Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 11, No. 2, June 2025, pp. 290-304
ISSN: 2338-3070, DOI: 10.26555/jiteki.v11i2.30687

Depression Detection on Social Media X Using Hybrid Deep Learning CNN-BiGRU with Attention Mechanism and FastText Feature Expansion

I Wayan Abi Widiarta, Erwin Budi Setiawan
Faculty of Informatics, Telkom University, Jl. Terusan Buah Batu, Bandung 40257, Indonesia

ARTICLE INFO

Article history:

Received February 10, 2025
Revised March 29, 2025
Accepted June 20, 2025

Keywords:

Mental Health;
CNN;
BiGRU;
FastText;
Attention Mechanism

ABSTRACT

Depression is a global mental health disorder affecting over 280 million people, with significant challenges in identifying sufferers due to societal stigma. In Indonesia, the National Adolescent Mental Health Survey in 2022 revealed that 17.95 million adolescents experience mental health disorders, with a portion of them suffering from depression. Social media platform X offers an alternative for individuals to share their mental health status anonymously, bypassing societal stigma. This study proposes a hybrid deep learning model combining CNN and BiGRU with an attention mechanism, TF-IDF for feature extraction, and FastText for feature expansion to detect depression in Indonesian tweets. The dataset comprises 50,523 Indonesian tweets, supplemented by a similarity corpus of 151,117 data. To optimize model performance, five experimental scenarios were conducted, focusing on split ratios, n-gram configurations, maximum features, feature expansion, and attention mechanisms. The main contribution of this research is the novel integration of FastText for feature expansion and the attention mechanism within a CNN-BiGRU hybrid model for depression detection. The results demonstrate the effectiveness of this combination, with the BiGRU-ATT-CNN-ATT model achieving an accuracy of 84.40%. However, challenges such as handling noisy, ambiguous social media data and addressing out-of-vocabulary words remain. Future research should explore additional feature expansion techniques, optimization algorithms, and approaches to handle noisy data, improving model robustness for real-world applications in mental health detection.

Corresponding Author:

Erwin Budi Setiawan, Telkom University, Jl. Terusan Buah Batu, Bandung 40257, Indonesia
Email: erwinbudisetiawan@telkomuniversity.ac.id

1. INTRODUCTION

Depression is a mental health disorder characterized by feelings of sadness, hopelessness, sleep disturbances, and decreased appetite [1] . Depression has been recognized as a significant global mental health problem and the leading cause of more than two-thirds of suicides each year [2]. According to the World Health Organization (WHO), approximately 280 million people worldwide suffer from depression [3]. A survey in 2022 called the Indonesian National Adolescent Mental Health Survey (I-NAMHS) reported that 17.95 million Indonesian adolescents were diagnosed with mental health problems. Among them, 1.0% had major depressive illness [4]. Identifying individuals with depression remains a challenge due to the negative stigma in society associated with mental health problems that often prevent individuals from openly discussing their condition with family or close relatives [5]. However, the emergence of social media in society offers an alternative for individuals to communicate their mental health problems anonymously to avoid societal stigma [5]. Social media allows users to share text, images, audio, and video in a digital space. X is one of the most widely used social media platforms, offering users a place to publish original thoughts and feelings [6].

The use of social media data in mental health detection research requires careful attention to ethical considerations, such as the protection of user privacy and the potential negative impact on social stigma. According to the journal Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice [7], while social media offers opportunities to gain valuable insights into individuals' mental states, the use of this data also carries risks related to privacy and confidentiality. In this context, it is important to obtain explicit consent from users to use their data, as well as ensure that the data is not misused. In addition, the social stigma attached to mental health conditions may add to the ethical challenges in this research, as the use of social media data may put individuals at further risk of social stigma or discrimination. Therefore, research should be conducted with strict ethical principles, which ensure that the rights of individuals are respected and data is used safely and responsibly.

User-generated data on X has the potential to detect patterns related to mental health disorders like depression. To overcome societal stigma and leverage social media data, a reliable detection system is needed. Several studies have applied machine learning, deep learning, and hybrid deep learning methods to perform depression detection from social media. For Example, Research [5] used a hybrid deep learning method, CNN + LSTM, with Word2Vec and TF-IDF feature extraction. The datasets used were from YouTube, Twitter, Facebook, and additional data from Kaggle and GitHub. The CNN + LSTM model achieved the highest accuracy of 98.54% with TF-IDF and 99.02% with Word2Vec. However, this study did not incorporate feature expansion, which could potentially improve performance. Other research [8] compared the performance of CNN, RNN, and hybrid CNN-BiLSTM models, using Twitter data crawled via the provided API. The dataset was divided into three sub-sections: D1 (depressed: 292,564 tweets), D2 (non-depressed: over 10 billion tweets), and D3 (potentially depressed: over 35 million tweets) [9]. After preprocessing, feature extraction was performed using CNN-based techniques, and word embeddings were applied to calculate the numerical vectors for each data point. The CNN + BiLSTM hybrid model achieved the highest accuracy of 94.28%, outperforming CNN and RNN models. A limitation of this study is the use of outdated datasets (2009-2013), which may not reflect current social media trends and user behavior [10]. Research using hybrid deep learning has also been conducted by Bendebane, et al [11]. This research uses 6 hybrid deep learning models namely CNN-GRU, CNN-BiGRU, CNN-LSTM, CNN-BiLSTM, CNN RNN, and CNN-BiRNN to perform classification. Glove is used as a method for feature extraction and expansion. The dataset used is the result of crawling on Twitter using API using keywords that indicate depression and anxiety. The total dataset after pre-processing is 3,178,579 data. Based on the testing that has been done, CNN-BiGRU gets the highest accuracy with a value of 93.38%. This research can be further improved by comparing several feature extraction and feature expansion algorithms.

As far as the researchers know, while various models have been developed for depression detection through social media, there has been limited exploration of feature expansion techniques and attention mechanisms, which have shown potential to improve text classification accuracy. Additionally, some existing models rely on older datasets that may not fully capture current social media communication patterns, particularly in Indonesia. Existing models have primarily used traditional feature extraction methods such as TF-IDF, Word2Vec, and Glove, and have not extensively explored more advanced techniques like FastText, nor have they fully integrated attention mechanisms. This presents an opportunity to enhance depression detection performance by incorporating newer techniques and adapting models to better address the linguistic challenges in Indonesian. Thus, while much progress has been made, there is still a gap in utilizing new techniques to improve depression detection performance, as well as in adapting these models to handle linguistic challenges in Indonesian. This research aims to fill the gap by integrating the FastText feature expansion method and attention mechanism in a hybrid CNN-BiGRU model, which is expected to provide higher accuracy in detecting depression through text in Indonesian. Despite the potential of this research, there are several limitations to consider. First, the application of FastText may be impacted by the morphological complexity of the Indonesian language, which could affect the accuracy of word representation. Second, while the study uses more current datasets, variability in user demographics could still impact model performance, as social media data is highly diverse. Finally, the hybrid CNN-BiGRU model's complexity could lead to longer training times and greater computational resource demands.

The research contribution of this study is the development and evaluation of a hybrid deep learning model that combines CNN and BiGRU, integrated with the attention mechanism, TF-IDF for feature extraction, and FastText for feature expansion. FastText is used for feature expansion due to its effectiveness during training compared to other feature expansion models [12]. The attention mechanism is used because of its proven ability to improve accuracy in text classification [13], [14]. Based on the literature study that has been conducted, the combination of feature expansion and attention mechanism for depression detection has not been widely applied. Therefore, this research focuses on the application of the FastText feature expansion technique and

attention mechanism combined with the CNN-BiGRU hybrid model to detect depression in the Indonesian language on X social media. Additionally, this study incorporates the use of current, more relevant data reflecting modern social media communication patterns, which enhances the model's ability to identify depression in the context of recent social media interactions. This research seeks to provide valuable insights into the development of an accurate and efficient depression detection system by utilizing text-based datasets from X social media platforms.

2. METHODS

In this research, several stages are passed before obtaining prediction results. These stages begin with crawling data in X to collect datasets according to the research topic. When the dataset has been collected, then the labeling process is carried out. Data that has been labeled will go through the data pre-processing stage to perform data cleaning. Data that has been pre-processed will go to the feature extraction stage using TF-IDF, corpus creation, and feature expansion. Before the classification of the model, the data is separated into training data (train) and test data (test). Next, the training and classification process is carried out using a predetermined scenario. The results of the classification will then be evaluated using an evaluation matrix [15], [16]. In this study, the CNN-BiGRU hybrid model was selected due to its ability to handle both spatial and sequential features of text. CNN effectively extracts local patterns and keyword features, which is important for detecting specific terms related to depression, as shown in previous works [17], [18]. On the other hand, BiGRU is specifically designed to process sequential information, which is essential in understanding the context of a sentence, especially in language-based tasks [19]. Since social media text, such as tweets, often relies on context and word order to convey meaning, BiGRU's bidirectional processing ensures that both preceding and succeeding words are considered. This bidirectional processing is crucial for handling the nuanced and contextual nature of depression-related language [20]. This makes BiGRU ideal for capturing the subtleties of language in depression detection, where understanding both the beginning and end of a statement is crucial. The design of the system built in this study can be seen in Fig. 1.

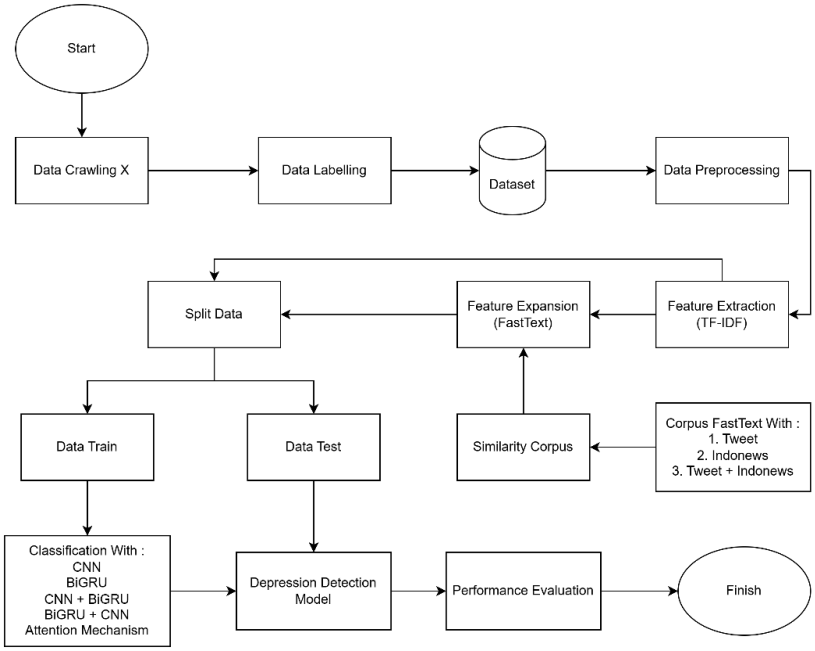

Fig. 1. Flowchart system

2.1. Data Collection and Data Labeling

The dataset used in this research is the result of crawling on social media X in Indonesian using APIs that are available online. The crawling process is carried out using keywords that have indications of depression. The crawling results managed to collect as much as 26,523 data. Labeling was done manually to further ensure accuracy and reliability due to the need for interpretation of the varied Twitter language styles [21]. The data that has been labeled then goes through a pre-processing process which aims to clean the data to avoid empty

values, unnecessary characters, and irrelevant data [22]. The number of each keyword to be used can be seen in Table 1.

Table 1. The amount of data for each dataset keyword

s	Total
capek	5,137
gelisah	10,023
lelah	4,180
putus asa	4,774
sedih	1,305
sengsara	610
stress	494
Total	**26,523**

Although these keywords were selected based on indications of depression, it is possible that some other contexts or terms related to depression were not included. In addition, the use of data from two languages (Indonesian and English) may also add language bias that affects the model's understanding of the depression context. The process of translating English tweets into Indonesian can introduce inaccuracies in sentiment understanding, so it is important to be aware of any imbalances or biases in the data representation. To address this, the datasets used have been checked to ensure diversity in representation, however, data imbalance between depression and non-depression classes can still affect the model results, which needs to be taken into account during evaluation. To minimize the risk of subjective interpretation during the manual labeling process, a standardized guideline was used to categorize tweets as either depressive or non-depressive. This ensures that labeling decisions are consistent across the dataset. Additionally, a second reviewer was involved in verifying the labels for a random sample of the dataset to further minimize inconsistencies and biases. Any disagreements between the primary and secondary reviewers were resolved through discussion to ensure that the labels reflect a fair and accurate classification. These steps help to reduce the impact of subjective interpretation in the labeling process and ensure the reliability of the dataset.

In this study, the dataset used consists of 26,523 tweet data in the Indonesian language from crawling and an additional 24,000 data from GitHub containing tweets from English-language social media X which are then translated. Thus, the total data used in this study is 50,523 data. This research uses binary classification, with Depression (1) and Non-Depression (0) classes. The percentage distribution of data in this study can be seen in Table 2.

Table 2. Percentage of data distribution

Label	Class	Total	Percentage
Depression	1	25,281	50.08%
Non- Depression	0	25,242	49.92 %
Total		50,523	100%

2.2. Pre-Process Data

In this study, data pre-processing is conducted to enhance the classification model's performance by minimizing noise in the collected data. [23]. The pre-processing process includes several important steps that are integrated into one seamless workflow. First, data cleaning is performed to clean the text from unnecessary elements such as numbers, symbols, emojis, and excess spaces, as well as to ensure there are no unnecessarily repeated letters. Then, case folding is applied to convert all letters to lowercase, creating consistency and avoiding unnecessary duplication in the data [24]. Next, data normalization ensures that non-standard words are converted into standard form to avoid word recognition errors by the model [25]. Tokenizing separates the text into tokens based on spaces, and stopword removal involves the elimination of frequently occurring words that contribute minimal semantic value to the classification process [26]. Finally, stemming is implemented to reduce words to their base form, helping the model recognize and process word variations more efficiently [27]. This process as a whole aims to provide a cleaner and more structured dataset for effective classification model training. The pre-process stages in this study can be seen in Table 3, using one example of a tweet resulting from crawling.

During the data pre-processing stage, the main challenge is dealing with noisy and ambiguous data. Social media data often contains irrelevant information, such as hashtags that have nothing to do with the topic, unnecessary characters, or spelling mistakes. To reduce these influences, the pre-processing stage involves cleaning the data by removing unnecessary characters and correcting any spelling errors that may be present.

In addition, tweets that are ambiguous or difficult to categorize as depressed or not, such as general complaints without clear indications of depression, are dealt with by the use of manual labels to ensure accuracy. Handling these ambiguities is important for the model to classify the data more precisely, given the complexity of the language used on social media.

Table 3. Example of Data Pre-Process Stages

Pre-processing Stage	Output
Raw Data	Aku udah lelah disini aku muak dengan semua yang terjadi... Jemput aku pulang ya Allah. Rumahku disini udah nggak berbentuk lagi udh hancur banget... Ayuk aku lelah disini ya Allah. Pengen pulang ya Allah.
Data cleaning	Aku udah lelah disini aku muak dengan semua yang terjadi Jemput aku pulang ya Allah Rumahku disini udah nggak berbentuk lagi udh hancur banget Ayuk aku lelah disini ya Allah Pengen pulang ya Allah
Case Folding	aku udah lelah disini aku muak dengan semua yang terjadi jemput aku pulang ya allah rumahku disini udah nggak berbentuk lagi udh hancur banget ayuk aku lelah disini ya allah pengen pulang ya allah
Data normalization	aku sudah lelah disini aku muak dengan semua yang terjadi jemput aku pulang ya allah rumahku disini sudah enggak berbentuk lagi sudah hancur banget ayo aku lelah disini ya allah pengin pulang ya allah
Tokenization	['aku', 'sudah', 'lelah', 'disini', 'aku', 'muak', 'dengan', 'semua', 'yang', 'terjadi', 'jemput', 'aku', 'pulang', 'ya', 'allah', 'rumahku', 'disini', 'sudah', 'enggak', 'berbentuk', 'lagi', 'sudah', 'hancur', 'banget', 'ayo', 'aku', 'lelah', 'disini', 'ya', 'allah', 'pengin', 'pulang', 'ya', 'allah']
Stopword Removal	['aku', 'lelah', 'aku', 'muak', 'jemput', 'aku', 'pulang', 'allah', 'rumah', 'bentuk', 'hancur', 'banget', 'ayo', 'aku', 'lelah', 'allah', 'pengin', 'pulang', 'allah']
Stemming	['aku', 'lelah', 'aku', 'muak', 'jemput', 'aku', 'pulang', 'allah', 'rumah', 'bentuk', 'hancur', 'banget', 'ayo', 'aku', 'lelah', 'allah', 'pengin', 'pulang', 'allah']

2.3. Feature Extraction

The next process after pre-processing the data is feature extraction. This research uses TF IDF feature extraction. TF-IDF has the advantage of recognizing the most significant words in a document because it gives higher values to words that appear frequently in certain documents, but rarely appear in other documents [28]. The text will be represented in vector form and each word which is the smallest unit of the vector will be given a weight. The weighting process will be done based on predetermined words and using certain methods. One method that can be used for word weighting is Term Frequency- Inverse Document Frequency (TF-IDF). TF-IDF is a weighting method that will calculate the importance value of a word [11]. TF-IDF is a combination of the Term Frequency (TF) concept which serves to calculate the frequency of occurrence of words and the Inverse Document Frequency (IDF) concept which serves to calculate the number of documents that have that word [29].

In this research, tweet data will go through TF-IDF analysis to give weight to each word. The weight of a word is determined by its frequency in the document and its rarity in other documents, with higher weights attributed to words that show high frequency in the document but rarely appear in other documents [30]. The formula of TF-IDF is described in formula (1)

$$TFIDF = tfi \times idf(i) \tag{1}$$

TF is a function of assigning weight values to words in a text document. n_i is the total number of words i in the document, and N denotes the total number of words in the text corpus. The weight value for word i is calculated based on formula (2)

$$tf_i = \log(\frac{n_i}{N}) \tag{2}$$

IDF measures how unique a word is by comparing the total number of documents (|D|) to the number of documents containing that word D_i. Words that appear infrequently have a high IDF value, while words that appear frequently have a low IDF value [31].

$$idf(i) = \log(\frac{|D|}{D_i}) \tag{3}$$

2.4. Corpus Building and Feature Expansion

In this study, the corpus-building process was carried out using FastText. The corpus was built using tweet data, indonews, and a combination of tweet+indonews. The total data from the corpus used can be seen in Table 4.

Table 4. Number of Corpus Data

Corpus	Total
Tweets	50,523
Indonews	100,594
Tweets + Indonews	151,117

In this study, 3 corpus similarity ranks will be used, namely Top 1, Top 5, and Top 10. Table 4 is an illustration of the similarity results of the word "sad" using the tweet corpus with the Top 10 ranking. The ranking system is based on the level of word similarity from highest to lowest, as shown in Table 5.

Table 5. Top 10 Corpus similarity tweets of the word "sedih"

Kata	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5
sedih	pedih	hampa	sedu	sedikit	Rasai
	Rank-6	**Rank-7**	**Rank-8**	**Rank-9**	**Rank-10**
	sengir	hampir	rasa	frasa	marah

When the similarity corpus has been formed, the next step is to perform the feature expansion process. This research uses FastText feature expansion to recognize words that have not been detected. Feature expansion aims to overcome the problem of vocabulary mismatch, words that in the vector have a value of 0 will be replaced with similar words based on the similarity ranking in the corpus. Thus, words that previously had a value of 0 and become worth 1 and more representative [32], [33].

2.5. Data Splitting

In this research, data splitting is used to split the dataset into two partitions, specifically the dataset is divided into training data and testing data. The training data is utilized to train the model, while the testing data is employed to evaluate its predictive performance. The proportion of training and testing data is arranged in three ratios, namely 90:10 (with 90% for training data and 10% for testing), 80:20 (with 80% being training data and 20% testing data), and 70:30 (with 70% training data and 30% testing data) [34].

2.6. *Convolutional Neural Network* (CNN)

Convolutional Neural Network (CNN) is a form of feed-forward neural network. In the CNN model, the output of one layer will be used as input to the next layer [17]. The layer consists of an input layer, several hidden layers, and an output layer [18]. In this CNN model research consists of several layers, starting from the Conv1D layer with 32 filters and a kernel of size 3, as well as the ReLU activation function. MaxPooling1D layer to reduce feature dimensions without losing important information [35], [36]. The flattened layer is used to flatten the features to match the input of the fully connected layer. Next, a Dense layer with 32 units and a ReLU activation function is used for the classification process, ending with a Dense output layer that uses a sigmoid activation function to produce a binary output, as shown in Fig. 2.

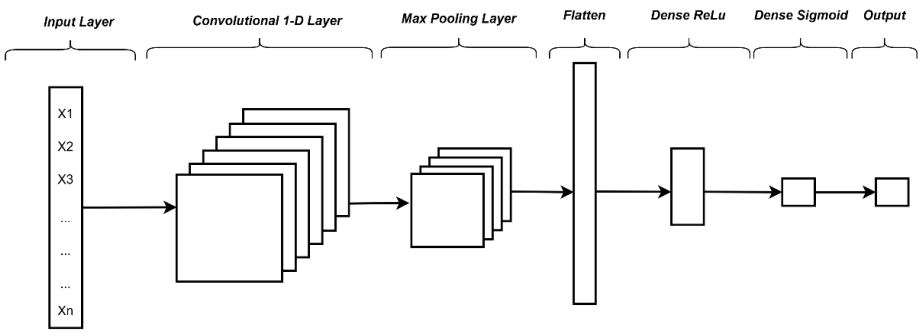

Fig. 2. Design of depression detection system from social media X

2.7. *Bidirectional Gated Reccurent Unit* (BiGRU)

The Bidirectional Gated Reccurent Unit (BiGRU) is a development of the Gated Reccurent Unit (GRU) model. BiGRU has two layers with opposite directions, this model has excellent performance in recognizing patterns in sentences. The model consists of two layers, a forward layer that sequentially handles word processing, starting from the first word and moving to the last, and a backward layer that processes words in the opposite direction, beginning with the last word and moving to the first [19]. In this study, the BiGRU model architecture is used which consists of several main layers, namely Bidirectional GRU, GlobalMaxPooling1D layer to reduce data dimensions by taking the maximum value of each feature, and two Dense layers as fully connected layers for the classification process [37]. The hidden layer employs the ReLU activation function, whereas the output layer utilizes the sigmoid activation function to produce binary predictions [38], [39], as shown in Fig. 3.

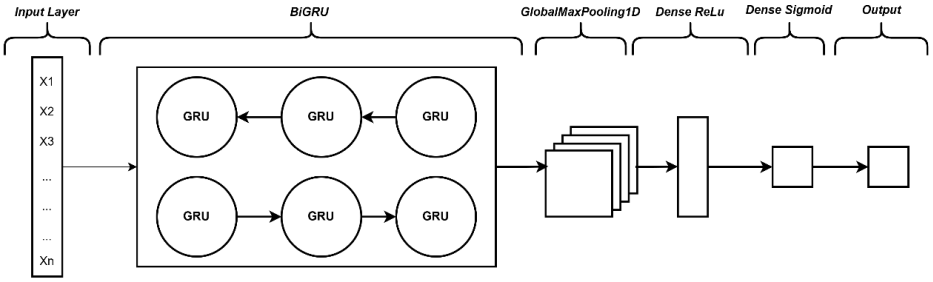

Fig. 3. Proposed BiGRU Architecture

2.8. Hybrid CNN-BiGRU

This research uses a combined approach between Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (BiGRU) to form a hybrid deep learning model CNN-BiGRU and BiGRU-CNN. By leveraging the advantages of each model, this research aims to optimize the processing and understanding capabilities of text content [20]. Through the integration of CNN, the model can extract spatial features from the text, while BiGRU allows the model to effectively account for sequential context. The combination of these two approaches is expected to produce a more powerful model in text data modeling and analysis [40]. The architecture of the CNN-BiGRU model is described in Fig. 4 and the architecture of the BiGRU-CNN model in Fig. 5.

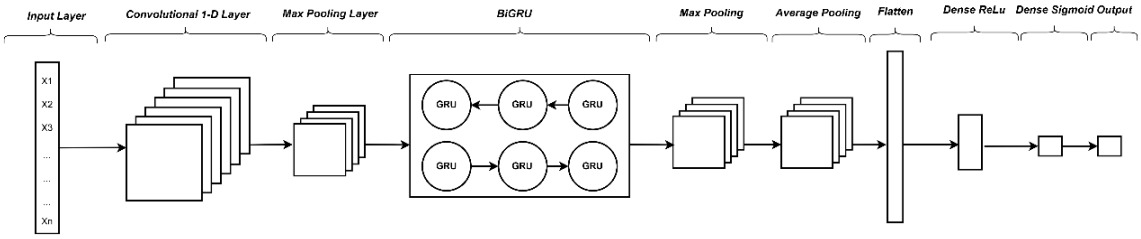

Fig. 4. Proposed CNN-BiGRU architecture

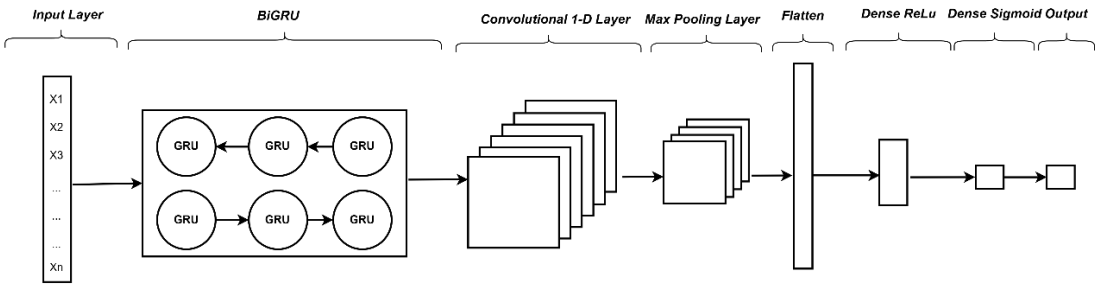

Fig. 5. Proposed BiGRU-CNN Architecture

2.9. Attention Mechanism

The attention mechanism is a mechanism that revolutionizes deep learning. It has the concept of processing words in parallel rather than sequentially. This mechanism emphasizes the crucial aspects of the input and captures the underlying correlations [41]. Research [42] found that the attention mechanism can improve the performance of the hybrid deep learning model marked by an increase in the F1-Score value. In this research, the attention mechanism will be implemented to obtain higher classification accuracy by combining the attention mechanism in CNN-BiGRU hybrid deep learning. In this research, eight hybrid approaches have been tested by integrating two forms of variations of the CNN and BiGRU models as follows:

1. CNN- ATT
2. CNN-BiGRU-ATT
3. CNN-ATT-BiGRU
4. CNN- ATT -BiGRU-ATT
5. BiGRU- ATT
6. BiGRU-CNN- ATT
7. BiGRU-ATT -CNN
8. BiGRU-ATT-CNN-ATT

2.10. Evaluation (*Confussion Matrix*)

In developing a classification system, system performance plays a crucial role. A well-performing system enhances classification accuracy, ensuring reliable and precise outcomes [43]. Confusion matrix is one of the well-known methods in measuring the performance of a classification system. Confusion matrix can be used for binary classification systems or multi-class classification [44]. The confusion matrix will represent the sum of the predicted and actual values of the results of running a classification system. The predicted and actual values will be represented in 4 combinations of metrics [45]. The four combinations of metrics can be seen in Fig. 6 which includes True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

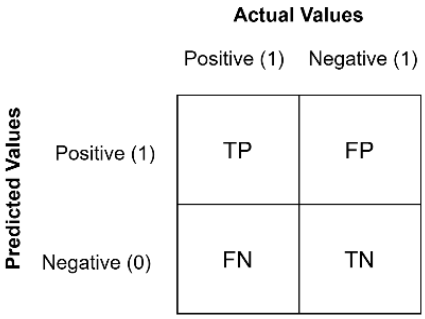

Fig. 6. Confusion matrix

The True Positive (TP) value shows the amount of positive data detected correctly, the True Negative (TN) value shows the amount of negative data detected correctly, False Positive (FP) shows the amount of negative data detected as positive and False Negative (FN) is the amount of positive data detected as negative. Based on 4 combinations of the metrics used, the values of Accuracy, Precision, Recall, and F1-Score can be generated. Accuracy is the ratio of the amount of data that is predicted correctly to the entire data [46]. Accuracy can be obtained using equation 3 as follows:

$$Accuracy = \frac{TP \ + TN}{TP + FP + FN + TN} \tag{4}$$

Precision is the amount of positive data that is correctly classified as positive data based on the total data classified as positive [47]. Precision can be obtained using equation 4 as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall is the percentage of positive data that has been classified correctly [48]. Recall can be obtained using equation 5 as follows:

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

F1-Score is the value obtained from the comparison of the average value of Recall and Precision [49]. F1-Score value can be obtained using equation 6 as follows:

$$F1\ Score = \frac{2 * precision * recall}{precision + recall} \quad (7)$$

3. RESULTS AND DISCUSSION

3.1. Testing Results

This research was conducted through multiple scenarios designed to evaluate and optimize classification models using CNN, BiGRU, and a hybrid approach to achieve the highest accuracy. The list of scenarios is outlined in Table 6.

Table 6. List of Test Scenarios

Scenario	Description	Objective
1	The data is split between 90:10, 80:20, and 70:30 ratios. The ratio with the best accuracy will be used in the next scenario.	Baseline Determination
2	Apply unigram, bigram, trigram, unigram+bigram, and unigram+bigram+trigram n-gram configurations	Obtain the best n-gram configuration
3	Comparing the maximum number of features between 5000, 10,000, and 15,000	Obtain the best maximum features
4	Application of feature expansion with a corpus that has been generated using FastText with Top 1, Top 5, and Top 10 rankings	Obtain the best corpus and ranking for each model
5	Testing the model by applying the attention mechanism to the model results with the best accuracy in the previous scenario	Obtain the attention model with the best accuracy.

The first scenario is baseline determination by applying data splitting using 90:10, 80:20, and 70:30 ratios. The model uses TF-IDF as feature extraction with unigram n-grams and 10,000 features. In this scenario, an early stopping mechanism is implemented to automatically terminate the training process when no significant improvement in performance is observed, preventing overfitting and optimizing computational efficiency. The use of early stopping aims to avoid overfitting and save the necessary resources. Table 7 shows that the data split ratio that produces the highest accuracy is the 90:10 ratio for all models, with the highest accuracy value produced by the BiGRU and BiGRU-CNN models at 82.90%. A data split ratio of 90:10 will be used in the next scenario.

Table 7. Results from testing scenario 1

Split Ratio	Accuracy (%)			
	CNN	***BiGRU***	***CNN-BiGRU***	***BiGRU-CNN***
90:10	**82.73**	**82.90**	**82.80**	**82.90**
80:20	82.54	82.72	82.76	82.75
70:30	81.96	82.48	82.54	82.51

The second scenario is the application of different n-gram configurations when performing feature extraction using TF-IDF. In this scenario, a data split ratio of 90:10 and 10,000 features will be used based on the best results from the previous scenario. Table 8 shows that the n-gram configuration can affect the accuracy of the model. In this scenario, the unigram+bigram n-gram configuration provides the most significant improvement among other n-gram configurations with all models having an increase in accuracy [50]. CNN experienced an increase of 0.37% to 83.10%, BiGRU experienced an increase of 0.44% to 83.34%, CNN-BiGRU experienced an increase of 0.50% to 83.30% and BiGRU-CNN experienced an increase of 0.41% to 83.31%. The highest improvement is achieved by the CNN-BiGRU model, with an increase of 0.50%, while the highest accuracy is attained by the BiGRU model, reaching 83.34%.

The third scenario is model testing by comparing the maximum features during the feature extraction process. The maximum features used are 5000, 10,000 (baseline), and 15,000. Table 9 shows that in the CNN model, the maximum feature of 5000 provides the most significant improvement, while the BiGRU, CNN-BiGRU, and BiGRU-CNN models experienced the greatest increase in accuracy with a maximum feature of 15,000 [51]. CNN experienced an increase of 0.07% to 83.17%, while BiGRU, CNN-BiGRU, and BiGRU-

CNN experienced a decrease of 0.26%, 0.22%, and 0.20%, respectively, with a final accuracy of 83.08%, 83.08%, and 83.11%. At the maximum feature value of 15,000, CNN decreased by 0.26% to 82.84%, while BiGRU increased by 0.24% to 83.58%, CNN-BiGRU increased by 0.33% to 83.63%, and BiGRU-CNN increased by 0.15% to 83.46%. The highest improvement was achieved by CNN-BiGRU by 0.33%, while the highest overall accuracy was achieved by BiGRU with a value of 83.58%.

Table 8. Results from testing scenario 2

N-gram	Accuracy (%)			
	CNN	***BiGRU***	***CNN-BiGRU***	***BiGRU-CNN***
Unigram (baseline)	82.73	82.90	82.80	82.90
Bigram	78.15 (-4.58)	79.14 (-3.76)	79.17 (-3.63)	79.30 (-3.60)
Trigram	74.19 (-8.54)	73.92 (-8.98)	73.99 (-8.81)	73.90 (-9.00)
Unigram+Bigram	**83.10 (+0.37)**	**83.34 (+0.44)**	**83.30 (+0.50)**	**83.31 (+0.41)**
Allgram	82.88 (+0.15)	83.08 (+0.18)	83.06 (+0.26)	83.15 (+0.25)

Table 9. Results from testing scenario 3

Max-features	Accuracy (%)			
	CNN	***BiGRU***	***CNN-BiGRU***	***BiGRU-CNN***
10000 (baseline)	83.10	83.34	83.30	83.31
5000	**83.17 (+0.07)**	83.08 (-0.26)	83.08 (-0.22)	83.11 (-0.20)
15000	82.84 (-0.26)	**83.58 (+0.24)**	**83.63 (+0.33)**	**83.46 (+0.15)**

The fourth scenario is the application of the expansion feature using the corpus that has been built with FastText. In this scenario, each model will apply the expansion feature with the Top 1, Top 5, and Top 10-word similarity ratings on the Tweet, News, and Tweet+Indonews corpus. Table 10 shows that the corpus that provides the most significant accuracy improvement is the Tweet corpus, where CNN has the highest improvement in the Top 5. BiGRU, CNN-BiGRU, and BiGRU-CNN experienced the highest improvement in the Top 1 with the tweet corpus. In the Top 1 category, BiGRU achieved the highest accuracy with 84.34% for Tweet data (+0.76), 83.92% for Indonews data (+0.34), and 84.20% for combined Tweet-Indonews (+0.62). Meanwhile, CNN-BiGRU has the second-best accuracy in the Top 1 category with 84.06% for Tweet data (+0.43), followed by BiGRU-CNN with 84.19% for Tweet data (+0.73). In the Top 5 category, BiGRU-CNN achieved the highest accuracy with 83.84% for Tweet data (+0.38), 84.10% for Indonews (+0.64), and 84.14% for Tweet-Indonews (+0.68), followed by BiGRU with slightly lower accuracy. In the Top 10 category, BiGRU-CNN remained ahead with 82.80% for Tweet data (-0.66) and 83.85% for Indonews (+0.39), although accuracy values across all models tended to decrease. Overall, BiGRU and BiGRU-CNN showed consistent performance with high accuracy across various categories [52].

Table 10. Results from testing scenario 4

Model	Rank	Accuracy (%)		
		Tweet	***Indonews***	***Tweet Indonews***
CNN	Top 1	83.53 (+0.36)	83.46 (+0.29)	83.33 (+0.16)
	Top 5	**83.68 (+0.51)**	83.16 (-0.01)	83.38 (+0.21)
	Top 10	82.67 (-0.50)	82.24 (-0.93)	82.83 (-0.34)
BiGRU	Top 1	**84.34 (+0.76)**	83.92 (+0.34)	84.20 (+0.62)
	Top 5	83.74 (+0.16)	84.18 (+0.60)	84.28 (+0.70)
	Top 10	82.79 (-0.79)	84.02 (+0.44)	83.48 (-0.10)
CNN-BIGRU	Top 1	**84.06 (+0.43)**	83.68 (+0.05)	83.91 (+0.28)
	Top 5	83.50 (-0.13)	83.68 (+0.05)	83.63 (0.00)
	Top 10	82.79 (-0.84)	83.27 (-0.36)	83.53 (-0.10)
BiGRU-CNN	Top 1	**84.19 (+0.73)**	83.99 (+0.53)	83.97 (+0.51)
	Top 5	83.84 (+0.38)	84.10 (+0.64)	84.14 (+0.68)
	Top 10	82.80 (-0.66)	83.85 (+0.39)	83.54 (+0.08)

The fifth scenario assesses the performance of CNN and BiGRU-based models by incorporating an attention mechanism to improve accuracy. Table 11 shows the results obtained by applying the attention mechanism. CNN has increased accuracy by 0.45% to 84.13%, BiGRU has increased accuracy by 1.03% to 84.37. For the hybrid model, CNN-BiGRU with CNN-ATT-BiGRU structure increased 0.21% to 84.27%.

BiGRU-CNN experienced an increase of 0.21% and achieved the highest accuracy of the other models at 84.40%.

Table 11. Results from testing scenario 5

Model	Accuracy (%)
CNN-ATT	84.13 (+0.45)
CNN-BiGRU-ATT	84.26 (+0.2)
CNN-ATT- BiGRU	**84.27 (+0.21)**
CNN-ATT- BiGRU -ATT	84.05 (-0.01)
BiGRU-ATT	84.37 (+1.03)
BiGRU-CNN-ATT	84.31 (+0.12)
BiGRU-ATT-CNN	84.28 (+0.09)
BiGRU-ATT-CNN-ATT	**84.40 (+0.21)**

3.2. Analysis of Test Results

Based on Fig. 7, each scenario resulted in an increase in accuracy from the previous scenario, with the highest accuracy observed in Scenario V. CNN experienced an improvement of 1.4% from the baseline, and BiGRU experienced an improvement of 1.47%. The CNN-BiGRU hybrid model achieved an improvement of 1.47% compared to the baseline. In addition, the BiGRU-CNN hybrid model achieved an increase of 1.5% from the baseline and became the model with the highest increase in accuracy and the model with the highest accuracy, namely 84.40% accuracy. The accuracy improvement of each model from baseline to scenario 5 is shown in Table 12.

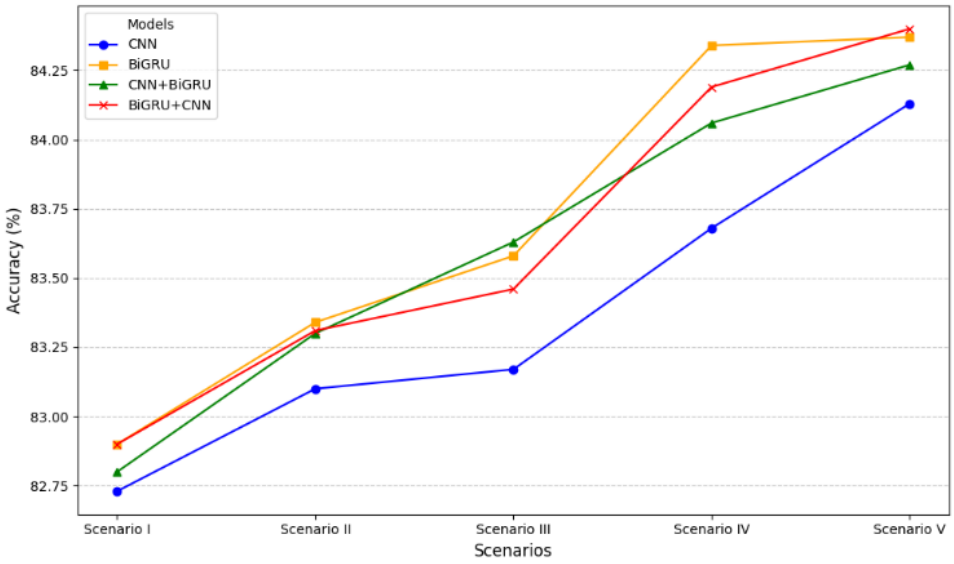

Fig. 7. Improved accuracy from scenario 1 to scenario 5

Table 12. Total improvement in model accuracy from baseline to scenario 5

Model	Accuracy (%)
CNN	84.13 (+1,4)
BiGRU	84.37 (+1.47)
CNN-BiGRU	84.27 (+1,47)
BiGRU-CNN	84.40 (+1,5)

Based on the results from all five tested scenarios, it is clear that the use of FastText and the attention mechanism significantly improves the model's accuracy. In Scenario 4, the application of FastText for feature expansion led to the highest accuracy for the BiGRU-CNN model, particularly in the Top 1, Top 5, and Top 10 categories (Table 10), demonstrating that FastText enhances the model's ability to handle out-of-vocabulary words. Additionally, in Scenario 5, implementing the attention mechanism improved accuracy by 0.21% to 0.40% (Table 11), helping the model focus on the most relevant words in the context of depression. Overall, these techniques resulted in the highest accuracy of 84.40% for BiGRU-CNN in Scenario 5, proving that

FastText and the attention mechanism play a crucial role in boosting performance in depression detection. The combination of these techniques resulted in the highest accuracy of 84.40% for BiGRU-CNN in Scenario 5, demonstrating the crucial role of both FastText and the attention mechanism in boosting performance for depression detection. This improvement is not just incremental, but it showcases the strength of these techniques in handling the inherent challenges of social media data, such as noisy, ambiguous text, and informal language usage. For instance, the ability of FastText to manage out-of-vocabulary words directly addresses one of the primary obstacles in NLP for social media texts, where vocabulary is continuously evolving [12], [29]. Moreover, the attention mechanism helps the model focus on emotionally significant words, which are often context-dependent and subtle, thus improving the model's capacity to identify depression-related content more effectively [13], [41].

Compared to previous studies, this research shows a clear advantage. Study [5] achieved 99.02% accuracy using CNN + LSTM with Word2Vec and TF-IDF but did not incorporate feature expansion. In contrast, our model, with FastText, outperformed previous studies in various ranking categories (Top 1, Top 5, Top 10) as shown in Table 10. Study [8] achieved 94.28% with CNN + BiLSTM, but our CNN-BiGRU model reached 83.58%-83.69%, improving to 84.40% after applying the attention mechanism. Similarly, Research [11] using CNN-BiGRU achieved 93.38%, but our BiGRU-CNN, utilizing FastText and the attention mechanism, achieved 84.40%. This study demonstrates that FastText, which handles missing words, and the attention mechanism, which focuses on relevant terms, are essential for improving depression detection in social media contexts, especially when considering the challenges of noisy and ambiguous data [26].

4. CONCLUSION

This research focuses on detecting depression on social media platform X using a hybrid deep learning model with an attention mechanism that combines CNN and BiGRU, utilizing TF-IDF for feature extraction and FastText for feature expansion. This research uses a text dataset sourced from Indonesian and translated tweets, consisting of 50,523 data, along with the English translation data. Furthermore, a similarity corpus with 151,117 data was built to improve the performance of the model. To achieve the best accuracy, five experimental scenarios were conducted to optimize the model performance. Scenario 1 achieved the best accuracy at a data ratio of 90:10, scenario 2 with unigram+bigram n-grams, scenario 3 with a maximum of 5000 features for CNN and 15,000 for BiGRU, CNN-BiGRU, BiGRU-CNN. Scenario 4 shows that feature expansion can improve the accuracy of the model, especially using the Top 1 Tweet Corpus. Scenario 5 with the attention mechanism was able to achieve the highest accuracy of 84.40% on the BiGRU model. While this study shows a significant improvement with the application of FastText and attention mechanisms, challenges remain, such as handling noisy and ambiguous text often found on social media. These techniques help the model focus on relevant words, but real-world challenges like informal language and context shifts still pose difficulties in accurate depression detection. Despite these challenges, the integration of attention mechanisms and feature expansion significantly boosted model accuracy. This research opens the possibility for the application of depression detection in social media, enabling early detection and faster intervention for individuals in need. Based on the research findings, the integration of the attention mechanism and the implementation of feature expansion contribute to enhancing the accuracy of text-based depression detection models. Future research should explore comparative analyses involving various feature expansion techniques and optimization algorithms to further improve model performance and robustness.

REFERENCES

[1] H. Tufail, S. M. Cheema, M. Ali, I. M. Pires, and N. M. Garcia, "Depression Detection with Convolutional Neural Networks: A Step Towards Improved Mental Health Care," *Procedia Comput Sci*, vol. 224, pp. 544–549, 2023, https://doi.org/10.1016/j.procs.2023.09.079.

[2] Vandana, N. Marriwala, and D. Chaudhary, "A hybrid model for depression detection using deep learning," *Measurement: Sensors*, vol. 25, p. 100587, 2023, https://doi.org/10.1016/j.measen.2022.100587.

[3] "Depressive disorder (depression)." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[4] "Survei: 17,9 Juta Remaja Indonesia Punya Masalah Mental, Ini Gangguan yang Diderita." [Online]. Available: https://www.detik.com/edu/detikpedia/d-7150554/survei-17-9-juta-remaja-indonesia-punya-masalah-mental-ini-gangguan-yang-diderita.

[5] M. Ahmad Wani, M. A. Elaffendi, K. A. Shakil, A. Shariq Imran, and A. A. Abd El-Latif, "Depression Screening in Humans With AI and Deep Learning Techniques," *IEEE Trans Comput Soc Syst*, vol. 10, no. 4, pp. 2074–2089, 2023, https://doi.org/10.1109/TCSS.2022.3200213.

[6] S. Bengtsson and S. Johansson, "The Meanings of Social Media Use in Everyday Life: Filling Empty Slots, Everyday Transformations, and Mood Management," *Social Media and Society*, vol. 8, no. 4, 2022, https://doi.org/10.1177/20563051221130292.

[7] Aschbrenner, J. A. Naslund, A. Bondre, J. Torous, and K. A., "Social Media and Mental Health: Benefits, Risks, and Opportunities for Researhc and Practice," *J Technol Behav Sci*, vol. 5, no. 3, pp. 245–257, 2020, [Online]. Available: https://doi.org/10.1007/s41347-020-00134-x.
[8] H. Kour and M. K. Gupta, *An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM*, vol. 81, no. 17, 2022. https://doi.org/10.1007/s11042-022-12648-y.
[9] T. Srajan Kumar, "A Deep Learning Framework with a Hybrid Model for Automatic Depression Detection in Social Media Posts," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 4, pp. 3217–3231, Jun. 2024, [Online]. Available: https://www.ijisae.org/index.php/IJISAE/article/view/6816.
[10] S. Almutairi, M. Abohashrh, H. Hayder, Zulqarnain, A. Namoun, and F. Khan, "A Hybrid Deep Learning Model for Predicting Depression Symptoms From Large-Scale Textual Dataset," *IEEE Access*, p. 1, Feb. 2024, https://doi.org/10.1109/ACCESS.2024.3496741.
[11] L. Bendebane, Z. Laboudi, A. Saighi, H. Al-Tarawneh, A. Ouannas, and G. Grassi, "A Multi-Class Deep Learning Approach for Early Detection of Depressive and Anxiety Disorders Using Twitter Data," *Algorithms*, vol. 16, no. 12, 2023, https://doi.org/10.3390/a16120543.
[12] A. Amalia, O. S. Sitompul, E. B. Nababan, and T. Mantoro, "An Efficient Text Classification Using fastText for Bahasa Indonesia Documents Classification," *International Conference on Data Science, Artificial Intelligence, and Business Analytics, DATABIA 2020 - Proceedings*, pp. 69–75, 2020, https://doi.org/10.1109/DATABIA50434.2020.9190447.
[13] J. Teng, W. Kong, Y. Chang, Q. Tian, C. Shi, and L. Li, "Text Classification Method Based on BiGRU-Attention and CNN Hybrid Model," *ACM International Conference Proceeding Series*, pp. 614–622, 2021, https://doi.org/10.1145/3488933.3488970.
[14] W. Yan, L. Zhou, Z. Qian, L. Xiao, H. Zhu "Sentiment Analysis of Student Texts Using the CNN-BiGRU-AT Model," *Scientific Programming*, p. 8405623, 2021, https://doi.org/10.1155/2021/8405623.
[15] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans Assoc Comput Linguist*, vol. 12, pp. 820–836, Jun. 2024, https://doi.org/10.1162/tacl_a_00675.
[16] Y. Zhang, Y. Zhou, and J. Yao, "Feature extraction with TF-IDF and game-theoretic shadowed sets," in *Communications in Computer and Information Science*, *in Communications in computer and information science, Cham: Springer International Publishing,* pp. 722–733, 2020, https://doi.org/10.1007/978-3-030-50146-4_53.
[17] H. A. Taba and H. Suparwito, "Convolutional neural networks for text classification: A study on public activity restriction," *AIP Conf Proc*, vol. 3077, no. 1, p. 40016, Feb. 2024, https://doi.org/10.1063/5.0201145.
[18] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM Model for Document-Level Sentiment Analysis," *Mach Learn Knowl Extr*, vol. 1, no. 3, pp. 832–847, 2019, https://doi.org/10.3390/make1030048.
[19] L. Zhou and X. Bian, "Improved text sentiment classification method based on BiGRU-Attention," *J Phys Conf Ser*, vol. 1345, no. 3, 2019, https://doi.org/10.1088/1742-6596/1345/3/032097.
[20] Q. Wang, J. Tian, M. Li, and M. Lu, "Text Classification Based on CNN-BiGRU and Its Application in Telephone Comments Recognition," *Int J Comput Intell Appl*, vol. 22, Feb. 2023, https://doi.org/10.1142/S1469026823500219.
[21] C. and S. R. J. Jamali Ali Akbar and Berger, "Momentary Depressive Feeling Detection Using X (Formerly Twitter) Data: Contextual Language Approach," *JMIR AI*, vol. 2, p. e49531, Nov. 2023, https://doi.org/10.2196/49531.
[22] P. Sankar, N. Palanichamy, and K.-W. Ng, "Sentiment Analysis on Twitter Data for Depression Detection," *Journal of Logistics, Informatics and Service Science*, vol. 11, no. 3, pp. 21–36, 2024, https://doi.org/10.33168/JLISS.2024.0302.
[23] M. A. Palomino and F. Aider, "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," *Applied Sciences*, vol. 12, no. 17, 2022, https://doi.org/10.3390/app12178765.
[24] M. D. Samad, N. D. Khounviengxay, and M. A. Witherow, "Effect of Text Processing Steps on Twitter Sentiment Classification using Word Embedding," *arXiv preprint arXiv:2007.13027*, 2020. [Online]. Available: https://arxiv.org/abs/2007.13027.
[25] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, 2023, https://doi.org/10.1017/S1351324922000213.
[26] K. M. G. S. Karunarathna and R. Rupasingha, "Learning to Use Normalization Techniques for Preprocessing and Classification of Text Documents," *International Journal of Multidisciplinary Studies*, vol. 9, no. 2, pp. 67–82, 2022, [Online]. Available: https://journals.sjp.ac.lk/index.php/ijms/article/view/6429.
[27] S. Pradha, M. N. Halgamuge, and N. Tran Quoc Vinh, "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data," in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 2019, pp. 1–8. https://doi.org/10.1109/KSE.2019.8919368.
[28] K. Harmandini and K. L, "Analysis of TF-IDF and TF-RF Feature Extraction on Product Review Sentiment," *Sinkron*, vol. 8, pp. 929–937, Feb. 2024, https://doi.org/10.33395/sinkron.v8i2.13376.
[29] S. S. Shaker, D. Alhajim, A. A. T. Al-Khazaali, H. A. Hussein, and A. F. Athab, "Feature Extraction based Text Classification: A review," *J Algebr Stat*, vol. 13, no. 1, pp. 646–653, 2022, [Online]. Available: https://www.researchgate.net/publication/361226607_Feature_Extraction_based_Text_Classification_A_review.
[30] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, p. 30, 2019, https://doi.org/10.1186/s13673-019-0192-7.

[31] S. Robertson, "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF," *Journal of Documentation - J DOC*, vol. 60, pp. 503–520, Feb. 2004, https://doi.org/10.1108/00220410410560582.
[32] M. Umer *et al.*, "Impact of convolutional neural network and FastText embedding on text classification," *Multimed Tools Appl*, vol. 82, no. 4, pp. 5569–5585, 2023, https://doi.org/10.1007/s11042-022-13459-x.
[33] S. Khomsah, R. Ramadhani, and S. Wijaya, "The Accuracy Comparison Between Word2Vec and FastText On Sentiment Analysis of Hotel Reviews," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, pp. 352–358, Feb. 2022, https://doi.org/10.29207/resti.v6i3.3711.
[34] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: An ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, 2022, https://doi.org/https://doi.org/10.1002/sam.11583.
[35] A. Jarrahi, R. Mousa, and L. Safari, "SLCNN: Sentence-Level Convolutional Neural Network for Text Classification," *arXiv preprint arXiv:2301.11696*, 2023. [Online]. Available: https://arxiv.org/abs/2301.11696.
[36] Y. Zhu, "Research on News Text Classification Based on Deep Learning Convolutional Neural Network," *Wirel Commun Mob Comput*, vol. 2021, no. 1, p. 1508150, 2021, https://doi.org/https://doi.org/10.1155/2021/1508150.
[37] S. Liu, W. Lin, Y. Wang, D. Z. Yu, Y. Peng, and X. Ma, "Convolutional Neural Network-Based Bidirectional Gated Recurrent Unit–Additive Attention Mechanism Hybrid Deep Neural Networks for Short-Term Traffic Flow Prediction," *Sustainability*, vol. 16, no. 5, 2024, https://doi.org/10.3390/su16051986.
[38] A. Traoré and M. A. Akhloufi, "2D Bidirectional Gated Recurrent Unit Convolutional Neural Networks for End-to-End Violence Detection in Videos," in *Image Analysis and Recognition*, *Springer International Publishing*, pp. 152–160, 2020, https://doi.org/10.1007/978-3-030-50347-5_14.
[39] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark," *Neurocomputing*, vol. 503, pp. 92-108, 2022, [Online]. Available: https://arxiv.org/abs/2109.14545.
[40] Z. Gao, Z. Li, J. Luo, and X. Li, "Short Text Aspect-Based Sentiment Analysis Based on CNN + BiGRU," *Applied Sciences*, vol. 12, no. 5, 2022, https://doi.org/10.3390/app12052707.
[41] 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, 2023, https://doi.org/10.1109/TKDE.2021.3126456.
[42] G. Alfattni, N. Peek, and G. Nenadic, "Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries," *J Biomed Inform*, vol. 123, p. 103915, 2021, https://doi.org/10.1016/j.jbi.2021.103915.
[43] S. Visa, B. Ramsay, A. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection," *CEUR Workshop Proc*, vol. 710, pp. 120–127, 2011, https://openworks.wooster.edu/facpub/88/ .
[44] S. Sathyanarayanan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023–4031, Nov. 2024, https://doi.org/10.53555/AJBR.v27i4S.4345.
[45] K. Riehl, M. Neunteufel, and M. Hemberg, "Hierarchical confusion matrix for classification performance evaluation," *J R Stat Soc Ser C Appl Stat*, vol. 72, no. 5, pp. 1394–1412, Feb. 2023, https://doi.org/10.1093/jrsssc/qlad057.
[46] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, p. 6086, 2024, https://doi.org/10.1038/s41598-024-56706-x.
[47] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans Assoc Comput Linguist*, vol. 12, pp. 820–836, Jun. 2024, https://doi.org/10.1162/tacl_a_00675.
[48] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Sci. Rep.*, vol. 12, no. 1, p. 5979, Apr. 2022, https://doi.org/10.1038/s41598-022-09954-8.
[49] C. Cao, D. Chicco, and M. M. Hoffman, "The MCC-F1 curve: a performance evaluation technique for binary classification," 2020. [Online]. Available: https://arxiv.org/abs/2006.11278.
[50] Y. Setiawan, N. U. Maulidevi, and K. Surendro, "The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification," *Data Sci J*, May 2024, https://doi.org/10.5334/dsj-2024-031.
[51] A. Maiti, A. Abarda, and M. Hanini, "The impact of feature extraction techniques on the performance of text data classification models," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, p. 1041, Feb. 2024, https://doi.org/10.11591/ijeecs.v35.i2.pp1041-1052.
[52] S. Lamar, R. Fitting, and A. Jayasumana, "Cognitive Communications System for Ultra-Low Size, Weight and Power (SWAP) Attributable Platforms," *IEEE Access*, vol. 10, pp. 41381–41387, 2022, https://doi.org/10.1109/ACCESS.2022.3167039.

BIOGRAPHY OF AUTHORS

I Wayan Abi Widiarta, is currently undertaking a bachelor's degree in Computer Science at Telkom University, Indonesia.. Email, abiwidiarta@student.telkomuniversity.ac.id

Erwin Budi Setiawan, is an Associate Professor at the School of Computing, Telkom University, Bandung, Indonesia, with over a decade of experience in research and teaching in the field of Informatics. His academic interests encompass machine learning, people analytics, and social media analysis. Email, erwinbudisetiawan@telkomuniversity.ac.id