VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i4.6004

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

MODEL OF CYBERBULLYING DETECTION ON SOCIAL MEDIA USING MULTI-LABEL DEEP LEARNING: A COMPARATIVE STUDY

Lemi Iryani[1*]; Junaidi[2]; Paisal[3]; Mariana Purba[4], Nia Umilizah[5], Bakhtiar[6]; Nur Ani[7]

Department of Informatics[1,3,4,5,6]
Department of Law[2]
Universitas Sjakhyakirti, Palembang, Indonesia[1,3,4,5,6]
https://unisti.ac.id[1,3,4,5,6]
lemi_iryani@unisti.ac.id[1*], junaidi@unisti.ac.id[2], paisal@unisti.ac.id[3], mariana_purba@unisti.ac.id[4], nia_umilizah@unisti.ac.id[5], bakhtiar@unisti.ac.id[6]

Institute of Visual Informatics (IVI)[7]
Universiti Kebangsaan Malaysia, Malaysia[7]
https://www.ukm.my/portalukm[7]
p93828@siswa.ukm.edu.my[7]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Cyberbullying is the deliberate act of using technology to harm others. This study aims to analyze 400 Instagram comments obtained via API from previous research. The data were labeled into three classes: negative (containing cyberbullying), positive (non-bullying, supportive), and neutral (neither positive nor negative). The data for experiment was divided into 70% for training and 30% for testing. The research methodology consists of three main stages. The first stage is text preprocessing, which includes tokenization (splitting comments into tokens), filtering (removing unimportant words or stop-words), and stemming (converting words with affixes into their root forms). The second stage is classification analysis using BiLSTM, LSTM, RNN, and CNN-1D methods. The third stage is evaluation by comparing the model's classification results with manually labeled data using accuracy as the evaluation metric. The results show that the BiLSTM model performed the best, achieving an accuracy of 98.51% on the training data and 81.82% on the testing data. The BiLSTM method used in this study can be further adapted to enhance the effectiveness of cyberbullying detection in various applications.*

Keywords: *BiLSTM, CNN-1D, cyberbullying, LSTM, RNN.*

Intisari— *Cyberbullying adalah tindakan menggunakan teknologi untuk menyakiti orang lain secara sengaja. Penelitian ini bertujuan menganalisis 400 komentar Instagram yang diperoleh melalui API dari penelitian sebelumnya. Data dilabeli dalam tiga kelas: negatif (mengandung cyberbullying), positif (tidak mengandung bullying, cenderung mendukung), dan netral (tanpa makna positif atau negatif). Dataset dibagi menjadi 70% pada tahap pelatihan dan 30% pada tahap pengujian. Metodologi penelitian terdiri dari tiga tahap utama. Tahap pertama adalah pra-pemrosesan teks, yang meliputi tokenisasi (memotong komentar menjadi token), filtering (menghapus kata tidak penting atau stop-words), dan konversi kata berimbuhan ke kata dasar. Tahap kedua adalah analisis klasifikasi menggunakan metode BiLSTM, LSTM, RNN, dan CNN-1D. Tahap ketiga adalah evaluasi, dengan membandingkan hasil klasifikasi model dengan data manual menggunakan metrik akurasi. Hasil menunjukkan bahwa model BiLSTM memberikan performa terbaik, dengan akurasi 98,51% pada data pelatihan dan 81,82% pada data pengujian. Metode BiLSTM yang digunakan dapat diadaptasi lebih lanjut untuk meningkatkan efektivitas deteksi cyberbullying di berbagai aplikasi.*

Kata Kunci: *BiLSTM, CNN-1D, cyberbullying, LSTM, RNN.*

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6004

INTRODUCTION

Cyberbullying is an act of bullying through technological devices to intentionally hurt others. This action is usually done repeatedly because the perpetrator feels safe by hiding his identity so that he does not have time to see the victim's response directly [1]–[3]. The effects of cyberbullying play a huge role in mental health. The act of cyberbullying can cause negative affective disorder, loneliness, anxiety, depression and suicidal ideation in the victim [4]. The aim of this research is to identify and analyse comments that contain bullying meaning, with a special focus on interactions on social media [5].

This analysis is important to provide information on social media comments that are negative and have content of violations of the law so as to provide a clearer picture of the right legal action [6]. In addition, the urgency of this research is to recommend how social media applications develop new features in filtering comments that contain the meaning of bullying and legal images that occur against comments made by perpetrators [7]. More detailed research on social media comments is expected to obtain a contribution to the more effective bullying prevention strategies on social media platforms. To support the solution of this problem, the previous research used deep learning approaches [8].

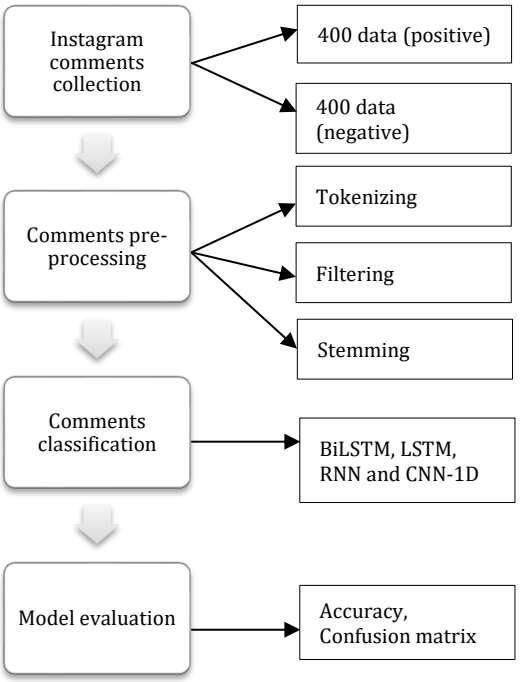
In problem of cyberbullying detection, each model of deep learning has its strengths and weaknesses. The Recurrent Neural Network or RNN is effective for processing dataset in sequential type like text but is prone to the vanishing gradient problem, which can affect performance on longer sequences [9]–[11]. The Long Short-Term Memory or LSTM addressed this limitation by incorporating long-term memory, allowing the model to better capture broader context in text, though it has higher computational complexity [12]–[14].

The Bidirectional LSTM or BiLSTM enhanced LSTM by analyzing forward and backward directions of dataset to obtaining linguistic patterns, making it suitable for identifying complex contexts such as cyberbullying [15]–[17]. CNN-1D (Convolutional Neural Network) is efficient at extracting local features from text through convolutional filters but is less effective at capturing long-range dependencies between words [18]–[20]. The result of this research demonstrated that BiLSTM achieved the best performance with high accuracy due to its bidirectional context-capturing capabilities, making it the preferred choice over other models for detecting cyberbullying in text.[11], [21]–[23] .

Previous research has only focused on multi-class data, this will classify the sentiment class (positive, neutral, negative) [9], [11], [21]–[25]. The results of the research are still not able to be utilized optimally to process non-crimes that are increasingly massive. This study will classify cyberbullying texts based on criminal law labels (insults, defamation, extortion and threats).

MATERIALS AND METHODS

The research methodology is divided into four phases. The phases are Instagram comments collection, comments pre-processing, comments classification and model evaluation. Every phase has different techniques of method to accomplish the goal. The detail of research methodology is depicted in **Figure 1**.



Source: (Research Results, 2025)
Figure 1. Research methodology

The dataset used in this experiment is secondary data. The amount of data used was 400 comments from Instagram users. Data in the form of comments from Instagram by scraping using application programming interface services that have been processed in previous studies [26]. Keywords used in retrieving comments are based on words that contain the meaning of mocking or vilifying an object. Labeling each comment is done by giving 3 classes, namely positive class, negative class, and neutral class. Negative class means Instagram comments that contain the meaning of bully and positive class means Instagram comments contain

meaning not bully (tends to the meaning of motivation or support).

In the second stage, there are three processes, namely tokenization, filtering and stemming. The tokenization process is carried out by cutting Instagram comments based on space characters into several pieces based on each word that makes up a comment. The result of tokenization called a token is a single word that will characterize the classification of Instagram comments. The second process is filtering. This process is carried out by taking important words from the results of the tokenization process. In this process, unimportant words (stop-words) will be eliminated to reduce the number of words that will be processed next. The third process is stemming. This process is done by converting affixes from filtered words to stem (root words).

The third stage is the use of the BiLSTM, LSTM, RNN and CNN-1D method to conduct an analysis of Instagram comment classification. Evaluation of Instagram comment classification is done by comparing between prediction data and actual data. Prediction data is in the form of comment classification results generated by the BiLSTM, LSTM, RNN and CNN-1D while actual data is in the form of comment classification results generated from manual labeling. In this study, the evaluation used is accuracy by comparing cases that are classified correctly with the number of all existing classification cases.

The hyperparameter selection and tuning process played a critical role in optimizing the performance of the classification model. For all models, the embedding dimension was set to 128, which effectively captured word semantics while balancing computational efficiency. The dropout rates, including SpatialDropout1D in BiLSTM, LSTM, and RNN models, were implemented to tackle problem of overfitting by reducing 20% of neurons in training phase. The number of units in recurrent layers varied: 196 for LSTM and RNN, and 392 for the BiLSTM, reflecting the bidirectional nature of the latter, which doubles the parameter count. For CNN-1D, the GlobalAveragePooling1D layer simplified feature extraction while preserving global information, followed by dense layers with ReLU activation and softmax for classification. Batch sizes and learning rates were tuned iteratively, with a batch size parameter with value 32 and optimizer parameter with value Adam for solving convergence problem. These hyperparameter choices were informed by grid search and empirical testing to maximize precision, recall, and F1-scores across all models, with BiLSTM ultimately achieving the best balance between these metrics.

RESULTS AND DISCUSSION

The first experiment was an experiment using BiLSTM. The behavior of the Bidirectional Long Short – Term Memory Neural Network (BiLSTM) from this experiment can be seen in **Table 1.**

Table 1. Architecture BiLSTM

Layer (type)	Output shape	Param#
Embedding (embedding)	(None, 98, 128)	2560000
Spatial_dropout1d (SpatialDropout1D)	(None, 98, 128)	0
Bidirectional (Bidirectiona1)	(None, 392)	509600
Dense (Dense)	(None, 2)	786

Source: (Research Results, 2025)

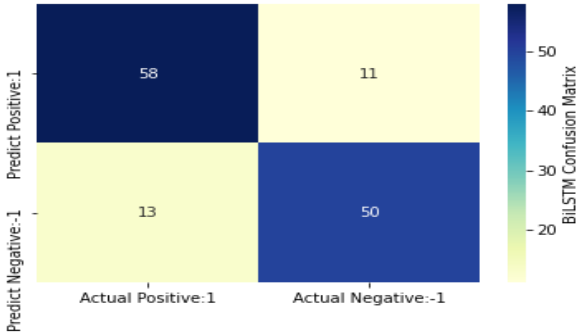
The results of the experiment for cyberbullying text detection using the BiLSTM model show that this model has excellent accuracy on the training data, which is 98.51%. However, the results of the experiment on the test data only obtained an accuracy of 81.82% (decreased). The precision of experiment, recall and f1-score values from the testing experiment is presented in **Table 2**.

Table 2. BiLSTM Performance Analysis

	Precision	Recall	F1-score	support
0	0.82	0.84	0.83	69
1	0.82	0.79	0.81	63

Source: (Research Results, 2025)

The difference in accuracy values in the training data could indicate that the BiLSTM model can detect cyberbullying text in training phase, but it was not generalize to the new cyberbullying dataset. This is often a problem in complex models such as BiLSTM if there is no appropriate regulatory technique in place. In detail, the test results is depicted in confusion matrix on **Figure 2**.



Source: (Research Results, 2025)

Figure 2. Confusion matrix BiLSTM

The second experiment is an experiment using LSTM. The behavior of the LSTM model for of this cyberbullying classification experiment is presented in **Table 3**.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6004

Table 3. LSTM Architecture

Layer (type)	Output shape	Param#
Embedding (embedding)	(None, 98, 128)	2560000
Spatial_dropout1d (SpatialDropout1D)	(None, 98, 128)	0
Lstm (LSTM)	(None, 196)	254800
Dense (Dense)	(None, 2)	394

Source: (Research Results, 2025)

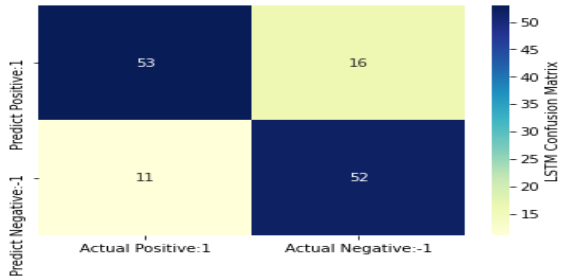
The results of the experiment for cyberbullying text detection using the LSTM model show quite good performance. The precision of experiment, recall and f1-score from the training stage is presented in **Table 4**.

Table 4. Performance Analysis LSTM

	Precision	Recall	F1-score	support
0	0.83	0.77	0.80	69
1	0.76	0.83	0.79	63

Source: (Research Results, 2025)

The training accuracy of 99.25% showed that LSTM managed to learn the patterns in the cyberbullying dataset at a very good stage. However, the accuracy of testing by LSTM for detecting cyberbullying datasets is only 79.55% (decreased). In detail of test results based on the confusion matrix is depicted in **Figure 3**.



Source: (Research Results, 2025)
Figure 3. Confusion matrix LSTM

The third experiment was an experiment using 1D CNN. The behavior of the CNN-1D from this experiment is presented in **Table 5.**

Table 5. Architecture of experiment using CNN-1D

Layer (type)	Output shape	Param#
Embedding (embedding)	(None, 98, 128)	2560000
Global_average_pooling1d (GlobalAveragePooling1D)	(None, 128)	0
Dense (Dense)	(None, 196)	25284
Dense_1 (Dense)	(None, 2)	394

Source: (Research Results, 2025)

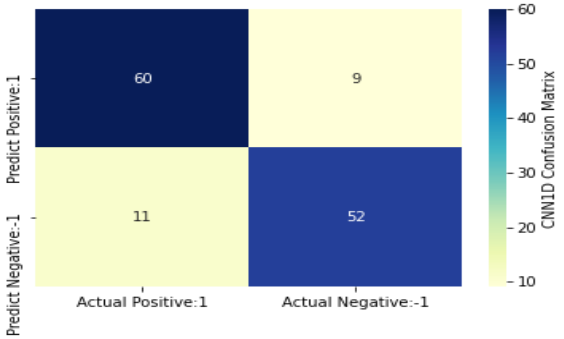
The results of experiments for cyberbullying text detection using the CNN-1D model show very good performance compared to other models. The CNN-1D model obtained a training accuracy of 93.66%, which showed that the model successfully learned the patterns and characteristics of the cyberbullying text in the training dataset. This shows the ability of CNN-1D to recognize important features in the context of cyberbullying text in the training dataset. The precision, recall and f1-score values from the training stage can be seen in **Table 6**.

Table 6 Performance Analysis CNN-1D

	Precision	Recall	F1-score	support
0	0.85	0.87	0.86	69
1	0.85	0.83	0.84	63

Source: (Research Results, 2025)

The CNN-1D model obtained a test accuracy accuracy of 84.85%, indicating that the model was able to generalize to new data quite well, although there was a decrease in training accuracy, but this decrease in accuracy was not significantly different. In detail, the confusion matrix of the test results can be depicted in **Figure 4**.



Source: (Research Results, 2025)
Figure 4. Confusion matrix CNN-1D

The fourth experiment is an experiment using RNNs. The behavior of the RNN from this experiment is presentedin **Table 7**.

Table 7. Architecture of RNN experiment

Layer (type)	Output shape	Param#
Embedding layer	(None, 98, 128)	2560000
Spatial_dropout1d (SpatialDropout1D)	(None, 98, 128)	0
Simple_rnn (SimpleRNN)	(None, 196)	63700
Dense_1 (Dense)	(None, 2)	394

Source: (Research Results, 2025)

The results of the experiment for cyberbullying text detection using the RNN model showed an atypical performance. The training accuracy of the RNN performance of 100% shows that the RNN model successfully recognizes all the patterns in the cyberbullying dataset at the training stage perfectly. However, this accuracy is very different from the results at the testing stage. The

precision of experiment and other evaluation values from the testing phase can be depicted in **Table 8**.

Table 8. RNN Performance Analysis

	Precision	Recall	F1-score	support
0	0.56	0.64	0.59	69
1	0.53	0.44	0.48	63

Source: (Research Results, 2025)

The test accuracy of RNN's performance was only 54.55%. These results are in stark contrast to the training results and show that the RNN model cannot generalize well to the existing cyberbullying datasets at the training stage so RNNs are not recommended for use for cyberbullying text detection. In detail, the confusion matrix of the test results can be seen in **Figure 5**.

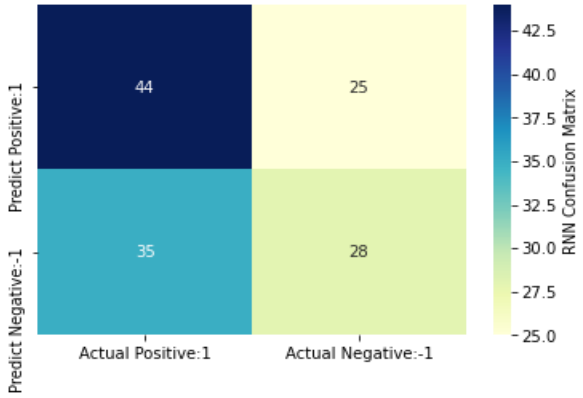


Figure 5. Confusion matrix

Based on the experiment of the training accuracy experiment: RNN has the highest training accuracy (100%) while the test accuracy: CNN 1D gets the highest test accuracy (84.85%). Although the RNN shows that the highly trained model is not yet the results in the test are irrelevant to the training accuracy. The LSTM and CNN 1D models show a good balance between training and testing accuracy. The BiLSTM model is the recommended model in this study. The model comparison of the values of the model can be depicted in **Table 9**.

Table 9 Comparison of the Performance of Text Mining Models

Model	Training	Testing
BiLSTM	98.51%	81.82%
LSTM	99.25%	79.55%
CNN 1D	93.66%	84.85%
RNN	100.00%	54.55%

Source: (Research Results, 2025)

The results of this study highlight significant implications for real-world applications, particularly in social media moderation and automated reporting systems. The BiLSTM model, with its ability to capture bidirectional contextual relationships, demonstrates strong potential for accurately detecting cyberbullying in nuanced and complex texts, making it suitable for integration into moderation tools that require high precision and recall. However, its lower generalization on test data compared to CNN-1D suggests a need for further tuning or regulatory mechanisms, such as dropout layers or additional training data, to reduce overfitting. CNN-1D, with its balanced performance and high test accuracy, is particularly promising for scalable systems requiring efficient processing of large datasets, as its simpler architecture enables faster inference without significant accuracy trade-offs. Conversely, the poor generalization of RNNs underscores the importance of avoiding models prone to overfitting in real-world settings, where data variability is high. These findings indicate that while BiLSTM and CNN-1D are suitable candidates for deployment, additional safeguards such as active learning, continuous model retraining, and human-in-the-loop systems can enhance reliability and adaptability in detecting cyberbullying across diverse linguistic and cultural contexts.

CONCLUSION

The dataset used consists of 400 comments taken from Instagram users. This data is collected in the form of comments through the API service. In this study, a comparative analysis was carried out with a division of 70% for training data and 30% for testing data. Based on the results of the training accuracy experiment: RNN had the highest training accuracy (100%), followed by LSTM (99.25%), BiLSTM (98.51%), and CNN 1D (93.66%) while test accuracy: CNN 1D got the highest testing accuracy (84.85%), followed by BiLSTM (81.82%), LSTM (79.55%), and RNN (54.55%). These findings indicate that while BiLSTM and CNN-1D are suitable candidates for deployment, additional safeguards such as active learning, continuous model retraining, and human-in-the-loop systems can enhance reliability and adaptability in detecting cyberbullying across diverse linguistic and cultural contexts.

ACKNOWLEDGMENT

We would like to express my gratitude to Universitas Sjakhyakirti and Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia through research funding with 104/E5/PG.02.00.PL/2024; 1129/LL2/KP/PL/2024; 03/01.02/LPPM/VI/2024

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6004

REFERENCE

[1] I. Yosep, R. Hikmat, and A. Mardhiyah, "Nursing intervention for preventing cyberbullying and reducing its negative impact on students: a scoping review," *J. Multidiscip. Healthc.*, pp. 261–273, 2023.

[2] E. Buçaj and F. Haziri, "Cyberbullying and cyberstalking: Parents' role to protect their children," *Multidiscip. Rev.*, vol. 7, no. 4, p. 2024079, 2024.

[3] A. Kintonova, A. Vasyaev, and V. Shestak, "Cyberbullying and cyber-mobbing in developing countries," *Inf. Comput. Secur.*, vol. 29, no. 3, pp. 435–456, 2021.

[4] S. I. Ali and N. B. Shahbuddin, "The relationship between cyberbullying and mental health among university students," *Sustainability*, vol. 14, no. 11, p. 6881, 2022.

[5] M. F. López-Vizcaíno, F. J. Nóvoa, V. Carneiro, and F. Cacheda, "Early detection of cyberbullying on social media networks," *Futur. Gener. Comput. Syst.*, vol. 118, pp. 219–229, 2021.

[6] A. Iileka, E. Kamati, and S. F. Nyalugwe, "Understanding the Influence of Cybercrime Law Absence on Cyberbullying in Higher Institutions of Learning: A Case of the International University of Management," *J. Inf. Syst. Informatics*, vol. 5, no. 4, pp. 1779–1792, 2023.

[7] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Comput. Sci.*, vol. 181, pp. 605–611, 2021.

[8] M. M. Agüero-Torales, J. I. A. Salas, and A. G. López-Herrera, "Deep learning and multilingual sentiment analysis on social media data: An overview," *Appl. Soft Comput.*, vol. 107, p. 107373, 2021.

[9] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimed. Syst.*, vol. 29, no. 3, pp. 1839–1852, 2023.

[10] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detection," *Futur. Internet*, vol. 15, no. 5, p. 179, 2023.

[11] S. M. Fati, A. Muneer, A. Alwadain, and A. O. Balogun, "Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction," *Mathematics*, vol. 11, no. 16, p. 3567, 2023.

[12] M. Purba *et al.*, "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022.

[13] S. F. Ahmed *et al.*, "Deep learning modelling techniques: current progress, applications, advantages, and challenges," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13521–13617, 2023.

[14] A. Khan, M. M. Fouda, D.-T. Do, A. Almaleh, and A. U. Rahman, "Short-term traffic prediction using deep learning long short-term memory: Taxonomy, applications, challenges, and future trends," *IEEE Access*, vol. 11, pp. 94371–94391, 2023.

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

[16] T. Mahmud, M. Ptaszynski, and F. Masui, "Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts," *Electronics*, vol. 13, no. 9, p. 1677, 2024.

[17] K. Ambareen, S. M. Sundaram, and P. Murugesan, "Cyberbullying in social media Using Bidirectional Long Short-Term and Feed Forward Neural Network," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 6, 2024.

[18] E. U. H. Qazi, A. Almorjan, and T. Zia, "A one-dimensional convolutional neural network (1D-CNN) based deep learning system for network intrusion detection," *Appl. Sci.*, vol. 12, no. 16, p. 7986, 2022.

[19] W. Agastya and S. Huda, "Multichannel Convolutional Neural Network Model to Improve Compound Emotional Text Classification Performance," *IAENG Int. J. Comput. Sci.*, vol. 50, no. 3, 2023.

[20] M. C. Aparna and M. N. Nachappa, "Automatic Author Profiling of Nobel Prize Winners Using 1D-CNN," in *International Conference on Intelligent Systems Design and Applications*, 2023, pp. 400–411.

[21] M. Alotaibi, B. Alotaibi, and A. Razaque, "A multichannel deep learning framework for cyberbullying detection on social media," *Electronics*, vol. 10, no. 21, p. 2664, 2021.

[22] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, "Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques," *Electronics*, vol. 10, no. 22, p. 2810, 2021.

[23] T. H. H. Aldhyani, M. H. Al-Adhaileh, and S. N. Alsubari, "Cyberbullying identification system based deep learning algorithms," *Electronics*, vol. 11, no. 20, p. 3273, 2022.

[24] V. L. Paruchuri and P. Rajesh, "CyberNet: a hybrid deep CNN with N-gram feature selection for cyberbullying detection in online social networks," *Evol. Intell.*, vol. 16, no. 6, pp. 1935–1949, 2023.

[25] P. K. Roy and F. U. Mali, "Cyberbullying detection using deep transfer learning," *Complex Intell. Syst.*, vol. 8, no. 6, pp. 5449–5467, 2022.

[26] S. D. Anggita and F. F. Abdulloh, "Optimasi Algoritma Support Vector Machine Berbasis PSO Dan Seleksi Fitur Information Gain Pada Analisis Sentimen," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 52–57, 2023.