Cyberbullying Detection on Social Networks Using a Hybrid Deep Learning Architecture Based on Convolutional and Recurrent Models

Aigerim Altayeva[1], Rustam Abdrakhmanov[2], Aigerim Toktarova[3], Abdimukhan Tolep[4]
International Information Technology University, Almaty, Kazakhstan[1]
International University of Tourism and Hospitality, Turkistan, Kazakhstan[2]
M. Auezov South Kazakhstan University, Shymkent, Kazakhstan[3]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[4]

***Abstract*—This research paper explores the development and efficacy of a hybrid deep learning architecture for cyberbullying detection on social media platforms, integrating Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. By leveraging the strengths of both CNNs and LSTMs, the model aims to enhance the accuracy and sensitivity of detecting cyberbullying incidents. The study systematically evaluates the performance of the proposed model through a series of experiments involving a diverse dataset derived from various social media interactions, categorized by sentiment and type of bullying. Results indicate that while the model achieves high accuracy in identifying cyberbullying, challenges such as overfitting and the need for better generalization to unseen data persist. The paper also discusses ethical considerations and the potential for bias in automated monitoring systems, stressing the importance of ethical AI practices in social media governance. The findings underscore the complexity of automated cyberbullying detection and highlight the necessity for advanced machine learning techniques that are robust, scalable, and aligned with ethical standards. This study contributes to the broader discourse on the application of artificial intelligence in enhancing digital safety and advocates for a multidisciplinary approach to address the socio-technical challenges posed by cyberbullying in the digital age.**

Keywords*—*Cyberbullying detection; deep learning; CNN; LSTM; social media monitoring; sentiment analysis; digital safety

I. Introduction

Cyberbullying has emerged as a significant social problem with the rise of digital communication platforms. Unlike traditional forms of bullying, cyberbullying allows perpetrators to extend their reach beyond physical spaces, affecting victims at any time and from any location. This form of harassment is especially prevalent among adolescents and young adults and has been linked to a range of negative outcomes, including depression, anxiety, and even suicidal thoughts [1]. The anonymous and pervasive nature of online interactions complicates the detection and intervention processes, making it imperative to develop automated tools that can effectively identify and mitigate instances of cyberbullying.

Recent advancements in machine learning and natural language processing have paved the way for more sophisticated approaches to monitoring online behavior. In particular, deep learning architectures have demonstrated significant potential in text classification tasks, which are central to identifying harmful content [2]. However, the complexity of language, including the use of slang, coded messages, and context-specific references, poses unique challenges that require robust and adaptable models.

Hybrid deep learning architectures that combine convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown promising results in various domains, such as image recognition and speech processing [3]. CNNs are adept at extracting hierarchical feature maps from spatial data, making them useful for capturing the textural patterns in language, whereas RNNs, particularly those using Long Short-Term Memory (LSTM) cells, excel in handling sequences with long-range dependencies, capturing the contextual nuances necessary for understanding the intent behind words [4].

The proposed research aims to leverage the strengths of both CNNs and LSTMs to develop a comprehensive model for cyberbullying detection on social networks. This approach is motivated by the hypothesis that a hybrid model can better accommodate the complex and dynamic nature of language used in online social platforms than models based solely on one type of network. By integrating the local feature extraction capabilities of CNNs with the sequence modeling strengths of LSTMs, the model is expected to perform with higher accuracy and sensitivity in detecting cyberbullying instances [5].

Additionally, the vast amount of data available on social networks provides a fertile ground for training deep learning models but also introduces challenges such as data imbalance, privacy concerns, and the need for models to perform well across diverse demographic groups. Addressing these challenges requires not only technical solutions but also an understanding of the social and ethical implications of deploying AI-driven monitoring tools [6].

The use of dropout layers and other regularization techniques in the architecture aims to prevent overfitting, ensuring that the model generalizes well to new, unseen data, which is critical in the ever-evolving landscape of social media language [7]. Moreover, the model's performance must be rigorously evaluated to ensure its efficacy and fairness,

preventing potential biases that could harm certain groups disproportionately [8].

This research contributes to the growing body of knowledge on AI applications in social good by providing a nuanced approach to detecting cyberbullying, which could be integrated into social networks as part of proactive measures to safeguard users. The outcome of this study has the potential to inform policies and practices not only for social media companies but also for educators and policymakers concerned with cyber welfare [9].

In summary, the increasing prevalence of cyberbullying and the limitations of current detection methodologies underscore the need for innovative solutions. This paper proposes a novel hybrid deep learning architecture that integrates the distinct advantages of CNNs and LSTMs, aiming to enhance the detection accuracy and operational efficiency of cyberbullying interventions on social platforms. By addressing both technical challenges and ethical considerations, this study seeks to contribute significantly to the safer utilization of digital spaces.

II. Related Work

The study of cyberbullying has garnered significant interest due to its profound impact on individuals and society. This section delves into the evolution of cyberbullying detection methodologies, tracing the trajectory from manual monitoring to sophisticated hybrid deep learning models. By examining past efforts and their limitations, this paper aims to underscore the novelty and necessity of the proposed approach in enhancing detection mechanisms on social media platforms [10].

A. Cyberbullying: Definitions and Impact

Cyberbullying is characterized by repeated behavior aimed at harming others using digital platforms. Definitions vary, but common elements include intent, repetition, and the use of electronic forms of contact [11]. As for its impact, studies have consistently shown that victims of cyberbullying exhibit higher levels of anxiety, depression, and even suicidal ideation, demonstrating the critical need for effective detection and prevention strategies [12].

B. Traditional Methods for Cyberbullying Detection

Initially, cyberbullying detection relied heavily on manual monitoring, where human moderators reviewed content to identify harmful behavior. However, this method is not scalable and is subject to human error and bias [13]. Subsequently, rule-based systems were developed, utilizing predefined keywords and patterns to automate detection. These systems, while faster, often failed to capture the context and complexity of human interactions, leading to high rates of false positives and negatives [14].

C. Machine Learning Approaches

The advent of machine learning heralded a new era in cyberbullying detection, with the development of sophisticated, feature-based models. Support Vector Machines (SVM) and Naive Bayes classifiers are prominent examples, employed extensively to analyze textual data. These models classify text by extracting and leveraging specific features such as word frequency and sentence structure [15]. Despite their initial success, these traditional machine learning approaches require extensive feature engineering to operate effectively. Moreover, their rigid frameworks often fail to adapt to the dynamic and evolving use of language typical of social media contexts. This limitation significantly restricts their practical applicability, as they cannot seamlessly accommodate new slang, code-switching, or emergent linguistic patterns without manual updates or retraining [16].

D. Deep Learning in Text Analysis

The limitations of traditional machine learning models led to the adoption of deep learning techniques, which can automatically learn features from data. Single-model architectures, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been widely used in text analysis due to their ability to extract spatial hierarchies and manage sequence dependencies, respectively [17]. CNNs are effective in extracting text patterns without the need for manual feature specification, while RNNs, especially those equipped with LSTM cells, excel in understanding context over longer text sequences [18]. However, each has limitations when used independently; CNNs can miss nuanced linguistic context, and RNNs can be computationally intensive and difficult to train on long sequences [19].

E. Hybrid Deep Learning Models

The concept of hybrid deep learning models, which integrate the strengths of CNNs and LSTMs, has started to gain traction. These models capitalize on CNNs' ability to process spatial data and LSTMs' proficiency in sequence processing, offering a robust framework for understanding complex text data. Such architectures have shown promise in fields like sentiment analysis and natural language processing, where both local features and global context are critical for accurate interpretation [20]. In the realm of cyberbullying detection, these hybrid models are proposed to enhance accuracy by effectively capturing both the textual nuances and the broader semantic context of interactions [21].

F. Challenges and Ethical Considerations

Implementing deep learning models for cyberbullying detection on social networks introduces several challenges. Data imbalance—where instances of bullying are vastly outnumbered by normal interactions—can skew model training and performance [22]. Privacy is another significant concern, as these models require access to personal data, potentially infringing on user confidentiality [23]. Moreover, the deployment of such models must be handled with care to avoid biases that could disproportionately affect certain demographics, necessitating continuous evaluation to ensure fairness and ethical integrity [24].

G. Research Gaps and Opportunities

Despite the advancements in detection technologies, several research gaps remain. Current models often fail to consider the multi-modal nature of cyberbullying, which can include text, images, and videos. Moreover, the adaptability of models to new, evolving forms of language and bullying tactics is still limited. There is a pressing need for models that can dynamically learn and adapt from incoming data streams without requiring frequent retraining [25]. Furthermore, there is a lack of comprehensive studies that integrate ethical

considerations into the model development process, an area ripe for further exploration [26-28].

III. Materials and Methods

A. Proposed Model

The proposed hybrid deep learning architecture integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to exploit the strengths of both in detecting cyberbullying within textual data. This section elaborates on the configuration, interplay, and computational aspects of the model. Fig. 1 demonstrates the proposed model for cyberbullying detection.

The model comprises several layers, each tailored to perform specific functions critical to processing and classifying textual inputs:

Input Layer: This layer accepts pre-processed textual data, where each input unit represents a tokenized word encoded as a vector in a high-dimensional space.

Dropout Layer: Positioned immediately after the input layer, this layer randomly omits a fraction p of the input units (neurons) during training to prevent overfitting, where p is the dropout rate set to 0.5 as a starting point.

LSTM Layer: The LSTM layer processes the sequence data, preserving long-term dependencies within texts. The LSTM cells are formulated as follows:

$$i_t = \sigma\left(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}\right) \quad (1)$$

$$f_t = \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}\right) \quad (2)$$

$$g_t = \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}\right) \quad (3)$$

$$o_t = \sigma\left(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}\right) \quad (4)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \quad (5)$$

$$h_t = o_t \otimes \tan g\left(c_t\right) \quad (6)$$

Where, I, f, g, and o are the input, forget, cell, and output gates, respectively; W and b represent weights and biases; σ denotes the sigmoid function; and ⊙ indicates element-wise multiplication.

Convolutional Layer: Following the LSTM, a convolutional layer is applied to extract local feature patterns from the output sequences of the LSTM, enhancing the model's ability to detect nuanced language patterns indicative of cyberbullying. The convolutional operation is defined as:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \quad (7)$$

where x is the input from the LSTM, w is the kernel, and s(t) is the convolved feature at time t.

Flatten Layer: This layer flattens the multi-dimensional output of the convolutional layer into a single-dimensional array for subsequent processing.

Output Layer: The final layer is a fully connected layer that maps the flattened features to the output classes, which are 'cyberbullying' and 'non-cyberbullying'. The activation function used here is the softmax function, given by:

$$Soft\max(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (8)$$

where z_i are the inputs to the output layer from the flatten layer.

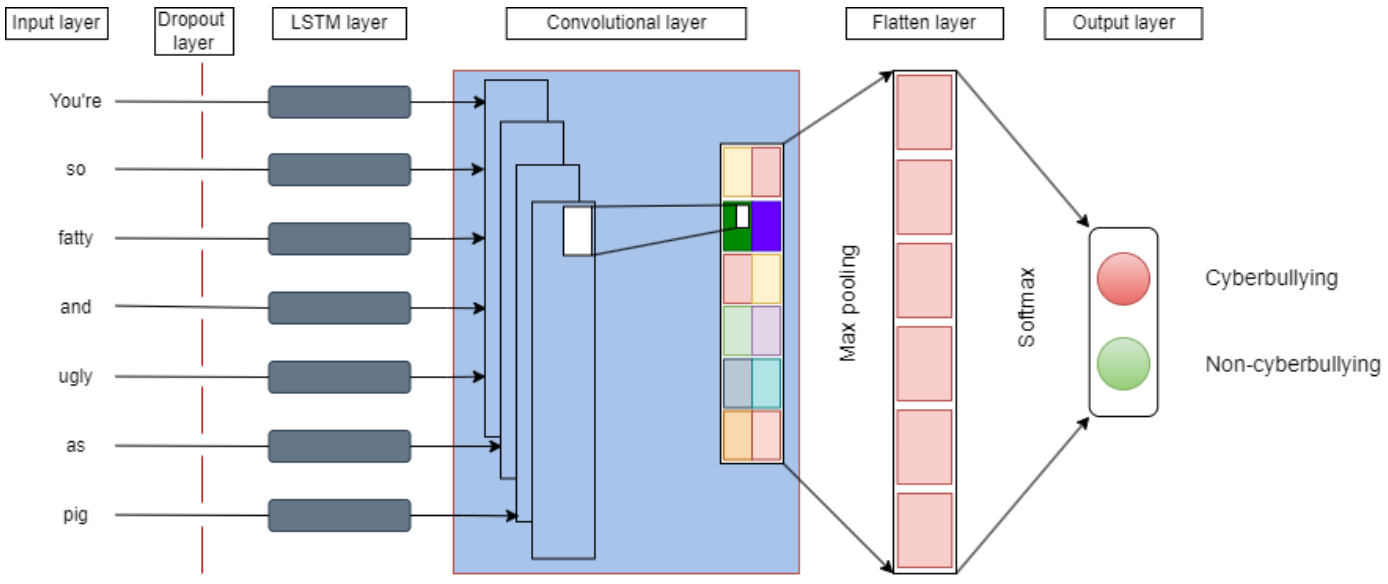

Fig. 1. Architecture of the proposed model.

The proposed model is trained using a cross-entropy loss function, which is suitable for binary classification problems like cyberbullying detection. The loss is calculated as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \quad (9)$$

Where, N is the number of training samples, y_i is the actual label, and $\hat{y}_i$ is the predicted probability of the sample belonging to the 'cyberbullying' class.

By combining LSTM and CNN layers, the architecture harnesses both temporal dynamics and spatial feature extraction, making it robust against the complex variations in textual data typical of cyberbullying contexts. This model promises not only improved accuracy but also better generalizability across different social media platforms.

B. Dataset

The dataset utilized in this study is sourced from Kaggle and consists of structured data specifically tailored for cyberbullying detection tasks. This dataset is comprised of textual data collected from various social media platforms, providing a diverse representation of linguistic styles and contexts. It includes examples labeled with categories indicative of different forms of cyberbullying, facilitating the training and evaluation of machine learning models designed to identify and classify cyberbullying instances.

The data are categorized into distinct classes based on the nature and severity of the bullying behavior. This classification aids in fine-tuning the model's sensitivity to various forms of cyberbullying, from subtle to overt aggression. The dataset's structured format and comprehensive labeling make it an ideal resource for developing robust detection systems using advanced machine learning techniques such as the proposed hybrid deep learning architecture.

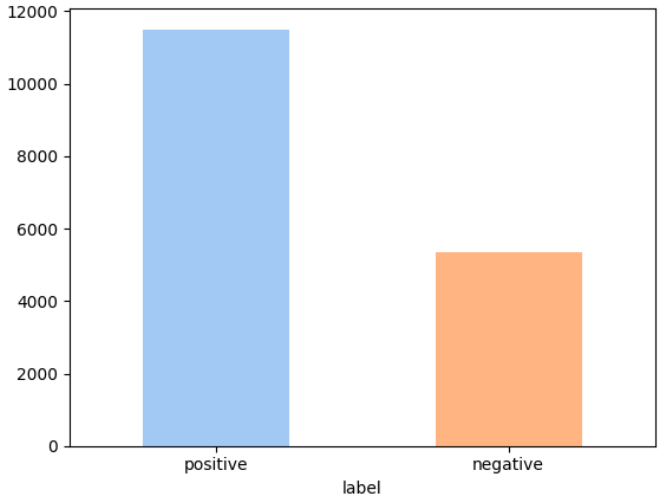

Fig. 2. Class distribution in the cyberbullying dataset.

This dataset not only enhances the model's ability to learn from real-world examples but also provides a benchmark for assessing the effectiveness of different architectural configurations and learning algorithms in accurately detecting cyberbullying across diverse digital environments.

In Fig. 2, the dataset depicted in the provided bar chart is characterized by a significant imbalance between two classes: "positive" and "negative". The "positive" class, which presumably represents instances identified as cyberbullying, substantially outnumbers the "negative" class, indicative of non-cyberbullying instances. This distribution highlights a common challenge in machine learning applications, where class imbalance can significantly impact the learning algorithm's ability to accurately generalize to new data. Such disparities necessitate the implementation of specialized techniques in data preprocessing, such as oversampling the minority class or undersampling the majority class, to ensure that the model trained on this dataset does not exhibit a bias towards the more frequently represented class. This imbalance also underscores the importance of employing robust evaluation metrics that can appropriately measure model performance in the presence of skewed class distributions.

IV. Results

A. Evaluation Parameters

In the context of this research paper on cyberbullying detection using a hybrid deep learning architecture, the evaluation of the model's performance is critical. This involves several key parameters that offer insights into the effectiveness and reliability of the model across various dimensions:

Accuracy measures the proportion of total correct predictions (both positive and negative) over all predictions made by the model. This metric is straightforward but can be misleading in datasets where class imbalance exists, as it might reflect the prevalence of the majority class rather than the model's ability to identify all classes accurately.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

Precision assesses the correctness of positive predictions. In the context of cyberbullying, it measures the proportion of true cyberbullying instances among all instances flagged by the model as cyberbullying. High precision is crucial in this domain to minimize the number of non-bullying content mistakenly classified as bullying, which can have significant social and psychological impacts on users.

$$preision = \frac{TP}{TP + FP} \quad (11)$$

Recall, also known as sensitivity, recall quantifies the model's ability to identify all actual positives. It answers the question: "Of all the true cyberbullying instances, how many did the model correctly identify?" High recall is essential for ensuring that the model does not overlook cyberbullying instances, thus providing a safe online environment.

$$recall = \frac{TP}{TP + FN} \quad (12)$$

The F-score or F1-score is the harmonic mean of precision and recall. It is particularly useful when the balance between precision and recall is important to the application. The F-score provides a single metric that summarizes model performance in terms of both the precision and the recall, which is valuable in scenarios where both false positives and false negatives have serious implications.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (13)$$

These metrics provide a comprehensive view of the model's performance, allowing for balanced optimization in scenarios where making accurate and reliable predictions is crucial. In cyberbullying detection, where the cost of false positives (non-bullying content flagged as bullying) and false negatives (failing to detect actual bullying) can be high, these measures help in fine-tuning the model to achieve optimal performance.

B. Data Preparation Results

Fig. 3 illustrates a word cloud generated from a topic modeling analysis on the cyberbullying dataset, depicting the most frequently occurring words within a specific cluster of the data labeled as "Topic 1." The prominent terms, including derogatory and offensive language, highlight the aggressive and harmful nature of the communications that are commonly identified in instances of cyberbullying. Words such as "dumb," "racist," and more severe pejorative terms point to the type of content that algorithms need to effectively identify and address. This visualization not only underscores the challenges of automatically detecting such harmful content but also serves as a stark reminder of the harsh reality faced by victims of cyberbullying. The word cloud thus provides a visual summary of the key themes and language used in cyberbullying, guiding the development of more nuanced detection mechanisms.

Fig. 3. Word cloud from cyberbullying dataset.

Fig. 4 presents the sentiment distribution across various topics derived from the cyberbullying dataset, segmented into negative, neutral, and positive sentiments. The vertical stacked bar graph illustrates a notable predominance of negative sentiments across all topics, underscoring the pervasive nature of negative expressions in discussions related to cyberbullying. Topic 0 exhibits a particularly high concentration of negative sentiment, dwarfing the other sentiments and suggesting a significant alignment with more aggressive or harmful content. In contrast, Topics 1 through 3 display a more balanced distribution but still with a visible skew towards negative sentiments. This visualization highlights the critical need for effective sentiment analysis tools in cyberbullying detection systems, as it underscores the varying emotional contexts that can pervade different discussion topics, aiding in more nuanced understanding and classification of potentially harmful content.

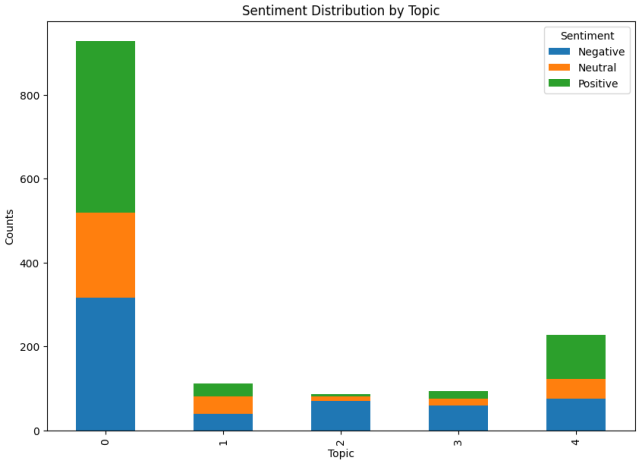

Fig. 4. Distribution of classes in the dataset.

Fig. 5 depicts the sentiment distribution across five different topics extracted from a cyberbullying dataset, with sentiments categorized as negative, neutral, and positive. The heat map highlights the count of sentiments within each topic, revealing a complex landscape of emotional expressions. Notably, Topic 0 is predominantly characterized by positive sentiments, contrasting sharply with other topics where negative sentiments prevail. This suggests that Topic 0 might encompass discussions or interactions of a different nature or context compared to the others. Topics 1 through 4 show a higher prevalence of negative sentiments, indicative of the adversarial or harmful content typically associated with cyberbullying. This distribution aids in understanding the emotional underpinnings of the topics discussed within the dataset and underscores the importance of sentiment analysis in contextualizing the interactions captured in social media data for cyberbullying detection.

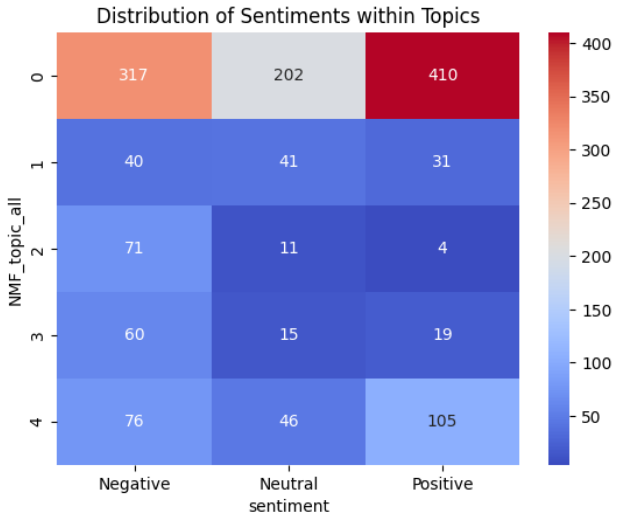

Fig. 5. Distribution of sentiments within topics.

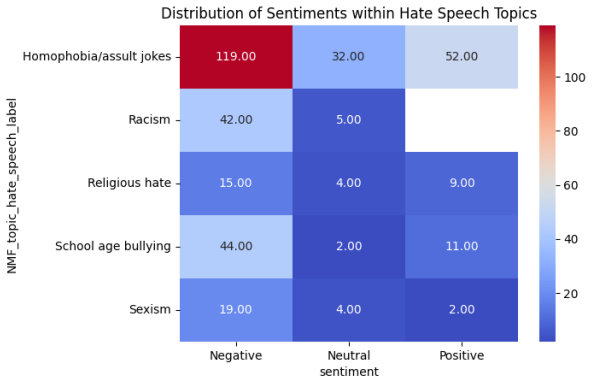

Fig. 6. Distribution of sentiments within hate speech topics.

Fig. 6 presents a heat map illustrating the distribution of sentiments categorized as negative, neutral, and positive across various hate speech topics derived from social media data. This visualization reveals that homophobia and assault jokes have a notably high incidence of positive sentiments relative to their negative sentiments, suggesting complex interactions that might include endorsements or trivialization of serious issues. In contrast, topics like racism, religious hate, school-age bullying, and sexism predominantly exhibit negative sentiments, aligning with the adversarial nature of the content. The presence of neutral sentiments across all topics indicates discussions that potentially include factual statements or undetermined stances. This detailed sentiment breakdown provides valuable insights into the nature of discussions within each hate speech category, underscoring the importance of nuanced analysis in automated detection systems to differentiate between contexts and intentions behind the words used.

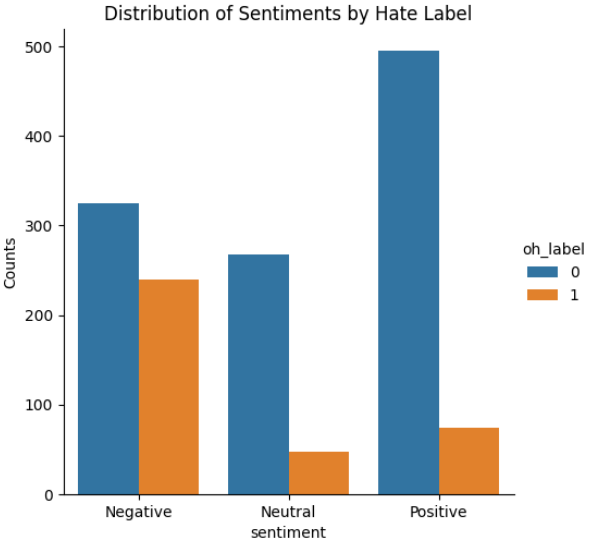

Fig. 7. Distribution of sentiments by hate label.

Fig. 7 displays a bar chart representing the distribution of sentiments categorized as negative, neutral, and positive across two distinct hate labels, identified here as "0" and "1". The labels likely denote the presence (1) or absence (0) of hate speech within the sentiments expressed. Notably, there is a predominant concentration of negative sentiments associated with the label "1", which underscores the alignment of negative sentiments with recognized hate speech. In contrast, sentiments labeled as "0", suggesting non-hate speech contexts, show a more balanced distribution among negative, neutral, and positive sentiments. This distribution highlights the complex relationship between expressed sentiments and their classification as hate speech, illustrating the challenge in distinguishing hate speech from non-hate speech based on sentiment alone. The data underscores the importance of sophisticated models that can interpret not only the sentiment but the context and intent behind expressions to effectively moderate content on social platforms.

The analysis presented in this subsection highlights the intricate relationship between sentiment and hate speech within digital communications. The visualizations and data distributions discussed reveal not only the prevalence of negative sentiments associated with confirmed instances of hate speech but also the more nuanced interplay of sentiments in texts labeled as non-hate speech. These insights underscore the challenges in deploying automated systems for the detection and moderation of hate speech, emphasizing the necessity for models that can adeptly discern context, intent, and sentiment nuances. Future research should focus on enhancing the accuracy and sensitivity of detection algorithms by incorporating multidimensional data analysis and applying advanced machine learning techniques that can better understand and process the complexities of human language and interaction in digital platforms. This approach is crucial for developing more effective and ethical tools for managing online safety and promoting positive digital communication environments.

C. Experimental Results

Fig. 8 presents a confusion matrix for a sentiment classification model applied to cyberbullying detection, displaying the numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. In the matrix, the horizontal axis represents the actual labels of the data ('Negative' and 'Positive'), while the vertical axis represents the predicted labels. The matrix shows that the model correctly identified 1498 negative cases and 531 positive cases as true negatives and true positives, respectively. Conversely, there were 228 cases where positive sentiments were incorrectly predicted as negative (false negatives), and 271 cases where negative sentiments were incorrectly labeled as positive.

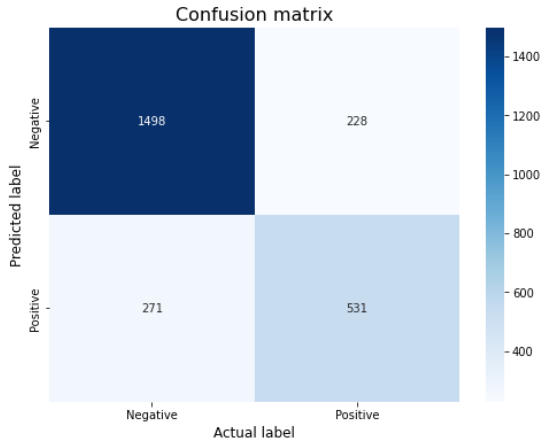

Fig. 8. Confusion matrix of sentiment classification model.

The significant number of true positives and true negatives indicates that the model is fairly effective in identifying both negative and positive sentiments accurately. However, the presence of false positives and false negatives highlights areas for potential improvement. The false negatives, in particular, are concerning because they represent instances where the model failed to detect positive sentiments, which could be crucial in contexts where affirming positive communication is as important as detecting negative or harmful content. Enhancing the model's sensitivity and specificity, possibly through more advanced feature engineering, optimization of hyperparameters, or the incorporation of context-aware algorithms, could further improve its performance. This analysis not only serves as a basis for evaluating the effectiveness of the current model but also guides future modifications and enhancements to better handle the complexities of sentiment analysis in cyberbullying contexts.

Fig. 9 displays the training and validation accuracy and loss of a deep learning model over several epochs. The left graph illustrates a steady increase in training accuracy from approximately 82.5% to 95.4% across six epochs, suggesting that the model is effectively learning from the training data. However, the validation accuracy starts around 90% and shows a declining trend to stabilize near 87.5%, indicating potential issues with overfitting as the model becomes too specialized to the training data and performs less effectively on unseen validation data. This divergence between training and validation performance is a common challenge in machine learning, highlighting the need for techniques like regularization or dropout to mitigate overfitting.

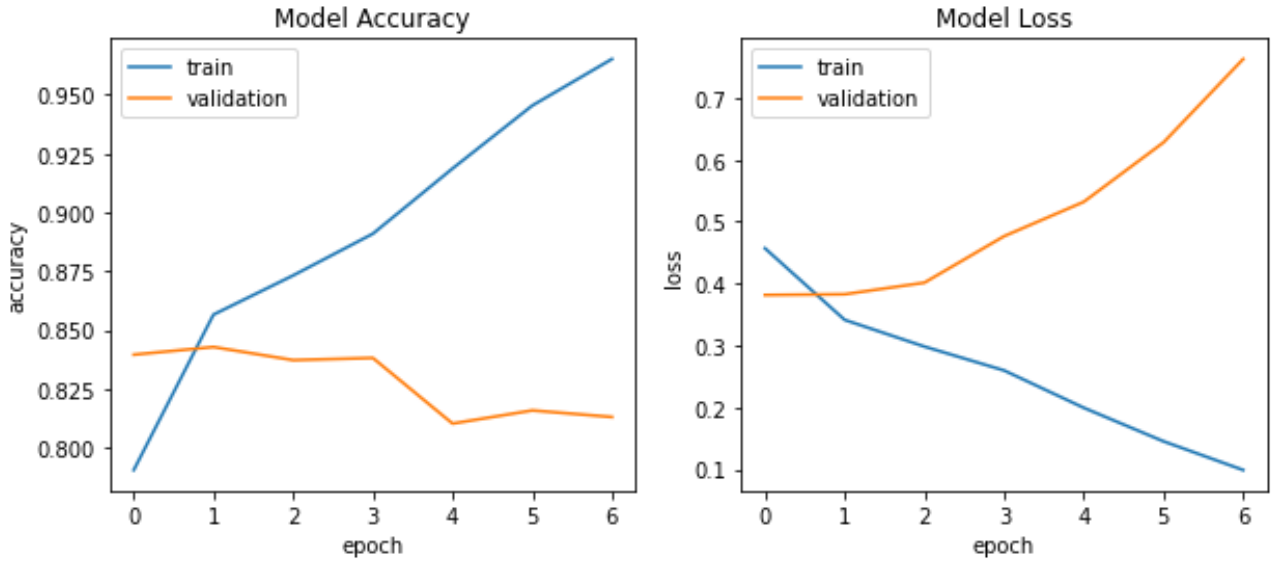

Fig. 9. Model accuracy and loss.

The right graph of model loss shows an inverse relationship to the accuracy graph. The training loss decreases sharply from about 0.7 to under 0.1, while the validation loss initially decreases but then begins to rise after the first epoch, reaching about 0.6 by the sixth epoch. This increase in validation loss alongside a decrease in validation accuracy further supports the possibility of overfitting. The early plateau and subsequent increase in validation loss could also suggest that the model's capacity is either insufficient or excessively tuned to nuances in the training data that do not generalize well. Addressing this might involve adjusting model parameters, introducing early stopping in training, or enriching the dataset to enhance the model's generalization capabilities. These adjustments are crucial for improving model robustness and ensuring reliable predictions across diverse data sets.

V. Discussion

The development and evaluation of a hybrid deep learning model to detect cyberbullying across social media platforms have yielded significant insights and highlighted several areas of concern and opportunity that warrant further discussion. The integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks in the proposed model aimed to leverage the spatial and temporal processing strengths of these architectures to enhance detection accuracy.

A. Model Performance and Overfitting

The proposed system demonstrated significant accuracy in the training and validation results, as illustrated in Fig. 9, demonstrate a clear disparity between training accuracy and validation accuracy, suggesting a tendency towards overfitting. This phenomenon, where the model performs exceptionally well on training data but less so on unseen validation data, is indicative of the model's inability to generalize beyond the training set [29-31]. Overfitting is a common challenge in machine learning, particularly in complex models such as deep neural networks. Strategies such as introducing dropout layers, increasing data augmentation, and employing regularization techniques have been recommended to mitigate this issue [32-34].

B. Sentiment and Hate Speech Analysis

The analysis of sentiments and their correlation with hate speech, as discussed through Fig. 5, 6, and 7, reveals a complex interaction between sentiment classifications and the

identification of hate speech. The prevalence of negative sentiments in recognized hate speech categories underscores the critical role of sentiment analysis in contextualizing content. However, the presence of positive and neutral sentiments within these categories also highlights the complexity of human communication, where not all harmful or aggressive communications are overtly negative [35-37].

C. Ethical Considerations and Bias

The deployment of AI in monitoring and moderating online content raises substantial ethical questions, particularly concerning privacy, consent, and the potential for bias. The model's susceptibility to biases, whether in data representation or algorithmic predilections, can lead to disproportionate flagging of content from specific demographics or linguistic groups [38-40]. It is crucial that continuous efforts be made to audit and refine these systems to ensure fairness and equity in automated content moderation [41].

D. Future Directions

Looking forward, the research highlights several pathways for improvement and further study. First, enhancing the model's generalizability through a more diverse and extensive dataset could help in reducing overfitting and improving the model's robustness across various social media contexts [42]. Second, exploring alternative hybrid architectures or newer approaches like transformer models, which have shown promise in other NLP tasks, might offer new avenues for increasing both accuracy and efficiency in processing [43-45].

Additionally, the integration of multimodal data, including images and videos alongside text, could provide a more holistic view of content, given that cyberbullying often transcends textual communication [46-47]. Finally, more nuanced understanding and classification frameworks are necessary to better differentiate between types of cyberbullying and their severities, which could aid in more targeted and effective interventions.

This research contributes to the ongoing discourse on AI's role in social media safety and the technical challenges associated with developing effective cyberbullying detection systems. While the results are promising, they also reflect the need for a cautious and considered approach to implementing such technologies. The balance between technological advancements and ethical considerations remains a critical frontier in the research and application of AI in societal contexts.

VI. Conclusion

The research undertaken in this paper has highlighted the pivotal role that hybrid deep learning architectures can play in the detection and analysis of cyberbullying across social media platforms. The integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks demonstrated potential in enhancing the accuracy of identifying harmful content by leveraging both spatial and sequential data processing capabilities. Despite this, challenges such as model overfitting and the need for greater generalization point to areas requiring further refinement. Ethical considerations around the implementation of AI in public domains, particularly concerning privacy, bias, and fairness, underscore the necessity for rigorous standards and continuous oversight in deploying such technologies. Future research directions should focus not only on improving the technical robustness of detection models through advanced machine learning techniques and diversified data sets but also on exploring the socio-technical impacts of AI interventions in human interactions. The balance between technological innovation and ethical responsibility remains a critical frontier in the application of AI, advocating for a multidisciplinary approach to ensure that advancements in AI contribute positively to societal welfare. This research contributes to a growing body of knowledge that seeks to harness the power of AI for social good, emphasizing the importance of aligning technological developments with human values and ethical standards.

Acknowledgment

This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP23488900 "Automatic detection of cyberbullying among young people in social networks using artificial intelligence".

References

[1] Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. IEEE Access, 10, 25857-25871.

[2] Sultan, D., Mendes, M., Kassenkhan, A., & Akylbekov, O. (2023). Hybrid CNN-LSTM Network for Cyberbullying Detection on Social Networks using Textual Contents. International Journal of Advanced Computer Science and Applications, 14(9).

[3] Albayari, R., Abdallah, S., & Shaalan, K. (2024). Cyberbullying Detection Model for Arabic Text Using Deep Learning. Journal of Information & Knowledge Management, 2450016.

[4] Omarov, B., Narynov, S., & Zhumanov, Z. (2023). Artificial Intelligence-Enabled Chatbots in Mental Health: A Systematic Review. Computers, Materials & Continua, 74(3).

[5] Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2023). Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems, 29(3), 1839-1852.

[6] Balaji, P. G., Katariya, P. P., Sruthi, S., & Venugopalan, M. (2024, April). Cyberbullying Detection on Multiclass Data Using Machine Learning and A Hybrid CNN-BiLSTM Architecture. In 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS) (Vol. 1, pp. 1-6). IEEE.

[7] Kumar, A. S., Kumar, N. S., Devi, R. K., & Muthukannan, M. (2024). Analysis of Deep Learning-Based Approaches for Spam Bots and Cyberbullying Detection in Online Social Networks. AI-Centric Modeling and Analytics, 324-361.

[8] Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M. A., Yaseen, Q., & Gupta, B. B. (2024). Image cyberbullying detection and recognition using transfer deep machine learning. International Journal of Cognitive Computing in Engineering, 5, 14-26.

[9] Hasan, M. T., Hossain, M. A. E., Mukta, M. S. H., Akter, A., Ahmed, M., & Islam, S. (2023). A review on deep-learning-based cyberbullying detection. Future Internet, 15(5), 179.

[10] 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.

[11] Paruchuri, V. L., & Rajesh, P. (2023). CyberNet: a hybrid deep CNN with N-gram feature selection for cyberbullying detection in online social networks. Evolutionary Intelligence, 16(6), 1935-1949.

[12] Suhas Bharadwaj, R., Kuzhalvaimozhi, S., & Vedavathi, N. (2022). A novel multimodal hybrid classifier based cyberbullying detection for

social media platform. In Proceedings of the Computational Methods in Systems and Software (pp. 689-699). Cham: Springer International Publishing.Murshed, B. A. H., Suresha, Abawajy, J., Saif, M. A. N., Abdulwahab, H. M., & Ghanem, F. A. (2023). FAEO-ECNN: cyberbullying detection in social media platforms using topic modelling and deep learning. Multimedia Tools and Applications, 82(30), 46611-46650.

[13] Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., & Choo, K. K. R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. Concurrency and Computation: Practice and Experience, 32(23), e5627.

[14] Dutta, S., Neog, M., & Baruah, N. (2024, February). Towards Safer Social Spaces: LSTM, Bi-LSTM and Hybrid Approach for Cyberbullying Detection in Assamese language on Social Networks. In 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE) (pp. 1-6). IEEE.

[15] 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.

[16] 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.

[17] 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.

[18] Roy, P. K., & Mali, F. U. (2022). Cyberbullying detection using deep transfer learning. Complex & Intelligent Systems, 8(6), 5449-5467.

[19] Shibly, F. H. A., Sharma, U., & Naleer, H. M. M. (2022). Detection of Cyberbullying in Social Media to Control Users' Mental Health Issues Using Recurrent Neural Network Architectures. Journal of Pharmaceutical Negative Results, 434-441.

[20] 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.

[21] Nasution, M. A. S., & Setiawan, E. B. (2023). Enhancing Cyberbullying Detection on Indonesian Twitter: Leveraging Fast Text for Feature Expansion and Hybrid Approach Applying CNN and BiLSTM. Revue d'Intelligence Artificielle, 37(4), 929-936.

[22] 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.

[23] Teng, T. H., & Varathan, K. D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. IEEE Access, 11, 55533-55560.

[24] Rodela, A. T., Nguyen, H. H., Farid, D. M., & Huda, M. N. (2023, October). Bangla Social Media Cyberbullying Detection Using Deep Learning. In International Conference on Intelligent Systems and Data Science (pp. 170-184). Singapore: Springer Nature Singapore.

[25] Omarov, B., Anarbayev, A., Turyskulov, U., Orazbayev, E., Erdenov, M., Ibrayev, A., & Kendzhaeva, B. (2020). Fuzzy-PID based self-adjusted indoor temperature control for ensuring thermal comfort in sport complexes. J. Theor. Appl. Inf. Technol, 98(11), 1-12.

[26] Pericherla, S., & Ilavarasan, E. (2024). Transformer network-based word embeddings approach for autonomous cyberbullying detection. International Journal of Intelligent Unmanned Systems, 12(1), 154-166.

[27] Jing, Y., Haowei, M., Ansari, A. S., Sucharitha, G., Omarov, B., Kumar, S., ... & Alyamani, K. A. (2023). Soft computing techniques for detecting cyberbullying in social multimedia data. ACM Journal of Data and Information Quality, 15(3), 1-14.

[28] Al Duhayyim, M., Mohamed, H. G., Alotaibi, S. S., Mahgoub, H., Mohamed, A., Motwakel, A., ... & Eldesouki, M. I. (2022). Hyperparameter Tuned Deep Learning Enabled Cyberbullying Classification in Social Media. Computers, Materials & Continua, 73(3).

[29] Omarov, B., Altayeva, A., Suleimenov, Z., Im Cho, Y., & Omarov, B. (2017, April). Design of fuzzy logic based controller for energy efficient operation in smart buildings. In 2017 First IEEE International Conference on Robotic Computing (IRC) (pp. 346-351). IEEE.

[30] Omarov, B. (2017, October). Development of fuzzy based smart building energy and comfort management system. In 2017 17th International Conference on Control, Automation and Systems (ICCAS) (pp. 400-405). IEEE.

[31] Obaid, M. H., Guirguis, S. K., & Elkaffas, S. M. (2023). Cyberbullying detection and severity determination model. IEEE Access.

[32] 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.

[33] Ghosh, T., Chowdhury, A. A. K., Banna, M. H. A., Nahian, M. J. A., Kaiser, M. S., & Mahmud, M. (2022, October). A hybrid deep learning approach to detect bangla social media hate speech. In Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021 (pp. 711-722). Singapore: Springer Nature Singapore.

[34] Ahmed, M. T., Akter, N., Rahman, M., Das, D., & AZM T, R. G. (2023). Multimodal cyberbullying meme detection from social media using deep learning approach. Int J Comput Sci Inf Technol (IJCSIT), 15, 27-37.

[35] Kumari, K., Singh, J. P., Dwivedi, Y. K., & Rana, N. P. (2021). Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. Future Generation Computer Systems, 118, 187-197.

[36] Omarov, B. (2017, October). Exploring uncertainty of delays of the cloud-based web services. In 2017 17th International Conference on Control, Automation and Systems (ICCAS) (pp. 336-340). IEEE.

[37] Kadiri, P., Arjun, U., Sravani, N., Jyothi, N. S., Mahesh, P., & Naik, S. J. (2024, May). Detecting Cyberbullying through social media: A Deep Learning Approach. In 2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE) (pp. 1-7). IEEE.

[38] Altayeva, A. B., Omarov, B. S., Aitmagambetov, A. Z., Kendzhaeva, B. B., & Burkitbayeva, M. A. (2014). Modeling and exploring base station characteristics of LTE mobile networks. Life Science Journal, 11(6), 227-233.

[39] Bădiță, V. N. (2023). Cyberbullying Detection. International Journal of Information Security and Cybercrime (IJISC), 12(1), 37-44.

[40] Mahmud, T., Ptaszynski, M., Eronen, J., & Masui, F. (2023). Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. Information Processing & Management, 60(5), 103454.

[41] Omarov, B., Orazbaev, E., Baimukhanbetov, B., Abusseitov, B., Khudiyarov, G., & Anarbayev, A. (2017). Test battery for comprehensive control in the training system of highly Skilled Wrestlers of Kazakhstan on national wrestling" Kazaksha Kuresi". Man In India, 97(11), 453-462.

[42] Akhter, M. P., Jiangbin, Z., Naqvi, I. R., AbdelMajeed, M., & Zia, T. (2022). Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimedia Systems, 28(6), 1925-1940.

[43] Azzeh, M., Alhijawi, B., Tabbaza, A., Alabboshi, O., Hamdan, N., & Jaser, D. (2024). Arabic cyberbullying detection system using convolutional neural network and multi-head attention. International Journal of Speech Technology, 1-17.

[44] Omarov, B., Altayeva, A., & Cho, Y. I. (2017). Smart building climate control considering indoor and outdoor parameters. In Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Bialystok, Poland, June 16-18, 2017, Proceedings 16 (pp. 412-422). Springer International Publishing.

[45] Yi, P., & Zubiaga, A. (2023). Session-based cyberbullying detection in social media: A survey. Online Social Networks and Media, 36, 100250.

[46] 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.

[47] Bhowmik, S., Sultana, S., Sajid, A. A., Reno, S., & Manjrekar, A. (2024). Robust multi-domain descriptive text classification leveraging conventional and hybrid deep learning models. International Journal of Information Technology, 16(5), 3219-3231.