Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

Predicting Mental Health Status using a Fine-Tuned CNN-LSTM Hybrid Model

Agustin*[1], Junadhi[2], Susi Erlinda[3], Triyani Arita Fitri[4], Lusiana Efrizoni[5]

[1,2,3,4,5]Department of Informatic Engineering, Faculty of Engineering and Informatics Universitas Sains dan Teknologi Indonesia, Pekanbaru, Indonesia

Email: [1]agustin@usti.ac.id

Received: Apr 21, 2026; Revised: May 4, 2026; Accepted: May 5, 2026; Published: Jun 15, 2026

Abstract

Mental health has become a critical global concern in the digital era, particularly as social media platforms increasingly serve as spaces where users express psychological conditions, emotions, and personal struggles. This study aims to predict mental health status from Twitter text using a fine-tuned hybrid CNN–LSTM deep learning model. A total of 12,214 tweets were collected, cleaned, and labeled into five categories: Normal, Stress, Anxiety, Depression, and High-Risk Condition. The dataset was split using stratified sampling into 70% training, 15% validation, and 15% testing portions. Text was transformed into numerical representations through tokenization, padding, and 100-dimensional word embeddings. The hybrid CNN–LSTM architecture combines the CNN's ability to extract local linguistic features with the LSTM's strength in capturing long-term contextual dependencies, supported by dropout, early stopping, and hyperparameter fine-tuning. Experimental results show that the hybrid model achieves superior performance compared to standalone CNN and LSTM architectures, obtaining an overall accuracy of 0.892, macro precision of 0.874, macro recall of 0.861, and a macro F1-score of 0.865. Class-wise evaluation indicates that the Normal category achieves the highest accuracy (0.960), followed by Anxiety (0.884) and High-Risk Condition (0.808). Meanwhile, Stress (0.751) and Depression (0.745) show lower accuracies due to semantic overlap in linguistic expressions commonly found on social media. The training process demonstrates stable convergence without significant overfitting, confirming the effectiveness of the selected architecture and training strategy. Overall, this study highlights the effectiveness of the hybrid CNN–LSTM model for early mental health detection based on text data. The findings provide a strong foundation for developing scalable and data-driven mental health monitoring systems in digital environments and contribute to advancing natural language processing approaches for mental health analysis.

Keywords: Mental Health Classification, CNN–LSTM, Deep Learning, Twitter Data, Text Mining

1. INTRODUCTION

Mental health has emerged as an increasingly urgent global issue over the past decade. Recent reports from the World Health Organization (2022) indicate that more than one billion individuals worldwide experience mental health disorders, with the highest prevalence occurring among younger populations. This growing psychological burden is exacerbated by social, economic, and digital factors, particularly intensive interactions on social media, which often reflect users' emotional patterns and mental states [1]. Recent studies further confirm that linguistic expressions on social media contain psychological signals that can be leveraged for early detection of mental health risks [2].

With the advancement of artificial intelligence technologies, Natural Language Processing (NLP) has become an effective approach for identifying mental health conditions through textual data. Deep learning methods have demonstrated superior capability in capturing semantic features and emotional context compared to classical machine learning techniques such as Support Vector Machines or Logistic

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

Regression [3]. Among deep learning architectures, Convolutional Neural Networks (CNNs) excel in extracting local features and emotional phrase patterns, while Long Short-Term Memory (LSTM) networks effectively model long-term dependencies and emotional dynamics within word sequences [4].

Although CNN and LSTM models perform well individually, recent research highlights the advantages of hybrid CNN–LSTM architectures, which integrate CNN's robust local feature extraction with LSTM's comprehensive temporal modeling. Contemporary studies show that such hybrid models achieve higher accuracy on mental health classification tasks, particularly for detecting depression, anxiety, stress, emotional distress, and other high-risk conditions [5], [6], [7]. This combined capability is particularly relevant for complex natural languages, where emotional meaning depends not only on individual words but also on sentence structure and temporal context.

Furthermore, emerging research trends emphasize the importance of fine-tuning hybrid architectures to enhance generalization and reduce predictive bias. Hyperparameter optimization including the number of CNN filters, kernel size, number of LSTM units, dropout rates, and learning rates has been shown to significantly improve model accuracy in multi-class classification settings [8], [9]. In this research context, the system is designed to classify mental health conditions into five categories (e.g., normal, stress, anxiety, depression, and high-risk conditions), thereby providing a more comprehensive representation of users' emotional states [10], [11].

In Indonesia, mental health issues among adolescents and university students have become increasingly concerning [12], [13]. Recent national studies report a significant rise in negative expressions, self-blame, and emotional distress across digital platforms, underscoring the need for AI-based early detection systems to support timely intervention [14]. However, despite the growing adoption of deep learning approaches for mental health classification, existing studies still face significant challenges in accurately distinguishing closely related emotional categories. In particular, expressions of stress, anxiety, and depression often exhibit substantial semantic overlap in social media text, leading to reduced classification performance in multi-class settings [15], [16]. Moreover, many prior studies primarily focus on binary classification or limited class scenarios and do not sufficiently address the complexity of five-class mental health categorization in short-text environments.

To address these limitations, this study proposes a fine-tuned hybrid CNN–LSTM model designed to capture both local linguistic patterns and long-range contextual dependencies in social media text. The proposed approach aims to improve classification robustness across overlapping emotional categories by optimizing model architecture and training strategies. Specifically, this study contributes by: (1) developing a five-class mental health classification framework based on Twitter data, (2) enhancing contextual representation through a hybrid CNN–LSTM architecture, and (3) improving classification performance through systematic hyperparameter fine-tuning. These contributions are expected to strengthen the role of natural language processing in scalable and data-driven mental health monitoring systems.

2. METHOD

The research methodology outlines a comprehensive set of stages employed in developing the CNN–LSTM model for five-class mental health classification using social media text data. The workflow consists of seven main phases: (1) data collection, (2) text cleaning and preprocessing, (3) dataset partitioning, (4) text representation and embedding, (5) CNN–LSTM architectural design, (6) fine-tuning, and (7) model performance evaluation. The complete stages of the proposes research methodology are illustrated in Figure 1.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

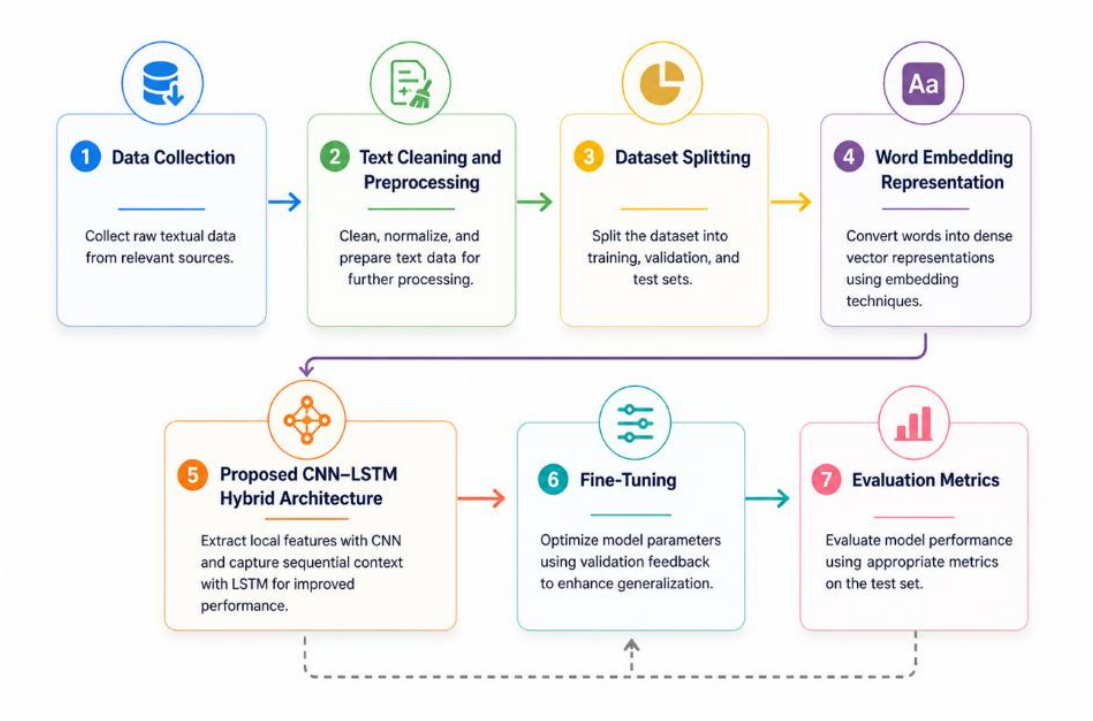

Figure 1. Research Framework

2.1. Data Collection

The data collection phase was conducted by gathering textual posts from the Twitter platform using the official Twitter API and compliant web-scraping techniques aligned with public data usage policies. The data were collected over a specific period, from January to March 2024, to ensure temporal consistency and relevance of the collected content. A set of predefined keywords related to mental health conditions such as “stress”, “anxiety”, “depression”, “hopeless”, and “mental exhaustion” was used to retrieve relevant tweets. Only English-language tweets were included in this study to maintain linguistic consistency and reduce noise during text processing [17], [18]. A total of 12,687 tweets were initially collected. These tweets were selected based on the premise that Twitter serves as a rich medium for spontaneous emotional expression and contains meaningful indicators of users’ psychological conditions. Prior to further processing, all data were anonymized by removing user identifiers and sensitive information to ensure compliance with ethical considerations in handling mental health-related data [19]. The collected tweets were subsequently annotated into five mental health categories: Normal, Stress, Anxiety, Depression, and High-Risk Condition, following a multi-class classification framework commonly adopted in social media-based mental health research. The annotation process was conducted manually by multiple annotators using predefined labeling guidelines to ensure consistency. Inter-annotator agreement was measured using Cohen’s Kappa, achieving a score above 0.80, which indicates strong agreement. Irrelevant entries, duplicate tweets, spam content, and tweets lacking meaningful emotional signals were removed to ensure dataset quality prior to the preprocessing stage. After this filtering process, 12,214 tweets were retained as the final dataset for further analysis. Previous studies have demonstrated that Twitter is among the most effective data sources for mental health text mining due to its real-time, open, and emotionally dense communication characteristics.

2.2. Text Cleaning and Preprocessing

The text preprocessing stage aims to prepare the data in accordance with the requirements of deep learning models. The process begins with case folding to normalize all tokens into lowercase, followed by noise removal to eliminate URLs, mentions, hashtags, excessive emojis, and other irrelevant

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

characters that may interfere with downstream analysis. Tokenization is then performed to segment the text into word-level units using modern tokenizers such as spaCy. Subsequently, stop-word removal is applied to eliminate common words that do not contribute meaningful emotional information. Lemmatization or stemming is carried out to reduce words to their base forms, thereby producing more stable and concise semantic representations. These procedures align with modern NLP practices that have been shown to improve model performance in text-based mental health classification tasks [20], [21].

2.3. Dataset Splitting

The cleaned dataset was subsequently divided into three subsets using a stratified split technique: 70% for training, 15% for validation, and 15% for testing. Stratification was essential to maintain consistent distribution across the five mental health categories, ensuring that the model did not become biased toward particular classes. Recent studies highlight that stratified splitting enhances predictive stability, particularly in multi-class classification scenarios. The validation subset was kept separate to ensure that the model did not overfit the training data [22].

2.4. Word Embedding Representation

Text representation into numerical form was performed using pre-trained GloVe embeddings with 100 dimensions. Each word token was mapped into a dense vector space that captures semantic meaning and contextual relationships between words. The embedding layer was initialized with pre-trained GloVe weights and further fine-tuned during the training process to adapt to domain-specific linguistic patterns found in social media text. Word embeddings play a crucial role in transforming discrete textual data into continuous vector representations, enabling deep learning models to effectively capture emotional nuances and semantic relationships. Previous studies have demonstrated that GloVe embeddings are effective in representing linguistic features relevant to mental health analysis, particularly in short-text environments such as social media [23], [24]. By combining pre-trained semantic knowledge with task-specific fine-tuning, the embedding representation enhances the model's ability to distinguish subtle emotional expressions across different mental health categories, including those with overlapping linguistic characteristics such as stress and depression.

2.5. CNN–LSTM Hybrid Architecture

The CNN–LSTM model was designed to leverage the strengths of both deep learning architectures. The CNN component extracts local emotional features such as phrases indicative of depression, anxiety, or emotional distress. Kernel sizes of 3–5 have been shown to be effective in capturing short linguistic patterns commonly found in social media text [25], [26]. The convolution operation in the Conv1D layer is defined as follows:

$$h_i = f\left(\sum_{j=0}^{k-1} w_j \; x_{i+j} + b\right) \quad (1)$$

where x represents the input word embeddings, w denotes the convolutional kernel, k is the kernel size, b is the bias term, and $f(cdot)$ is a non-linear activation function such as ReLU. This operation enables the model to capture local n-gram features that are critical for identifying emotional cues embedded within short text segments. The feature maps generated by the CNN layer are subsequently fed into the LSTM layer to model long-term dependencies and contextual relationships across the sequence. The internal mechanism of the LSTM is governed by a series of gating functions:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (2)$$

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = tanh\,(W_c[h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (5)$$

$$h_t = o_t \cdot \tanh\,(C_t) \quad (6)$$

where f_t, i_t, and o_t denote the forget, input, and output gates, respectively, and C_t represents the cell state. These gating mechanisms regulate the flow of information, allowing the model to retain relevant contextual features while discarding irrelevant information across time steps. In the context of emotional text classification, the LSTM layer plays a crucial role in capturing semantic dependencies and contextual nuances that are not explicitly expressed in isolated words but emerge from the sequence as a whole. This is particularly important in distinguishing closely related emotional categories such as stress and depression, which often exhibit overlapping linguistic patterns.

The integration of CNN and LSTM enables the model to jointly exploit local feature extraction and global contextual modeling. This hybrid architecture is particularly effective in addressing semantic overlap in emotional text by combining phrase-level representation learning with sequence-level dependency modeling. Previous studies have demonstrated that CNN–LSTM models consistently outperform single-model approaches in multi-class mental health classification tasks [27], [28]. The detailed configuration of the proposed CNN–LSTM model is presented in Table 1.

Table 1. Compared Model Architectures

Model	Embedding	CNN Layers	LSTM Layers	Output Layer
CNN	Embedding 100-dim, vocab 20.000	Conv1D (128 filter, kernel 5) + MaxPooling1D	–	Dense (64, ReLU) → Dense (5, Softmax)
LSTM	Embedding 100-dim, vocab 20.000	–	Bidirectional LSTM (64 units, dropout 0.3, recurrent 0.3)	Dense (64, ReLU) → Dense (5, Softmax)
CNN–LSTM Hybrid	Embedding 100-dim, vocab 20.000	Conv1D (128 filter, kernel 5) + MaxPooling1D	Bidirectional LSTM (64 units, dropout 0.3, recurrent 0.3)	Dense (64, ReLU) + Dropout (0.3) → Dense (5, Softmax)

2.6. Fine-Tuning

The fine-tuning stage was conducted to optimize key hyperparameters, including the number of CNN filters, kernel sizes, LSTM units, dropout rates, learning rates, batch sizes, and input sequence length. The optimization process was performed using a grid search strategy, where multiple combinations of hyperparameters were systematically evaluated based on validation performance [29], [30].

The search space included learning rates $\{0.001, 0.0005\}$, batch sizes $\{32, 64\}$, CNN filter sizes $\{64, 128\}$, kernel sizes $\{3, 5\}$, and LSTM units $\{32, 64\}$. The optimal configuration was selected based on the highest macro F1-score on the validation set. The best-performing model utilized a learning rate of 0.001, batch size of 64, 128 CNN filters with a kernel size of 5, and 64 LSTM units.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

Fine-tuning plays a critical role in maximizing model accuracy and stability, particularly in multi-class classification tasks, which are inherently more complex than binary classification problems [25]. The use of grid search ensures a transparent and reproducible optimization process, allowing systematic exploration of parameter combinations while avoiding arbitrary selection. The overall fine-tuning workflow is illustrated in Figure 2.

Figure 2. Fine-Tuning Workflow for the CNN–LSTM Hybrid Model

The fine-tuning workflow diagram illustrates the optimization stages of the CNN–LSTM model, beginning from data processing to final model selection. The process starts by loading and cleaning the dataset, followed by partitioning it into training, validation, and testing subsets using a stratified splitting method. Fine-tuning is then applied across multiple model components, including hyperparameter configurations such as learning rate, batch size, and number of epochs, as well as adjustments to the CNN and LSTM architectures to identify the most optimal configuration. The training procedure employs Early Stopping and Model Checkpoint mechanisms to prevent overfitting and ensure that the best-performing weights are preserved. Once the model reaches stable convergence, its performance is evaluated using accuracy, precision, recall, and F1-score metrics. This sequence of stages results in a fully fine-tuned CNN–LSTM model ready for predicting mental health status from social media text.

2.7. Evaluation

The evaluation phase was conducted to measure the model's performance in classifying the five mental health categories using standard metrics for multi-class classification, namely Accuracy, Precision, Recall, F1-score, and Confusion Matrix analysis [31]. All metrics were computed based on the model's predictions on the test set, which consisted of 1,834 tweets.

1. Accuracy

Accuracy is the proportion of correct predictions compared to the entire sample [32].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

In the multi-class case, accuracy is calculated by summing all correct predictions for all classes.

$$Accuracy_{multi-class} = \frac{\sum_{i=1}^{K} TP_i}{Total\ Samples} \quad (8)$$

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

2. Precision

Precision measures the model's accuracy in providing positive predictions for each class. For class I [9].

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (9)$$

Macro Precision is calculated by averaging the precision of all classes

$$Precision_{macro} = \frac{1}{K}\sum_{i=1}^{K} Precision_i \quad (10)$$

3. Recall

Recall shows the model's ability to find all positive samples from each class [33].

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (11)$$

Macro recall:

$$Recall_{macro} = \frac{1}{K}\sum_{i=1}^{K} Recall_i \quad (12)$$

4. F1-score

F1-score is the harmonic mean between Precision and Recall. For class i [34].

$$F1_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (13)$$

Macro F1-score:

$$F1_{macro} = \frac{1}{K}\sum_{i=1}^{K} F1_i \quad (14)$$

5. Confusion Matrix

The confusion matrix is used to view the distribution of predictions for each class and the types of errors the model makes. The general form of a multi-class confusion matrix [35].

$$CM = \begin{bmatrix} TP_1 & FP_{12} & \cdots & FP_{1C} \\ FP_{21} & TP_2 & \cdots & FP_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ FP_{C1} & FP_{C2} & \cdots & TP_C \end{bmatrix} \quad (15)$$

where $\boldsymbol{TP_i}$ represents the number of correctly classified instances for class $\boldsymbol{i}$, and $\boldsymbol{FP_{ij}}$ denotes the number of instances from class $\boldsymbol{i}$ that are misclassified as class $\boldsymbol{j}$. This matrix not only provides a detailed breakdown of classification performance but also enables a deeper error analysis by revealing patterns of misclassification between closely related emotional classes, such as stress and depression, thereby offering insights into semantic overlap and model limitations in distinguishing nuanced emotional expressions.

3. RESULT

3.1. Data Description and Preprocessing Results

The initial dataset consisted of 12,687 tweets collected from Twitter using keywords related to stress, anxiety, depression, mental fatigue, and expressions of psychological well-being. After

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

undergoing data-cleaning procedures including removal of duplicates, empty text, spam, and non-relevant content the number of usable entries decreased to 12,214 tweets. The text was then processed through several preprocessing steps: case folding, noise removal (elimination of URLs, mentions, hashtags, excessive emoticons, and random characters), tokenization, stop-word removal, and lemmatization. Each tweet was manually or semi-automatically labeled into one of five mental health categories: normal, stress, anxiety, depression, and high-risk condition. The class distribution after preprocessing was as follows: normal (4,800 tweets), stress (2,900), anxiety (2,100), depression (1,600), and high-risk condition (814). The average tweet length after preprocessing ranged from approximately 22 to 25 tokens per text. This distribution indicates a moderate class imbalance, particularly for the high-risk condition category, which contains the fewest samples and therefore requires careful consideration in result analysis and model interpretation. The overall data collection and preprocessing pipeline is illustrated in Figure 3.

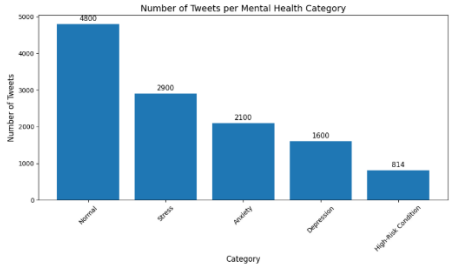

Figure 3. Distribution of Mental Health Class Dataset

3.2. Model Training Results

The complete configuration of the hybrid CNN–LSTM architecture used in this study is presented as follows. The model begins with a 100-dimensional Embedding layer that transforms text into vector representations with a maximum vocabulary size of 20,000 words. The CNN component consists of a Conv1D layer with 128 filters and a kernel size of 5, which extracts local textual patterns, followed by a MaxPooling layer to reduce feature dimensionality. This is followed by a Bidirectional LSTM layer with 64 units and a dropout rate of 0.3 to capture bidirectional sequential context while mitigating overfitting. The fully connected layer employs a Dense layer with 64 units and ReLU activation, along with an additional dropout layer. The output layer consists of 5 Softmax neurons corresponding to the five mental health categories. The model is trained using the Adam optimizer with a learning rate of 0.001, a batch size of 64, and up to 20 epochs, equipped with EarlyStopping to halt training when validation performance no longer improves. The detailed model configuration is summarized in Table 2.

Table 2. Model Architecture Configuration

Komponen	Configuration
Embedding Layer	100-dim, vocab size = 20.000
CNN Layer	Conv1D, 128 filters, kernel size = 5, activation = ReLU
MaxPooling Layer	Pool size = 2
LSTM Layer	Bidirectional LSTM, 64 units, dropout = 0.3, recurrent dropout = 0.3
Fully Connected Layer	Dense (64 units, ReLU) + Dropout (0.3)
Output Layer	Dense (5 units, Softmax)
Optimizer	Adam (learning rate = 0.001)
Batch Size	64
Maximum Epoch	20
Early Stopping	Patience = 3, restore_best_weights = True

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

3.3. Performance Evaluation

The performance comparison of the CNN, LSTM, and CNN–LSTM models is presented in Table 3, evaluated using four primary metrics: Accuracy, Macro Precision, Macro Recall, and Macro F1-score. The results show that the hybrid CNN–LSTM model outperforms the other architectures across all evaluation metrics. The hybrid model achieves the highest accuracy of 0.892, surpassing the CNN model (0.842) and the LSTM model (0.865). This finding demonstrates that the combined architecture leverages the strengths of both components: CNN effectively extracts local textual patterns, whereas LSTM captures long-term dependencies and emotional context within sentences. In terms of precision, the hybrid model attains the highest score of 0.874, indicating superior capability in producing accurate positive predictions compared to the other models. The hybrid model also achieves the highest recall (0.861), reflecting its ability to correctly identify a larger proportion of true samples across all categories. Meanwhile, the F1-score of 0.865 indicates a more balanced performance between precision and recall relative to the CNN and LSTM models.

Tabel 3. Performance Comparison of CNN, LSTM, and CNN–LSTM Models

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-score (Macro)
CNN	0.842	0.830	0.818	0.824
LSTM	0.865	0.852	0.846	0.849
CNN–LSTM Hybrid	**0.892**	**0.874**	**0.861**	**0.865**

To provide a clearer illustration of the effectiveness of each architecture, the following chart presents a performance comparison of the CNN, LSTM, and Hybrid CNN–LSTM models based on four key metrics: Accuracy, Macro Precision, Macro Recall, and Macro F1-score. This visualization aims to quantitatively demonstrate the contribution of each model and highlight which architecture achieves the most optimal performance in predicting the five mental health categories. By examining the chart, one can directly observe the performance improvements achieved by the hybrid model compared to the two single architectures. The comparative performance visualization is shown in Figure 4.

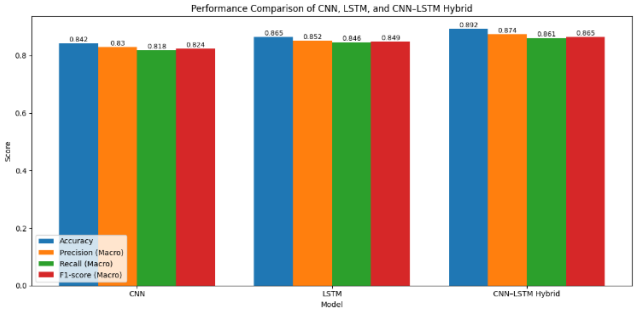

Figure 4. Performance Comparison of CNN, LSTM, and CNN–LSTM Hybrid

The model's accuracy for each of the five mental health classes—Normal, Stress, Anxiety, Depression, and High-Risk Condition—is summarized in Table 4. The highest accuracy is achieved in the Normal category with a score of 0.960, indicating that the model is highly effective in identifying texts that reflect a normal psychological state. This also suggests that linguistic features in normal texts tend to be more consistent and easier to distinguish from emotionally charged categories. The Anxiety class achieves an accuracy of 0.884, demonstrating strong performance in detecting anxiety-related expressions based on linguistic patterns in the text. The High-Risk Condition category also shows solid performance (0.808), despite having fewer samples compared to other classes. In contrast, the lowest accuracies are observed in the Depression (0.745) and Stress (0.751) categories. This reflects semantic

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

overlap among expressions of depression, stress, and anxiety in social media texts, making them more challenging for the model to differentiate. The Macro Average Accuracy of 0.829 indicates that when accuracy is averaged across all classes without considering class size, the model maintains strong and stable performance. Overall, Table 4 demonstrates that the model performs effectively on both majority and minority classes, although challenges remain in distinguishing semantically overlapping emotional categories.

Tabel 4. Class-wise Accuracy

Class	Accuracy
Normal	0.960
Stress	0.751
Anxiety	0.884
Depression	0.745
High-Risk Condition	0.808
Macro Average	**0.829**

To further understand the model's ability to differentiate each mental health category, the following Class-wise Accuracy chart presents the accuracy values for the five analyzed classes: Normal, Stress, Anxiety, Depression, and High-Risk Condition. This visualization provides a more specific overview of which classes are most easily recognized by the model and which ones still lead to misclassification due to similarities in linguistic patterns. Thus, the chart plays an essential role in identifying the model's strengths and limitations at a granular, per-class level. The class-wise accuracy distribution is illustrated in Figure 5.

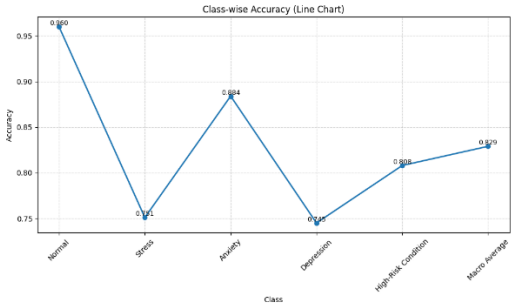

Figure 5. Class-wise Accuracy

3.4. Confusion Matrix

The confusion matrix, as shown in Figure 6, illustrates the performance of the CNN–LSTM model in distinguishing the five mental health categories: Normal, Stress, Anxiety, Depression, and High-Risk Condition. Overall, the model demonstrates strong classification capability, particularly for the Anxiety, High-Risk Condition, and Stress classes, which achieve 245, 210, and 145 correct predictions, respectively. The high number of correct predictions in these three categories indicates that the model is able to recognize distinctive linguistic patterns present in texts expressing anxiety, severe emotional distress, and stress.

For the Depression class, the model yields 167 correct predictions; however, misclassifications still occur, particularly with High-Risk Condition and Anxiety. This may be attributed to the semantic overlap between depressive expressions and those related to anxiety and severe emotional distress, which often appear in similar forms of psychological complaints. For instance, tweets containing expressions of hopelessness or mental exhaustion can be ambiguous and may be predicted as High-Risk Condition by the model.

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

The Normal class obtains 146 correct predictions, with some misclassifications into Stress and Anxiety. This suggests that a small portion of normal tweets contain words or expressions that may superficially resemble emotional states, leading to incorrect predictions. Conversely, the Stress class achieves strong performance with 145 correct predictions, although some misclassifications occur toward Anxiety, Depression, and High-Risk Condition—categories that share similar linguistic characteristics. Distinguishing these conditions accurately requires deeper contextual understanding.

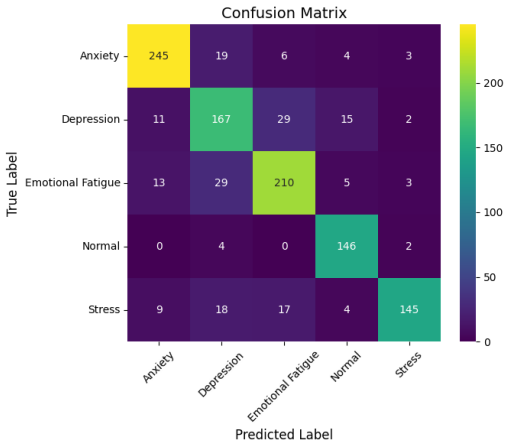

Figure 6. Confusion Matrix

Overall, the confusion matrix presented in Figure 6 indicates that the model performs exceptionally well on majority classes such as Anxiety and High-Risk Condition but still faces challenges in classes with overlapping linguistic patterns, particularly among Depression, Anxiety, and Stress. These findings align with the psychological characteristics commonly observed in social media, where emotional expressions often blend and are not always explicitly distinguishable. This further highlights the limitation of the model in handling subtle contextual differences in short-text emotional expressions.

3.5. Training and Validation Curves

The Training and Validation Accuracy graph illustrates a consistent improvement in the model's performance throughout the training process. Training accuracy increases sharply from approximately 0.35 in the first epoch to over 0.90 by the fourth epoch, indicating that the model rapidly learns the linguistic patterns associated with the five mental health categories. Meanwhile, validation accuracy also shows a significant rise during the first two epochs, thereafter stabilizing in the range of 0.80–0.85. This stability suggests that the model generalizes well to unseen data and does not exhibit strong indications of overfitting.

In the Training and Validation Loss graph, the training loss decreases steadily from around 1.4 to approximately 0.25 by the fourth epoch. This consistent decline signifies that the model successfully minimizes prediction errors on the training data. Conversely, the validation loss drops sharply from about 1.0 in the first epoch to around 0.5 in the second epoch, but then experiences minor fluctuations in subsequent epochs, remaining around 0.7. Although there is a slight increase in validation loss after the second epoch, the variation is relatively small and remains within an acceptable range, indicating no significant signs of overfitting. Overall, both graphs demonstrate that the CNN–LSTM training process progresses effectively. The model is able to learn quickly and stably while maintaining strong generalization capability on the validation data. The use of dropout, the optimized hyperparameter configuration obtained through fine-tuning, and the hybrid CNN–LSTM architecture collectively

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

contribute to the model's performance stability during training. The training and validation loss trends are illustrated in Figure 7.

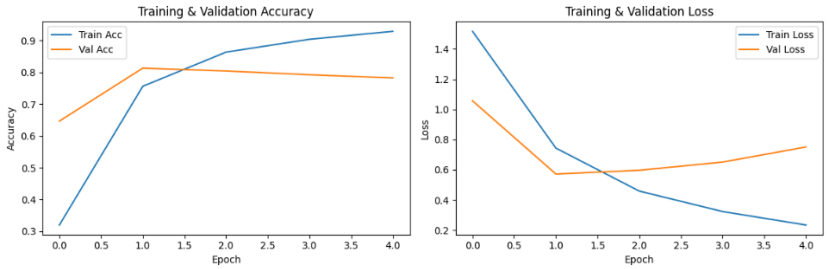

Figure 7. Plot Training & Validation Curves

3.6. Discussion

The results of this study demonstrate that the hybrid CNN–LSTM model delivers superior performance compared to the two single architectures, CNN and LSTM. As shown in Table 3, the CNN–LSTM model achieves an accuracy of 0.892—higher than CNN (0.842) and LSTM (0.865). This finding confirms that the combined architecture effectively leverages the strengths of both methods: CNN excels at extracting local linguistic features, while LSTM captures long-term contextual dependencies. The integration of these components enables the model to better represent complex emotional patterns in social media text.

These findings are consistent with previous studies reporting that hybrid CNN–LSTM architectures outperform single-model approaches in text classification tasks, particularly in multi-class scenarios. Prior research in mental health detection from social media has reported F1-scores in the range of 0.80–0.85. In comparison, the proposed model achieves a macro F1-score of 0.865, indicating improved capability in capturing complex emotional representations. This improvement can be attributed to the combination of local feature extraction and contextual sequence modeling, supported by systematic hyperparameter fine-tuning. At the class level, the results presented in Table 4 show that the model performs exceptionally well for the Normal (0.960) and Anxiety (0.884) categories.

This indicates that linguistic expressions in these classes tend to be more consistent and easier to distinguish. In contrast, the Stress and Depression categories yield lower accuracies (0.751 and 0.745, respectively), reflecting the inherent difficulty of distinguishing closely related emotional states in short-text environments. A deeper error analysis reveals that misclassification between Stress, Depression, and High-Risk Condition is primarily driven by semantic ambiguity in social media text. Expressions such as "feeling tired", "mentally exhausted", and "losing motivation" can represent multiple psychological conditions depending on context. In short-text formats like tweets, the lack of extended contextual information further increases this ambiguity, making it difficult for the model to capture fine-grained emotional distinctions.

This limitation is clearly reflected in the confusion matrix (Figure 6), where overlapping predictions frequently occur among these categories. The High-Risk Condition class achieves an accuracy of 0.808, which, although relatively strong, still indicates some instability due to class imbalance. The limited number of samples in this category reduces the model's ability to learn representative patterns, highlighting the importance of data balancing strategies. This observation aligns with prior studies showing that imbalanced datasets significantly affect classification performance in multi-class mental health detection tasks. From the perspective of training behavior, the accuracy and loss curves demonstrate stable learning without significant overfitting. The relatively small gap between training and validation performance indicates that the use of dropout, early stopping, and

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

hyperparameter optimization effectively enhances generalization capability. This stability is crucial for real-world applications where models must perform reliably on unseen data.

Overall, the findings confirm that the hybrid CNN–LSTM architecture is an effective approach for text-based mental health classification, particularly in handling short-text data with high semantic ambiguity. However, several limitations remain, including class imbalance and difficulty in distinguishing overlapping emotional categories. Future research may address these challenges by incorporating more advanced contextual representation techniques, such as transformer-based models, as well as data augmentation strategies to improve class balance. From a computer science perspective, this study contributes to advancing natural language processing approaches for mental health analysis by demonstrating how hybrid deep learning models can enhance robustness in multi-class classification scenarios.

4. CONCLUSIONS

This study demonstrates that the fine-tuned CNN–LSTM hybrid model provides the most effective approach for predicting mental health status from social media text, outperforming standalone CNN and LSTM architectures. Through a series of training, evaluation, and comparative experiments, the hybrid model achieved the highest overall performance with an accuracy of 0.892, supported by strong macro precision, recall, and F1-score values. These results indicate that the integration of CNN's ability to extract local linguistic patterns with LSTM's capacity to capture long-term contextual dependencies leads to a more robust representation of psychological expressions in text. Class-wise evaluation further highlights the model's strengths and limitations. The model performed exceptionally well in identifying Normal and Anxiety categories, while exhibiting lower accuracy for Stress and Depression, which often share overlapping semantic patterns. Meanwhile, the High-Risk Condition class achieved relatively good accuracy, although its performance remains constrained by limited training instances. Despite these challenges, the overall macro accuracy of 0.829 reflects the model's balanced capability in handling all five mental-health categories. The training process exhibited stable convergence with no significant overfitting, confirming the effectiveness of the selected hyperparameters, dropout mechanisms, and early stopping strategies. Overall, this research contributes a reliable and efficient deep-learning approach for early detection of mental-health indicators from textual data. Future work may focus on addressing class imbalance, enriching linguistic features, and integrating transformer-based models to further enhance predictive performance in real-world mental-health monitoring applications.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENT

The authors gratefully acknowledge Universitas Sains Dan Teknologi Indonesia (USTI) for the support and facilities provided throughout the completion of this research

REFERENCES

[1] World Health Organization, "Over a billion people living with mental health conditions – services require urgent scale-up." Accessed: Sep. 12, 2025. [Online]. Available: https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up?utm_source=chatgpt.com

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

[2] M. E. Aragón, A. P. López-Monroy, M. Montes-y-Gómez, and D. E. Losada, "Online expressions, offline struggles: Using social media to identify depression-related symptoms," *Online Soc. Netw. Media*, vol. 50, p. 100338, 2025, doi: https://doi.org/10.1016/j.osnem.2025.100338.

[3] A. Younus and A. Al-Allawee, "A Comparative Study between Logistic Regression and SVM for Resource Management in Network Slicing," *sinkron*, vol. 9, pp. 1924–1934, Nov. 2025, doi: 10.33395/sinkron.v9i4.15222.

[4] M. Krichen and A. Mihoub, "Long Short-Term Memory Networks: A Comprehensive Survey," *AI*, vol. 6, p. 215, Nov. 2025, doi: 10.3390/ai6090215.

[5] H. Asnal, K. Andesa, and F. Erlin, "Hybrid Machine Learning Model for Risk Prediction and Action Recommendation Based on Artificial Mental Systems," *415 TEKNIKA*, vol. 14, no. 3, pp. 415–423, 2025, doi: 10.34148/teknika.v14i13.1357.

[6] Y. Zhang *et al.*, "Employing Machine Learning and Deep Learning Models for Mental Illness Detection," *Computation*, vol. 13, no. 8, 2025, doi: 10.3390/computation13080186.

[7] J. Aina, O. Akinniyi, M. M. Rahman, V. Odero-Marah, and F. Khalifa, "A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition," *IEEE Access*, vol. 12, pp. 91410–91425, 2024, doi: 10.1109/ACCESS.2024.3421376.

[8] A. A. Maulani, S. Winarno, J. Zeniarja, R. T. E. Putri, and A. N. Cahyani, "Comparison of Hyperparameter Optimization Techniques in Hybrid CNN-LSTM Model for Heart Disease Classification," *Sinkron*, vol. 9, no. 1, pp. 455–465, Jan. 2024, doi: 10.33395/sinkron.v9i1.13219.

[9] M. Wojciuk, Z. Swiderska-Chadaj, K. Siwek, and A. Gertych, "Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization," *Heliyon*, vol. 10, no. 5, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26586.

[10] C. Chatzaki and M. Tsiknakis, "An Overview of Stress Analysis Based on Physiological Signals: Systematic Review of Open Datasets and Current Trends," *Sensors*, vol. 25, no. 23, 2025, doi: 10.3390/s25237108.

[11] N. Holzapfel, "A Depression, Anxiety, and Stress Scale (DASS-42) Study on the Mental Health Conditions of Japanese Employees," *Japanese Psychological Research*, vol. n/a, no. n/a, Feb. 2025, doi: https://doi.org/10.1111/jpr.12587.

[12] M. D. Pham *et al.*, "Mental Health Problems Among Indonesian Adolescents: Findings of a Cross-Sectional Study Utilizing Validated Scales and Innovative Sampling Methods," *Journal of Adolescent Health*, vol. 75, no. 6, pp. 929–938, Dec. 2024, doi: 10.1016/j.jadohealth.2024.07.016.

[13] M. D. Pham *et al.*, "Mental Health Problems Among Indonesian Adolescents: Findings of a Cross-Sectional Study Utilizing Validated Scales and Innovative Sampling Methods," *Journal of Adolescent Health*, vol. 75, no. 6, pp. 929–938, 2024, doi: https://doi.org/10.1016/j.jadohealth.2024.07.016.

[14] Q. Cheng *et al.*, "Determinants of healthcare utilization under the Indonesian national health insurance system – a cross-sectional study," *BMC Health Serv. Res.*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12913-024-11951-8.

[15] A. Pandya *et al.*, "Evaluating current trends in stress and depression detection using artificial intelligence and machine learning," *Intelligent Hospital*, vol. 2, no. 1, p. 100047, 2026, doi: https://doi.org/10.1016/j.inhs.2025.100047.

[16] Z. S. Chen, P. (Param) Kulkarni, I. R. Galatzer-Levy, B. Bigio, C. Nasca, and Y. Zhang, "Modern views of machine learning for precision psychiatry," *Patterns*, vol. 3, no. 11, Nov. 2022, doi: 10.1016/j.patter.2022.100602.

[17] M. Garg, "Mental Health Analysis in Social Media Posts: A Survey," *Archives of Computational Methods in Engineering*, vol. 30, no. 3, pp. 1819–1842, Apr. 2023, doi: 10.1007/s11831-022-09863-z.

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

[18] T. Zhang, K. Yang, S. Ji, and S. Ananiadou, "Emotion fusion for mental illness detection from social media: A survey," *Information Fusion*, vol. 92, pp. 231–246, 2023, doi: https://doi.org/10.1016/j.inffus.2022.11.031.

[19] H. Asnal, K. Andesa, and F. Erlin, "Hybrid Machine Learning Model for Risk Prediction and Action Recommendation Based on Artificial Mental Systems," *415 TEKNIKA*, vol. 14, no. 3, pp. 415–423, 2025, doi: 10.34148/teknika.v14i13.1357.

[20] M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, "Natural language processing for mental health interventions: a systematic review and research framework," Dec. 01, 2023, *Springer Nature*. doi: 10.1038/s41398-023-02592-2.

[21] J. Golec and T. Hachaj, "Ten Natural Language Processing Tasks with Generative Artificial Intelligence," *Applied Sciences*, vol. 15, no. 16, 2025, doi: 10.3390/app15169057.

[22] H. Bichri, A. Chergui, and H. Mustapha, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *International Journal of Advanced Computer Science and Applications*, vol. 15, Nov. 2024, doi: 10.14569/IJACSA.2024.0150235.

[23] Y. Hu, C. Xu, B. Lin, W. Yang, and Y. Y. Tang, "Medical multimodal large language models: A systematic review," *Intelligent Oncology*, vol. 1, no. 4, pp. 308–325, 2025, doi: https://doi.org/10.1016/j.intonc.2025.09.005.

[24] E. Oro, F. M. Granata, and M. Ruffolo, "A Comprehensive Evaluation of Embedding Models and LLMs for IR and QA Across English and Italian," *Big Data and Cognitive Computing*, vol. 9, no. 5, 2025, doi: 10.3390/bdcc9050141.

[25] S. K. Sharma, A. R. Khan, G. G. Tejani, D. Bassir, and S. Tripathi, "MultiMindNet: AI-based mental health analysis using hybrid deep learning approach and Hybrid Ant-Grey Wolf Optimization (HAGWO) algorithm," *PLOS Digital Health*, vol. 5, no. 4 April, Apr. 2026, doi: 10.1371/journal.pdig.0001158.

[26] P. Ta, N. Tran, H. Nguyen, and H. D. Nguyen, "Detecting signs of depression on social media: A machine learning analysis and evaluation," *Sustainable Futures*, vol. 10, p. 100827, 2025, doi: https://doi.org/10.1016/j.sftr.2025.100827.

[27] T. Aris and C. Ningning, "Integration of CNN and LSTM Networks for Behavior Feature Recognition: An Analysis," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, pp. 1793–1799, Nov. 2024, doi: 10.18517/ijaseit.14.5.10116.

[28] M. Dwirizqy Wimbassa, T. Marsyah Noor, S. Yasara, and T. Muhammad Arsyah, "Emotional Text Detection dengan Long Short Term Memory (LSTM) 1," *Jurnal Format*, vol. 12, 2023.

[29] W. Hussain *et al.*, "Ensemble genetic and CNN model-based image classification by enhancing hyperparameter tuning," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-024-76178-3.

[30] N. Alkaabi, S. Shakya, and R. Mizouni, "A hyperparameter optimization framework for transformer-based time series forecasting using evolutionary algorithms," *Neural Comput. Appl.*, vol. 38, no. 7, p. 236, 2026, doi: 10.1007/s00521-026-11959-7.

[31] A. Lubis, Irawan Yuda, Junadhi, and Defit Sarjon, "Leveraging K-Nearest Neighbors with SMOTE and Boosting Techniques for Data Imbalance and Accuracy Improvement," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1625–1638, Dec. 2024, doi: 10.47738/jads.v5i4.343.

[32] Y. Deng, M. R. Eden, and S. Cremaschi, "Metrics for Evaluating Machine Learning Models Prediction Accuracy and Uncertainty," in *33rd European Symposium on Computer Aided Process Engineering*, vol. 52, A. C. Kokossis, M. C. Georgiadis, and E. Pistikopoulos, Eds., in Computer Aided Chemical Engineering, vol. 52. , Elsevier, 2023, pp. 1325–1330. doi: https://doi.org/10.1016/B978-0-443-15274-0.50211-0.

[33] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Eng. Appl. Artif. Intell.*, vol. 136, p. 108972, 2024, doi: https://doi.org/10.1016/j.engappai.2024.108972.

Vol. 7, No. 3, June 2026, Page. 2978-2993
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5882

[34] H. Handoko, A. Asrofiq, J. Junadhi, and A. S. Negara, "Sentiment Analysis of Sirekap Tweets Using CNN Algorithm," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 8, no. 2, pp. 312–329, Aug. 2024, doi: 10.29407/intensif.v8i2.23046.

[35] H. Asnal, K. Andesa, F. Erlin, and Junadhi, "Hybrid Machine Learning Model for Risk Prediction and Action Recommendation Based on Artificial Mental Systems," *Teknika*, vol. 14, no. 3, pp. 415–423, 2025, doi: 10.34148/teknika.v14i3.1357.