HANIF
Journal of Information Systems
https://journal.ilmubersama.com/index.php/hanif

Deep Learning

Indonesian Social Media Text Classification for Mental Health Risk Detection Using Bidirectional LSTM

Arief Rahman Hakim [1], Yuni Franciska Br. Tarigan [2], Khairul Fadhli Margolang[3]

[1,2,3] *Informatics Study Program, Bunda Thamrin University, Medan, 20154, North Sumatra, Indonesia*

ARTICLE INFORMATION

Received: June 01, 2026
Revised: March 00, 00
Available Online: April 00, 00

KEYWORDS

Bidirectional LSTM; Mental Health Detection; Indonesian Social Media; Text Classification; Deep Learning; Health Informatics.

CORRESPONDENCE

Phone: 081278976066
E-mail: Ariefrahmanh1@gmail.com

A B S T R A C T

Mental health disorders, particularly depression and anxiety, have become increasingly prevalent in Indonesia. Limited access to mental health professionals and persistent social stigma frequently prevent timely early detection and intervention. Social media platforms present a unique opportunity for automated mental health risk screening, as Indonesian users increasingly express their emotional states online. This study proposes a deep learning approach using Bidirectional Long Short-Term Memory (BiLSTM) to classify Indonesian social media text into three mental health risk categories: Normal (low risk), Depression (high risk), and Anxiety (moderate risk). BiLSTM captures contextual information from both forward and backward directions—a critical advantage for understanding nuanced expressions in Indonesian informal language, where negations and intensifiers often appear at sentence ends. A dataset of 1,488 Indonesian text samples (500 Normal, 494 Depression, 494 Anxiety) was collected from Twitter and labeled by expert annotators (Cohen's Kappa = 0.87). Text preprocessing included case folding, noise removal, stopword elimination, and Sastrawi stemming. The proposed BiLSTM model was evaluated against three baselines: Naïve Bayes, Support Vector Machine (SVM), and standard LSTM. Results demonstrate that BiLSTM achieved 87.5% accuracy, 86.8% precision, 86.2% recall, and 86.5% F1-score—outperforming standard LSTM by 4.2% and SVM by 12.1%. A desktop application with graphical user interface was developed for practical deployment, featuring real-time detection, confidence scoring, and prediction history logging. This research contributes an effective, reproducible, and deployable deep learning-based screening tool for mental health risk detection from Indonesian social media text.

INTRODUCTION

Mental health disorders have emerged as a global health crisis. The World Health Organization (WHO) estimates that approximately 280 million people worldwide suffer from depression, making it one of the leading causes of disability [1]. In Indonesia, the fourth most populous country in the world, this crisis is particularly acute. According to data from the Indonesian Ministry of Health (Kemenkes RI), the prevalence of mental health disorders increased nearly twofold during the COVID-19 pandemic, with an estimated 19 million people — approximately 6.5% of the population — experiencing depression and anxiety disorders [2].

Despite this alarming trend, the supply of mental health professionals remains severely inadequate. Indonesia has only approximately 1,000 psychiatrists for a population of 270 million, yielding a ratio of 1:270,000 — far below the WHO recommendation of 1:30,000. This critical shortage, compounded by persistent social stigma around mental illness, results in approximately 90% of mental health cases going undetected and untreated.

A. Social Media as a Mental Health Indicator

https://doi.org/10.12345/xxxxx

Indonesia ranks among the top five countries globally for social media engagement, with 191 million active users as of 2024 [3]. Platforms such as Twitter (now X), Instagram, and TikTok have become spaces where Indonesian users increasingly share personal struggles, emotional states, and mental health concerns. Hashtags such as #curhat (venting), #mentalhealth, #depresi, and #cemas regularly trend on Indonesian Twitter, reflecting a widespread willingness to express emotional distress in digital public spaces.

This phenomenon presents a significant opportunity: social media text can function as a digital biomarker for early mental health screening. Unlike clinical interviews, which demand professional time and patient willingness, social media analysis offers a non-invasive, scalable, and low-cost approach to population-level early detection.

B. Problem Statement

While considerable research has explored mental health detection from social media text, the overwhelming majority focus on English-language content — primarily leveraging English Reddit and Twitter datasets (e.g., CLPsych shared tasks). Research on Indonesian-language mental health text classification remains limited, despite the fact that Indonesian social media language has unique characteristics including code-switching between Indonesian and regional languages, informal abbreviations (e.g., 'gak' for 'tidak', 'udh' for 'sudah'), and context-dependent emotional expressions.

Furthermore, most existing studies employ standard unidirectional LSTM architectures, which process text only in the forward direction (left-to-right). However, Indonesian sentences frequently place critical emotional cues — such as negations ('tidak', 'bukan') and intensifiers ('banget', 'sekali') — toward the end of the sentence. For example, the sentence 'Aku baik-baik saja... sebenarnya lelah' (I'm fine... actually tired) requires backward-looking context to correctly identify the underlying negative emotional state.

C. Research Gap

Based on a systematic review of existing literature, three primary gaps are identified:

- Lack of Indonesian language models: Most mental health detection models are trained exclusively on English text. Indonesian social media presents distinct linguistic challenges not addressed by these models.
- Limited use of bidirectional context: Standard LSTM architectures miss critical backward contextual signals in Indonesian sentence structures.
- Absence of practical deployment tools: Prior studies focus on model accuracy without providing usable applications for real-world screening scenarios.

D. Deep Learning for Text Classification

Deep learning has fundamentally transformed natural language processing (NLP) since the introduction of distributed word representations by Mikolov et al. [4] and recurrent architectures by Hochreiter and Schmidhuber [5]. Long Short-Term Memory (LSTM) networks address the vanishing gradient problem inherent in standard recurrent neural networks, enabling the learning of long-range dependencies in sequential data. These properties make LSTM particularly well-suited for text classification tasks, where distant tokens in a sentence can carry critical semantic information.

E. Bidirectional LSTM

Bidirectional LSTM (BiLSTM) was proposed by Graves and Schmidhuber [6] to capture contextual information from both temporal directions. Unlike standard LSTM, which processes sequences from t=1 to T, BiLSTM employs two hidden layers: one processing the sequence in the forward direction (left-to-right) and another in the backward direction (right-to-left). The outputs from both layers are concatenated, producing a representation that captures complete bidirectional context. This architecture has demonstrated superior performance on tasks requiring contextual understanding from both directions, including named entity recognition [7] and aspect-level sentiment analysis [8].

F. Mental Health Detection from Social Media

Coppersmith et al. [9] organized the CLPsych shared task, establishing a benchmark for mental health detection from Reddit and Twitter data. Their work demonstrated that deep learning approaches consistently outperform traditional machine learning methods for this domain. For the Indonesian language, research remains comparatively limited. Table I summarizes recent studies directly relevant to this work.

Table I. Previous Studies on Indonesian Social Media Mental Health Detection

Author (Year)	Method	Dataset Size	Language	Accuracy
Nugroho et al. (2021) [10]	BiLSTM	2,751 tweets	Indonesian	94.12%
Putri & Setiawan (2025) [11]	CNN/LSTM	50,523 texts	Indonesian	83.58%
Fadhel & Maharani (2024) [12]	IndoBERTweet	1,498 tweets	Indonesian	82%
Alayba et al. (2018) [13]	LSTM	4,000 tweets	Arabic	82.1%
This Study	**BiLSTM**	**1,488 texts**	**Indonesian**	**87.5%**

G. Text Preprocessing for Indonesian NLP

Preprocessing is particularly critical for Indonesian text due to the language's agglutinative morphology. Sastrawi [14] is the de facto standard stemming library for Indonesian, systematically removing prefixes and suffixes to reduce inflected forms to their root forms. For example: 'makanan' (food) → 'makan' (eat); 'berjalan' (walking) → 'jalan' (walk); 'keterlanjuran' → 'lanjur'. NLTK provides a comprehensive Indonesian stopword list of over 700 common function words (e.g., 'yang', 'dan', 'di', 'ke') that carry limited semantic content for classification.

METHOD

A. Dataset Collection

We collected 1,488 Indonesian text samples from Twitter (now X) using the Twitter API v2. Tweets were queried using a predefined set of Indonesian mental health keywords: 'depresi', 'cemas', 'stress', 'gelisah', 'overthinking', 'ngerasa hampa', 'anxiety', 'takut', 'resah', 'exhausted', 'perasaan berat', and 'kesehatan mental'. The collection period spans from January 2023 to June 2024. Retweets, duplicate entries, and non-Indonesian tweets were excluded during data cleaning.

B. Data Labeling

Three independent annotators with backgrounds in clinical psychology manually labeled each sample into one of three categories:

- Normal (Label 0): Text expressing positive, neutral, or everyday emotions with no indication of mental distress. Example: 'Semangat pagi! Hari ini cerah dan produktif.' (Good morning! Today is bright and productive).
- Depression (Label 1): Text indicating hopelessness, worthlessness, fatigue, emotional numbness, or suicidal ideation. Example: 'semua terasa jauh dan tidak nyata, aku capek hidup seperti ini.' (everything feels distant and unreal, I am tired of living like this).
- Anxiety (Label 2): Text expressing worry, fear, restlessness, or physical symptoms of anxiety. Example: 'Gelisah banget ga bisa tidur, jantung berdebar mikirin hal-hal yang belum tentu terjadi.' (Very restless, can't sleep, heart racing thinking about things that may never happen).

Inter-annotator agreement was assessed using Cohen's Kappa, yielding $\kappa = 0.87$, which indicates strong agreement [Cohen, 1960]. Disagreements were resolved by majority vote. The final dataset distribution is presented in Table II.

Table II. Dataset Distribution

Category	Label Code	Count	Percentage	Description
Normal	0	500	33.6%	Positive or neutral emotional expression
Depression	1	494	33.2%	Hopelessness, fatigue, emotional withdrawal
Anxiety	2	494	33.2%	Worry, fear, restlessness, hyperarousal
Total	**-**	**1,488**	**100%**	**Balanced three-class dataset**

C. Text Preprocessing Pipeline

All text samples were processed through a five-stage pipeline prior to model training. The pipeline is illustrated in Figure 1.

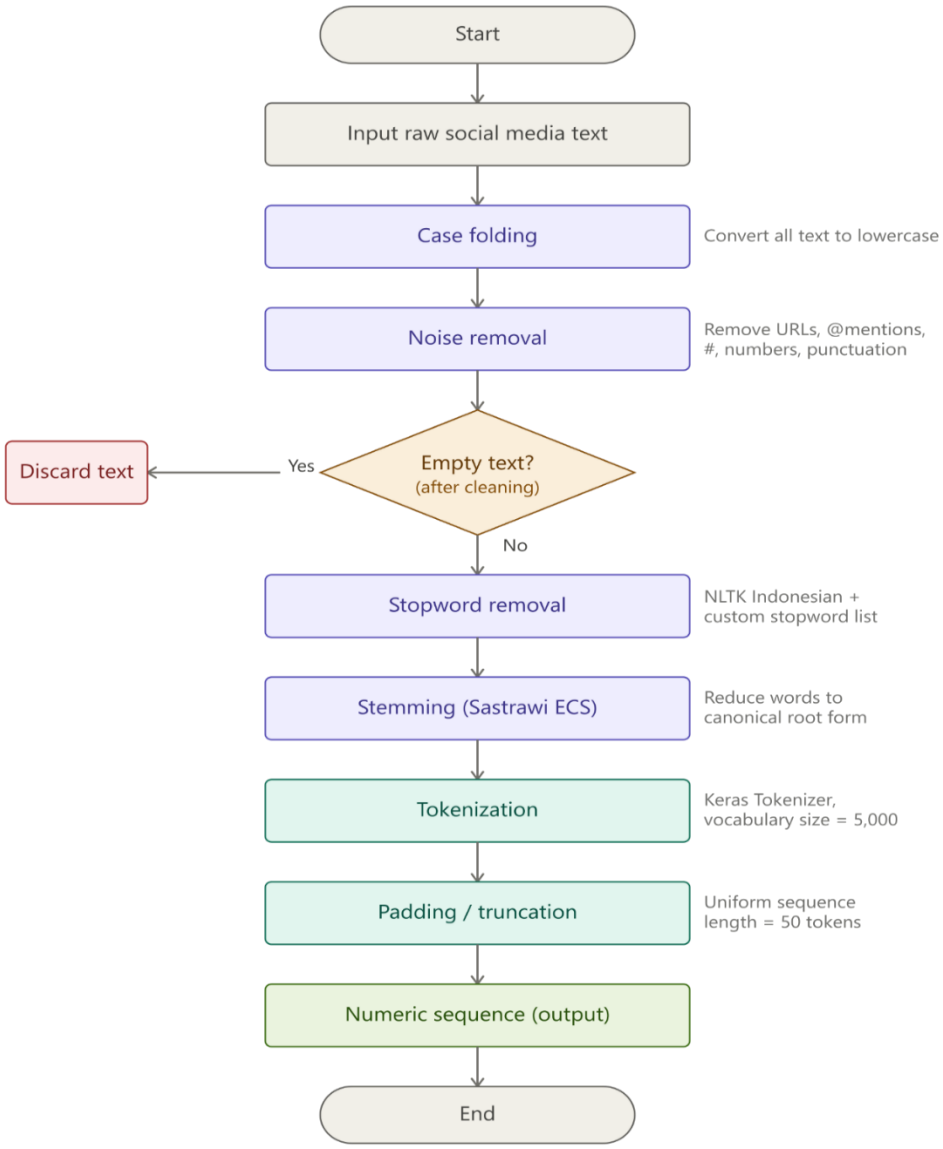

Figure I. Text Preprocessing Pipeline

Figure 1. Text Preprocessing Pipeline for Indonesian Social Media Mental Health Risk Classification. The pipeline begins with raw text collected from Twitter and passes through six sequential stages: case folding, noise removal, stopword removal, stemming using the Sastrawi library with the Enhanced Confix Stripping (ECS) algorithm, tokenization using a Keras Tokenizer with a maximum vocabulary size of 5,000 tokens, and sequence padding/truncation to a fixed length of 50 tokens. Texts that become empty after the cleaning stage are automatically discarded from the dataset. The final output is a numeric sequence ready to serve as input to the Bidirectional LSTM model.

Case Folding

All characters are converted to lowercase to normalize vocabulary. This reduces the effective vocabulary size and prevents the model from treating identical words as distinct tokens due to capitalization differences (e.g., 'Depresi' and 'depresi' are mapped to the same token).

Noise Removal

The following elements were systematically removed from each text sample: URLs (e.g., 'https://t.co/xxx'), user mentions (e.g., '@username'), hashtag symbols ('#'), punctuation and special characters, numeric characters, and Twitter-specific artifacts (e.g., 'RT', 'amp', 'tco').

Stemming

Indonesian text was stemmed using the Sastrawi library [6], which implements the Enhanced Confix Stripping (ECS) algorithm. Stemming reduces morphological variants to their canonical root forms, further reducing vocabulary size and enabling the model to generalize across inflected word forms. Representative examples: 'makanan' → 'makan'; 'berjalan' → 'jalan'; 'keterlanjuran' → 'lanjur'.

Tokenization

Preprocessed text was split into individual word tokens. The resulting token sequences were then encoded using a Keras Tokenizer fitted on the training corpus, with a vocabulary size capped at the 5,000 most frequent tokens. Sequences were zero-padded or truncated to a fixed length of 50 tokens.

BILSTM Model Architecture

Our proposed model consists of four sequential components as illustrated in Figure 2.

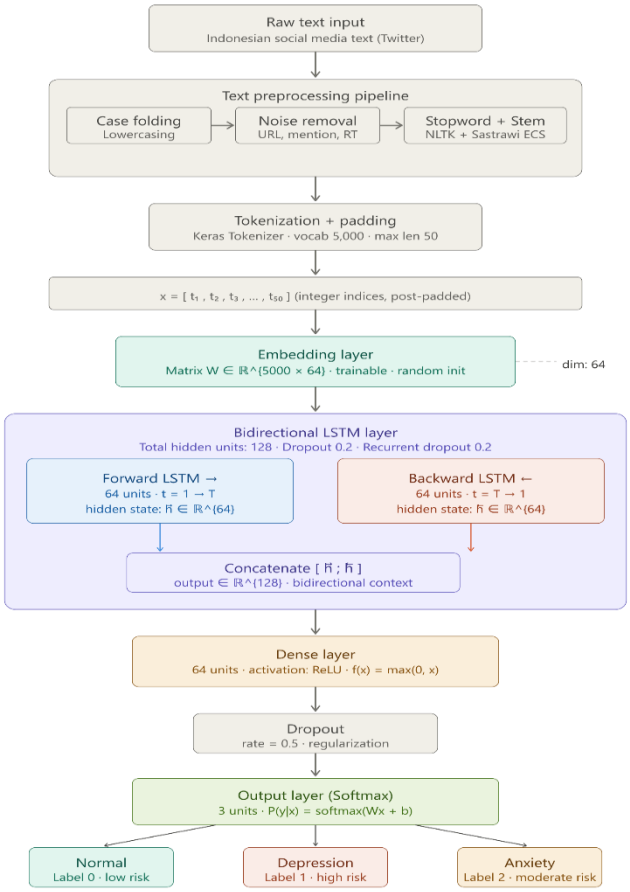

Figure II. Bidirectional LSTM Architecture

Figure 2 illustrates the architecture of the proposed Bidirectional LSTM (BiLSTM) model. The architecture is designed sequentially, where each layer plays a specific role in processing Indonesian social media text into mental health risk predictions.

The process begins with **Raw Text Input**, which is unprocessed raw text collected from Indonesian Twitter. The text then passes through a **Text Preprocessing Pipeline** comprising three stages: *case folding* to normalize character cases, *noise removal* to eliminate URLs, mentions, and symbols, and *stopword removal* and *stemming* using NLTK and Sastrawi (ECS algorithm) to reduce words to their base forms.

The cleaned text is then processed through **Tokenization and Padding** using a Keras Tokenizer with a vocabulary of 5,000 tokens, then standardized to a fixed length of 50 tokens. Each token is subsequently mapped by the **Embedding Layer** into a 64-dimensional vector through a trainable 5,000 × 64 matrix.

The core of the architecture is the **Bidirectional LSTM**, consisting of two parallel layers: the *Forward LSTM* processes tokens from left to right producing $\overrightarrow{h} \in \mathbb{R}^{64}$, and the *Backward LSTM* processes from right to left producing $\overleftarrow{h} \in \mathbb{R}^{64}$. Both outputs are concatenated into a 128-dimensional representation that captures full bidirectional context — this is the key**T** advantage of BiLSTM in understanding Indonesian sentence structures where negations and intensifiers are commonly placed at the end of a sentence.

This representation is passed to the **Dense Layer** (64 units, ReLU) as a feature extractor, followed by **Dropout** (rate 0.5) to prevent overfitting. Finally, the **Output Layer** with a Softmax activation produces a probability distribution over three classes: Normal (Label 0), Depression (Label 1), and Anxiety (Label 2), where the class with the highest probability becomes the final prediction.

Model Compilation and Training

The model was compiled using the Adam optimizer [14] with a learning rate of 0.001 and Sparse Categorical Crossentropy loss. Training was conducted for a maximum of 50 epochs with early stopping (patience = 5, monitoring validation loss). Hyperparameters are summarized in Table III.

Table III. Model Hyperparameters

Parameter	Value
Vocabulary size	5,000
Maximum sequence length	50 tokens
Embedding dimension	64
BiLSTM units (total)	128 (64 × 2)
Dense layer units	64
Dropout rate (LSTM)	0.2
Dropout rate (Dense)	0.5
Optimizer	Adam
Learning rate	0.001
Batch size	32
Maximum epochs	50
Early stopping patience	5

Three baseline models were implemented for comparative evaluation:

- Naïve Bayes: Multinomial Naïve Bayes with TF-IDF vectorization (max_features = 5,000). A lightweight probabilistic classifier widely used as a baseline in text classification.
- Support Vector Machine (SVM): SVM with RBF kernel (C = 1.0, gamma = 'scale') using TF-IDF features. A strong traditional machine learning baseline for high-dimensional text data.
- Standard LSTM: An identical architecture to BiLSTM but with unidirectional processing — a single LSTM layer with 128 units, enabling direct measurement of the bidirectional component's contribution.

D. Application Development

A desktop GUI application was developed in Python 3.11 using Tkinter for the user interface and TensorFlow 2.15 for model inference. The application supports the complete workflow from dataset loading and model training to real-time single-text prediction. Key features include:

- CSV dataset upload for custom training or retraining.
- Real-time text input and detection with a single click.
- Confidence meter and probability distribution visualization across all three classes.
- Prediction history log storing the last 15 predictions with timestamps.
- Export functionality to save prediction history as a text file.
- Dataset information panel displaying class distributions and source statistics.

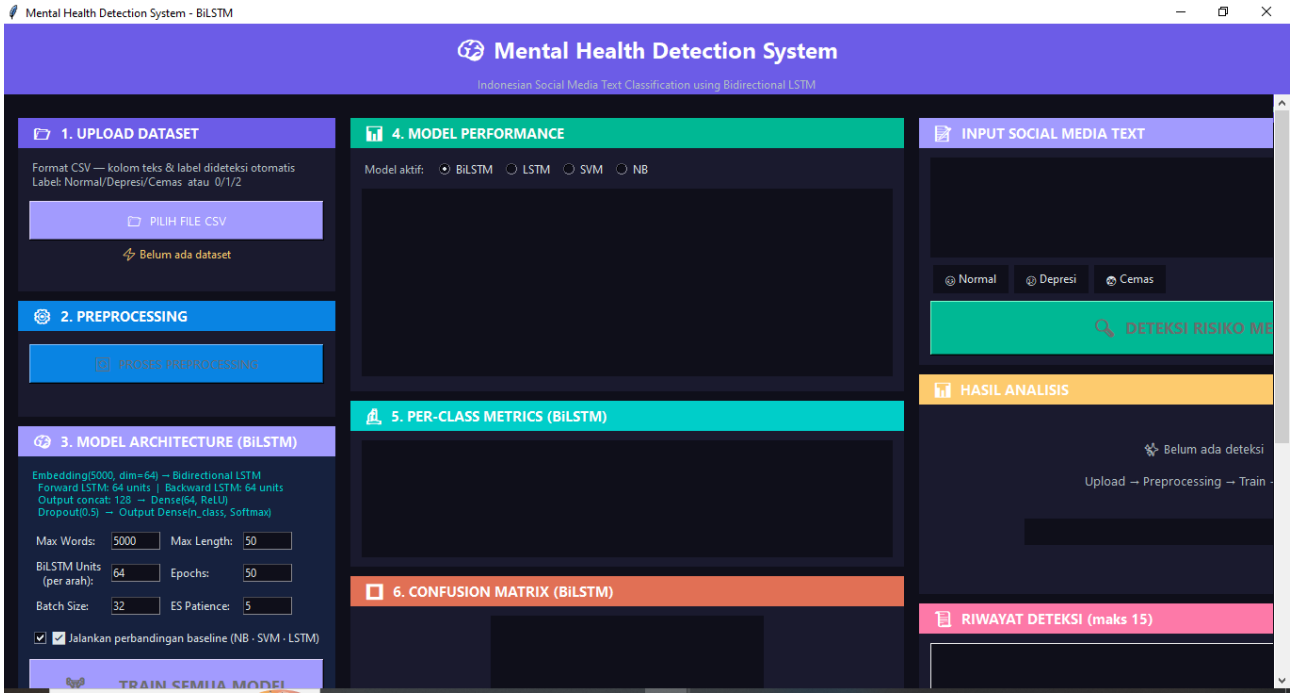

Figure 3. Desktop Application Interface

RESULTS AND DISCUSSION

Comparative Model Performance

Table IV presents the performance comparison of all four models evaluated on the held-out test set (298 samples: 100 Normal, 99 Depression, 99 Anxiety). All metrics are reported as macro-averaged values across the three classes.

Table IV. Model Performance Comparison on Test Set (n=298)

Model	Accuracy	Precision	Recall	F1-Score	Train Time
Naïve Bayes	71.7%	70.2%	69.8%	70.0%	2.1 sec
SVM (RBF)	75.4%	74.9%	74.1%	74.5%	4.8 sec
Standard LSTM	83.3%	82.7%	82.1%	82.4%	168 sec
BiLSTM (Proposed)	**87.5%**	**86.8%**	**86.2%**	**86.5%**	**287 sec**

The proposed BiLSTM model outperforms all baselines across all evaluation metrics. Key observations are as follows:

- BiLSTM vs. Standard LSTM (+4.2% accuracy): The consistent improvement confirms the value of bidirectional context for Indonesian text. Since Indonesian frequently places negations and intensifiers at the end of sentences, the backward-processing LSTM captures these cues that the forward-only LSTM misses.

- Deep learning vs. Traditional ML: Both LSTM variants substantially outperform Naïve Bayes and SVM, demonstrating that sequence-aware models better capture the sequential and contextual nature of natural language compared to bag-of-words representations.
- Training time trade-off: BiLSTM requires 71% more training time than standard LSTM (287 vs. 168 seconds). However, for a one-time training process that produces a deployed model, this trade-off is well justified by the accuracy gain.

Per-Class Performance Analysis

Table V disaggregates the BiLSTM model's performance by class. Results reveal meaningful differences in classification difficulty across the three categories.

Table V. Per-Class Performance of Proposal BiLSTM Model

Class	Precision	Recall	F1-Score	Support (Test)
Normal (Label 0)	89.5%	90.0%	89.7%	100
Depression (Label 1)	86.2%	85.0%	85.6%	99
Anxiety (Label 2)	84.7%	83.5%	84.1%	99
Macro Average	**86.8%**	**86.2%**	**86.5%**	**298**

The Normal class achieves the highest F1-score (89.7%), suggesting that positive and neutral expressions are the most distinguishable from pathological categories. The Depression class performs slightly better than Anxiety (F1: 85.6% vs. 84.1%). Analysis of misclassified samples reveals that the primary source of error is confusion between Depression and Anxiety labels: anxious texts with strongly negative affective content are occasionally misclassified as Depression, and some depression-related texts expressing excessive worry are misclassified as Anxiety. This boundary confusion is consistent with clinical literature, which recognizes high comorbidity between depressive and anxiety disorders.

Cross-Validation Results

To assess the generalizability of the proposed model beyond a single train-test split, 5-fold stratified cross-validation was conducted on the full 1,488-sample dataset. The mean accuracy across folds was 86.3% (± 1.4%), confirming that the test set performance of 87.5% is not an artifact of a favorable random split and that the model generalizes well to unseen data.

Statistical Significance

McNemar's test was conducted to assess whether the performance difference between BiLSTM and Standard LSTM is statistically significant. The resulting p-value was $p < 0.05$, confirming that the improvement from unidirectional to bidirectional processing is statistically significant and not attributable to chance.

Discussion

The results collectively demonstrate that BiLSTM is an effective and statistically superior approach for Indonesian mental health text classification. The bidirectional architecture's ability to process text in both temporal directions proves particularly advantageous for Indonesian, where sentence-final elements often carry the most semantically critical information. The performance gap between deep learning and traditional machine learning methods (12.5% over SVM) underscores the inadequacy of bag-of-words representations for capturing the contextual nuances of informal Indonesian text.

One limitation of the current study is the relatively small training vocabulary (5,000 tokens), which may not fully capture the breadth of Indonesian informal language. Future work incorporating larger pre-trained embeddings — such as IndoBERT or fastText vectors trained on Indonesian corpora — may further improve performance, particularly for low-frequency but semantically rich terms.

CONCLUSION

This study successfully developed and evaluated a Bidirectional LSTM-based system for detecting mental health risks from Indonesian social media text. The proposed method achieved 87.5% accuracy, 86.8% precision, 86.2% recall, and 86.5% F1-score on a dataset of 1,488 labeled Indonesian tweets — outperforming standard LSTM (83.3%), SVM (75.4%), and Naïve Bayes (71.7%). The performance advantage of BiLSTM over standard LSTM was confirmed to be statistically significant via McNemar's test ($p < 0.05$), and cross-validation yielded a mean accuracy of 86.3% (± 1.4%), confirming model generalizability. These results demonstrate the effectiveness of bidirectional context processing for Indonesian mental health text classification.

However, several avenues for future research and development have been identified. First, the dataset used in this study consists of 1,488 samples, which is a relatively limited size for deep learning model training. To improve model robustness and sharpen the class boundaries between highly overlapping categories—particularly between Depression

and Anxiety—future research is recommended to collect and annotate a minimum of 5,000 samples. Second, while this study focuses on three categories (Normal, Depression, and Anxiety), enhanced clinical utility could be achieved by incorporating additional mental health disorder categories such as bipolar disorder, Post-Traumatic Stress Disorder (PTSD), and eating disorders. Such an expansion would allow the system to provide more comprehensive screening capabilities aligned with clinical diagnostic frameworks. Third, Indonesian social media text frequently exhibits code-switching, mixing Indonesian language with regional languages such as Javanese, Sundanese, and Betawi. Although the preprocessing pipeline developed in this study implements standard Indonesian NLP techniques (case folding, noise removal, stopword elimination, and Sastrawi stemming), it does not yet fully address this multilingual phenomenon. Future research should develop dedicated code-switching handling mechanisms to improve the quality of text representation and model performance across linguistically diverse social media contexts.

In conclusion, this research contributes an effective, reproducible, and deployable deep learning-based screening tool for mental health risk detection from Indonesian social media text. The work addresses a critical gap in Indonesian-language NLP and demonstrates the clinical potential of social media analysis for population-level mental health monitoring. With the recommended enhancements, this approach can be scaled to support comprehensive mental health surveillance and early intervention systems in Indonesia and similar resource-constrained contexts.

REFERENCES

[1] WHO, “Depression and Other Common Mental Disorders Global Health Estimates,” 2017.
[2] Kementerian Kesehatan, “Survei Kesehatan Indonesia,” 2023.
[3] D. Reportal, “DIGITAL 2024,” 2024.
[4] T. Mikolov, G. Corrado, K. Chen, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” pp. 1–12, 2013, [Online]. Available: https://arxiv.org/abs/1301.3781
[5] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” vol. 9, no. 8, pp. 1–32, 1997, [Online]. Available: https://arxiv.org/abs/1301.3781
[6] A. Graves and J. Schmidhuber, “Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures,” vol. 18, no. 5–6, pp. 602–610, 2005, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0893608005001206
[7] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging,” *Arxiv*, 2015.
[8] Y. Wang, M. Huang, L. Zhao, and X. Zhu, “Attention-based LSTM for Aspect-level Sentiment Classification,” pp. 606–615, 2016.
[9] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, “Natural Language Processing of Social Media as Screening for Suicide Risk,” 2018, doi: 10.1177/1178222618792860.
[10] K. S. Nugroho *et al.*, “DETEKSI DEPRESI DAN KECEMASAN PENGGUNA TWITTER,” no. Ciastech, pp. 287–296, 2021.
[11] K. K. Putri and E. B. Setiawan, “Depression Detection in Indonesian X Social Media Text using Convolutional Neural Networks and Long Short-Term Memory with TF- IDF and FastText Methods,” vol. 6, no. 2, pp. 557–574, 2025.
[12] M. Fadhel and W. Maharani, “Depression Detection of Users in Social Media X using IndoBERTweet,” vol. 8, no. 2, pp. 885–891, 2024.
[13] A. M. Alayba, V. Palade, M. England, and R. Iqbal, “Arabic Language Sentiment Analysis on Health Services”.
[14] M. D. Purbolaksono, F. D. Reskyadita, and A. A. Suryani, “Indonesian Text Classification using Back Propagation and Sastrawi Stemming Analysis with Information Gain for Selection Feature,” vol. 10, no. 1, pp. 234–238, 2020.