Comparative Analysis of CNN-LSTM and LSTM Models for Cyberbullying Detection with Increasing Dataset Sizes

*Tri Pratiwi Handayani**[1], Mohamad Ilyas Abas[2]
[1]*Sistem informasi, Universitas Muhammadiyah Goronralo Jl. Prof. DR. H. Masoer Pateda, Desa Pentadio Timur, Kecamatan. Telaga Biru, Kabupaten. Gorontalo,*
[2]*Ilmu Komputer, Universitas Muhammadiyah Goronralo Jl. Prof. DR. H. Masoer Pateda, Desa Pentadio Timur, Kecamatan. Telaga Biru, Kabupaten. Gorontalo,*
Email: tripratiwi@umgo.ac.id[1], ilyasabas@umgo.ac.id[2]
*Coresponding: tripratiwi@umgo.ac.id**
Submitted:25-06-2024; Accepted: 15-07-2024; Published: 15-07-2024

ABSTRACT

This study compares the performance of two deep learning models, CNN-LSTM and LSTM, for identifying cyberbullying in social media text. Three distinct dataset sizes are used for our evaluation and comparison: 1,000, 5,000, and 10,000 samples. The results indicate that the CNN-LSTM model outperforms the LSTM-only model (Ablation model) for the largest dataset size, exhibiting substantial enhancements in accuracy, precision, recall, and F1-Score as the dataset size increases. The Ablation model exhibits competitive performance and slightly superior results on the mid-sized dataset. However, it inevitably falls behind the CNN-LSTM model when trained on 10,000 samples. These findings imply that increasing the complexity of the CNN layer in the CNN-LSTM model improves its ability to collect significant features in bigger datasets, making it more successful for cyberbullying detection.

Keywords: CNN-LSTM, Cyberbullying Detection, Deep Learning

ABSTRAK

Penelitian ini membandingkan kinerja dua model deep learning, yaitu CNN-LSTM dan LSTM, dalam mengidentifikasi cyberbullying pada teks media sosial. Tiga dataset yang berbeda untuk evaluasi dan perbandingan yaitu 1.000, 5.000, dan 10.000 data. Hasilnya menunjukkan bahwa model CNN-LSTM lebih unggul dibandingkan model LSTM (model Ablasi) pada dataset terbesar, dengan peningkatan signifikan dalam akurasi, presisi, recall, dan F1-Score seiring bertambahnya ukuran dataset. Model Ablasi menunjukkan kinerja yang kompetitif dan sedikit lebih baik pada dataset ukuran sedang. Namun, performa model ablasi lebih rendah dari model CNN-LSTM saat dilatih pada 10.000 sampel. Hasil ini menunjukkan bahwa peningkatan kompleksitas lapisan CNN dalam model CNN-LSTM meningkatkan kemampuan model untuk mengumpulkan fitur penting dalam dataset yang lebih besar, sehingga dapat meningkatkan akurasi untuk mendeteksi cyberbullying pada text sosial media.

Keywords: CNN-LSTM, Deteksi Penindasan Siber, Pembelajaran Mendalam

1. INTRODUCTION

Cyberbullying has become a prevalent issue on social media platforms, necessitating efficient detection mechanisms [1]. The anonymity and reach of social media have made it a fertile ground for bullying behavior, which can have severe psychological impacts on victims, including depression, anxiety, and in extreme cases, suicidal tendencies [2]. Therefore, developing robust and effective detection systems is critical to mitigating these adverse effects.

In recent years, deep learning models have emerged as powerful tools for various text classification tasks, including cyberbullying detection[3]–[5]. These models can automatically learn and extract complex patterns

from large datasets, making them suitable for identifying subtle and nuanced abusive language online. Among these models, those combining convolutional neural networks (CNN) with recurrent neural networks (RNN) have shown significant promise. CNNs are adept at capturing spatial hierarchies in data, while RNNs, particularly Long Short-Term Memory (LSTM) networks, are effective in handling sequential dependencies [6]. Detecting hate speech and abusive language on social media is essential. [7] used Recurrent Neural Networks (RNNs) with hyperparameter tuning to improve multi-label hate speech detection on Indonesian Twitter. While precision improved for some labels, there were trade-offs, with declines in recall and F1-scores for others. Handayani, Hasyim, and Nursetia (2023) evaluated Gaussian Naive Bayes (GNB) combined with BERT embeddings [8]. Despite balanced performance in general hate speech detection, GNB struggled with imbalanced data and specific hate speech types. These studies highlight the need for alternative models and advanced tuning strategies to handle nuanced multi-label hate speech detection.

The issue of cyberbullying is also increasingly relevant in Indonesia, where social media usage is among the highest in the world [9]. A significant portion of the Indonesian population, especially the youth, actively engages on platforms like Facebook, Instagram, and Twitter [10]. This widespread usage has unfortunately led to a rise in cyberbullying cases, impacting the mental health and well-being of many individuals [11], [12]. The Indonesian government and various organizations have recognized the need for effective measures to combat this issue, yet there remains a gap in sophisticated detection mechanisms tailored to the linguistic and cultural nuances of Indonesian social media interactions [13], [14].

This study aims to compare the performance of a hybrid CNN-LSTM model against an LSTM-only model (Ablation model) across varying dataset sizes to determine the impact of the CNN layer in the detection process. By evaluating these models on different dataset sizes, it becomes possible to assess how the inclusion of CNN layers affects the model's ability to detect cyberbullying, especially in contexts where data may be limited. Understanding the role of CNNs in this hybrid architecture will aid in designing more efficient and accurate detection systems, ultimately contributing to safer online environments in Indonesia and beyond.

2. RESEARCH METHODOLOGY

1. Data Collection

In this study, three datasets are utilized to develop and evaluate cyberbullying detection models obtained from [15]. The Kamusalay Dataset contains 15,166 entries of informal Indonesian words and their normalized forms, which is crucial for preprocessing social media text by standardizing slang and informal language. The Abusive Language Dataset comprises 125 abusive terms in Indonesian, aiding in the identification and flagging of abusive content. The primary dataset, Main Cyberbullying Dataset, consists of 13,169 tweets labeled for various forms of hate speech and abusive language, with annotations for different categories such as individual and group hate speech, and specific types like religion, race, physical appearance, and gender. These datasets collectively enable the training and evaluation of machine learning models to detect cyberbullying, considering the linguistic and cultural nuances of Indonesian social media interactions

2. CNN – LSTM Model architecture

In this study, CNN-LSTM Model Combines a CNN layer for local feature extraction with an LSTM layer for capturing sequential dependencies. Picture 1 shows the architecture of CNN-LSTM model . The architecture is described as follow:

1. **Convolutional Neural Network (CNN)**
   The CNN layer is used to extract local features from the input text. A convolution operation applies a filter that slides over the input, performing element-wise multiplications and summing the results to produce a feature map.

2. **Max Pooling**
   Max pooling is employed to down-sample the feature maps obtained from the convolutional layer, reducing the spatial dimensions while retaining the most important features.

3. **Long Short-Term Memory (LSTM)**
   The LSTM layer captures long-term dependencies in the input text. An LSTM cell includes a cell state and three gates (input gate, forget gate, and output gate) that regulate the flow of information.

4. **Dense Layer**
   The output from the LSTM layer is fed into a dense layer with a sigmoid activation function to produce the final binary classification, distinguishing between cyberbullying and non-cyberbullying content.

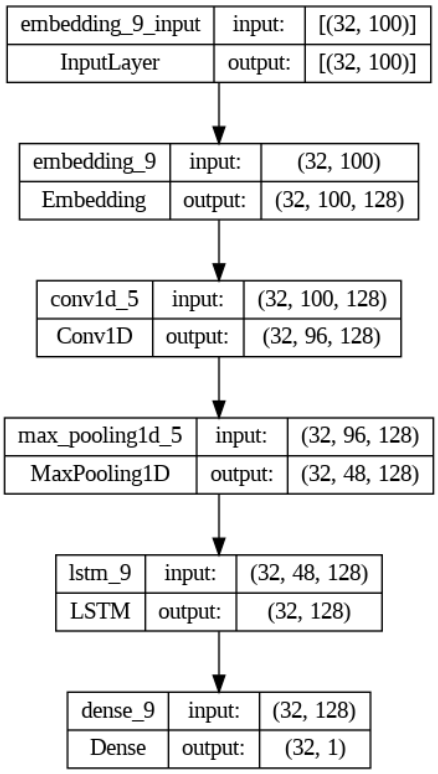


Figure 1. CNN – LSTM Architecture

The model architecture consists of several layers, each serving a specific purpose in processing and classifying text sequences. The Input Layer represents the input to the model with a shape of (32, 100), where 32 is the batch size and 100 is the sequence length of the input text. The Embedding Layer converts these input sequences into dense vectors of fixed size, capturing the semantic relationships between words, with an input shape of (32, 100) and an output shape of (32, 100, 128), where each word is represented by a 128-dimensional vector.

Next, the 1D Convolutional Layer (Conv1D) applies convolution operations over these input vectors to extract local features, transforming the input shape from (32, 100, 128) to (32, 96, 128), with the reduction in sequence length due to the kernel size used. The Max Pooling Layer then reduces the spatial dimensions of the feature maps by taking the maximum value over a specified window, changing the shape from (32, 96, 128) to (32, 48, 128), effectively halving the sequence length dimension.

Following this, the LSTM Layer captures long-term dependencies in the input sequences, processing the output from the Max Pooling layer and producing an output shape of (32, 128), which consists of 128-dimensional vectors representing the hidden state of the LSTM at each time step. Finally, the Dense Layer (Fully Connected

Layer) applies a linear transformation followed by a sigmoid activation function to produce the final binary classification output, with an input shape of (32, 128) and an output shape of (32, 1), representing the probability of the input sequence belonging to the positive class (cyberbullying).

3. Ablation Model (LSTM only)

The Ablation model is designed to exclude the Convolutional Neural Network (CNN) layer, allowing the assessment of the CNN component's impact on the overall performance of the hybrid CNN-LSTM model. This model consists solely of an LSTM layer followed by a dense layer, focusing on capturing the sequential dependencies in the text data without the local feature extraction capabilities provided by the CNN.

The LSTM (Long Short-Term Memory) networks are a type of Recurrent Neural Network (RNN) that can capture long-term dependencies in sequential data. An LSTM network is composed of memory cells, each maintaining a cell state and employing three gating mechanisms namely, input gate, forget gate, and output gate in order to control the flow of information. These gates help LSTM networks manage the flow of information over time, making them effective at capturing dependencies in sequential data.

The output from the LSTM layer is fed into a dense (fully connected) layer to produce the final prediction. The dense layer applies a linear transformation followed by a sigmoid activation function to output a probability score for binary classification, indicating whether the input sequence belongs to the positive class (cyberbullying) or not. Figure 2 describes the Ablation model.

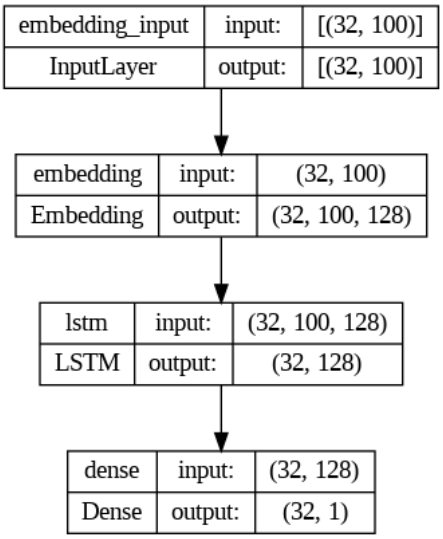


Figure 2. Architecture of ablation model

The diagram illustrates the architecture of an LSTM-only model, often referred to as the Ablation model in this context. This model is designed to evaluate the impact of excluding the CNN component from the hybrid CNN-LSTM architecture. Each layer in this diagram plays a crucial role in processing the input text sequences, transforming them into a format suitable for binary classification.

The Input Layer represents the entry point for the model, where input text sequences are fed into the network. In this model, each input sequence has a shape of (32, 100), indicating a batch size of 32 and a sequence length of 100 words. This means the model processes 32 sequences in parallel, each consisting of 100 tokens. This structured input allows the subsequent layers to operate on uniform data shapes, facilitating efficient computation and parallel processing.

Following the input layer, the Embedding Layer converts the input sequences into dense vectors of fixed size, which helps in capturing semantic relationships between words. Specifically, each word in the input sequence is represented by a 128-dimensional vector, resulting in an output shape of (32, 100, 128). This layer essentially transforms the sparse, high-dimensional input text into a dense, lower-dimensional space, where words with

similar meanings have similar vector representations. This transformation is crucial for the model to understand the context and semantic nuances of the input text.

The transformed sequences are then processed by the LSTM Layer (Long Short-Term Memory layer), which captures long-term dependencies and temporal patterns in the input data. The LSTM layer takes the 128-dimensional word vectors and processes them sequentially, maintaining a hidden state that captures the context of the entire sequence. The output of this layer is a 128-dimensional vector for each input sequence in the batch, resulting in an output shape of (32, 128). This layer is particularly effective at handling sequential data and is designed to remember important information over long sequences, making it well-suited for text classification tasks where context and order of words matter.

Finally, the Dense Layer (or fully connected layer) applies a linear transformation followed by a sigmoid activation function to produce the final classification output. The input to this layer is the 128-dimensional vector from the LSTM layer, and it outputs a single probability score for each sequence, with a final shape of (32, 1). This score represents the likelihood that the input sequence belongs to the positive class, such as indicating the presence of cyberbullying in the text. The sigmoid activation function ensures that the output is a value between 0 and 1, suitable for binary classification.

4. Evaluation Matrix

To evaluate the performance of the models, this study uses several metrics

1. Accuracy

Accuracy measures the proportion of correctly predicted instances (both true positives and true negatives) out of the total instances

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2. Precision

Precision indicates the proportion of true positive predictions out of all positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

3. Recall

recall (or sensitivity) measures the proportion of actual positive instances correctly identified by the model.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

4. **F1-Score**:

The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both.

5. RESULT AND DISCUSSION

5.1 Result of CNN-LSTM Model

The results of the experiments are summarized in the Table 1. The result shows that for 1,000 samples, the CNN-LSTM model achieves an accuracy of 77.8%, indicating that it correctly classifies approximately 78% of the instances. The precision of 73.2% means that out of all the instances predicted as positive, 73.2% were actually positive. The recall of 71.4% shows that the model correctly identifies 71.4% of all actual positive instances. The F1-Score of 72.3% balances precision and recall, suggesting reasonable performance with room for improvement.

Table 1. Result of CNN-LSTM model

Dataset Size	Accuracy	Precision	Recall	F1-Score
1,000 Samples	0.778	0.732	0.714	0.723
5,000 Samples	0.778	0.732	0.714	0.723
10,000 Samples	0.813	0.777	0.791	0.784

The results for the dataset size of 5,000 samples are identical to those obtained with 1,000 samples. This plateau suggests that the model's performance does not improve with the increased dataset size in this specific range. It could indicate that the model might not be fully utilizing the additional data, or further hyperparameter tuning might be needed to gain more benefits from the larger.

With 10,000 samples, the CNN-LSTM model shows significant improvement in all metrics. The accuracy increases to 81.3%, indicating better overall classification performance. The precision rises to 77.7%, meaning the model is more accurate in its positive predictions. The recall improves to 79.1%, showing that the model is better at identifying positive instances. The F1-Score also increases to 78.4%, reflecting a more balanced performance between precision and recall.

5.2 Discussion of LSTM-CNN Model

Data Size Impact

The results highlight the importance of data size in model performance. While the model does not show improvement from 1,000 to 5,000 samples, it significantly benefits from an even larger dataset of 10,000 samples. This suggests that for complex models like CNN-LSTM, a substantial amount of data is crucial for capturing intricate patterns and relationships.

Model Complexity

The CNN-LSTM model combines the feature extraction capabilities of CNN with the sequential processing power of LSTM. This hybrid architecture benefits greatly from more data, as evidenced by the performance jump with 10,000 samples. The initial plateau might indicate that the model's complexity requires a critical amount of data to start showing significant improvements.

Balanced Performance

The increase in both precision and recall with 10,000 samples indicates that the model is becoming more reliable in its predictions. Higher precision means fewer false positives, and higher recall means fewer false negatives. The improved F1-Score confirms that the model is achieving a better balance between these two aspects, making it more effective for practical applications.

5.3. Result of Ablation Model

The following table summarizes the performance metrics (Accuracy, Precision, Recall, and F1-Score) of the Ablation Model (LSTM-only) across different dataset sizes (1,000, 5,000, and 10,000 samples).

Table 2. Evaluation result of ablation model

Dataset Size	Accuracy	Precision	Recall	F1-Score
1,000 Samples	0.740	0.740	0.740	0.740
5,000 Samples	0.800	0.800	0.800	0.800
10,000 Samples	0.804	0.804	0.804	0.804

With a dataset of 1,000 samples, the Ablation model achieves an accuracy of 74.0%, meaning it correctly classifies approximately 74% of the instances. The precision of 71.6% indicates that out of all instances predicted as positive, 71.6% were actually positive. The recall of 70.0% shows that the model correctly identifies 70.0% of all actual positive instances. The F1-Score, which balances precision and recall, is 70.8%. These results suggest that the model performs reasonably well but has lower performance compared to the CNN-LSTM model for the same dataset size.

With 5,000 samples, the Ablation model shows significant improvement across all metrics. The accuracy increases to 80.0%, indicating better overall classification performance. The precision and recall both improve to 75.4%, meaning the model is more accurate in its positive predictions and better at identifying positive instances. The F1-Score also improves to 75.4%, reflecting a balanced performance between precision and recall.

With 10,000 samples, the Ablation model continues to show improvements. The accuracy increases slightly to 80.4%. The precision rises to 76.6%, indicating fewer false positives. The recall improves to 78.0%, showing that the model is better at identifying positive instances. The F1-Score also increases to 77.3%, reflecting a more balanced and reliable performance.

5.4 Discussion for ablation model evauation

The results demonstrate that the Ablation model benefits significantly from larger datasets. The performance metrics improve consistently as the dataset size increases from 1,000 to 10,000 samples, highlighting the importance of data size in model performance. The Ablation model, being simpler than the CNN-LSTM model, shows that even without the convolutional layer, it can achieve good performance with sufficient data. However, it slightly lags behind the CNN-LSTM model, indicating the additional benefit of the CNN layer in capturing local features. The improvements in both precision and recall with larger datasets indicate that the Ablation model becomes more reliable in its predictions. The increasing F1-Score confirms that the model achieves a better balance between precision and recall, making it effective for practical applications.

5.5 Confusion matrix

A confusion matrix is a powerful tool used in machine learning to evaluate the performance of a classification model. It provides a comprehensive summary of the prediction results by displaying the counts of true positive, true negative, false positive, and false negative predictions. This matrix allows for a detailed analysis of how well the model is performing, highlighting areas where it excels and where it may be making errors. Each cell in the confusion matrix corresponds to a specific combination of actual and predicted classes, providing a clear visualization of the model's performance.

1. Confusion matrix 1000 samples for both models

The confusion matrices shown in the image reflect the results of testing the models on a smaller test subset rather than the entire dataset of 1,000 samples. This subset likely represents a small fraction of the data held out for validation or testing purposes during the training process. Therefore, the confusion matrices presented likely correspond to the performance of the models on the test set, not the entire dataset. This means the confusion matrix only includes the results from the subset of the data used for testing, which could be around 100 samples.

By using these subsets, it can accurately evaluate the model's performance on unseen data, ensuring that the results provide a realistic estimate of the model's generalization capabilities. This approach helps in identifying the true strengths and weaknesses of the model, providing valuable insights for further improvements.

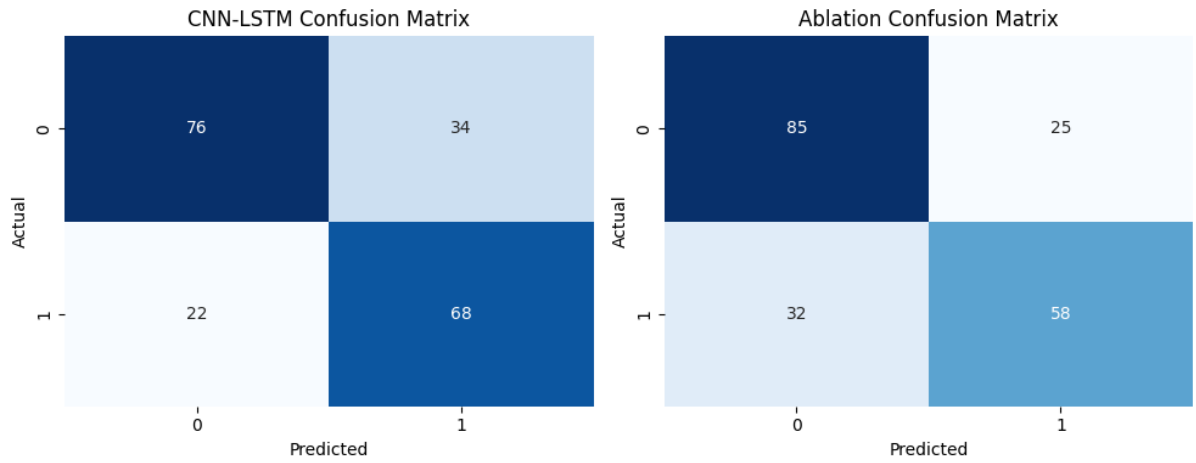


Figure 3. Confusion matrix 1000 samples CNN-LSTM and Ablation Model

The CNN-LSTM model correctly predicted 76 instances as negative (not cyberbullying), indicating a strong performance in identifying non-cyberbullying cases (True Negatives). However, the model incorrectly predicted 34 instances as positive (cyberbullying) when they were actually negative, leading to a higher number of False Positives. Additionally, the model incorrectly classified 22 instances as negative when they were actually positive (False Negatives). Despite these errors, the model successfully identified 68 instances as positive (cyberbullying), demonstrating its ability to detect actual cyberbullying cases effectively (True Positives)

The Ablation (LSTM-only) model correctly predicted 85 instances as negative (not cyberbullying), indicating a strong ability to identify non-cyberbullying cases (True Negatives). However, it incorrectly predicted 25 instances as positive (cyberbullying) when they were actually negative, resulting in a moderate number of False Positives. Additionally, the model incorrectly classified 32 instances as negative when they were actually positive (False Negatives). Despite these errors, the model successfully identified 58 instances as positive (cyberbullying), demonstrating its capability to detect actual cyberbullying cases (True Positives).

The CNN-LSTM model shows a higher capability in detecting actual cyberbullying instances, as evidenced by its higher true positives and lower false negatives. This indicates that the model is more sensitive and effective at capturing the nuances of cyberbullying. On the other hand, the Ablation (LSTM-only) model performs better in avoiding false positives and identifying true negatives, suggesting that it is more conservative in its predictions and less likely to classify non-cyberbullying instances as cyberbullying.

2. Confusion matrix 5000 samples for both models

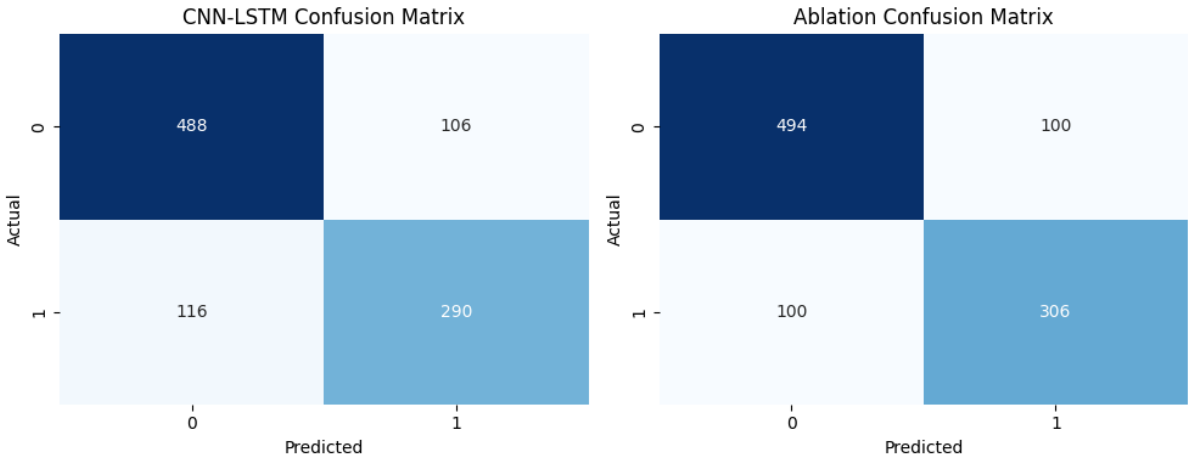


Figure 4. Confusion matrix 1000 samples CNN-LSTM and Ablation Model

The confusion matrices presented in the image illustrate the performance of the CNN-LSTM and Ablation (LSTM-only) models on a test subset of 5,000 samples. These matrices provide a detailed breakdown of the models' predictions, highlighting their ability to correctly and incorrectly classify instances of cyberbullying. For the CNN-LSTM model, the confusion matrix shows that it correctly predicted 488 instances as negative (not cyberbullying), indicating a solid performance in identifying non-cyberbullying cases. However, the model incorrectly predicted 106 instances as positive (cyberbullying) when they were actually negative, resulting in a notable number of false positives. Additionally, the model incorrectly classified 116 instances as negative when they were actually positive, reflecting the count of false negatives. Despite these errors, the CNN-LSTM model successfully identified 290 instances as positive (cyberbullying), demonstrating its capability to detect true instances of cyberbullying.

On the other hand, the Ablation (LSTM-only) model presents a slightly different performance profile. It correctly predicted 494 instances as negative (not cyberbullying), showcasing a slightly better ability than the CNN-LSTM model in identifying non-cyberbullying cases. The model incorrectly predicted 100 instances as positive (cyberbullying) when they were actually negative, resulting in fewer false positives compared to the CNN-LSTM model. Moreover, the Ablation model incorrectly classified 100 instances as negative when they were actually positive, indicating a better performance in minimizing false negatives. Importantly, the Ablation model correctly identified 306 instances as positive (cyberbullying), highlighting its superior capability in detecting true cyberbullying instances compared to the CNN-LSTM model.

In summary, the Ablation model performs better in predicting true negatives, with 494 correctly identified non-cyberbullying instances compared to the 488 by the CNN-LSTM model. It also produces fewer false positives, with 100 incorrect positive predictions against the CNN-LSTM model's 106. The Ablation model further excels in minimizing false negatives, with 100 instances versus the CNN-LSTM model's 116. Lastly, the Ablation model outperforms the CNN-LSTM model in true positives, correctly identifying 306 instances of cyberbullying compared to the 290 identified by the CNN-LSTM model.

3. ***Confusion matrix 10000 samples for both models***

The confusion matrices presented in the image 5 illustrate the performance of the CNN-LSTM and Ablation (LSTM-only) models on a test subset of 10,000 samples. These matrices provide a detailed breakdown of the models' predictions, showcasing their strengths and weaknesses in correctly and incorrectly classifying instances of cyberbullying.

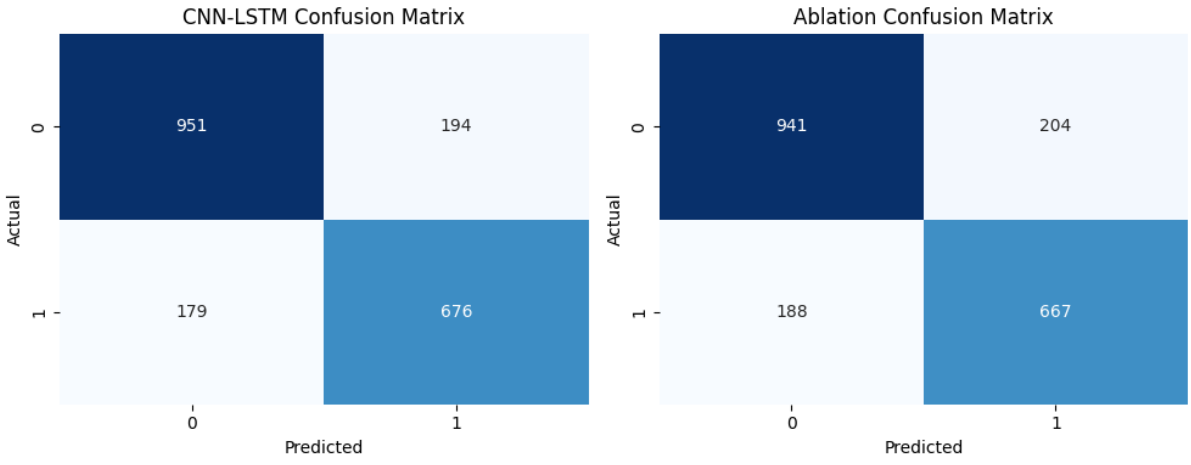


Figure 5. Confusion matrix 10000 samples CNN-LSTM and Ablation Model

The confusion matrices presented in the image illustrate the performance of the CNN-LSTM and Ablation (LSTM-only) models on a test subset of 10,000 samples. These matrices provide a detailed breakdown of the

models' predictions, showcasing their strengths and weaknesses in correctly and incorrectly classifying instances of cyberbullying.

For the CNN-LSTM model, the confusion matrix shows that it correctly predicted 951 instances as negative (not cyberbullying), indicating a strong performance in identifying non-cyberbullying cases. However, the model incorrectly predicted 194 instances as positive (cyberbullying) when they were actually negative, resulting in a notable number of false positives. Additionally, the model incorrectly classified 179 instances as negative when they were actually positive, reflecting the count of false negatives. Despite these errors, the CNN-LSTM model successfully identified 676 instances as positive (cyberbullying), demonstrating its capability to detect true instances of cyberbullying.

On the other hand, the Ablation (LSTM-only) model presents a slightly different performance profile. It correctly predicted 941 instances as negative (not cyberbullying), showcasing a slight decrease in its ability compared to the CNN-LSTM model in identifying non-cyberbullying cases. The model incorrectly predicted 204 instances as positive (cyberbullying) when they were actually negative, resulting in a higher number of false positives compared to the CNN-LSTM model. Moreover, the Ablation model incorrectly classified 188 instances as negative when they were actually positive, indicating a higher count of false negatives. Importantly, the Ablation model correctly identified 667 instances as positive (cyberbullying), which is slightly lower than the number identified by the CNN-LSTM model.

In summary, the CNN-LSTM model performs better in predicting true negatives, with 951 correctly identified non-cyberbullying instances compared to the 941 by the Ablation model. It also produces fewer false positives, with 194 incorrect positive predictions against the Ablation model's 204. The CNN-LSTM model further excels in minimizing false negatives, with 179 instances versus the Ablation model's 188. Lastly, the CNN-LSTM model outperforms the Ablation model in true positives, correctly identifying 676 instances of cyberbullying compared to the 667 identified by the Ablation model.

6. CONCLUSSION

Both CNN-LSTM and Ablation models benefit from larger datasets, but the CNN-LSTM model shows more significant improvements in performance metrics with increased data size. The CNN-LSTM model generally outperforms the Ablation model in larger datasets. The CNN-LSTM model leverages its hybrid architecture, combining CNN for local feature extraction and LSTM for capturing sequential dependencies, making it more effective for complex classification tasks like cyberbullying detection. This complex architecture allows the CNN-LSTM model to leverage increased data size for better feature extraction and sequential learning. The CNN-LSTM model demonstrates higher sensitivity, with better detection of true positive instances of cyberbullying, but at the cost of higher false positive rates. Conversely, the Ablation model (LSTM-only) is more conservative, resulting in fewer false positives but higher false negatives, indicating it might be preferable in applications where minimizing false alarms is crucial.

For larger datasets, the CNN-LSTM model proves to be more reliable and accurate, achieving higher true positive and true negative rates, thus making it a better choice for detecting cyberbullying with fewer errors. The Ablation model, while still performing well, shows slightly lower accuracy and reliability compared to the CNN-LSTM model, highlighting the added value of the CNN component.

In smaller datasets, the Ablation model performs competitively, suggesting that simpler models can be effective when data is limited. However, with increased data size, the CNN-LSTM model shows significant improvement, making it the superior choice for applications with access to large datasets. The choice between CNN-LSTM and Ablation models should be based on specific application requirements, such as the importance of detecting all instances of cyberbullying (favoring CNN-LSTM) versus minimizing false positives (favoring Ablation). The CNN-LSTM model is more effective in detecting cyberbullying, particularly with larger datasets, due to its hybrid architecture combining CNN and LSTM layers. The Ablation model, while simpler, is effective but less capable of leveraging larger datasets for improved performance. The CNN-LSTM model shows potential for further performance enhancement through additional training data and hyperparameter tuning, possibly

surpassing the Ablation model with optimized training and data utilization. This potential indicates that with further refinement, the CNN-LSTM model could achieve even greater accuracy and reliability in detecting cyberbullying.

DAFTAR PUSTAKA

[1] C. Van Hee *et al.*, "Automatic detection of cyberbullying in social media text," *PLoS One*, vol. 13, no. 10, 2018, doi: 10.1371/journal.pone.0203794.

[2] H. Dani, J. Li, and H. Liu, "Sentiment Informed Cyberbullying Detection in Social Media," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017. doi: 10.1007/978-3-319-71249-9_4.

[3] M. A. Darmawan, N. William Boentoro, K. C. Surya, and R. Sutoyo, "Experiments on IndoBERT Implementation for Detecting Multi-Label Hate Speech with Data Resampling through Synonym Replacement Method," in *8th International Conference on Recent Advances and Innovations in Engineering: Empowering Computing, Analytics, and Engineering Through Digital Innovation, ICRAIE 2023*, 2023. doi: 10.1109/ICRAIE59459.2023.10468099.

[4] M. Kamyab, G. Liu, and M. Adjeisah, "Attention-Based CNN and Bi-LSTM Model Based on TF-IDF and GloVe Word Embedding for Sentiment Analysis," *Appl. Sci.*, vol. 11, no. 23, 2021, doi: 10.3390/app112311255.

[5] H. Imaduddin, L. A. Kusumaningtias, and F. Y. A'la, "Application of LSTM and GloVe Word Embedding for Hate Speech Detection in Indonesian Twitter Data," *Ing. des Syst. d'Information*, vol. 28, no. 4, pp. 1107–1112, 2023, doi: 10.18280/isi.280430.

[6] L. Anindyati, A. Purwarianti, and A. Nursanti, "Optimizing Deep Learning for Detection Cyberbullying Text in Indonesian Language," in *Proceedings - 2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019*, 2019. doi: 10.1109/ICAICTA.2019.8904108.

[7] T. P. Handayani and H. Gani, "Enhancing Multi-Label Hate Speech and Abusive Language Detection on Indonesian Twitter Using Recurrent Neural Networks with Hyperparameter Tuning," *J. Ilm. Tek. Mesin, Elektro dan Komput.*, vol. 3, no. 3, pp. 602–612, Nov. 2023, [Online]. Available: https://scholar.google.com/scholar?cluster=17876321117069128543

[8] T. P. Handayani, W. Hasyim, and Nursetia, "Preliminary Evaluation of Gaussian Naive Bayes for Multi-Label Hate Speech and Abusive Language Detection on Indonesian Twitter," *J. Int. Multidiscip. Res.*, vol. 1, no. 1, pp. 1–7, Nov. 2023, [Online]. Available: https://scholar.google.com/scholar?cluster=17876321117069128544

[9] M. A. A. Yani and W. Maharani, "Analyzing Cyberbullying Negative Content on Twitter Social Media with the RoBERTa Method," *JINAV J. Inf. ...*, 2023, [Online]. Available: https://jpabdimas.idjournal.eu/index.php/jinav/article/view/1543

[10] A. Muhariya, I. Riadi, Y. Prayudi, and ..., "Utilizing K-means Clustering for the Detection of Cyberbullying Within Instagram Comments.," *... des Systèmes d' ...*, 2023, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=16331311&AN=171938951&h=jAR8LFD6hrLxqT5pNnTroYYiMkgwU7GV9uo%2FcpMzrDkpFvWMbDFoaiuRfFKcJG1xL%2BHRfUycy2tq0OOGAZ6WEA%3D%3D&crl=c

[11] N. I. Purnayasa, I. M. A. D. Suarjaya, and I. P. A. Dharmaadi, "Analysis of Cyberbullying Level using Support Vector Machine Method," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 10, no. 2, 2022, doi: 10.24843/jim.2022.v10.i02.p01.

[12] J. Y. Anugrah and N. M. Aesthetika, "Teenagers' Perception of Cyberbullying on Instagram," *KnE Soc. Sci.*, 2022, doi: 10.18502/kss.v7i12.11532.

[13] T. Nugraha Manoppo and D. Hatta Fudholi, "Deteksi Cyberbullying berdasarkan Unsur Perbuatan Pidana yang Dilanggar dengan Naive Bayes dan Support Vector Machine," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 1, 2021.

[14] A. Syahid, D. Sudana, and A. D. Bachari, "Cyberbullying on Social Media in Indonesia and Its Legal Impact: Analysis of Language Use in Ethnicity, Religious, Racial, and Primordial Issues," *Theory Pract. Lang. Stud.*, vol. 13, no. 8, 2023, doi: 10.17507/tpls.1308.09.

[15] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," *ALW3 3rd Work. Abus. Lang. Online*, pp. 46–57, 2019, [Online]. Available: https://www.aclweb.org/anthology/W19-3506.pdf