Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 655-672
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4321

Enhancing Cyberbullying Detection on Platform 'X' Using IndoBERT and Hybrid CNN-LSTM Model

Annisaa Alya Hafiza*[1], Erwin Budi Setiawan[2]

[1,2] Informatics, Telkom University, Indonesia

Email: [1]anisalyahfza@student.telkomuniversity.ac.id

Received : Jan 15, 2025; Revised : Feb 9, 2025; Accepted : Feb 11, 2025; Published : Apr 26, 2025

Abstract

Cyberbullying on social media platforms has become widespread in society. Cyberbullying can take many forms, including hate speech, trolling, adult content, racism, harassment, or rants. One social media platform that has many cyberbullies is Twitter, which has been renamed 'X'. The anonymous nature of this 'X' platform allows users from all over the world to commit cyberbullying as they can freely share their thoughts and expressions without having to account for their identity. This research aims to explore the influence of IndoBERT's semantic features on hybrid deep learning models for cyberbullying detection while integrating TF-IDF feature extraction and FastText feature expansion to enhance text classification performance. Specifically, this study examines how IndoBERT's semantic capabilities affect the hybrid deep learning model in detecting cyberbullying on platform 'X'. This study has 30,084 tweets with a hybrid deep learning approach that combines CNN and LSTM. In the IndoBERT scenario, IndoBERT features were first combined with TF-IDF, then expanded using FastText before being applied to the hybrid deep learning model. The test results produced the highest accuracy rate by: CNN (80.69%), LSTM (80.67%), CNN-LSTM (81.18%), CNN-LSTM-IndoBERT (82.05%). This research contributes to informatics by integrating hybrid deep learning (CNN-LSTM) with IndoBERT and TF-IDF, demonstrating its effectiveness in improving cyberbullying detection in Indonesian text. Future research can explore the use of other transformer-based models such as RoBERTa or ALBERT to enhance contextual understanding in cyberbullying classification.

Keywords : *Convolutional Neural Network (CNN), Cyberbullying Detection, Fasttext, IndoBERT, Long Short-Term Memory (LSTM), TF-IDF.*

1. INTRODUCTION

Social media is now used by everyone. One of the most popular social media is Twitter. Recently Twitter has changed its name to 'X'. 'X' is a social media that has an anonymous nature [1]. The anonymous nature of the 'X' platform makes it easier for people to commit cyberbullying because users from different parts of the world can freely express their opinions and expressions [2].

Cyberbullying is a serious problem that arises in cyberspace, Cyberbullying can be in the form of hate speech, trolling, adult content, racism, harassment, or profanity [3]. The impact of cyberbullying is often far worse than the physical or verbal bullying that occurs in the real world, as the rapid and accessible spread of information can exacerbate the trauma experienced by the victim [4]. In addition, the presence of perpetrators with hidden identities makes the reporting and prosecution process more difficult[5].

The interaction in real time and rapid content-sharing features of 'X' increase the visibility and harm caused by cyberbullying occurrences [6]. This anonymity allows abusers to avoid accountability, providing substantial hurdles in tackling the problem [7]. According to research, victims of cyberbullying frequently have severe emotional repercussions such as anxiety, depression, and, in extreme situations, suicide impulses [8]. Given 'X's global reach and importance, tackling cyberbullying

on this platform is crucial not only for individual protection, but also for creating a safer and more inclusive online environment [9].

This research employs the hybrid deep learning CNN-LSTM technique to detect cyberbullying because previous research has demonstrated that this method is more effective than non-hybrid methods. The Convolutional Neural Network (CNN) model is effective in extracting local features from text and identifying significant patterns and structures in a document [10]. Furthermore, the Long Short-Term Memory (LSTM) model can effectively analyze complex and large textual contexts [11]. This model, which combines CNN and LSTM, can identify key text characteristics and understand contextual relationships in sequential data. This method has proven to be more effective than traditional methods for detecting cyberbullying [12].

Building on this foundation, this research takes a novel approach by exploring the influence of semantic features, namely IndoBERT on cyberbullying detection and integrating TF-IDF feature extraction and FastText feature expansion. To see the effect of semantic features on the hybrid deep learning CNN-LSTM model, this research combines semantic features with TF-IDF in detecting cyberbullying. Based on our knowledge, there is no research that discusses the influence of semantic features on cyberbullying detection with hybrid deep learning CNN-LSTM on social media.

Some previous studies show the potential of using deep learning models for cyberbullying detection. Research by Sultan et al. [13], which combines LSTM and CNN with 97.52% accuracy, 96.87% precision, 98.96% recall, 98.28% F-measure, and 98.67% AUC-ROC, but does not utilize feature extraction and expansion, which may limit the model's ability to detect more complex cyberbullying text [14]. This limitation hampers the model's capacity to completely assess nuanced contextual meanings in text, which is crucial for detecting subtle signs of cyberbullying [15]. As a result, the model's effectiveness in dealing with different and complex cyberbullying scenarios is dramatically diminished [16]. In contrast, our research takes a holistic strategy, incorporating three major techniques: TF-IDF for statistical feature extraction, IndoBERT semantic embeddings for contextual nuances, and FastText feature expansion to enrich text data representation [17]. These combined techniques offer a strong and comprehensive foundation for evaluating and diagnosing cyberbullying in a variety of complicated settings.

Another research by Andika et al. which combines LSTM and CNN with 84% F1-score, although it was limited to YouTube comments, so the lack of variety in data sources could affect the generalization of the model [18]. Their model does not use the semantic IndoBERT feature or feature extraction and feature expansion. Our research addresses this issue by employing TF-IDF and FastText for extraction and feature extraction, as well as IndoBERT for contextual semantics resulting in a model that is more generalizable across a variety of data sets [19].

Research by Asqolani et al. demonstrated the effectiveness of a hybrid CNN-LSTM model with Word2Vec feature expansion, achieved the highest accuracy of 79.48% but did not utilize TF-IDF, which is important in text feature extraction [20]. However, this study did not use TF-IDF, which is significant in text feature extraction because it captures word frequencies as well as contextual relationships in the text [21]. Our research addresses these constraints by combining TF-IDF, IndoBERT semantic embedding, and FastText feature expansion to improve the accuracy and robustness of cyberbullying detection.

Research by Anggraeni et al. utilized IndoBERT in a hybrid LSTM-CNN model for tweets about dengue fever, achieving an accuracy of 91% accuracy, 89% recall, 91% precision, and 90% F1-score [22]. While their analysis proved IndoBERT's usefulness, it was confined to tweets concerning dengue sickness and did not use feature extraction techniques like TF-IDF or feature expansion methods like FastText, as supported by research demonstrating the significance of statistical and semantic factors in strengthening text classification models [23]. In contrast, our study overcomes these constraints by

combining IndoBERT semantic characteristics with statistical feature extraction via TF-IDF and feature expansion with FastText. This is consistent with the observation that incorporating several feature extraction strategies can improve model adaptability and robustness in difficult circumstances. This combination improves model performance and flexibility to different and complicated cyberbullying circumstances .

Another study by Fabillah et al. [24], using IndoBERT, found 92.6% accuracy, 91.5% recall, 92% precision, and 91.7% F1-score, but there is a risk of overfitting in more complex models, as evidenced by studies showing that relying solely on transformer embeddings can lead to overfitting in less diverse datasets[25]. Their dependence simply on IndoBERT embeddings without TF-IDF and FastText renders their model less robust and more prone to overfitting, a problem emphasized by research emphasizing the need of combining statistical and semantic variables for enhanced robustness. Our research addresses this by mixing different feature extraction strategies, resulting in a balanced and thorough representation of text data, consistent with previous findings that a hybrid strategy improves generalization and minimizes overfitting risks. This combination enhances generalization and performance, which is consistent with studies promoting feature diversity for complex text categorization tasks [26].

Based on previous research related to cyberbullying detection, the application of hybrid deep learning models obtained higher accuracy values compared to models that did not use hybrid deep learning [13]. However, previous research has not investigated the combination of IndoBERT semantic characteristics with feature extraction TF-IDF and FastText for feature expansion in hybrid deep learning CNN-LSTM, especially for cyberbullying detection on Indonesian social media. This limitation indicates a research gap in which combining statistical, semantic, and contextual information could improve model robustness and generalizability.

To address this gap, this research proposes a novel approach by integrating semantic features IndoBERT with feature extraction TF-IDF and FastText for feature expansion in hybrid deep learning, which has not been explored in previous research on cyberbullying detection in Indonesian social media. By utilizing these feature extraction techniques, this study improves the performance of cyberbullying detection models and provides more accurate insights into understanding the semantic relationships within Indonesian social media content [27].

2. METHOD

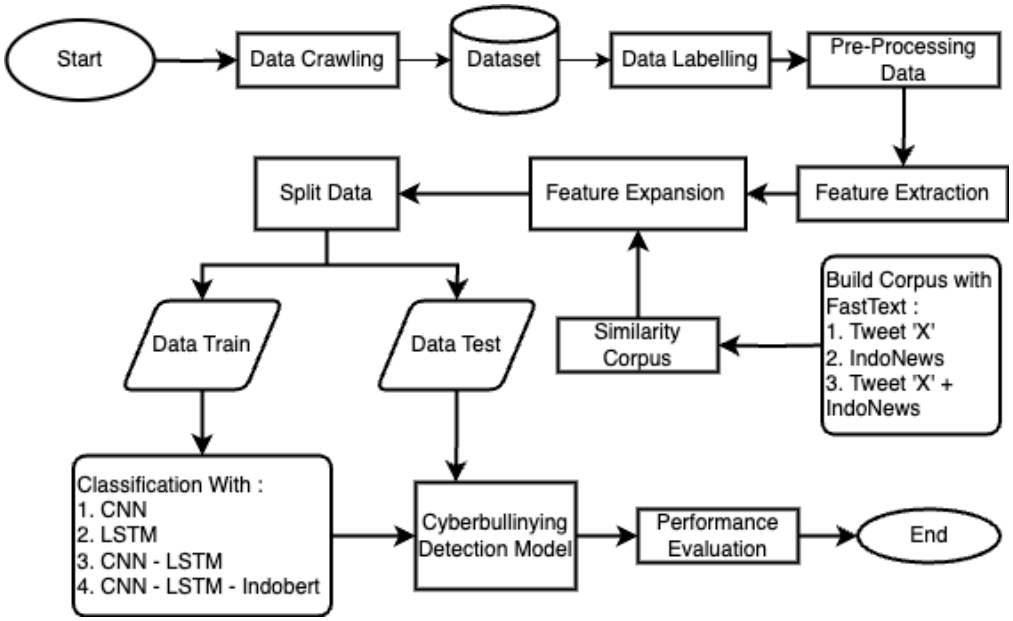

Figure 1. Cyberbullying Detection System

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4321

The system built to detect Indonesian language cyberbullying on platform 'X' using a hybrid deep learning model, TF-IDF feature extraction, FastText feature expansion, and IndoBERT semantic features can be seen in Figure 1. The steps of the cyberbullying detection system include data crawling, data labeling, data preprocessing, TF-IDF feature extraction, FastText feature expansion, and data splitting into test and train data. And data classification with five models: CNN, LSTM, CNN-LSTM, CNN-LSTM-IndoBERT, and IndoBERT. Finally, the performance of the built system will be evaluated.

2.1. Crawling Data

Collecting data through crawling is a way to retrieve information from various sources, such as blogs, social media, or other sites [28]. This research takes data from platform 'X', which uses Indonesian language through the crawling process. Platform 'X' provides an API to facilitate the data crawling process. The data crawling process is done by focusing on tweets that may contain one type of cyberbullying. Cyberbullying includes the use of abusive and blunt language in text messages, as well as other behaviors such as making disrespectful comments, spreading gossip, and threatening online violence. These behaviors were the focus of the cyberbullying identification process.

In this data collection process, words such as rants and other types of hate speech were collected from references from various relevant journals. Previous studies, such as the one conducted by Irfan Ahmad et al [20]. showed that the use of abusive words, hate speech, and negative comments are important components in identifying cyberbullying. In addition, references related to these keywords come from literature that identifies language patterns often used in aggressive communication on social media. These words are also referred to as key indicators in detecting cyberbullying through text analysis. The keywords used in data collection in this study can be seen in Table 1.

Table 1. Tweet Keyword

Keyword	Total
J*lek	3.663
B*nci	2.855
L*nte	3.610
B*doh	2.367
T*lol	6.874
G*ndut	1.630
B*go	926
G*blok	6,572
B*ngsat	634
K*ntol	953
Total	30,084

2.2. Data Labelling

Labeling is done on the data that has been collected before going to the classification stage. This labeling aims to facilitate the data classification process [29]. In this labeling, data that does not meet the cyberbullying criteria is manually labeled with the value “0,” and data that meets the cyberbullying criteria is given a value of "1". Therefore, label "0" indicates non-cyberbullying, and label "1" indicates cyberbullying. This manual labeling procedure is reliable because it is carefully reviewed and validated by 5 people, which helps to reduce errors and preserve consistency throughout the labeling process with majority vote. An example of data labeling results can be seen in Table 2.

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4321

Table 2. Labeling Example

Tweet	Label
@TimpalBali Baru buka twitter udh nemu orang t*lol kek lo...ya u aja yg pindah dasar id*ot	1
@bdngfess Alay jg seru, rada t*lol dikit	0
@memefess @jawafess So keras t*lol rasis maling g*blok b*ngsat	1

After the labeling process, the distribution of cyberbullying and non-cyberbullying classes can be seen in Table 3. The table shows balanced data, with the cyberbullying label having a data count of 15.005 and non-cyberbullying having a data count of 15.079. This balanced distribution is required to keep the model from becoming biased towards one class, ensuring that it learns to correctly categorize both cyberbullying and non-cyberbullying events.

Table 3. Distribution of Labeling Data

Tweet	Label
1	15.005
0	15.079
Sum	30.084

2.3. Data Pre-Processing

The data taken is unstructured raw data and tends to have a lot of noise. Therefore, to prepare the data to be more easily processed in the classification stage, a series of pre-processing stages are needed [30]. This preprocessing consists of six stages. First, data cleaning is the process of removing elements or symbols that interfere with text data, such as symbols, numbers, hashtags (#), usernames (@), spaces, URLs, and emoticons. Second, case folding involves converting the entire text from uppercase to lowercase, aiming to equalize words with similar meanings so that they are not considered different due to variations in uppercase and lowercase usage. Third, normalization is the process of changing the words in the text from non-standard words to standard words to conform to the rules for the use of standard or common words. Fourth, tokenization is the stage where text is divided into separate units of tokens or words. Fifth, stopword removal is the process of eliminating words that are less meaningful or irrelevant. Lastly, stemming aims to convert words into shorter versions by removing affixes.

2.4. Feature Extraction TF-IDF

Feature extraction in this research is the process of calculating the weight of each word and converting the words in the text into vector form. This stage is the first step in the text classification process. The feature extraction used is TF-IDF. Term Frequency-Inverse Document Frequency (TF-IDF) is used to assign vectors to words (terms) in a document [31]. In a document Term Frequency (TF) aims to calculate how much a particular word appears, and Inverse Document Frequency (IDF) calculates unique words. The following is the formula for calculating vectors in the TF-IDF method:

$$W_{ij} = tf_{ij} \times idf_j \quad (1)$$

$$idf_j = \log \frac{D}{df_j} \quad (2)$$

$$W_{ij} = tf_{ij} \times \log \frac{D}{df_j} \quad (3)$$

The weight of the i-th document on the j-word (Wij) indicates the importance of the j-word in the i-th document. The value of tfij is the frequency of occurrence of the jth word in the document the

more often it appears, the higher the value. Conversely, the idfj indicates the rarity of the jth word in the corpus across documents. The rarer the word appears, the higher the value. D denotes the total documents, and dfj denotes the total documents that have the word (tj).

This research uses TfidfVectorizer for text feature extraction. TfidfVectorizer uses the N-gram range parameter to determine the length of N-grams as features in the extraction process. N-grams, which consist of a sequence of n- words in the text, help the model capture word patterns more accurately. This research applies five types of n-grams, namely Unigram (1,1), Bigram (2,2), Trigram (3,3), Uni-Bigram (1,2), and Uni-Trigram (1,3). Table 4 shows an example of using the N-gram parameter.

Table 4. TF-IDF N-gram Examples

N-Gram	Tweet
Unigram	[alay], [seru], [t*lol]
Bigram	[alay, seru], [seru,alay]
Trigram	[alay, seru, alay]
Unigram + Bigram	[alay], [seru], [t*lol], [alay, seru], [seru,alay]
Unigram + Trigram	[alay], [seru], [t*lol], [alay, seru, alay]

2.5. Semantic Feature (IndoBERT)

IndoBERT is a specially designed Indonesian language model. IndoBERT is built using BERT (Bidirectional Encoder Representations from Transformers) architecture [8]. This model uses a transformer architecture that understands a word in a sentence by considering the words before and after it (bidirectional) [17]. One of the advantages of IndoBERT is the ability to detect the semantic meaning of words in Indonesian text.

This research uses IndoBERT to generate semantic representations of text used for classification models. The process of using IndoBERT starts with text tokenization using the IndoBERT tokenizer, namely 'indobenchmark/IndoBERT-base-p' which can convert text into numeric tokens through the model. The output of this process will store the semantic information of the sentence that will be used for the classification model. Table 5 demonstrates an example of the semantic representation of the processed sentences.

Table 5. Example of Semantic Representation

Sentence	D-0	D-1	…	D-768
0	-0.752607	0.464577	...	0.130451
1	-0.246513	0.399060	...	0.396180
...	...	...	...	...
30084	-0.031182	0.299258	...	0.725595

2.6. TF-IDF combined with IndoBERT

TF-IDF feature extraction in this research is combined with semantic features, namely IndoBERT, to add features used in the classification model. To improve the semantic representation of the text, the features obtained from IndoBERT can capture the context and deeper meaning of the text, which cannot be reached by TF-IDF.

The merging process starts after TF-IDF produces features based on N-grams that assess how important each word in the sentence is. The output of TF-IDF is then combined with the semantic representation generated by IndoBERT, which has been processed previously. The process of merging the two methods integrates different approaches by combining the advantages of each method to obtain a more stable text representation. The merging process can be seen in Figure 2.

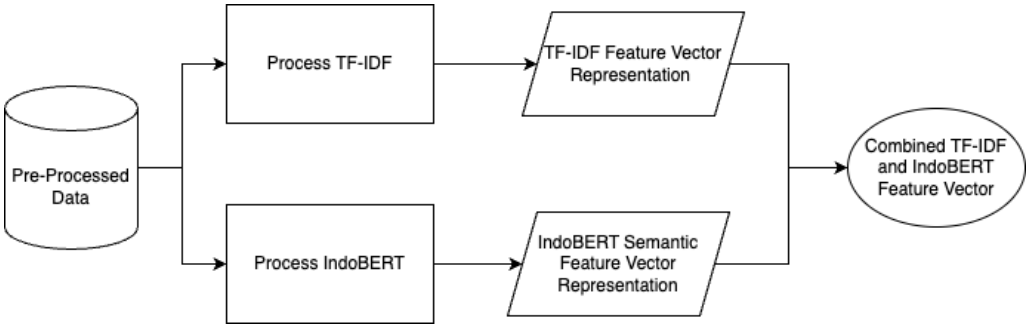

Figure 2. Process of Merging TF-IDF and IndoBERT

2.7. Feature Expansion FastText

FastText is a text classification model developed by Facebook AI Research. FastText is used in this research to measure the degree of similarity between words, utilizing Facebook's proprietary library that provides up to 600 billion word vectors [32]. In using FastText, text is converted into vectors and words into n-grams. The main purpose of using FastTesxt is to find and fill empty word vectors using similarity words in order to get the maximum value [33]. This research applies three types of corpus, namely Indonews, tweets, and a combination of Indonews and tweets. Table 6 shows the total data used in creating the corpus from each source.

Table 6. FastText Build Corpus

Corpus	Sum
Indonews	127.580
Tweet	30.084
Indonews + Tweet	157.664

Table 7 shows the top 10 words with similar meaning to the word "B*go" from the similarity corpus created using FastText.

Table 7. Similarity Words of 'b*go'

Rank	Similar Word	Value
1	G*blok	0.6309276819229126
2	T*lol	0.6227126717567444
3	B*doh	0.5940006971359253
4	J*lek	0.5793975591659546
5	K*nyol	0.5601352453231812
6	K*cak	0.5350925922393799
7	B*nci	0.5085372924804688
8	Egois	0.5005489587783813
9	Takut	0.4952262043952942
10	Ceroboh	0.4911308288574219

2.8. Classification Alghoritm

There are four classification models used in this research, namely Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), hybrid deep learning CNN-LSTM and CNN-LSTM-IndoBERT models. Convolutional Neural Network (CNN) is one of the networks that is frequently used in the field of natural language processing (NLP), especially for tasks involving text data [13]. CNNs utilize convolutional layers to extract vectors from input data, which involves the use of filters or kernels that move over the input data to detect vectors [34].

This CNN architecture starts with an input layer that receives data in vector format. A Conv1D layer with 32 filters, a kernel of 5, and a ReLU activation function is used to extract important text features. A MaxPooling1D layer of size 5 then performs down-sampling. To prevent overfitting, a Dropout layer of size 0.3 is used. After the convolution and pooling outputs are flattened by flattening, two density layers are used to process them. The first layer has 32 neurons and ReLU, and the second layer has 2 neurons and sigmoid for binary classification. The model uses a binary crossentropy loss function and Adam optimizer, with EarlyStopping used to prevent overfitting. Figure 3 shows the architecture of CNN.

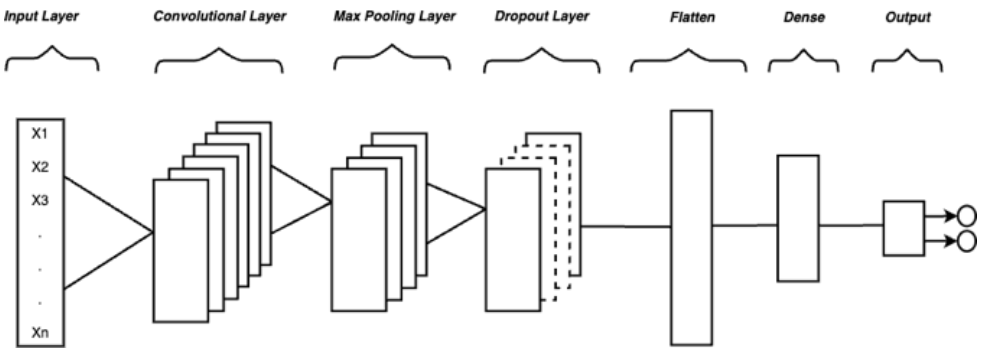

Figure 3. CNN Architecture

Recurrent Neural Network (RNN) has several types, one of which is Long Short-Term Memory (LSTM), which can detect long-term dependencies in sequenced data [35]. The LSTM method is designed by replacing the standard neurons in the RNN with LSTM cells. The LSTM architecture has 3 gates, namely Forget Gate, Input Gate, and Output Gate [36]. Forget gates function to control how much old information in cell memory needs to be forgotten. Input gates function to regulate how much new information is obtained in cell memory, and the last output gates function to determine the information taken from cell memory to produce output at a certain time.

This study's LSTM model contains a sequential data input layer, a 128-unit LSTM layer, and a recurrent_dropout of 0.2 to minimize overfitting. Non-linear relationships are then captured using an LSTM layer with 32 neurons using ReLU activation. A Dense output layer with one neuron and sigmoid activation was employed for binary classification. The Adam optimizer, which has a learning rate of 0.001, and binary_crossentropy as the loss function were used to construct the model. EarlyStopping is used to end training early if val_loss does not improve. Figure 4 shows the architecture of LSTM.

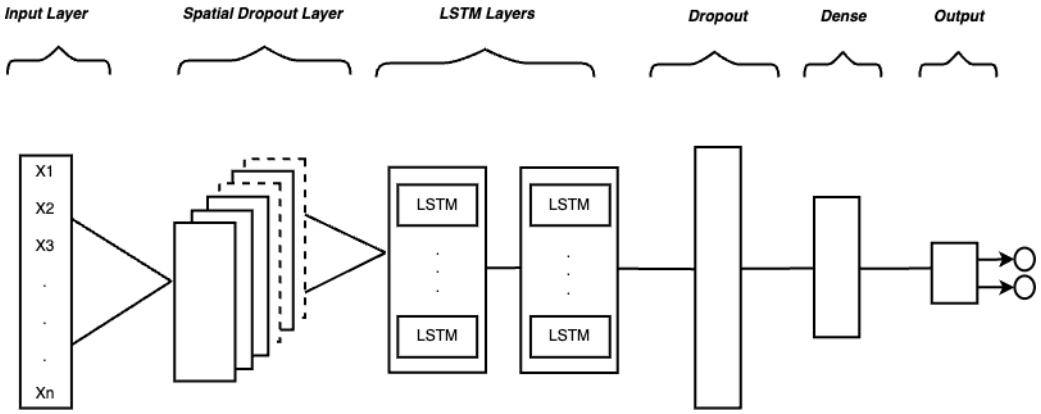

Figure 4. LSTM Architecture

The hybrid deep learning model in this research combines the architecture of CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory) models. The main advantage of CNN-LSTM lies in its ability to combine CNN in extracting local features of text, such as letter patterns, words, and phrases [37]. While LSTM's ability to understand context and sequence information. This combination results in a richer and more informative representation of the text, allowing the model to better capture dependencies between words and sentences [38]. Not only that, but this research also combines the IndoBERT model with hybrid deep learning.

In this study, the CNN-LSTM hybrid model consists of two LSTM layers with 128 units, Flatten, Dense with 128 units and ReLU activation, Conv1D with 128 filters and kernel size 1, SpatialDropout1D with rate 0.1, MaxPooling1D with pool size 1, Dropout with rate 0.5, and Dense output layer with 1 unit and sigmoid activation. The architecture of the hybrid model in this research can be seen in Figure 5.

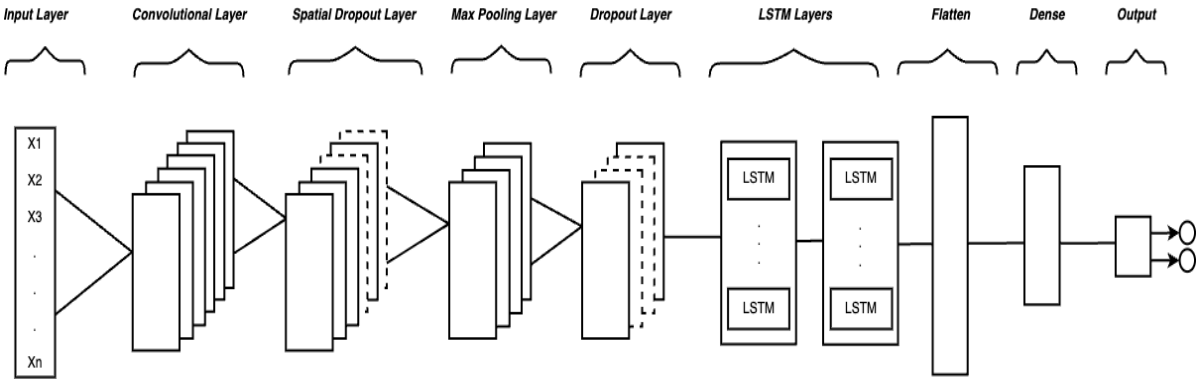

Figure 5. Hybrid Deep Learning CNN- LSTM Architecture

2.9. Performance Evaluation

A confusion matrix was used in this research to evaluate performance. A confusion matrix can estimate how well the model distinguishes between classes by comparing the actual classification and prediction of the model [39]. A confusion matrix has four combinations that describe the classification results, namely True Positive (TP) for correctly predicted positive data, False Positive (FP) for incorrectly predicted positive data, True Negative (TN) for correctly predicted negative data, and False Negative (FN) for incorrectly predicted negative data [40]. The results of the confusion matrix obtain values to evaluate how well the performance of the classification model, namely Accuracy, Presicion, Recall, and F1-score are weighted. The following is the formula used to calculate these four values:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (4)$$

$$Presicion = \frac{TP}{(TP+FP)} \quad (5)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

$$F1-score = 2 \times \frac{Recall \times Presicion}{Recall \times Presicion} \quad (7)$$

3. RESULT

This research will undertake five scenarios to detect cyberbullying in Indonesia using multiple classification approaches, including CNN, LSTM, hybrid deep learning CNN-LSTM and CNN-LSTM-IndoBERT. Each scenario builds on the prior one (for example, scenario 3 mentions scenarios 1 and 2). This study evaluates five situations. Table 8 provides a description of each scenario.

Table 8. Description of Each Scenario

Scenario	Description
1	Exploring the effect of training and testing data ratios to achieve the best configuration for optimal model performance.
2	Exploring the maximum number of feature vectors in TF-IDF at a baseline splitting ratio.
3	Testing the use of N-Gram parameters on TF-IDF with a focus on unigram, bigram, and their combinations.
4	Applying feature expansion using FastText on baseline models to improve performance.
5	Integrating TF-IDF feature extraction, IndoBERT semantic features, and FastText expansion in hybrid deep learning models to achieve optimal accuracy.

3.1. Exploring Data Ratios for Model Performance

This research has a first scenario that uses a spilitting ratio that determines the best baseline model. CNN uses parameters such as filter 32, kernel size 5, batch size 16, and epoch 30. LSTM uses parameters of unit 128, dropout 0.5, spatial dropout 0.25, learning rate 0.001, and epoch 30. This research uses data sharing ratios of 70:30, 80:20, and 90:10. The data sharing ratio of 70:30 means 70% of the data is for research and 30% for testing. Table 9 shows the test results of each model in the first scenario. The results indicate that both the CNN model, with an accuracy of 79.41%, and the LSTM model, with an accuracy of 78.77%, both demonstrate superior performance with the 90:10 splitting ratio compared to the other ratios. These findings highlight the efficiency of the 90:10 ratio in maximizing model performance. As a result, the results from the previous scenario, particularly with the 90:10 ratio, will be applied in the following test scenario. The results in the first scenario will be used in the next test scenario.

Table 9. Result of The First Scenario Test

Splitting Ratio	Accuracy (%)	
	CNN	LSTM
70:30	79.38	78.22
80:20	79.37	78.09
90:10	79.41	78.77

3.2. Investigating Optimal Feature Vectors in TF-IDF

Table 10. Results of The Second Scenario Test

Max Feature	Accuracy (%)	
	CNN	LSTM
2500	79.41	78.77
3000	79.47	78.75
3500	79.49	79.25
4000	79.39	78.57

The second scenario is to apply the baseline splitting ratio of 90:10 with feature vectors 2,500, 3,000, 3,500, 4,000. Table 10 shows the best accuracy results using 3,500 feature vectors with 79.49% CNN and 79.25% LSTM models. The higher accuracy achieved with 3,500 feature vectors demonstrates

the efficacy of this feature size in both models. The following test scenario will include 3,500 characteristics because they produced the best results.

3.3. Analyzing N-Gram Parameters in TF-IDF

The third scenario is conducted using TF-IDF for word weighting with N-grams. This scenario produces accuracy that will be compared to determine the best results from using several combinations of N-Gram parameters. Table 11 shows results of the third scenario, which has the best accuracy on the combined use of TF-IDF and Unigram-Bigram with 79.65% accuracy on the CNN model and 79.46% accuracy on the LSTM model. The results highlight the strength of the TF-IDF Unigram-Bigram model in improving accuracy over other combinations. Based on these findings, the TF-IDF Unigram-Trigram model will be used in the next test scenario.

Table 11. Results of The Third Scenario Test

N-Gram	Accuracy (%)	
	CNN	LSTM
Unigram	79.49	79.25
Bigram	69.49	68.86
Trigram	52.94	52.17
Unigram - Bigram	79.65	79.46
Unigram - Trigram	79.56	77.86

3.4. Enhancing Models with FastText Feature Expansion

The fourth scenario is to apply feature expansion to the baseline model. The feature expansion used in this research is FastText. Testing is done by selecting the top features from the corpus. There are three corpus used, namely Indonews, Tweet, and a combination of Tweet with Indonews. The top features used are top 1, top 5, and top 10. The results in this table highlight the effectiveness of the combined Tweet+Indonews corpus in improving the performance of both CNN and LSTM models. The best accuracy in the Tweet+Indonews corpus ranks top 1 can be seen in Table 12.

Table 12. Results of The Fourth Scenario Test

Model	Rank	Accuracy (%)			
		Baseline	Corpus Tweet	Corpus News	Corpus Tweet + News
CNN	Top 1	79.65	80.24 (+0.59)	80.51 (+0.86)	80.69 (+1.04)
	Top 5		79.71 (+0.06)	80.29 (+0.64)	80.17 (+0.52)
	Top 10		79.16 (-0.49)	79.97 (+0.32)	80.53 (0.88)
LSTM	Top 1	79.46	79.85 (+0.39)	80.12 (+0.66)	80.67 (+1.21)
	Top 5		79.46 (+0.00)	80.17 (+0.71)	80.11 (+0.65)
	Top 10		79.43 (-0.03)	79.23 (-0.23)	79.71 (+0.25)

3.5. Integrating TF-IDF, IndoBERT, and FastText in Hybrid Models

The fifth scenario is to apply feature expansion to the hybrid model and TF-IDF feature extraction and create a word similarity corpus built using FastText. The hybrid model is built with CNN and LSTM for the baseline. While the CNN-LSTM-IndoBERT model was built with TF-IDF feature extraction combined with IndoBERT and then continued with FastText feature expansion. Testing is done by selecting the top features from the corpus. This research uses three corpus, namely Indonews, Tweet, and combination of Indonews with Tweet. Table 13 shows that the highest accuracy in this study is achieved by the CNN-LSTM model, which ranks first in the Indonews corpus, and the CNN-LSTM-IndoBERT model, which ranks first in the combined Tweet+Indonews corpus. This demonstrates that

combining TF-IDF feature extraction with IndoBERT semantic features and feature expansion in the hybrid deep learning model significantly improves accuracy compared to the baseline.

Table 13. Results of The Fifth Scenario Test

Model	Rank	Accuracy (%)		
		Corpus Tweet	Corpus News	Corpus Tweet + News
CNN+ LSTM	Top 1	80.76	81.18	80.39
	Top 5	80.49	80.72	79,96
	Top 10	79.75	80.05	79.36
CNN+ LSTM+ IndoBERT	Top 1	80.68	81.28	82.05
	Top 5	80.95	80.93	81.18
	Top 10	80.15	80.12	80.68

3.6. Comparison of baseline model accuracy using FastText

The comparison of baseline model accuracy shows a significant increase in performance. The highest accuracy was obtained in the CNN+LSTM model with the addition of a combination of TF-IDF and IndoBERT, and FastText at 82.05% which had an accuracy increase of 2.64% in the CNN model and an accuracy increase of 3.28 in the LSTM model. The comparison of the accuracy of the baseline model using FastText can be seen in table 14.

Table 14. Comparison of Baseline Model Accuracy Using FastText

Model	Accuracy (%)		
	Baseline	CNN + LSTM	CNN+LSTM +IndoBERT
CNN	79.41	81.18 (+1.77)	82.05 (+2.64)
LSTM	78.77	81.18 (+2.41%)	82.05 (+3.28)

4. DISCUSSIONS

Testing scenarios were conducted to select the best baseline model through spiltting ratio, feature vectors, combining N-gram values with TF-IDF, as well as corpus with top rank in applying FastText feature expansion. In the first scenario, the test results get the best baseline using a splitting ratio of 90:10, whose accuracy value is 79.41% and LSTM accuracy value is 78.77%.

In the second scenario, the use of 3,500 feature vectors improved the accuracy of the model. In this scenario, the CNN model achieved 79.49% accuracy, while the LSTM model achieved 79.25%, both using a splitting ratio of 90:10. The increase in accuracy compared to the baseline in the first scenario is 0.08% for the CNN model and 0.48% for the LSTM model.

The third scenario shows a more significant improvement by combining TF-IDF with Unigram-Bigram. The use of a richer combination of n-grams helps the model capture the context and relationship between words in the text. The use of n-grams expands the feature representation, helping the model handle language variations and text data complexity, thus improving the accuracy of the model. The CNN model recorded 79.65% accuracy and the LSTM model recorded 79.46% accuracy. Compared to the baseline in the first scenario, the accuracy improvement was 0.24% for the CNN model and 0.69% for the LSTM model.

The fourth scenario shows that further accuracy improvement is achieved with the use of the top-ranked corpus (Tweet+Indonews top 1). This is due to a higher quality corpus, which has more relevant and representative information, which supports better model performance. The accuracy of the CNN model increased by 1.28% and the LSTM model increased by 1.90% compared to the baseline in the first scenario.

The fifth scenario demonstrates the most significant improvement, primarily due to the integration of CNN-LSTM and IndoBERT models. IndoBERT, which leverages customized language embeddings, enhances text feature representation and significantly increases accuracy. The results show that the CNN-LSTM hybrid model achieved the highest accuracy in the Indonews corpus, ranking first with 81.18%, while the CNN-LSTM-IndoBERT model ranked first in the combined Tweet+Indonews corpus with 82.05%. The accuracy improvement of the CNN-LSTM hybrid deep learning model over the baseline was 1.53% for the CNN model and 1.72% for the LSTM model. Meanwhile, the CNN-LSTM-IndoBERT hybrid model outperformed the baseline CNN by 2.40% and the baseline LSTM by 2.59%. The increase in accuracy value of all test scenarios is shown in Figure 6.

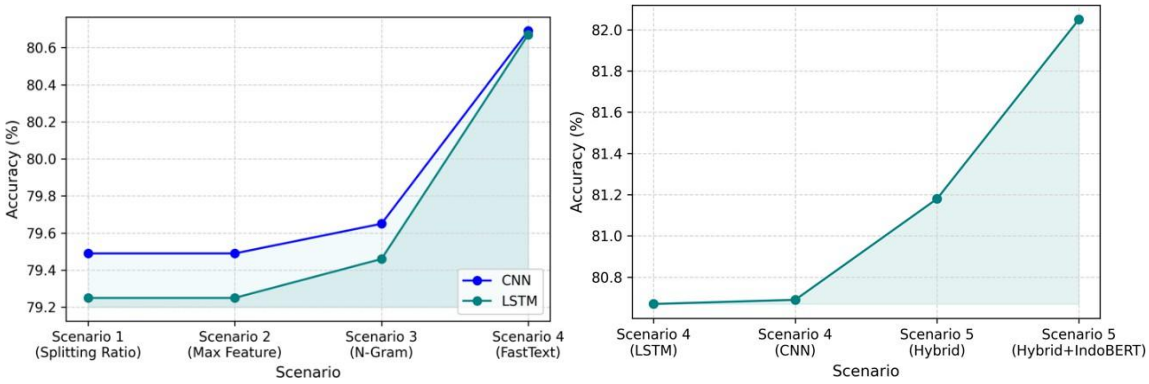

Figure 6. Increased Accuracy of All Scenarios

This study also validated the improvement in accuracy between the test scenarios using statistical significance tests. The statistical significance of the accuracy differences between the test scenarios was assessed using the concepts of P-Value and Z-Value. A P-Value < 0.05 and a Z-Value > 1.96 indicate a significant improvement in accuracy between the test scenarios. The improvement in accuracy between the results of scenario 1 (baseline) and scenario 5 (hybrid deep learning) is statistically significant, as shown in Table 15.

Table 15. Statistical Accuracy Significant

Parameters	Scenario				
	S1→S2	S2→S3	S3→S4	S4→S5	S1→S5
P-Value	0.413	0.227	0.0	0.032	0.0
Z-Value	0.819	1.209	11.777	2.146	4.675
Significant?	False	False	True	True	True

This research has conducted trials on related datasets using the method proposed in the previous research by Yudi Widhiyasana et al. [38]. The results of the trial resulted in an accuracy of 74.51%. This proves that the C-LSTM model in the previous study is not effective in detecting cyberbullying because it does not use feature extraction with IndoBERT and feature expansion. The understanding of semantic relationships between words is limited in previous research because it uses general techniques for text encoding without feature extraction with IndoBert and feature expansion. For a more contextual representation of language, this study used IndoBERT, which improved the performance of the model by 7.54% when compared to the C-LSTM base model in the previous study. In addition, TF-IDF creates a more organized representation of words than raw text, while FastText extends the meaning of words based on context similarity. The combination of these three strategies allows the model to capture text patterns more accurately than previous studies.

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4321

The accuracy of cyberbullying detection has been demonstrated to increase with the inclusion of CNN-LSTM hybrid model with TF-IDF, IndoBERT, and FastText. In order to make the TF-IDF extraction results more significant and pertinent, IndoBERT enhances the feature representation by more thoroughly capturing the semantic context of the Indonesian language. This result is proven by the comparison of accuracy, precision, recall, and f1-score as shown in Table 16.

Table 16. Comparison of The Best Models

Model	Acc	Precision	Recall	F1-Score
CNN	80.69	80.62	80.60	80.61
LSTM	80.67	81.00	83.19	82.08
CNN+LSTM	81.18	81.61	83.46	82.52
CNN+LSTM+IndoBERT	82.05	83.66	82.57	83.11

5. CONCLUSION

In this study, the researchers successfully developed a hybrid deep learning model for detecting cyberbullying in Indonesian tweets. This hybrid deep learning model uses TF-IDF feature extraction combined with IndoBERT and FastText as expansion features. This research uses an Indonesian dataset from platform "X" consisting of 30,084 tweets, with 15,005 tweets labeled as cyberbullying and 15,079 tweets labeled as non-cyberbullying. This dataset is analyzed using five classification models namely CNN, LSTM, Hybrid deep learning CNN-LSTM, and CNN-LSTM-IndoBERT.

In CNN, LSTM, and hybrid deep learning CNN-LSTM models, the feature extraction used is TF-IDF to produce a text vector representation. In comparison, the CNN-LSTM-IndoBERT model combines TF-IDF feature extraction with semantic features from IndoBERT. TF-IDF feature extraction combined with IndoBERT semantic features is done to see the influence of IndoBERT in cyberbullying classification in the hybrid deep learning model.

In addition, the FastText corpus consists of three dataset sources, namely Indonews (127,580 data), Tweets (30,084 data), and a mixture of both (157,664 data). This combination increases the ability to capture semantic patterns that cannot be found by TF-IDF and IndoBERT individually.

The best test results can be seen in the CNN-LSTM-IndoBERT hybrid model with a high accuracy of 82.05% which shows an increase in accuracy of 2.64% against the CNN initial baseline and 3.28% against the LSTM initial baseline so that the TF-IDF feature combined with IndoBERT is able to increase accuracy in detecting cyberbullying compared to other models. This study shows that the use of IndoBERT together with TF-IDF can effectively improve the classification performance in detecting cyberbullying.

This research makes a significant contribution to the development of a deep learning-based cyberbullying detection system in Indonesian. This research contributes to informatics by integrating hybrid deep learning (CNN-LSTM) with IndoBERT and TF-IDF, demonstrating its effectiveness in improving cyberbullying detection in Indonesian text. The integration of IndoBERT as a semantic representation has enriched the quality of the features, while FastText expands the scope of relevant semantic patterns. Thus, this research broadens the insight on how semantic features and hybrid techniques can be used to improve text classification performance.

However, this research has several limitations, including the potential bias in the dataset coming from platform X, which might not fully represent the variety of linguistic situations and variances found in the actual world. Furthermore, the effectiveness of cyberbullying identification in these kinds of texts may be impacted by the model's poor ability to handle linguistic variances like slang or mixed codes that are frequently used in tweets. For future research, it is recommended to explore other pre-trained transformer models, conduct real-world deployment testing, and integrate additional linguistic and contextual features to improve cyberbullying detection performance.

REFERENCES

[1] R. Rosemary, A. B. Wardhana, H. M. Syam, and N. Susilawati, "The Relationship Between Anonymity and Cyber Sexual Harassment by Twitter Users: A Cross-Sectional Study," *Journal of Community Mental Health and Public Policy*, vol. 6, no. 2, pp. 95–104, Apr. 2024, doi: 10.51602/cmhp.v6i2.131.

[2] A. P. Riyadisty and E. Fauziati, "Hate Expression Found on Twitter as a Response to Meghan Markle," *Indonesian Journal of English Language Studies (IJELS)*, vol. 8, no. 1, pp. 45–51, Mar. 2022, doi: 10.24071/ijels.v8i1.4421.

[3] S. Khan *et al.*, "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4335–4344, Jul. 2022, doi: 10.1016/j.jksuci.2022.05.006.

[4] F. Husain and O. Uzuner, "A Survey of Offensive Language Detection for the Arabic Language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 12, Apr. 2021, doi: 10.1145/3421504.

[5] A. Law School, "Legal Challenges of Cyberbullying and Online Harassment: A Comparative Analysis Shashank Mittal," vol. 6, no. 2, Mar. 2024, doi: https://doi.org/10.36948/ijfmr.2024.v06i02.19295.

[6] D. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi, "Cyberbullying on social media under the influence of COVID-19," *Global Business and Organizational Excellence*, vol. 41, no. 6, pp. 11–22, Sep. 2022, doi: 10.1002/joe.22175.

[7] G. Ray, C. D. McDermott, and M. Nicho, "Cyberbullying on Social Media: Definitions, Prevalence, and Impact Challenges," Sep. 01, 2024, *Oxford University Press*. doi: 10.1093/cybsec/tyae026.

[8] A. Candra, Wella, and A. Wicaksana, "Bidirectional encoder representations from transformers for cyberbullying text detection in indonesian social media," *International Journal of Innovative Computing, Information and Control*, vol. 17, no. 5, pp. 1599–1615, Oct. 2021, doi: 10.24507/ijicic.17.05.1599.

[9] S. Ge, L. Cheng, and H. Liu, "Improving cyberbullying detection with user interaction," in *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, Association for Computing Machinery, Inc, Apr. 2021, pp. 496–506. doi: 10.1145/3442381.3449828.

[10] D. Upadhyay, H. Singhdev, and N. Mohd, "Text Classification Using CNN and CNN-LSTM," *Webology*, vol. 18, 2021, doi: 10.29121/web/v18i4/149.

[11] I. Tabassum and V. Nunavath, "A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data," *Applied Sciences (Switzerland)*, vol. 14, no. 24, Dec. 2024, doi: 10.3390/app142412007.

[12] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," vol. 21, no. 6, Nov. 2024, doi: 10.34028/iajit/21/6/9.

[13] D. Sultan, M. Mendes, A. Kassenkhan, and O. Akylbekov, "Hybrid CNN-LSTM Network for Cyberbullying Detection on Social Networks using Textual Contents," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023, doi: 10.14569/IJACSA.2023.0140978.

[14] 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," *MDPI journals*, vol. 15, no. 5, May 2023, doi: 10.3390/fi15050179.

[15] C. Emmery *et al.*, "Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity," *Lang Resour Eval*, vol. 55, no. 3, pp. 597–633, Sep. 2021, doi: 10.1007/s10579-020-09509-1.

[16] L. Cheng, D. Hall, and H. Liu, "Session-based Cyberbullying Detection: Problems and Challenges," *IEEE Internet Comput*, vol. 25, no. 2, Oct. 2020, doi: 10.1109/MIC.2020.3032930.

[17] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *Proceedings of the 28th International*

Conference on Computational Linguistics, pp. 757–770, Dec. 2020, doi: 10.18653/v1/2020.coling-main.66.

[18] A. J. Andika, Y. Kristian, and E. I. Setiawan, "Detection of Cyberbullying Comments on Youtube Social Media Using Convolutional Neural Network – Long Short Term Memory Network (CNN-LSTM) Method," *Teknika*, vol. 12, no. 3, pp. 183–188, Oct. 2023, doi: 10.34148/teknika.v12i3.677.

[19] D. Y. Yefferson, V. Lawijaya, and A. S. Girsang, "Hybrid model: IndoBERT and long short-term memory for detecting Indonesian hoax news," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 1911–1922, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1913-1924.

[20] I. A. Asqolani and E. B. Setiawan, "A Hybrid Deep Learning Approach Leveraging Word2Vec Feature Expansion for Cyberbullying Detection in Indonesian Twitter," *Ingenierie des Systemes d'Information*, vol. 28, no. 4, pp. 887–895, Aug. 2023, doi: 10.18280/isi.280410.

[21] A. Jalilifard, V. F. Caridá, A. F. Mansano, R. S. Cristo, and F. P. C. da Fonseca, "Semantic Sensitive TF-IDF to Determine Word Relevance in Documents," Jan. 2020, doi: 10.1007/978-981-33-6977-1.

[22] W. Anggraeni, M. F. A. Kusuma, E. Riksakomara, R. P. Wibowo, Pujiadi, and S. Sumpeno, "Combination of BERT and Hybrid CNN-LSTM Models for Indonesia Dengue Tweets Classification," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 1, pp. 813–826, 2024, doi: 10.22266/ijies2024.0229.68.

[23] S. A. Sazan, M. H. Miraz, and A. B. M. Muntasir Rahman, "Enhancing Depressive Post Detection in Bangla: A Comparative Study of TF-IDF, BERT and FastText Embeddings," *Annals of Emerging Technologies in Computing*, vol. 8, no. 3, pp. 34–49, Jul. 2024, doi: 10.33166/AETiC.2024.03.003.

[24] D. Fabillah, R. Auliarahmi, S. D. Setiarini, and T. Gelar, "The Investigation of Convolution Layer Structure on BERT-C-LSTM for Topic Classification of Indonesian News Headlines," *Journal of Software Engineering, Information and Communication Technology (SEICT)*, vol. 4, no. 2, pp. 105–116, 2021, doi: 10.17509/seict.v4i2.63742.

[25] F. Baharuddin and M. F. Naufal, "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, Oct. 2023, doi: 10.20473/jisebi.9.2.253-263.

[26] Z. Ahanin, M. A. Ismail, N. S. S. Singh, and A. AL-Ashmori, "Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages," *Sustainability (Switzerland)*, vol. 15, no. 16, Aug. 2023, doi: 10.3390/su151612539.

[27] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[28] D. A. Komara and A. Hadiapurwa, "AUTOMATING TWITTER DATA COLLECTION: A RAPIDMINER-BASED CRAWLING SOLUTION," *PUBLIS JOURNAL*, vol. 6, Nov. 2022, doi: 10.24269/pls.v6i2.6326.

[29] A. Zhdanovskaya, D. Baidakova, and D. Ustalov Toloka, "Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions," vol. 37, no. 13, 2023, doi: https://doi.org/10.1609/aaai.v37i13.26886.

[30] D. Rifaldi, Abdul Fadlil, and Herman, "PREPROCESSING TECHNIQUES IN TEXT MINING: 'MENTAL HEALTH' TWEET DATA," *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 3, no. 2, pp. 161–171, Apr. 2023, doi: 10.51454/decode.v3i2.131.

[31] W. Yulita, M. C. Untoro, M. Praseptiawan, I. F. Ashari, A. Afriansyah, and A. N. Bin Che Pee, "Automatic Scoring Using Term Frequency Inverse Document Frequency Document Frequency and Cosine Similarity," *Scientific Journal of Informatics*, vol. 10, no. 2, pp. 93–104, Apr. 2023, doi: 10.15294/sji.v10i2.42209.

[32] S. D. Lestari and E. B. Setiawan, "Sentiment Analysis Based on Aspects Using FastText Feature Expansion and NBSVM Classification Method," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 469–477, Sep. 2022, doi: 10.47065/josyc.v3i4.2202.

[33] A. Raihan and E. B. Setiawan, "Aspect Based Sentiment Analysis with FastText Feature Expansion and Support Vector Machine Method on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 591–598, Aug. 2022, doi: 10.29207/resti.v6i4.4187.
[34] M. A. S. Nasution and E. B. Setiawan, "Enhancing Cyberbullying Detection on Indonesian Twitter: Leveraging FastText for Feature Expansion and Hybrid Approach Applying CNN and BiLSTM," *Revue d'Intelligence Artificielle*, vol. 37, no. 4, pp. 929–936, Aug. 2023, doi: 10.18280/ria.370413.
[35] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information*, vol. 15, no. 9, p. 517, Aug. 2024, doi: 10.3390/info15090517.
[36] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," vol. 404, Mar. 2020, doi: 10.1016/j.physd.2019.132306.
[37] H. Elzayady, K. M. Badran, and G. I. Salama, "Arabic Opinion Mining Using Combined CNN - LSTM Models," *International Journal of Intelligent Systems and Applications*, vol. 12, no. 4, pp. 25–36, Aug. 2020, doi: 10.5815/ijisa.2020.04.03.
[38] Y. Widhiyasana, T. Semiawan, I. Gibran, A. Mudzakir, and M. R. Noor, "Convolutional Long Short-Term Memory Implementation for Indonesian News Classification," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi* |, vol. 10, no. 4, pp. 354–361, Nov. 2021, doi: 10.22146/jnteti.v10i4.2438.
[39] E. Helmud, E. Helmud, F. Fitriyani, and P. Romadiana, "Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 1, pp. 92–97, Feb. 2024, doi: 10.32736/sisfokom.v13i1.1985.
[40] D. Widyawati, A. Faradibah, and P. L. L. Belluano, "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indonesian Journal of Data and Science*, vol. 4, no. 2, pp. 78–87, Jul. 2023, doi: 10.56705/ijodas.v4i2.76.

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 655-672
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4321