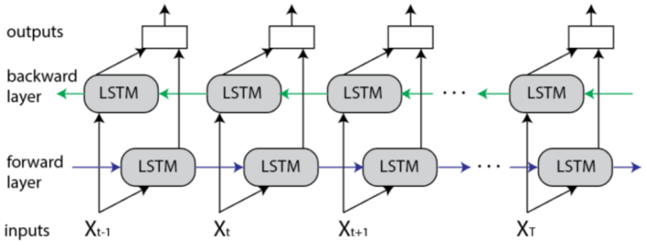

Fig. 4. BiLSTM for hate speech detection.

BiLSTMs extend the LSTM architecture by processing sequences in both forward and backward directions, allowing them to capture bidirectional dependencies (see Fig. 4). In hate speech detection, BiLSTMs are particularly effective in understanding contextual nuances and capturing relationships between words in both preceding and succeeding contexts.

IV. Experimental Results

A. 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 are 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.

B. Results

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

Confusion matrices, as depicted in Fig. 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.

Fig. 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.

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.

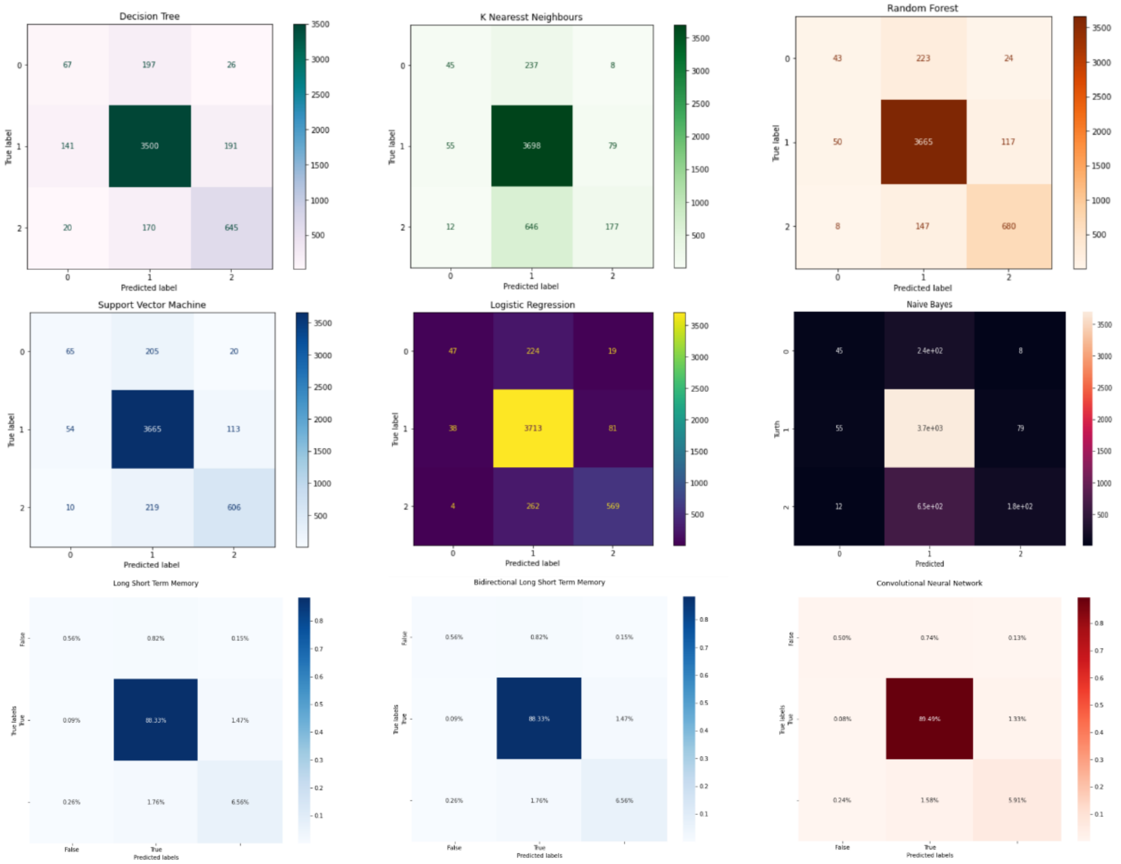

Fig. 5. Confusion matrices results in hate speech detection.

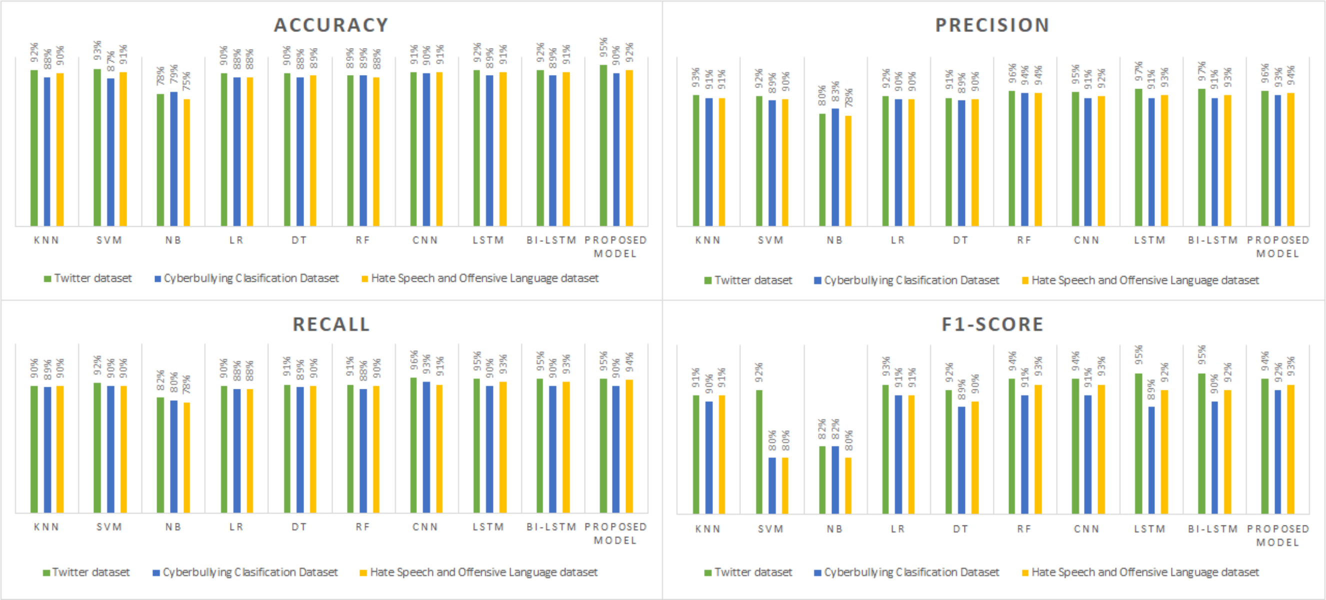

Fig. 6. Results in hate speech detection.

V. Discussion

The integration of machine learning (ML) and deep learning (DL) methodologies into the detection of hate speech on social media platforms marks a pivotal advancement in computational linguistics and artificial intelligence. While our results, as well as those reported in the literature, demonstrate high efficacy in detecting hate speech, this discussion aims to dissect the broader implications, inherent challenges, and the road ahead for these technologies in practical applications.

One of the primary strengths of ML and DL models, as highlighted in our findings, is their ability to adapt to the evolving nature of language used in hate speech. This adaptability is critical given the dynamic and ever-changing lexicon that characterizes online hate speech [26]. However, the dependency on large, annotated datasets for training these models raises significant concerns regarding the representativeness and bias of the data [27]. Models trained on datasets that are not representative of the diverse forms of speech and languages globally may exhibit biased or underperforming results when deployed in different demographic or linguistic contexts [28].

Furthermore, the ethical implications of deploying automated systems for hate speech detection cannot be overlooked. Concerns about privacy, freedom of speech, and the potential for over-surveillance are paramount [29]. The risk of false positives—where benign content is mistakenly classified as hate speech—poses a threat to free expression and could result in unwarranted censorship [30]. The balance between effectively moderating content and safeguarding user rights is a delicate one that requires ongoing scrutiny and adjustment of algorithms [31].

Another critical aspect is the scalability of these technologies. As social media platforms continue to grow, the volume of content that needs to be monitored for hate speech expands exponentially. While ML and DL models offer scalability, their computational demands and the need for continuous retraining with new data pose logistical and financial challenges [32]. The integration of these systems into existing social media infrastructure must be managed with careful consideration of these factors [33].

The transparency and interpretability of ML and DL models also present significant challenges. The often "black box" nature of these models, particularly those involving complex deep learning architectures, makes it difficult for practitioners to understand and explain how decisions are made [34]. This lack of transparency can be problematic, especially when decisions have significant consequences for users [35]. Efforts to develop more interpretable models are crucial to ensure that stakeholders can review and audit the processes involved in hate speech detection [36].

The international context further complicates the deployment of automated hate speech detection systems. Legal and cultural differences in the definition and perception of hate speech across countries necessitate a customizable approach to algorithm development [37]. Additionally, the multilingual nature of global platforms requires that models be effective across different languages, which is currently a significant limitation for many existing systems [38].

Technological responses to hate speech must also consider the human aspect. The integration of human moderators in the loop is essential not only for the training and fine-tuning of ML and DL models but also for handling cases where the algorithm's decision is unclear or disputed [39]. This hybrid approach could help mitigate some of the challenges associated with fully automated systems, offering a balance between human intuition and algorithmic efficiency [40].

In conclusion, while ML and DL methodologies have shown promise in addressing the scourge of hate speech on social media, their deployment is not without challenges. Issues of bias, ethical implications, scalability, transparency, and the need for international and multilingual capabilities must be addressed. Future research should focus on enhancing the representativeness of training datasets, developing interpretable models, and creating robust systems that can adapt to legal and cultural variations globally [41-43]. As this field evolves, it is imperative that technological advancements go hand in hand with ethical considerations to ensure that the fight against hate speech does not inadvertently harm the very individuals and freedoms it seeks to protect.

VI. 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 multifaceted 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.

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] D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova et al., "A review of machine learning techniques in cyberbullying detection," Computers, Materials & Continua, vol. 74, no.3, pp. 5625–5640, 2023.

[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] Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M., & Omarov, B. (2021, October). Chatbots and Conversational Agents in Mental Health: A Literature Review. In 2021 21st International Conference on Control, Automation and Systems (ICCAS) (pp. 353-358). IEEE.

[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] Saumya, S., Kumar, A., & Singh, J. P. (2024). Filtering offensive language from multilingual social media contents: A deep learning approach. Engineering Applications of Artificial Intelligence, 133, 108159.

[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] Altayeva, A., Omarov, B., & Im Cho, Y. (2017, December). Multi-objective optimization for smart building energy and comfort management as a case study of smart city platform. In 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 627-628). IEEE.

[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] L. Jayakumar, R. Jothi Chitra, J. Sivasankari, S. Vidhya, Laura Alimzhanova, Gulnur Kazbekova, Bakhytzhan Kulambayev, Alma Kostangeldinova, S. Devi, Dawit Mamiru Teressa, "QoS Analysis for Cloud-Based IoT Data Using Multicriteria-Based Optimization Approach", Computational Intelligence and Neuroscience, vol. 2022, Article ID 7255913, 12 pages, 2022. https://doi.org/10.1155/2022/7255913.

[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] Kulambayev, B., Nurlybek, M., Astaubayeva, G., Tleuberdiyeva, G., Zholdasbayev, S., & Tolep, A. (2023). Real-Time Road Surface Damage Detection Framework based on Mask R-CNN Model. International Journal of Advanced Computer Science and Applications, 14(9).

[22] S. Gupta, N. Mohan, P. Nayak, K. C. Nagaraju and M. Karanam, “Deep vision-based surveillance system to prevent train–elephant collisions,” Soft Computing, vol. 26, no. 8, pp. 4005–4018, 2022.

[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] B. Thuraisingham, “Trustworthy Machine Learning,” IEEE Intelligent Systems, vol. 37, no.1, pp. 21–24, 2022.

[25] V. Rupapara, F. Rustam, H. Shahzad, A. Mehmood, I. Ashraf et al., “Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model,” IEEE Access, vol. 9, no. 1, pp. 78621–78634, 2021.

[26] Saumya, S., Kumar, A., & Singh, J. P. (2024). Filtering offensive language from multilingual social media contents: A deep learning approach. Engineering Applications of Artificial Intelligence, 133, 108159.

[27] Yuan, L., Wang, T., Ferraro, G., Suominen, H., & Rizoiu, M. A. (2023). Transfer learning for hate speech detection in social media. Journal of Computational Social Science, 6(2), 1081-1101.

[28] Khan, A. A., Iqbal, M. H., Nisar, S., Ahmad, A., & Iqbal, W. (2023). Offensive language detection for low resource language using deep sequence model. IEEE Transactions on Computational Social Systems.

[29] Balakrishnan, V., Govindan, V., & Govaichelvan, K. N. (2023). Tamil Offensive Language Detection: Supervised versus Unsupervised Learning Approaches. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 1-14.

[30] Saumya, S., Kumar, A., & Singh, J. P. (2021, April). Offensive language identification in Dravidian code mixed social media text. In Proceedings of the first workshop on speech and language technologies for Dravidian languages (pp. 36-45).

[31] Khairy, M., Mahmoud, T. M., & Abd-El-Hafeez, T. (2021). Automatic detection of cyberbullying and abusive language in Arabic content on social networks: a survey. Procedia Computer Science, 189, 156-166.

[32] Roy, P. K., Bhawal, S., & Subalalitha, C. N. (2022). Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. Computer Speech & Language, 75, 101386.

[33] Fha, S., Sharma, U., & Naleer, H. M. M. (2023). Development of an efficient method to detect mixed social media data with tamil-english code using machine learning techniques. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(2), 1-19.

[34] Omarov, B., & Altayeva, A. (2018, January). Towards intelligent IoT smart city platform based on OneM2M guideline: smart grid case study. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 701-704). IEEE.

[35] Pillai, A. R., & Arun, B. (2024). A feature fusion and detection approach using deep learning for sentimental analysis and offensive text detection from code-mix Malayalam language. Biomedical Signal Processing and Control, 89, 105763.

[36] Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., & Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. Theoretical Computer Science, 943, 203-218.

[37] Sreelakshmi, K., Premjith, B., Chakravarthi, B. R., & Soman, K. P. (2024). Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages using Cost-Sensitive Learning Approach. IEEE Access.

[38] Khan, A., Ahmed, A., Jan, S., Bilal, M., & Zuhairi, M. F. (2024). Abusive Language Detection in Urdu Text: Leveraging Deep Learning and Attention Mechanism. IEEE Access.

[39] Omarov, B., Batyrbekov, A., Suliman, A., Omarov, B., Sabdenbekov, Y., & Aknazarov, S. (2020, November). Electronic stethoscope for detecting heart abnormalities in athletes. In 2020 21st International Arab Conference on Information Technology (ACIT) (pp. 1-5). IEEE.

[40] Quadri, S. M. K. (2024). Hate Speech Detection on Social Media using Machine Learning and Deep Learning: A review. Grenze International Journal of Engineering & Technology (GIJET), 10(1).

[41] Akhter, A., Acharjee, U. K., Talukder, M. A., Islam, M. M., & Uddin, M. A. (2023). A robust hybrid machine learning model for Bengali cyber bullying detection in social media. Natural Language Processing Journal, 4, 100027.

[42] Shah, S. M. A., & Singh, S. (2022, September). Hate Speech and Offensive Language Detection in Twitter Data Using Machine Learning Classifiers. In International Conference on Innovations in Computer Science and Engineering (pp. 221-237). Singapore: Springer Nature Singapore.

[43] Mohapatra, S. K., Prasad, S., Bebarta, D. K., Das, T. K., Srinivasan, K., & Hu, Y. C. (2021). Automatic hate speech detection in english-odia code mixed social media data using machine learning techniques. Applied Sciences, 11(18), 8575.