Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1113-1130
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4176

Enhancing Cyberbullying Detection with a CNN-GRU Hybrid Model, Word2Vec, and Attention Mechanism

Kaysa Azzahra Adriana*[1], Erwin Budi Setiawan[2]

[1,2]Informatics, School of Computing, Telkom University, Bandung, Indonesia
Email: [1]kaysaazzhra@student.telkomuniversity.ac.id

Received : Dec 5, 2024; Revised : Jan 4, 2025; Accepted : Jan 6, 2025; Published : Jun 10, 2025

Abstract

Cyberbullying is an act of violence commonly committed on online platforms such as social media X, often causing psychological effects for victims. Despite prevention efforts, traditional methods for detecting cyberbullying show limited effectiveness due to the complexity of language and diversity of expressions, leading to suboptimal performance. This study aims to enhance detection accuracy by applying Convolutional Neural Networks (CNN) and Gated Recurrent Unit (GRU) with an attention mechanism to analyze textual data from tweets. The model uses Term Frequency-Inverse Document Frequency (TF-IDF) for extracting important words and Word2Vec for expanding text representation. A total of 30,084 labeled datasets from tweets on social media X were utilized. Results indicate the hybrid CNN-GRU model with attention achieved the highest accuracy of 80.96%, outperforming stand-alone CNN and GRU models. Additionally, TF-IDF and Word2Vec significantly improved model performance, with the CNN-GRU combination proving most effective for detecting cyberbullying. This study contributes to computer science by proposing a novel approach that integrates CNN, GRU, and attention mechanisms with advanced feature extraction techniques, providing a more reliable detection system for online platforms. It also highlights the potential for integrating multimodal data to further enhance future performance.

Keywords : Attention Mechanism, Convolutional Neural Network (CNN), Cyberbullying Detection, Gated Recurrent Unit (GRU), Word2Vec

1. INTRODUCTION

Social media is used by almost everyone today. Whether in the field of education, work, or entertainment, almost everyone from various parts of the world chooses social media as the medium they use [1]. With the development of time, the use of social media has become increasingly prevalent in everyday life. One of the popularly used social media is X, formerly known as Twitter. Based on data from Statista, X users in Indonesia reached 24 million users and is the country with the most users in fourth place [2]. On X social media, users can make posts commonly called posts about their thoughts or comments in the form of text, images, videos, GIFs, and others [3]. X has many benefits but some people misuse X to attack, intimidate, and other things related to cyberbullying. Cyberbullying refers to bullying, frightening or mistreating someone indirectly or through digital platforms [4]. Cyberbullying actions can be in the form of sending texts containing abusive messages, spreading personal information, bullying on online platforms or sending threatening messages. This problem must be addressed immediately, one form of problem solving that can be done is to create a cyberbullying detection system.

Cyberbullying detection research has been developed based on previous studies related to similar methods and research objects. Previous research on cyberbullying detection has used hybrid deep learning methods as has been done by Nur Wakhidah Fitri Amalia who uses the CNN-GRU method, as well as TF-IDF feature extraction and GloVe feature expansion in its detection [5]. In this study, a comparison of several methods such as CNN, GRU, CNN-GRU combination, and GRU-CNN

combination was carried out. The four methods produce accuracy values that are not much different, but the highest accuracy is obtained by the GRU method with an accuracy value of 80.58%.

In addition, research related to cyberbullying detection was also conducted by Yudi Setiawan and his colleagues using SVM and KNN machine learning algorithms [6]. The main focus of this research is to utilize the combination of n-grams on TF-IDF to improve accuracy. The results showed that the application of a combination of machine learning algorithms with TF-IDF succeeded in increasing accuracy to 95.5%, which showed its effectiveness in detecting cyberbullying.

While cyberbullying detection focuses more on identifying unfavorable actions, sentiment analysis and other fields use similar approaches in improving model performance and relevance of extracted features. One of the studies related to sentiment analysis on movie reviews used Word2Vec combined with LDA. The main focus in this research is to use the skip-gram model in Word2Vec to improve the feature dictionary previously created by LDA analysis [7]. The results obtained show that the application of Word2Vec successfully increases the relevance of the extracted words, and can produce better and more targeted sentiment analysis.

In addition to feature extraction techniques, attention mechanisms have also proven effective in improving classification accuracy in various fields, including hate content detection and stock price prediction. Research on the recognition of hateful content in Arabic text was conducted by Abeer Aljohani and his colleagues using Convolutional Neural Network (CNN) and attention mechanism [8]. The main focus in this research is to utilize CNN in extracting features in the text and applying attention mechanisms to increase the accuracy value. The accuracy result obtained was 97.83%, indicating that the application of the combined model proved effective in increasing the accuracy of the model.

Another research using attention mechanism was conducted by Qingyang Liu and his colleagues to predict stock prices [9]. Attention mechanism in this study was combined with the LSTM model to create a more optimal model. The results obtained show that the ATT-LSTM model is able to achieve lower Mean Absolute Error, Mean Absolute Percentage Error, and Root Mean Square Error values compared to the LSTM model without attention, which indicates that the combination of the model is effective in improving model accuracy.

The combination of attention-mechanism with GRU and ResNet was also done by Gaurav and Pratistha Mathur for automatic image captioning [10]. In that study, attention-mechanism was used to create a proposed model that achieved a higher BLEU score compared to other models that used LSTM as a decoder. This finding shows that the combination of these models is effective in improving accuracy in generating image descriptions.

Although the application of techniques such as Word2Vec and attention mechanism is proven to be effective in improving model accuracy in other fields, the application of their combination in cyberbullying detection remains unexplored. While the CNN-GRU hybrid model has been applied in this field, its performance has not been able to surpass other approaches, which suggests there is still potential for improvement.

The main contribution of this study is to enhance the CNN-GRU hybrid model by applying attention mechanism and Word2Vec for feature expansion, which has been proven to produce good results in previous studies in other fields. This approach aims to improve the sensitivity and overall performance of the model in detecting patterns related to cyberbullying behavior, while also contributing to the development of more reliable detection systems for online platforms and paving the way for integrating multimodal data to further improve performance in the future.

2. METHOD

This study went through several stages of a structured research flow. These stages can be seen in Figure 1.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1113-1130
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4176

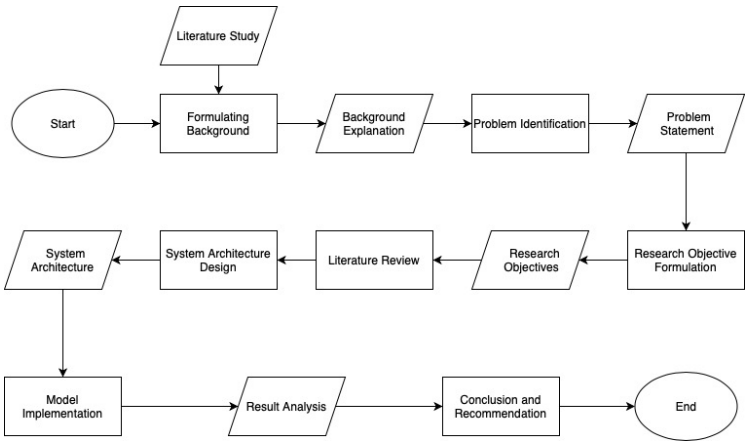

Figure 1. Stages of the research flow

This study applies a hybrid combination of Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) models with the application of attention mechanisms to the detection of cyberbullying on X social media. The data is obtained through the crawling process on social media X, which then goes through the labeling and preprocessing stages. Feature extraction uses TF-IDF to capture word representation based on occurrence weight, while feature expansion with Word2Vec is applied to enrich semantic context. Model performance evaluation is performed using confusion matrix to calculate accuracy, precision, recall, and F1-score. The system architecture steps taken in this study are shown in Figure 2.

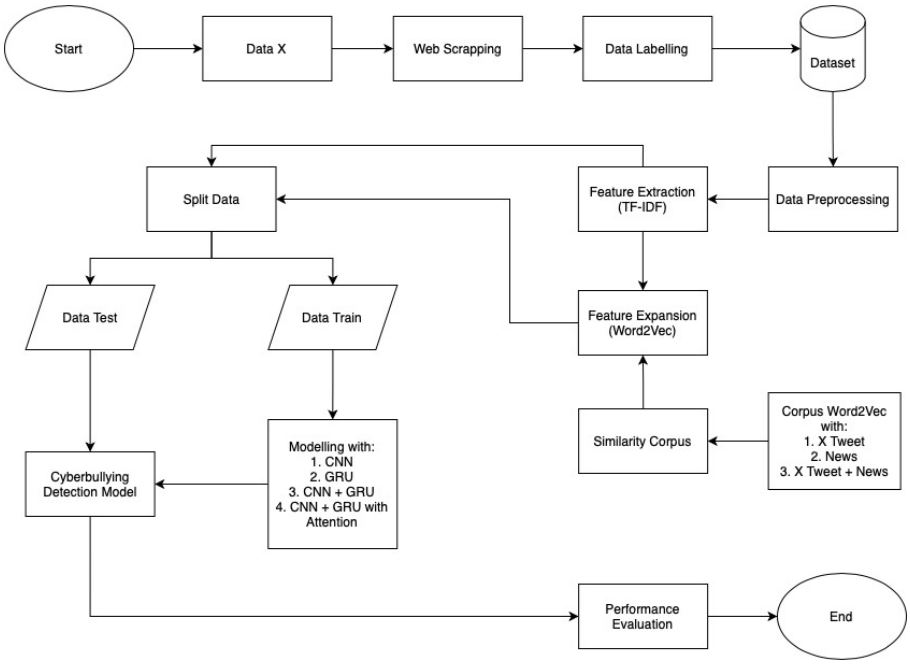

Figure 2. Architecture of the cyberbullying tweet detection system

2.1. Data Crawling

Data crawling is a commonly used method to collect data from various social media platforms. In this study, data is collected from social media X (previously known as Twitter) using data collection

techniques by utilizing the Application Programming Interface (API) of the X application. This process was carried out to collect text data in Indonesian relevant to the research topic, with crawling performed around July to August 2024. The total data collected amounted to approximately 30,000 entries, which will be used as the main dataset in the analysis and testing of the cyberbullying detection model. Table 1 shows the amount of data for each keyword.

Table 1. List of keywords in the dataset

Keyword	Total
Bangsat	634
Goblok	6,572
Jelek	3,663
Tolol	6,874
Banci	2,855
Kontol	953
Gendut	1,630
Tolol	6,874
Lonte	3,610
Bego	926
Jumlah	30,084

2.2. Data Labelling

In this process, data labeling is carried out on the data from the previous crawling results. Data is classified into two classes, namely data that includes cyberbullying and data that does not include cyberbullying. Labeling will use a binary form with a value of 1 labeling data that includes cyberbullying anda value of 0 to label data that does not include cyberbullying. Table 2 is an example of the data labeling process.

Table 2. Example of data labelling

Tweet	Label
orang tolol goblok idiot dunia	1
tanya bodoh nonton paham	0

2.3. Data Pre-processing

The tweets taken from social media X are unstructured text data. Therefore, the crawled data that has been labeled is upgraded through several processing stages to improve accuracy in classification. The first stage is data cleaning, where the data is cleaned from elements or symbols such as emoticons, numbers, links,and usernames. The next stage is case folding, where the entire text is converted into lowercase letters to homogenize the sentence format and facilitate the classification process. Next, the data normalization process is carried out to normalize non-standard words into a form that is appropriate and easy to understand. This study uses the Literature normalization module. The next stage is tokenization, which is the process of breaking the text into smaller word units to help the model understand the context of the data better and allow exploration of each word in the sentence. After that, stopword removal is performed to eliminate words that are not needed in the detection process. The next stage is stemming, which is the process of removing affixes on words to leave only the base form or main word. The last stage is detokenization, where the tokenized word units are recombined into sentence form.

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4176

2.4. TF-IDF Feature Extraction

This study uses TF-IDF as a feature extraction method to measure important and common words in documents [11]. TF or term frequency calculates the frequency of occurrence of words in documents, while IDF or inversedocument frequency identifies unique words [12]. For sentence processing, this study applies N-Gram, with theparameter 'n' determining the length of the segment. The result of N-Gram splitting is then weighted using TF-IDF. The formula to determine the magnitude of the IDF value is shown in Equation (1).

$$IDF_j = \log\left(\frac{D}{df_j}\right) \quad (1)$$

The IDF formula measures the importance of a word by comparing the total number of documents D to the number of documents containing that word df_j. The less frequently a word appears in a document, the higher its IDF value, which indicates that the word has more specific information.

CNN and GRU models require a consistent document length, so in the feature extraction stage a limit onthe maximum number of features used is applied. The use of max features in TF-IDF aims to limit the number of words considered as features, so that the model is not too complex, reduces resource consumption, and prevents overfitting, especially when working with neural network-based models such as CNN and GRU.

2.5. Word2Vec Feature Expansion

Word2Vec is one of the effective word embedding techniques in presenting the relationship between words by converting words into vectors of a certain dimension [13]. It uses two main algorithms, namely Continuous Bag-of-Words (CBOW) and Skip-Gram [14], each of which aims to predict words based on their surrounding context. In this study, the Skip-Gram algorithm is used to predict surrounding words based on a single word, which aims to improve word detection accuracy by considering the meaning and relationship between words in a document [15].

To train Word2Vec, the corpus was processed using specific parameters: 5 window size to determine the word context, 3 minimum word frequency to filter out infrequent words, and 4 worker threads to optimize the computation process. In addition, Skip-Gram and negative sampling (hs=0) algorithms were applied to improve training efficiency on large vocabularies. Figure 3 shows the architecture of the Word2Vec Skip-Gram algorithm.

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4176

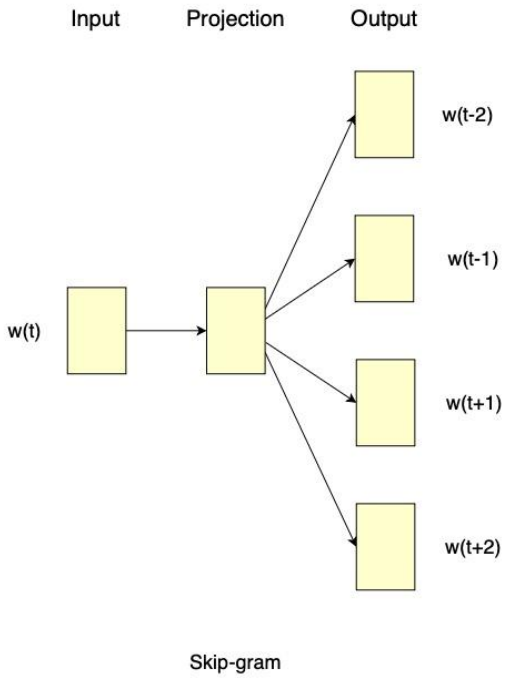

Figure 3. Word2Vec Skip-gram Architecture

For feature expansion, a corpus was created with three different data sources, namely Tweet data, Indonews data, and Tweet+Indonews data. This feature expansion also involves the use of N- Grams to calculate the similarity between sentences, which allows more precise and relevant identification ofword relationships in the context of the data used [16]. Table 3 shows the words that have the closest similarity to the word "b*go" in the Tweet+Indonews corpus.

Table 3. Semantic terms based on the word "b*go"

Rank 1	Rank 2	Rank 3
t*lol	g*blok	d*ngo

2.6. Data Splitting

In this study, the data is separated into two parts: test data and training data. Data is divided into three scenarios: 90:10, which consists of 90% training data and 10% test data; 80:20, which consists of 80% training data and 20% test data; and 70:30, which consists of 70% training and 30% test data. The 90:10 data ratio was used because based on the results of the tests conducted on the three scenarios, the best accuracy was obtained in the 90:10 data ratio.

2.7. Convolutional Neural Network (CNN)

One of the main models in this study is the Convolutional Neural Network (CNN). CNN in this study consists of several layers that are specifically designed to process sequential data. The CNN structure used can be seen in Figure 4 proposed by LeCun which consists of a convolutional layer, pooling layer, and fully connected layer [17].

Vol. 6, No. 3, Juni 2025, Page. 1113-1130
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4176

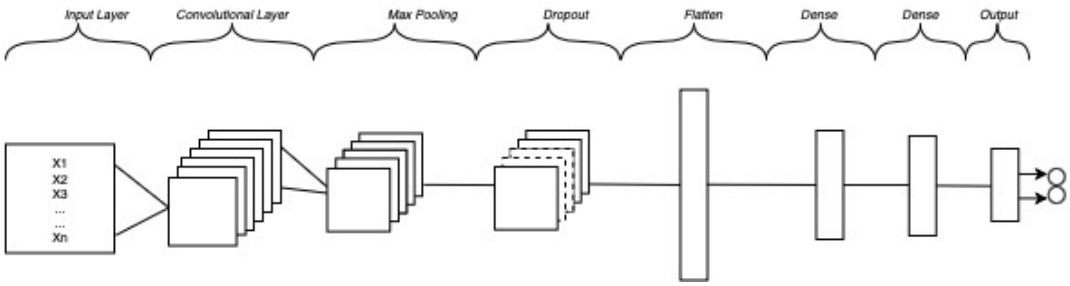

Figure 4. CNN architecture

Convolutional layer has the function of detecting important features in the input data. The pooling layer reduces the dimensionality of the output produced by the convolutional layer, reduces the number of parameters to be processed and also plays an important role in reducing computational complexity. [18]. Meanwhile, the dropout layer is used to reduce overfitting by randomly disabling neurons during training.

In CNN terms, a fully connected layer (or dense layer) refers to a layer where each neuron is connected to all neurons in the previous layer. In addition, the flatten layer, dense layer, and output layer are used to transform the data into a one-dimensional vector, capture complex patterns in the data, and generate the final prediction in the form of a binary classification, each with complementary functions and roles in the modeling process

2.8. Gated Recurrent Unit (GRU)

One type of artificial neural network developed to address the issues of gradient loss and short-term memory in sequential data modeling is the Gated Recurrent Unit (GRU) [19]. GRU has two main gates: the updategate and the reset gate [20]. The reset gate governs how much information from the previous step will be retained or forgotten [21]. The output value of the reset gate ranges between 0 and 1, where values close to 0 indicate that most of the previous information can be ignored, while values close to 1 indicate information that must be retained [22]. In this model, the ReLU activation function is used for the GRU main cell.

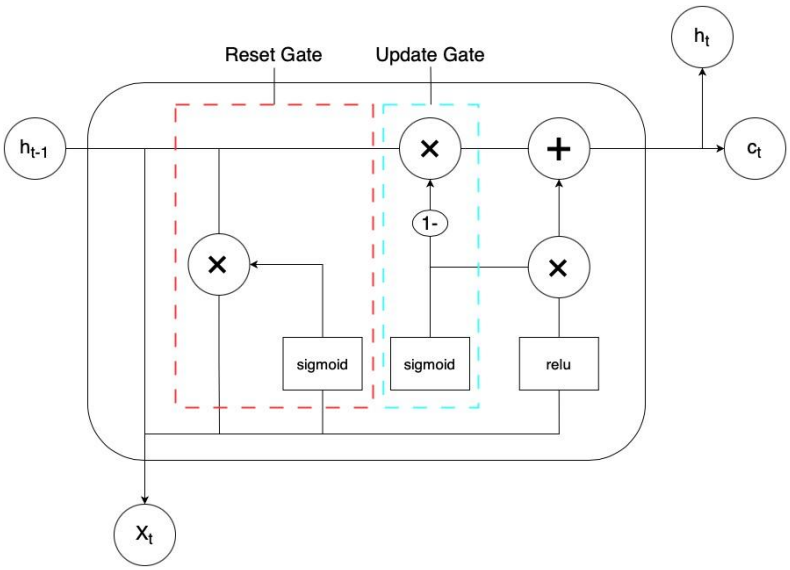

Figure 5. GRU architecture

This GRU model consists of three GRU layers, followed by a dense output layer with sigmoid activation togenerate binary classification predictions, as shown in Figure 5.

2.9. CNN-GRU Classification

The CNN-GRU framework used in this study is a model designed to detect cyberbullying on social mediaplatforms. This framework combines Convolutional Neural Networks and Gated Recurrent Units synergistically to capture spatial and temporal patterns of the input text, as illustrated in Figure 6.

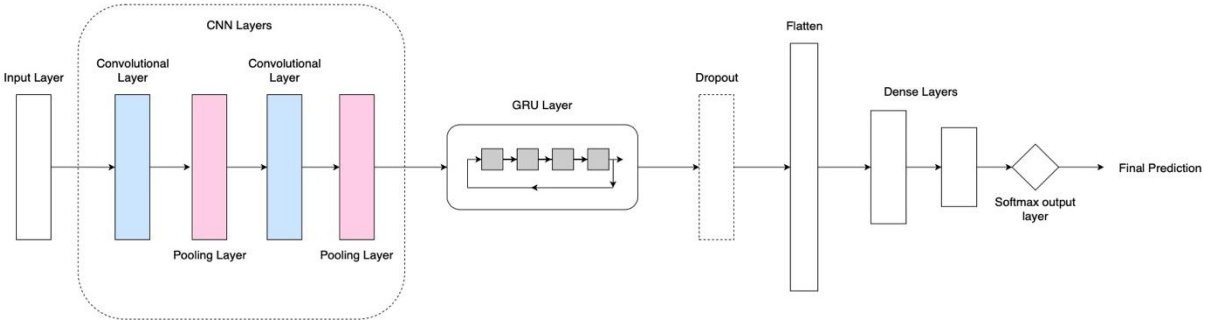

Figure 6. CNN-GRU hybrid architecture

The architecture in this study consists of several main components: CNN layer, GRU layer, and dense layer. The CNN layer has the function of extracting local features in the text, by processing them through a convolution layer and a pooling layer. Afterward, the results from the CNN layer are passed to the GRU layer, which is responsible for capturing the temporal context and relationships between words in the text sequence.

2.10. Attention Mechanism

Attention mechanisms have been used in recent years in medicine, chemical industry, and biotechnology [23]. Attention mechanisms are inspired by the human visual system which is an important component in neural architecture, especially for encoder-decoder based models that require high performance on long sequences [24]. This mechanism allows neural networks to focus on a specific subset of inputs and give greater weight to important words in context [25]. In addition, the attention mechanism also reduces the computational burden by selecting a relevant subset of the input, allowing the system to focus more on important information in the input data [26]. This study uses Attention which allows the model to give different weights to each part of the input data, according to its relevance in context. The attention mechanism formula is shown in Equation (2).

$$c_t = \sum_{i=1}^{T_x} \alpha_{t,i} h_i. \qquad (2)$$

With each hidden state hi is measured by αt, . The weights αt, of each hidden state hi is also called the alligment score.

2.11. Evaluation Matrix

Confusion Matrix displays the actual classification and predictions made by the model [27]. It consists of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) to show the number of correct or incorrect predictions in positive and negative categories [28]. Confusion Matrix evaluates model performance through several forms of metrics, namely, Accuracy, Precision, Recall, and F1-score [29].

The ratio of correct predictions (True Positives and True Negatives) is measured by Accuracy, indicating how often the model makes correct predictions overall. The formula for calculating accuracy is given in Equation (3), as follows:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (3)$$

Precision is the ratio of True Positives to total positive predictions. The formula for calculating precision is given in Equation (4), as follows:

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall is the ratio of True Positives to the total amount of data that is actually positive. The formula for calculating recall is given in Equation (5), as follows:

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

On the other hand, F1-score is a metric used to balance between Precision and Recall, which is useful for balancing between positive and negative classes. F1-score provides more accurate results in handling minority classes. The formula for calculating F1-score is given in Equation (6), as follows:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (6)$$

3. RESULT

The authors conducted experiments in six different scenarios to detect cyberbullying in Indonesian, with the aim of obtaining optimal model performance. Each scenario builds on the previous scenario (e.g., scenario 2 uses scenario 1 as a reference, and so on). There are six scenarios evaluated in this study. Table 4 presents the description of each scenario.

Table 4. Description of each scenario

Scenario	Description
1	Exploring the comparison of training data and test data to obtain the best configuration that results in optimal model performance.
2	Exploring the maximum number offeatures in TF-IDF.
3	Testing the use of N-Gram on TF-IDF parameters with focus on unigram, bigram, and trigram.
4	Testing N-Gram combinations on TF- IDF with a focus on unigram-bigram and unigram-trigram combinations, as well as starting testing the CNN-GRU hybrid model to find out which N-Gram combination gives the best performance.
5	Apply the expansion feature using Word2Vec to the model results from scenario 4 using the corpus that has been built.
6	Apply the attention technique to the CNN-GRU hybrid model that has the best performance from the previous scenario, with the aim of improving theaccuracy and effectiveness of the model.

3.1. Exploring Training and Test Data for Optimal Model Performance

In the first scenario, a Convolutional Neural Network (CNN) model with a Conv1D layer was applied for initial feature extraction. This CNN model consists of 32 filters with a kernel size of 5, followed by a MaxPooling1D layer with a pool size of 5, and a Dense layer with 32 units. This model was trained using a batch size of 16 for 35 epochs. In the Gated Recurrent Unit (GRU) model, three consecutive GRU layers wereused with the following configuration: first layer with 128 units, second layer with 64 units, and third layer with 32 units. This GRU model was also trained with a batch size of 16 for 35 epochs.

In the initial stage of this experiment, the basic parameters for feature extraction using the TF-IDF methodas well as the maximum feature value of 1000 were applied. The basic settings used at this stage involve the use of unigrams as the representation of text features.

Table 5. Basic accuracy with different data ratios

Model	Rasio Data	Accuracy (%)	F1-Score (%)
CNN	90:10	**79.25**	**79.24**
	80:20	79.22	79.22
	70:30	79.21	79.21
GRU	90:10	**79.48**	**78.82**
	80:20	78.57	78.32
	70:30	78.79	78.37

Table 5 shows the test results in the first scenario. From the comparison of three data ratio scenarios, it can be seen that the highest accuracy and F1-score values in each model are obtained at a ratio of 90:10. The CNN model obtained an accuracy value of 79.25% and an F1-score of 79.24%. Meanwhile, the GRU model produces an accuracy value of 79.48% and an F1-score of 78.82%. The highest results from eachof these models will be used as the basis for experiments in scenario 2 and beyond.

3.2. Exploration of Maximum Features in TF-IDF

In scenario two, a comparison of the maximum feature value between 1000 and 4000 was conducted. The test results show that the highest accuracy and F1-score values in each model are obtained at the maximumfeature value of 4000. The test results for each model are listed in Table 6.

Table 6. Comparison of maximum usage of features in TF-IDF

Model	Max Fitur	Accuracy (%)	F1-Score (%)
CNN	1000	79.25	79.24
	4000	**79.45**	**79.45**
GRU	1000	79.48	78,82
	4000	**80.10**	**79.52**

It shows that for the CNN model, the highest accuracy and F1-score values are obtained at the maximum number of features of 4000, with an accuracy value of 79.45% and F1-score of 79.45%. While in the GRU model, the highest accuracy and F1-score values are also obtained at the maximum number of features of 4000, with an accuracy value of 80.10% and F1-score of 79.52%.

3.3. Testing N-Gram Parameters in TF-IDF: Unigram, Bigram, and Trigram

In scenario three, tests were conducted using TF-IDF for word weighting, namely Unigram, Bigram, and Trigram. The purpose of this test is to compare the accuracy and F1-score obtained by using various types of N-Grams. The test results can be seen in Table 7. Based on the results obtained, it can be seen that the use of Bigram and Trigram has decreased the accuracy and F1-score values compared to Unigram.

Table 7. Accuracy value of the previous model tested using N-Gram

Model	N-Gram	Accuracy (%)	F1-Score (%)
CNN	Unigram	**79.45**	**79.45**
	Bigram	69.61	68.61
	Trigram	53.47	40.50
GRU	Unigram	**80.10**	**79.52**
	Bigram	69.44	62.63
	Trigram	53.28	13.20

It shows that for the CNN model, the use of Unigram provides the highest accuracyand F1-score values, while the use of Bigram and Trigram has decreased significantly. The same thing also happens in the GRU model, where the use of Unigram gives better results compared to Bigram and Trigram.

3.4. Evaluating N-Gram Combinations and Initiating CNN-GRU Hybrid Model Testing

In this scenario, a combination of N-Grams in TF-IDF was tested based on the results obtained from scenario 3. In addition, in scenario 4, the author also tested a combination of CNN-GRU hybrid models to determine the N-Gram combination that provides the best performance. The test results can be seen in Table 8. Based on these findings, the Unigram+Trigram N-Gram combination gives the best results in the CNN model, while in the GRU model, the Unigram+Bigram combination shows the best performance. For the CNN-GRU hybrid model, the Unigram+Trigram combination also gives the best results.

Table 8. Accuracy value of the tested model using N-Gram combination

Model	N-Gram	Accuracy (%)	F1-Score (%)
CNN	Unigram	79.45	79.45
	Unigram+Bigram	79.53	79.53
	Unigram+Trigram	**79.59**	**79.58**
GRU	Unigram	80.10	79.52
	Unigram+Bigram	**80.40**	**79.71**
	Unigram+Trigram	80.22	79.59
CNN-GRU	Unigram	80.25	80.25
	Unigram+Bigram	80.31	80.31
	Unigram+Trigram	**80.68**	**80.68**

It shows that in the CNN model, the Unigram+Trigram combination gives the best results with an accuracy value of 79.59% and F1-score of 79.58%. In the GRU model, the Unigram+Bigram combination provides the best performance with an accuracy of 80.40% and F1-score of 79.71%. While in the CNN-GRU hybrid model, the Unigram+Trigram combination produces the highest accuracy and F1-score values, which are 80.68% and 80.68% respectively.

3.5. Incorporating Word2Vec Features into All Models

In Scenario 5, CNN, GRU, and hybrid CNN-GRU models were evaluated by applying feature expansion using Word2Vec along with TF-IDF feature extraction. This test was conducted by selecting the top features from three corpora, namely Corpus Tweet, Corpus IndoNews, and a combination of the two, namely Corpus Tweet+IndoNews. The main purpose of this test is to analyze the impact of using expansion features on the accuracy of the classification model, in order to obtain more optimal performance. The application of Word2Vec feature expansion is expected to improve the accuracy of classification results on each model tested.

Table 9. The accuracy value of the model tested with the use of feature expansion

Model	Rank	Corpus Tweet		Corpus Indonews		Corpus Tweet+Indonews	
		Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)
CNN	Top 1	80.03	80.03	80.07	80.07	80.15	80.14
	Top 5	80.17	80.17	80.03	80.03	79.96	79.96
	Top 10	80.01	80.01	79.93	79.93	**80.19**	**80.19**
GRU	Top 1	80.54	80.00	80.10	79.33	80.63	80.03
	Top 5	79.45	78.97	79.81	78.97	**80.64**	**80.47**
	Top 10	79.31	78.87	79.50	78.70	78.96	78.75
CNN-GRU	Top 1	79.34	79.34	80.15	80.15	**80.85**	**80.85**
	Top 5	78.81	78.80	80.02	80.02	79.97	79.96
	Top 10	77.89	77.89	79.58	79.58	78.44	78.44

Based on the test results listed in Table 9, it can be concluded that the combination of Corpus Tweet+IndoNews gives the best results in each model tested. For the CNN model, the best results were obtained at rank Top 10 with an accuracy of 80.19% and F1-score of 80.19%. The GRU model produces the best performance at rank Top 5 with an accuracy of 80.64% and F1-score of 80.47%. As for the CNN-GRU hybridmodel, the best results were obtained at rank Top 10 with an accuracy of 80.85% and F1-score of 80.85%.

3.6. Enhancing Performance with Attention Mechanism on CNN-GRU Model

In the last scenario, the application of attention techniques to the CNN-GRU hybrid model was tested. The CNN-GRU hybrid model used in the previous scenario has achieved the highest accuracy and F1-score value, which is 80.85% for both on the combination of Corpus Tweet+IndoNews with rank Top 10. This test aims to analyze the effect of applying attention techniques on model performance. The test results shows in Table 10.

Table 10. Comparison of hybrid model test results from the previous scenario with the application of attention

Model	Attention	Accuracy (%)	F1-Score (%)
CNN-GRU	No	80.85	80.85
	Yes	**80.96**	**80.96**

3.7. Comparative Analysis of the Current CNN-GRU Hybrid Model with Prior Studies

Based on all the experiments conducted, the final result of the CNN-GRU hybrid model combination is 80.96%. Compared to previous research [5] that also uses CNN-GRU for cyberbullying detection, there is a significant increase in accuracy. The previous research obtained an accuracy of

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4176

80.41%, which means there was an increase of 0.55%. This improvement is obtained from the application of Word2Vec feature expansion and attention mechanism in the CNN-GRU hybrid model, which shows the effectiveness of the combination of the two approaches in cyberbullying detection. The comparison is shown in Table 11.

Table 11. Comparative performance of CNN-GRU hybrid model and previous studies

Model	Accuracy (%)	F1-Score (%)
CNN-GRU (Previous)	80.41	80.41
CNN-GRU (Current)	**80.96**	**80.96**

4. DISCUSSION

In the first scenario, Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) models were tested using basic parameters with TF-IDF feature extraction method of maximum 1000 features and Unigramas text representation. The test results show that the GRU model has the highest accuracy at a data ratio of 90:10 compared to the CNN model. In the second scenario, the maximum number of features in TF-IDF was increased to 4000, which resulted in improved accuracy and F1-score in both models, with the GRU model showing a moresignificant improvement. The third scenario tested the use of N-Gram types, namely Unigram, Bigram, and Trigram. Based on the test results, the use of Unigram provides the best performance in both models, while Bigram andTrigram actually decrease the classification performance. In the fourth scenario, the combination of N-Grams istested to determine the combination that provides optimal results. The combination of Unigram+Trigram gives the best performance in CNN and hybrid CNN-GRU models, while the combination of Unigram+Bigram excels in the GRU model. The fifth scenario applies feature expansion using Word2Vec with a combination of Tweet and IndoNews corpus. The test results show significant improvement in all models, with the hybrid CNN-GRU model achieving the highest accuracy of 80.85% on the Top 10 rank.

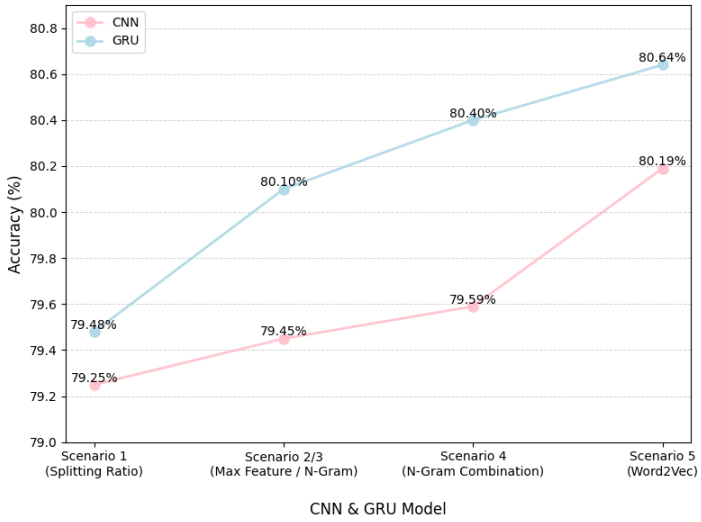

Figure 7. Comparison of CNN and GRU model accuracy values

Based on the test results shown in Figure 7, the comparison of scenarios in the CNN and GRU models showsthat the application of the TF-IDF feature extraction technique and Word2Vec feature expansion significantly improved the accuracy of each model. The CNN model achieved a final accuracy of 80.19%, which showsan improvement of 0.94% from the initial accuracy. Meanwhile, the GRU model obtained a final accuracy of 80.64%, with an increase of 1.16%.

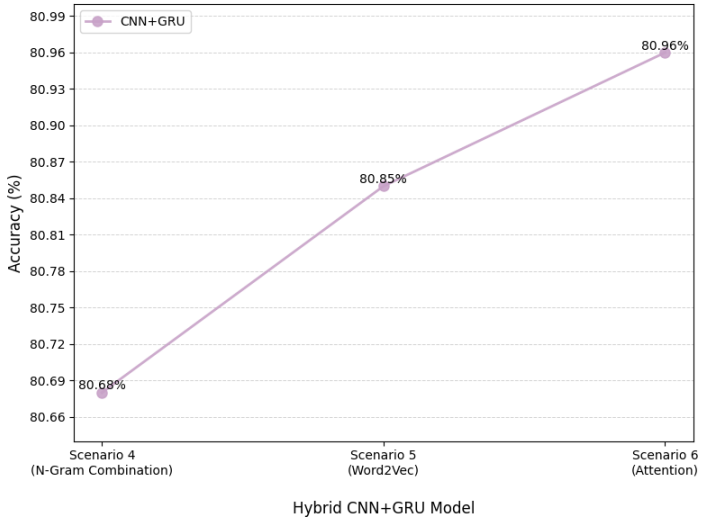

Figure 8. CNN-GRU hybrid model accuracy Comparison values

In Figure 8, which displays the accuracy comparison of the CNN-GRU hybrid model, it can be seen that the CNN-GRU hybrid model has increased accuracy with the use of feature expansion techniques and attention mechanisms. This improvement reaches 0.28%, resulting in the highest accuracy in the model.

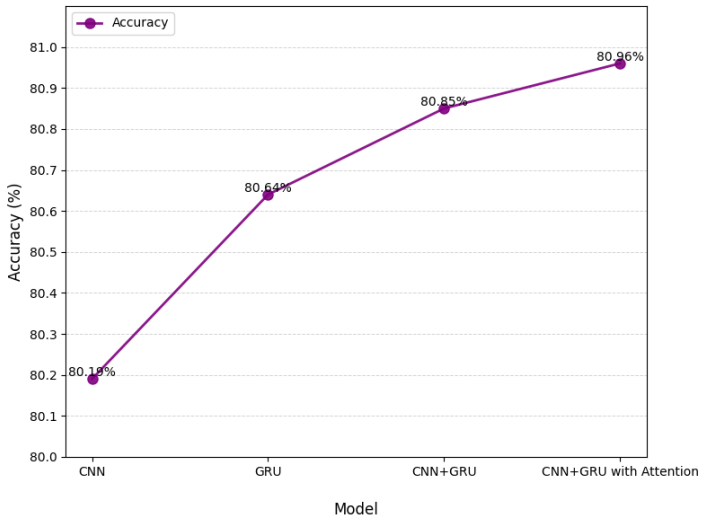

Figure 9. Comparison of the final accuracy value of each model

Figure 9 shows the accuracy comparison between CNN, GRU, and CNN-GRU hybrid models. The CNN-GRU model consistently outperforms CNN and GRU individually, with the application of attention mechanisms increasing its accuracy to 80.96%.

This study shows the effectiveness of the combination of CNN-GRU hybrid model combined with TF-IDF feature extraction, Word2Vec feature expansion, as well as the application of attention-mechanism on the Indonesian language cyberbullying text detection model. Previous research [5], which also used a combination of CNN-GRU hybrid models, has not shown significant advantages compared to stand-alone models, this study confirmed the superior performance of a hybrid combination of CNN-GRU models with additional feature extraction and expansion as well as attention-mechanism which enhances the model's ability to focus on important parts of the text.

However, this study is still limited to text data and does not pay attention to visual context such as images or videos which are often closely related to cyberbullying on social media. Future research is recommended to develop models that use more varied data, such as images and videos, to improve detection accuracy and relevance.

5. CONCLUSION

This research contributes to the field of informatics by introducing a CNN-GRU hybrid model equipped with an attention mechanism to detect cyberbullying. The model is developed using 30,084 Indonesian tweet data collected through API crawling from X social media and manually labeled. Feature extraction is performed with TF-IDF and extended with Word2Vec, which is proven to significantly improve the accuracy of the model. The test results show that the CNN-GRU hybrid model with attention mechanism outperforms CNN or GRU models individually, achieving an accuracy of 80.96%. This research successfully achieved the goal of producing a more reliable detection system suitable for online platforms. For future research, it is recommended to integrate more varied data, such as photos or videos, to improve accuracy and capture the variety of forms of cyberbullying. In addition, testing on various other social media platforms needs to be done to ensure the model is widely applicable.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest between the authors or with the research object in this paper

ACKNOWLEDGEMENT

The author expresses deep gratitude to all parties who have contributed to the completion of this study.

REFERENCES

[1] N. T. Martoredjo, "Social media as a learning tool in the digital age: A review," in *International Conference on Computer Science and Computational Intelligence*, 2023, pp. 534–539. doi: 10.1016/j.procs.2023.10.555.

[2] "X/Twitter: Countries with the largest audience 2024," Statista Research Department. Accessed: Dec. 04, 2024. [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[3] G. K. Nathanael, "Understanding recession response by Twitter users: A text analysis approach," *Heliyon*, vol. 10, no. 1, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23737.

[4] K. Hellfeldt, L. López-Romero, and H. Andershed, "Cyberbullying and psychological well-being in young adolescence: the potential protective mediation effects of social support from family, friends, and teachers," *Int J Environ Res Public Health*, vol. 17, no. 1, Jan. 2020, doi: 10.3390/ijerph17010045.

[5] N. W. F. Amalia, "Cyberbullying Detection on Twitter using Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU)," Telkom University, 2023.

[6] Y. Setiawan, N. U. Maulidevi, and K. Surendro, "The Use of Dynamic n-Gram to Enhance TF-IDF Features Extraction for Bahasa Indonesia Cyberbullying Classification," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Feb. 2023, pp. 200–205. doi: 10.1145/3587828.3587858.

[7] L. Ye, C. Wei, N. Heran, and Y. Yimeng, "Review Mining for Experiential Products Incorporating Word2vec and Review Sentiment Tendencies," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 1492–1499. doi: 10.1016/j.procs.2022.11.335.

[8] A. Aljohani, N. Alharbe, R. E. Al Mamlook, and M. M. Khayyat, "A hybrid combination of CNN Attention with optimized random forest with grey wolf optimizer to discriminate between

Arabic hateful, abusive tweets," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 2, Feb. 2024, doi: 10.1016/j.jksuci.2024.101961.

[9] Q. Liu, Y. Hu, and H. Liu, "Enhanced stock price prediction with optimized ensemble modeling using multi-source heterogeneous data: Integrating LSTM attention mechanism and multidimensional gray model," *J Ind Inf Integr*, vol. 42, Nov. 2024, doi: 10.1016/j.jii.2024.100711.

[10] Gaurav and P. Mathur, "An Attention Mechanism and GRU Based Deep Learning Model for Automatic Image Captioning," *International Journal of Engineering Trends and Technology*, vol. 70, no. 3, pp. 302–309, Mar. 2022, doi: 10.14445/22315381/IJETT-V70I3P234.

[11] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," in *CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies*, 2020, pp. 605–611. doi: 10.1016/j.procs.2021.01.207.

[12] M. M. Abedi and E. Sacchi, "A machine learning tool for collecting and analyzing subjective road safety data from Twitter," *Expert Syst Appl*, vol. 240, Apr. 2024, doi: 10.1016/j.eswa.2023.122582.

[13] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data," *Heliyon*, vol. 10, no. 16, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35945.

[14] M. Deja, Isto Huvila, G. Widén, and F. Ahmad, "Seeking innovation: The research protocol for SMEs' networking," *Heliyon*, vol. 9, no. 4, Apr. 2023, doi: 10.1016/j.heliyon.2023.e14689.

[15] R. Agrawal and R. Goyal, "Developing bug severity prediction models using word2vec," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 104–115, Jun. 2021, doi: 10.1016/j.ijcce.2021.08.001.

[16] H. Imaduddin and S. Fauziati, "Word Embedding Comparison for Indonesian Language Sentiment Analysis," in *International Conference of Artificial Intelligence and Information Technology*, 2019, pp. 426–430. doi: 10.1109/ICAIIT.2019.8834536.

[17] L. Shan, Y. Liu, M. Tang, M. Yang, and X. Bai, "CNN-BiLSTM hybrid neural networks with attention mechanism for well log prediction," *J Pet Sci Eng*, vol. 205, Oct. 2021, doi: 10.1016/j.petrol.2021.108838.

[18] M. N. I. Siddique, M. Shafiullah, S. Mekhilef, H. Pota, and M. A. Abido, "Fault classification and location of a PMU-equipped active distribution network using deep convolution neural network (CNN)," *Electric Power Systems Research*, vol. 229, Apr. 2024, doi: 10.1016/j.epsr.2024.110178.

[19] Y. Zhang and H. D. Fill, "TS-GRU: A Stock Gated Recurrent Unit Model Driven via Neuro-Inspired Computation," *Electronics (Basel)*, vol. 13, no. 23, p. 4659, Nov. 2024, doi: 10.3390/electronics13234659.

[20] T. Li, Y. Lin, B. Cheng, G. Ai, J. Yang, and L. Fang, "PU-CTG: A Point Cloud Upsampling Network Using Transformer Fusion and GRU Correction," *Remote Sens (Basel)*, vol. 16, no. 3, Feb. 2024, doi: 10.3390/rs16030450.

[21] Y. Yan *et al.*, "Hybrid GRU–Random Forest Model for Accurate Atmospheric Duct Detection with Incomplete Sounding Data," *Remote Sens (Basel)*, vol. 16, no. 22, Nov. 2024, doi: 10.3390/rs16224308.

[22] A. Zhang, S. Chun, Z. Cheng, and P. Zhao, "Predicting the core thermal hydraulic parameters with a gated recurrent unit model based on the soft attention mechanism," *Nuclear Engineering and Technology*, Mar. 2024, doi: 10.1016/j.net.2024.01.045.

[23] W. Jia, Y. Zhan, J. Zhang, and Y. Dai, "Robot assisted bone milling state classification network with attention mechanism," *Expert Syst Appl*, vol. 249, Sep. 2024, doi: 10.1016/j.eswa.2024.123726.

[24] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," May 02, 2021, *MDPI AG*. doi: 10.3390/electronics10101163.

[25] G. Brauwers and F. Frasincar, "A General Survey on Attention Mechanisms in Deep Learning," *IEEE Trans Knowl Data Eng*, vol. 35, no. 4, pp. 3279–3298, Apr. 2023, doi: 10.1109/TKDE.2021.3126456.

[26] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes," Aug. 01, 2022, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00521-022-07366-3.

[27] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf Sci (N Y)*, vol. 340–341, pp. 250–261, May 2016, doi: 10.1016/j.ins.2016.01.033.

[28] "Confusion Matrix." Accessed: Apr. 22, 2024. [Online]. Available: https://www.sciencedirect.com/topics/engineering/confusion-matrix

[29] G. Phillips *et al.*, "Setting nutrient boundaries to protect aquatic communities: The importance of comparing observed and predicted classifications using measures derived from a confusion matrix," *Science of the Total Environment*, vol. 912, Feb. 2024, doi: 10.1016/j.scitotenv.2023.168872.

Vol. 6, No. 3, Juni 2025, Page. 1113-1130
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4176