Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16294
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Depression Detection on Indonesian Social Media Using Fine-Tuned IndoBERT and SVM

Donny Amanullah Putra Rahman[1)*], Muhamad Akrom[2)], Muhammad Naufal[3)]
[1,2,3)]Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[1)]111202315426@mhs.dinus.ac.id, [2)]m.akrom@dsn.dinus.ac.id, [3)]m.naufal@dsn.dinus.ac.id

Submitted :May 29, 2026 | **Accepted** : Jul 27, 2026 | **Published** : Jul 31, 2026

Abstract: Depression has become a major mental health issue in Indonesia, where approximately 167 million of the country's 273 million citizens actively use social media platforms such as X (Twitter). The informal writing style, code-mixing, and linguistic variability in Indonesian tweets create significant challenges for automated depression detection systems. This study evaluates a fine-tuned IndoBERT model combined with a Support Vector Machine (SVM) classifier for detecting depression-related indications from Indonesian-language tweets. A total of 10,082 Indonesian tweets were collected and labeled into two categories: Terindikasi Depresi and Tidak Terindikasi; after deduplication, 3,874 unique tweets were used for modeling. Two scenarios were compared: (1) a fine-tuned IndoBERT model, and (2) fine-tuned IndoBERT CLS embeddings with a linear SVM classifier. The fine-tuned IndoBERT model achieved 73.20% accuracy (AUC-ROC = 0.8212), while the hybrid approach achieved a marginally higher 73.71% accuracy (AUC-ROC = 0.8088); a McNemar's test found this difference not statistically significant ($p = 0.86$). Both models outperformed five traditional TF-IDF-based baselines (best: 70.36%) on the same held-out test set. The hybrid model required only 0.01 MB of storage versus 475.24 MB for the full fine-tuned model. Given statistically equivalent accuracy, combining fine-tuned IndoBERT embeddings with SVM offers substantially lower storage requirements at no measurable cost in classification performance, making it a promising, resource-efficient approach for depression detection on Indonesian social media.

Keywords: Depression Detection, IndoBERT, Indonesian Language, Social Media, Support Vector Machine

INTRODUCTION

Depression is one of the most common mental health disorders worldwide, affecting an estimated 280 million people globally(Liu et al., 2020). In Indonesia, the increasing use of social media has created new opportunities for understanding public mental health conditions through online interactions. Among Indonesia's approximately 273 million citizens, around 167 million actively use social media platforms such as YouTube, Facebook, Instagram, and X (Twitter). These platforms, especially X, are frequently used by individuals to share personal thoughts, emotions, and daily experiences in real time (Segev, 2023). As a result, social media data has become a valuable source for identifying psychological conditions, including indications of depression (Guntuku, Yaden, Kern, Ungar, & Eichstaedt, 2017).

Conventional depression assessment methods, such as the Depression Anxiety Stress Scales (DASS-42), generally require direct clinical evaluation and professional interpretation (Sulak & Koklu, 2024). Although these approaches are widely accepted in mental health assessment, they are often time-consuming, costly, and difficult to apply on a large scale. In recent years, Natural Language Processing (NLP) and machine learning techniques have emerged as promising alternatives for automated depression detection using textual data from social media. Previous studies have shown that linguistic patterns, emotional expressions, and writing behavior in online posts can be utilized to identify depressive tendencies (Tejaswini, Sathya Babu, & Sahoo, 2024).

However, detecting depression from Indonesian-language tweets remains a challenging task. Indonesian social media users commonly employ informal writing styles, abbreviations, slang, mixed regional languages, code-switching, and non-standard spelling. These linguistic characteristics reduce the effectiveness of traditional text classification methods and create additional challenges for transformer-based language models. Several previous studies have attempted to address this issue using deep learning approaches. (Hidayat & Maharani, 2022)

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

reported an accuracy of only 51% using a standard BERT model, while (Devlin, Chang, Lee, & Toutanova, 2019) achieved 52% accuracy using a similar approach. Later, (Balasaranya & Ezhumalai, 2026) improved performance using XLNet-based methods. More recently, (Ridha, Nurjanah, & Rakha, 2024) achieved 82% accuracy using IndoBERTweet on a dataset collected from 184 users.

Despite these improvements, previous studies primarily focused on maximizing classification accuracy without considering model efficiency and deployment feasibility. Transformer-based models such as BERT and IndoBERT generally require large storage capacity and computational resources, making them less practical for lightweight or resource-constrained applications (Saxena & Santhanavijayan, 2026). This issue creates a research gap regarding how to maintain competitive classification performance while significantly reducing model size and computational requirements.

Therefore, this study proposes and evaluates two different approaches for depression detection in Indonesian tweets: a fine-tuned IndoBERT model (S1), and a hybrid model that combines fine-tuned IndoBERT CLS embeddings with a linear Support Vector Machine (SVM) classifier (S2). The main contributions of this study are threefold. First, this research provides a systematic evaluation of IndoBERT fine-tuning using the publicly available DatasetIndikasiDepresi dataset consisting of 10,082 raw tweets. Second, this study demonstrates that the hybrid IndoBERT-SVM model matches the fully fine-tuned model's classification performance (a statistically non-significant difference) while reducing storage requirements by approximately 47,524 times. Third, this research presents a comparative analysis of model performance using accuracy, AUC-ROC, confusion matrices, and storage efficiency to examine the trade-off between effectiveness and computational cost.

LITERATURE REVIEW

The use of Natural Language Processing (NLP) for mental health analysis on social media has grown significantly in recent years. One of the major breakthroughs in NLP was introduced by (Devlin et al., 2019) through Bidirectional Encoder Representations from Transformers (BERT), which established the transformer fine-tuning paradigm that is widely adopted in modern text classification tasks. Compared to traditional machine learning approaches, transformer-based models are capable of capturing contextual semantic information more effectively, leading to substantial improvements in sentiment analysis and psychological text classification. Several studies have applied transformer-based models to depression detection using social media data. (Poświata & Perełkiewicz, 2022) demonstrated that fine-tuned RoBERTa achieved 66.4% accuracy for depression detection tasks on social media datasets. Similarly, (Bokolo & Liu, 2023) reported that transformer-based architectures consistently outperformed conventional machine learning models in mental health text classification tasks due to their stronger contextual representation capabilities. These findings indicate that transformer models have become the dominant approach for automated depression detection.

In the Indonesian NLP domain, several pre-trained language models have been specifically developed to address linguistic characteristics unique to the Indonesian language. (Wilie et al., 2020) introduced IndoBERT and the IndoNLU benchmark, establishing state-of-the-art performance across various Indonesian NLP tasks. Later, (Koto, Lau, & Baldwin, 2021) developed IndoBERTweet, a transformer model pre-trained on Indonesian Twitter data, which achieved up to 90.4% accuracy on sentiment analysis tasks. (Cahyawijaya et al., 2021) further expanded Indonesian NLP resources through IndoNLG for text generation and language understanding tasks. In addition, (Saadah et al., 2022) compared IndoBERT, IndoBERTweet, and CNN-LSTM models for Indonesian COVID-19 opinion classification and found that transformer-based approaches generally provided superior performance.

Several studies have also specifically investigated depression detection in Indonesian social media content. (Darmawan, Joe, Kurniawan, & Afuan, 2023) applied Support Vector Machine (SVM) combined with TF-IDF features for depression sentiment analysis on Twitter data. (Amanat et al., 2022) utilized an LSTM-RNN architecture and achieved 87.3% accuracy in detecting depression-related text. (Rahayu, Fitria, Septhya, Rahmaddeni, & Efrizoni, 2023) implemented Random Forest and TF-IDF features to classify depression and anxiety tweets. Furthermore, (Bendebane et al., 2023) proposed a multi-class deep learning framework using BERT-based models for detecting depression and anxiety on Twitter data. More recently, (Rafi Jonathan Siger, Muhammad Naufal, & Farrikh Alzami, 2026) fine-tuned IndoBERT-base-p1 on a comparably sized Indonesian depression-indication dataset (3,863 tweets) and reported a baseline accuracy of 77.52%, while also showing that class-imbalance handling strategies such as class weighting and focal loss did not necessarily improve overall performance despite better class-level discrimination a finding directly relevant to the class-weighting strategy adopted in the present study. These studies demonstrate that deep learning and transformer-based approaches are increasingly effective for mental health classification tasks.

In addition to end-to-end transformer classification, several researchers have explored the use of transformer embeddings combined with classical machine learning classifiers. (Qasim, Bangyal, Alqarni, & Ali Almazroi, 2022) showed that fine-tuned BERT embeddings used as features for SVM classifiers could achieve competitive or even superior performance compared to fully fine-tuned transformer models, particularly when labeled datasets

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16294
e-ISSN : 2541-2019
p-ISSN : 2541-044X

are limited. (Porwal, Saritha, & Ahirwal, 2024) further demonstrated the effectiveness of a hybrid BERT+SVM framework for depression classification tasks. More recently, (Baydili, Tasci, & Tasci, 2025) combined pre-trained language model embeddings with SVM classifiers for depression and suicidal ideation detection across multiple social media datasets, reporting strong classification performance with improved computational efficiency.

Although previous studies have demonstrated promising results in depression detection using transformer-based models, most research primarily focuses on improving classification accuracy without considering storage efficiency and deployment feasibility. Large transformer models often require substantial computational resources and storage capacity, limiting their applicability in lightweight systems or resource-constrained environments. Furthermore, studies specifically examining the trade-off between model performance and storage efficiency for Indonesian depression detection remain limited. Therefore, this study addresses this gap by systematically comparing a fine-tuned IndoBERT model with a hybrid IndoBERT-SVM approach for Indonesian tweet-based depression detection while explicitly analyzing both classification performance and storage efficiency.

METHOD

This study implements a deep learning-based approach using IndoBERT to detect depression indications from tweets on the X platform. Fig. 1 shows the overall workflow of the proposed research methodology.

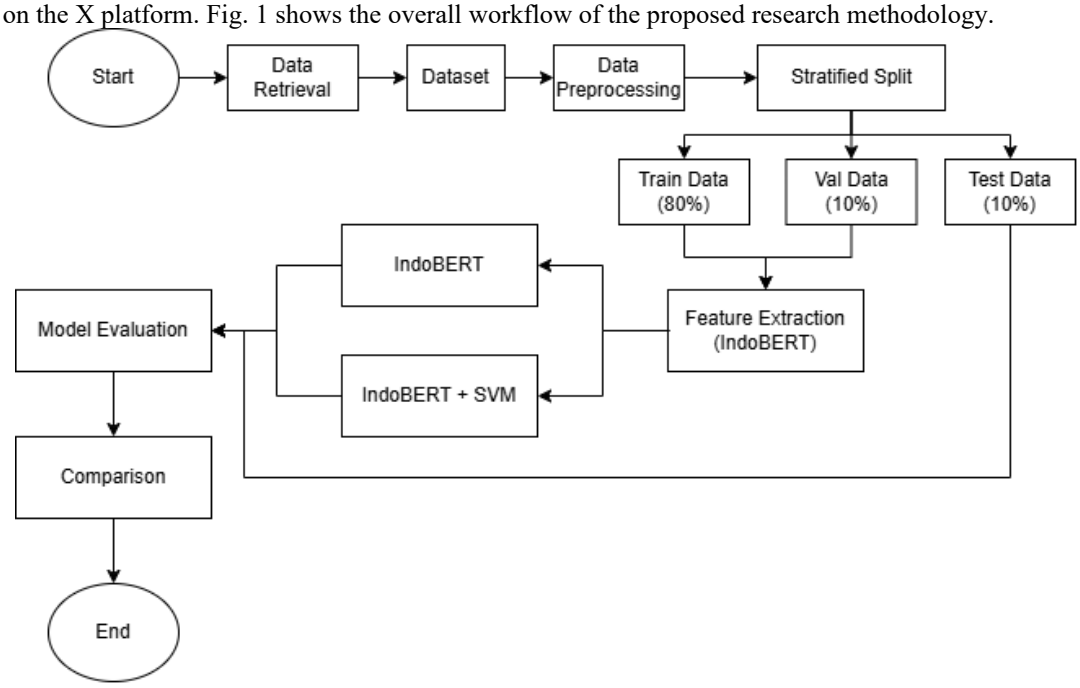

Fig. 1 Flowchart

Dataset

This study uses a publicly available dataset published by (Budiman, 2021) on GitHub (https://github.com/andrebudiman/DatasetIndikasiDepresi). The dataset contains 10,082 Indonesian-language tweets crawled from the Twitter platform via the Twitter API during the period 18 November to 7 December 2020. Data collection was geographically restricted to tweets originating from Java, Bali, and Sumatra, using carefully selected negative-sentiment keywords as crawling filters. The raw data includes the fields: Id_Tweet, Location, Username, Raw_Tweet, and a regex-cleaned version (deEmoji). Crucially, the initial annotation process involved a psychologist, which establishes the labeling credibility of the dataset. Labels were subsequently normalized manually to reduce noise, resulting in two categories: 'Terindikasi Depresi' (Depression Indicated) and 'Tidak Terindikasi' (Not Indicated). After applying the full preprocessing pipeline and removing 6,208 duplicate entries, the final cleaned dataset comprised 3,874 unique tweets used for model training and evaluation.

The raw dataset class distribution comprised 5,617 samples (55.7%) labeled Tidak Terindikasi and 4,465 samples (44.3%) labeled Terindikasi Depresi, indicating a moderate class imbalance that was subsequently addressed via class weighting during training (Table 2).

Word-cloud analysis of the preprocessed text, separated by class label, provides further qualitative support for this separability. For Tidak Terindikasi samples, dominant terms include orang (people), ya (yes), sakit (sick/pain), hidup (life), and tau (know). For Terindikasi Depresi samples, notably more emotionally charged terms appear: cinta (love/longing), sakit (pain), rindu (longing), sedih (sad), hati (heart), and kecewa (disappointed). This

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16294
e-ISSN : 2541-2019
p-ISSN : 2541-044X

qualitative distinction supports the semantic separability of the two classes and the effectiveness of the preprocessing pipeline.

Fig. 2 Word clouds after preprocessing for Tidak Terindikasi (left) and Terindikasi Depresi (right).

Data Preprocessing

The preprocessing stage followed a standard Indonesian NLP pipeline. First, case folding was applied to convert all text into lowercase format. Second, the cleaning process removed URLs, mentions, hashtags, punctuation, numbers, and special characters. Third, tokenization was performed to split each sentence into individual tokens. After tokenization, Indonesian stopwords were removed to eliminate common words with low semantic contribution. Finally, stemming was applied using the Sastrawi library to reduce words into their root forms. This preprocessing pipeline was intended to reduce textual noise and improve semantic consistency across tweets. For illustration, a representative raw tweet ("menyesal bahwa selama ini aku tidak pernah berani melangkah untuk memulai hal yang baru tapi aku sangat bersyukur") was reduced through this pipeline to its stemmed root-word form ("sesal berani lang syukur sama orang dorong berani maju lang").

Table 1 presents the full sequence of these preprocessing stages for this example tweet.

Table 1. Preprocessing Pipeline Example

Stage	Result of Pre-Processing Stages
Raw Tweet	menyesal bahwa selama ini aku tidak pernah berani melangkah untuk memulai hal yang baru tapi aku sangat bersyukur
Case Folding	menyesal bahwa selama ini aku tidak pernah berani melangkah ...
Tokenization	['menyesal','bahwa','selama','ini','aku','tidak','pernah','berani','melangkah', ...]
Stopword Removal	['menyesal','berani','melangkah','bersyukur','dibersamai','orang','mendorong', ...]
Stemming (Sastrawi)	sesal berani lang syukur sama orang dorong berani maju lang

Note on scope: the case-folding/stopword/stemming pipeline above was used for the exploratory text analysis and the word-cloud comparison described above. For the models reported in Table 3 (S1, S2) and Table 4 (traditional baselines), text is tokenized directly from the cleaned-but-unstemmed tweet (after de-duplication and label normalization only), consistent with standard practice for subword-tokenization models such as BERT and for TF-IDF n-gram features, where aggressive manual normalization can remove information the model would otherwise use.

Feature Extraction with IndoBERT

Feature extraction was performed using the IndoBERT Base-P1 model from the IndoNLU framework. IndoBERT was selected because it has been specifically pre-trained on large-scale Indonesian corpora, enabling better contextual representation for Indonesian-language text classification tasks.

Each tweet was tokenized using the IndoBERT tokenizer with a maximum sequence length of 128 tokens, including padding and truncation. The processed input was then passed into the TFBertModel architecture to obtain contextual embeddings. The [CLS] token representation from the final hidden layer, consisting of a 768-dimensional vector, was used as the sentence-level embedding representation for each tweet.

Model Training and Evaluation Protocol

Two classification scenarios were evaluated through a systematic ablation study. Scenario 1 (S1) Fine-tuned IndoBERT: A TFBertForSequenceClassification model was fine-tuned using selective layer unfreezing: the embedding layer and the bottom 8 of 12 encoder blocks were kept frozen, while the top 4 encoder blocks, the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16294
e-ISSN : 2541-2019
p-ISSN : 2541-044X

pooler, and the classification head remained trainable (approximately 23% of total parameters). Fine-tuning used the AdamW optimizer with learning rate 3×10^{-5}, a linear warm-up over the first 10% of training steps, weight decay of 0.01, hidden/attention dropout of 0.2, batch size 16, and early stopping with patience of 3 epochs monitoring validation loss over a maximum of 20 epochs. Class weights (computed on the training split only, using scikit-learn's balanced heuristic) were applied during training to counteract the residual class imbalance. The dataset was split 80:10:10 for train, validation, and test, stratified by class label so that all three partitions preserve the same class proportions; unlike the earlier draft, the validation and test partitions are non-overlapping and the test partition is used only once, for the final evaluation. The full hyperparameter and environment configuration is reported in Table 2.

Table 2. Hyperparameter and Environment Configuration

Parameter	Value
Random seed	42 (NumPy, TensorFlow, Python `random`)
Max sequence length	128 tokens
Batch size	16
Learning rate	3×10^{-5}
Optimizer	AdamW (transformers.create_optimizer), linear warm-up + linear decay
Warm-up ratio	10% of total training steps
Weight decay	0.01
Hidden-layer dropout	0.2
Attention-probs dropout	0.2
Frozen layers	Embedding layer + encoder blocks 0–7
Trainable layers	Encoder blocks 8–11, pooler, classification head (~23% of parameters)
Max epochs / early-stopping patience	20 / 3 (monitor: validation loss)
Class weighting	Balanced (inverse frequency), computed on the training split only
Training-set augmentation	Random single-word deletion applied to 10% of training tweets
SVM kernel / C search	Linear; grid search C ∈ {0.001, 0.01, 0.1, 1, 10, 100}, 3-fold CV
SVM feature scaling	StandardScaler fit on training embeddings only
Framework	TensorFlow + HuggingFace Transformers (indobenchmark/indobert-base-p1)
Train : Val : Test split	80 : 10 : 10, stratified by class, non-overlapping partitions

Note: training was carried out on CPU only, which is reported here for reproducibility; per-epoch wall-clock time was approximately 4–5 minutes on this hardware. During development, an initial training attempt was found to have inadvertently drawn the train/validation/test split from the pre-deduplication data (10,082 rows) rather than the deduplicated 3,874-row dataset, which would have caused substantial train/test text overlap, this was caught before use, and the split was corrected and verified (exact row counts: train=3,099, validation=387, test=388) before the S1/S2 results reported below were produced.

Scenario 2 (S2) Fine-tuned IndoBERT + SVM: After S1 convergence, CLS pooler output vectors (768-dimensional embeddings) were extracted from the frozen fine-tuned model for all samples. An SVM with a linear kernel was then trained on these embeddings. Hyperparameter C was selected via GridSearchCV with 3-fold cross-validation over C ∈ {0.001, 0.01, 0.1, 1, 10, 100}. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC-ROC). Accuracy measures overall prediction correctness, Precision evaluates the proportion of correctly predicted positive samples, Recall measures the model's ability to identify positive cases, and F1-Score represents the harmonic mean of Precision and Recall. In addition, AUC-ROC was used to assess the model's discrimination capability across different classification thresholds. Confusion matrices were also analyzed to visualize classification performance. Furthermore, model file size was recorded to evaluate deployment efficiency and storage requirements.

A confusion matrix is a table used to evaluate the performance of a classification model by comparing prediction results with the actual labels. The matrix consists of four components, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Jude Chukwura Obi, 2023). These components are then used to calculate several evaluation metrics to determine the overall quality of the classification model.

Accuracy measures the proportion of correctly classified data compared to the total number of data instances. The formula for calculating accuracy is as follows (Tharwat, 2021):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16294
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Precision represents the ability of the model to correctly identify positive predictions among all predicted positive instances. A higher precision value indicates fewer false positive predictions generated by the model. The precision formula is defined as (Jude Chukwura Obi, 2023):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall measures the model's ability to correctly identify all actual positive instances. This metric shows how effectively the model captures relevant positive data. The recall formula is expressed as (Jude Chukwura Obi, 2023):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score is a harmonic combination of precision and recall values. This metric is commonly used when there is an imbalance between classes because it provides a balanced evaluation of the model performance. The F1-score formula is given by (Jude Chukwura Obi, 2023):

$$F1\text{-}Score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad (4)$$

RESULT

Table 3 presents the comprehensive classification performance of both experimental scenarios overall accuracy, AUC-ROC, macro/weighted F1, model size, and per-class precision/recall/F1 on the held-out test set consisting of 388 samples. Overall, the hybrid Fine-tuned IndoBERT + SVM model (S2) achieved the best balance between classification performance and computational efficiency, making it the preferred model for deployment-oriented applications.

Scenario 1: Fine-Tuned IndoBERT

The Fine-tuned IndoBERT model (S1) was trained for seven epochs before the early stopping mechanism halted training to prevent overfitting. The best validation performance was achieved at epoch 4, with a validation loss of 0.5502 and validation accuracy of 74.16%. During training, model accuracy increased steadily and reached 90.43% by epoch 7, well above the validation accuracy, indicating the emergence of overfitting after the best epoch consistent with the early-stopping mechanism restoring the epoch-4 weights for final evaluation.

The confusion matrix for S1 is shown in Fig 3. The model correctly classified 127 depression-indicated tweets and 157 non-depression tweets, while producing 57 false positives (non-depression tweets misclassified as depression) and 47 false negatives (depression-indicated tweets missed). These results yielded an overall test accuracy of 73.20% and an AUC-ROC of 0.8212.

From the class-level evaluation, S1 shows fairly balanced recall between the two classes on the corrected split: *Tidak Terindikasi* recall is 0.734 and *Terindikasi Depresi* recall is 0.730, indicating no strong classification bias toward either class a more balanced picture than the previous (pre-correction) run suggested, consistent with the class-weighting strategy (Table 2) applied during training. Even so, the model still misses roughly one in four genuine depression-indicated tweets (47 of 174 in the test set); since a missed case (false negative) is generally more consequential for a screening tool than a false alarm, this residual error rate remains ethically relevant (see Ethical Considerations below and the qualitative patterns identified in Error Analysis).

Scenario 2: Fine-Tuned IndoBERT + SVM

In Scenario 2, CLS embeddings extracted from the fine-tuned IndoBERT model were used as input features for a linear Support Vector Machine (SVM) classifier. The extracted feature matrix had a dimensional shape of (3408, 768). Hyperparameter optimization using GridSearchCV identified C = 0.001 as the optimal parameter, with a 3-fold cross-validation accuracy of 89.79% on the training set (the gap between this cross-validation score and the 73.71% held-out test accuracy reported below is consistent with normal optimism of in-training cross-validation estimates and does not affect the test-set evaluation, which was computed once on data never used for model or hyperparameter selection).

The hybrid IndoBERT + SVM model achieved a test accuracy of 73.71%, a 0.51-percentage-point improvement over the fully fine-tuned IndoBERT model that a (McNemar, 1947) test found not to be statistically significant (see Statistical Significance Testing below). The confusion matrix shown in Fig. 3 indicates that the model correctly classified 126 depression-indicated tweets and 160 non-depression tweets, while generating 54 false positives (non-depression tweets misclassified as depression) and 48 false negatives (depression-indicated tweets missed), for an overall AUC-ROC of 0.8088.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16294
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Compared to S1, S2 shows a small, statistically non-significant shift in the precision-recall balance rather than a clear improvement in one class: recall for depression-indicated tweets is essentially unchanged (0.730 in S1 vs. 0.724 in S2), while recall for non-depression tweets improves modestly (0.734 in S1 vs. 0.748 in S2). Given that the overall accuracy difference between S1 and S2 is not statistically significant (see Statistical Significance Testing below), S2's main practical advantage over S1 is not classification behavior but efficiency: it matches S1's performance while requiring approximately 47,524× less storage (0.01 MB vs. 475.24 MB), making it considerably more practical for large-scale or resource-constrained deployment.

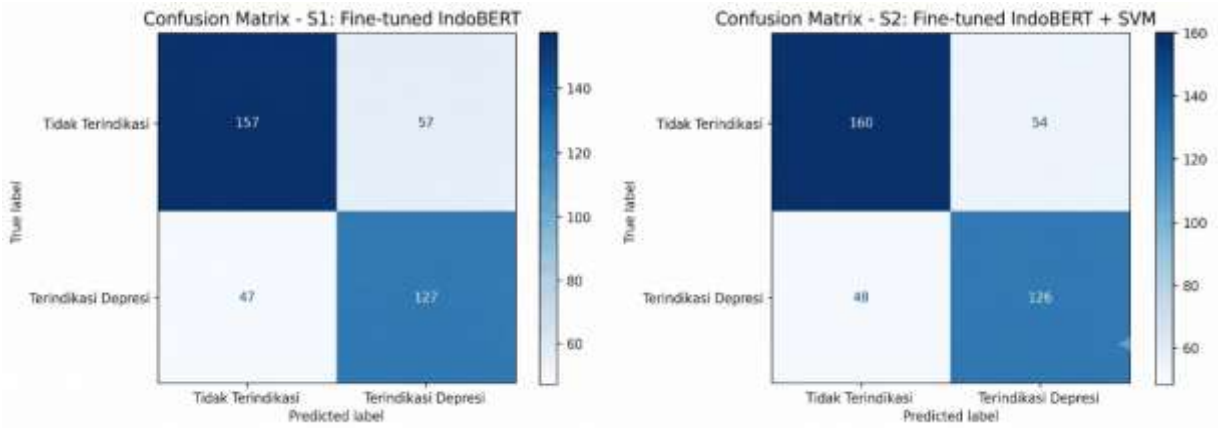

Fig. 3 Comparative Confusion Matrix of Fine-tuned IndoBERT (S1) and Fine-tuned IndoBERT + SVM (S2)

Table 3. Comprehensive Classification Performance Comparison S1 and S2

Scenario	Accuracy	Terindikasi Depresi Precision	Recall	F1	Tidak Terindikasi Precision	Recall	F1
S1: IndoBERT	0.732	0.690	0.730	0.710	0.770	0.734	0.751
S2: IndoBERT + SVM	0.737	0.700	0.724	0.712	0.769	0.748	0.758

Ablation Study: Scenario 1 vs Scenario 2

Table 3 summarises all evaluation metrics on the corrected split. S2 shows a marginally higher accuracy than S1 (+0.51 pp), a difference that is not statistically significant (see Statistical Significance Testing), and a similar macro F1 (0.735 vs 0.730). Critically, the S2 SVM model requires only 0.01 MB versus 475.24 MB for S1, a reduction factor of approximately 47,524×. ROC curves (Fig. 4) show S1 with AUC = 0.8212 and S2 with AUC = 0.8088 on the corrected split, indicating S1 retains marginally better threshold-independent discriminative capability even though S2 achieves a slightly higher accuracy at the default 0.5 threshold the same qualitative pattern observed in the original submission, now confirmed on leak-free data.

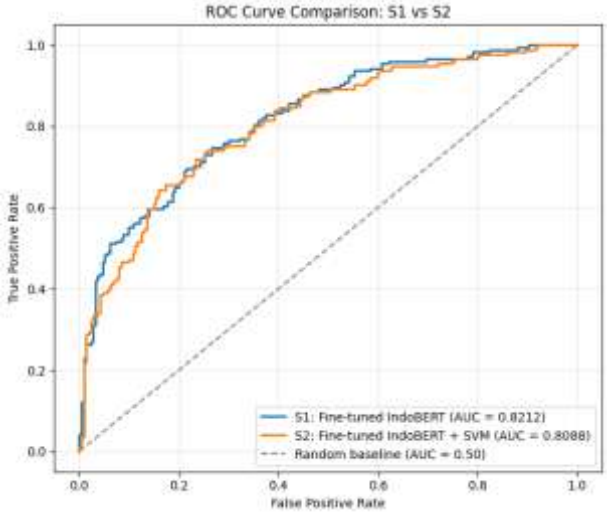

Fig. 4 Comparative ROC Curves Between S1 and S2

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16294
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison with Traditional Baselines

To contextualize the 71–74% accuracy range obtained by S1 and S2, five traditional machine-learning baselines were trained on TF-IDF features (unigrams + bigrams, 10,000 features): Linear SVM, Multinomial Naive Bayes, Random Forest, Decision Tree, and Gradient Boosting. All five baselines were trained and evaluated on the identical 80:10:10 stratified split (test n=388) used for S1 and S2 in Table 3, so all seven models are directly comparable on the same held-out test set.

Table 4. Traditional Baseline Performance

Model	Accuracy	AUC-ROC	Macro F1	Weighted F1
TF-IDF + SVM (linear)	0.7036	0.7828	0.6999	0.7034
TF-IDF + Naive Bayes	0.7036	0.7750	0.6933	0.6991
TF-IDF + Gradient Boosting	0.6753	0.6956	0.6673	0.6726
TF-IDF + Random Forest	0.6649	0.7368	0.6613	0.6649
TF-IDF + Decision Tree	0.5722	0.5697	0.5692	0.5729

The two strongest baselines, Linear SVM and Naive Bayes (both 70.36% accuracy; SVM achieves the higher AUC-ROC of the two, 0.7828 vs. 0.7750), remain 2.84–3.35 percentage points below both fine-tuned IndoBERT models (S1: 73.20%, S2: 73.71%), and no traditional baseline reaches 71%. This indicates that the transformer-based models provide a consistent, measurable advantage over classical TF-IDF baselines. At the same time, the fact that the best traditional baselines sit within a few points of IndoBERT suggests that the overall 71–74% range is still partly bounded by a ceiling set by the dataset itself most plausibly its semi-automated, non-clinical labels and the short, highly informal nature of tweet text rather than purely a limitation of model choice. This provides a data-driven answer to why the observed accuracy does not exceed 80% as in some prior studies that used larger or more strictly curated datasets. This pattern is consistent with (Rafi Jonathan Siger, Muhammad Naufal, & Farrikh Alzami, 2026), who fine-tuned IndoBERT on a comparably sized Indonesian depression dataset (3,863 tweets) and likewise found that imbalance-handling strategies (class weighting, focal loss) traded off overall accuracy against class-level sensitivity rather than improving both simultaneously, reinforcing the view that dataset scale and label quality, not model architecture alone, largely bound the achievable accuracy on datasets of this size.

Statistical Significance Testing

To assess whether accuracy differences between models are statistically meaningful rather than due to chance, (McNemar, 1947) test (chi-square with continuity correction, plus an exact binomial version) and a 10,000-resample paired bootstrap (Efron & Tibshirani, 1998) (5,000–10,000 resamples) both standard, widely-used non-parametric procedures for paired classifier comparison were applied to model pairs evaluated on the same held-out test set. Comparing S1 (73.20%) and S2 (73.71%) on the test set (n=388), the 0.51-percentage-point difference is not statistically significant: (McNemar, 1947) test gives χ^2=0.03 (p=0.855, exact binomial p=0.856), and the paired-bootstrap 95% confidence interval for the difference is [−2.32, +3.35] percentage points (p=0.813). In other words, S1 and S2 perform statistically equivalently in terms of raw accuracy on this test set. S2 nonetheless remains the preferred model in practice, since it matches S1's accuracy at a fraction of the storage cost (0.01 MB vs. 475.24 MB, a ≈47,524× reduction) and, as shown in Table 4, both transformer-based models outperform every traditional TF-IDF baseline evaluated on the same test set.

Error Analysis

To understand the models' failure modes beyond aggregate metrics, false positives (FP: non-depression tweets misclassified as depression-indicated) and false negatives (FN: depression-indicated tweets missed) were manually inspected on S1's predictions. Of 214 true non-depression test tweets, 57 (26.6%) were false positives; of 174 true depression-indicated test tweets, 47 (27.0%) were false negatives (S2: 54/214 and 48/174, respectively a very similar error profile). Representative examples (tweet text only; no usernames or identifiers are retained) include:

False Positive example: "Dihadapan senja aku masih saja memungut reruntuhan air mata yang disebabkan atas kepergianmu…" wistful, melancholic imagery about a past parting, but without content indicating an ongoing depressive state; the model appears to react to the sad tone rather than to clinically relevant content.

False Negative example: "Satu aja sih pesenku kalo gamau kehilangan aku jangan egois. Soalnya aku gak pernah takut buat hidup sendiri." the model appears to be misled by the explicit negation ("gak pernah takut", never afraid), which locally resembles the reassuring, negated phrasing common in non-depression tweets even though the broader context is a message about being left behind. A second example, "…yang berusaha untuk tidak akan pernah menyesali keputusan yg di pilih :)" ("…trying to never regret the decision I made"), shows the same negation pattern recurring independently in this run, strengthening the case that negation handling is a genuine, reproducible weak point rather than a one-off error.

*name of corresponding author

Across the false-negative set, several tweets contained explicit negation of distress (e.g., denying ever wanting to give up); such cases are particularly concerning from a screening-safety perspective, since these are exactly the tweets a screening tool should flag for human follow-up rather than dismiss (see Ethical Considerations). Overall, the pattern suggests that errors are driven more by lexical/negation ambiguity and emotional tone than by the presence or absence of genuinely clinical content a useful direction for future feature or model refinement.

DISCUSSIONS

The experimental results indicate that the hybrid Fine-tuned IndoBERT + SVM model (S2) achieved a slightly higher accuracy (73.71%) than the fully Fine-tuned IndoBERT model (S1) at 73.20%, though this difference is not statistically significant (see Statistical Significance Testing). S1 achieved a marginally higher AUC-ROC (0.8212 vs. 0.8088 for S2; Table 3, Fig. 4), indicating slightly better threshold-independent discrimination; even so, S2 remains the preferred choice overall because it provides significantly better storage efficiency (Porwal et al., 2024). These findings suggest that CLS embeddings generated by fine-tuned IndoBERT contain meaningful semantic representations that can be effectively classified using a lightweight linear SVM model.

The results are consistent with previous studies showing that hybrid transformer and classical machine learning approaches can achieve competitive performance with lower computational complexity. In addition, S2 produced a more balanced precision and recall across both classes, indicating better generalization capability for noisy Indonesian social media text containing slang, abbreviations, and informal sentence structures. Compared with previous Indonesian depression detection studies, the proposed approach remains competitive despite using a larger and more diverse dataset consisting of more than 10,000 tweets.

However, several limitations should be considered. The dataset labels were not independently validated by mental health professionals using standardized clinical instruments such as DASS-42 (Sulak & Koklu, 2024), and data coverage was limited to several Indonesian regions. Unlike an earlier draft of this study, class imbalance is explicitly addressed during training via class weighting (see “Model Training and Evaluation Protocol” and Table 2), so it is no longer listed here as an unaddressed limitation. Furthermore, this study only evaluated IndoBERT Base-P1 and relied solely on textual features without incorporating multimodal information. In addition, a dedicated explainability (XAI) analysis for example, using SHAP, LIME, or attention visualization to identify which words or features most influence individual predictions was not carried out in this study; this is discussed further as a specific future-work direction below. Nevertheless, the findings demonstrate that combining transformer embeddings with classical machine learning classifiers is a promising approach for efficient Indonesian-language depression detection systems.

Ethical Considerations

It is important to emphasize that the models developed in this study are intended as a preliminary, large-scale screening aid for identifying social-media text patterns associated with self-reported depressive expression, and must not be used as a substitute for clinical diagnosis by a qualified mental health professional. Both error types carry real consequences: a false positive may cause unwarranted distress or unwanted intervention for someone who is not experiencing depression, while a false negative as illustrated in the Error Analysis above, including cases involving explicit negation of distress risks overlooking someone who may benefit from support, which is generally the more consequential error for a screening application. Any real-world deployment (e.g., as a triage or early-warning tool) should route flagged content to a human reviewer or licensed professional rather than trigger automated action, and should be accompanied by informed consent and privacy safeguards for the individuals whose text is analyzed.

CONCLUSION

This study compared Fine-tuned IndoBERT (S1) and Fine-tuned IndoBERT + SVM (S2) for Indonesian tweet-based depression detection. Experimental results showed that the two approaches achieved statistically equivalent accuracy (73.20% for S1, 73.71% for S2; (McNemar, 1947) test $p = 0.86$), with S2 offering a marginally more balanced precision-recall profile between classes. In addition, the hybrid model reduced storage requirements from 475.24 MB to only 0.01 MB (≈47,524× smaller) while matching S1’s classification performance, making it the more practical choice for resource-constrained deployment. A comparison against five traditional TF-IDF-based baselines (best: 70.36% accuracy) further indicates that both fine-tuned IndoBERT models outperform classical machine-learning approaches by a clear margin, while the overall 71-74% accuracy range likely still reflects a ceiling set by the dataset's semi-automated labeling and short, informal text, rather than solely a modeling limitation. These findings indicate that combining transformer embeddings with lightweight classical classifiers can provide an effective balance between predictive performance and deployment efficiency for resource-constrained Indonesian NLP applications.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Future Work and Limitations

A further limitation of this study is that no explainability (XAI) analysis was performed to identify which words or features most influence the models' predictions. Building on this, Future Work and Limitations for this study include the following directions: applying explainable-AI methods, such as SHAP, LIME, or attention-based visualization, to interpret and validate model predictions this is identified as a limitation of the current study and is prioritized as the most immediate next step; multimodal depression detection that incorporates images, audio, or user metadata alongside text; continual learning strategies to keep the model current as language use evolves; domain adaptation across different Indonesian regions, dialects, and social media platforms; evaluation of large language models (LLMs) as classifiers or as data-augmentation tools for this task; and a clinically-supervised validation study, conducted with mental health professionals, to assess real-world screening utility and calibrate the model's decision threshold against clinical outcomes.

REFERENCES

Amanat, A., Rizwan, M., Javed, A. R., Abdelhaq, M., Alsaqour, R., Pandya, S., & Uddin, M. (2022). Deep Learning for Depression Detection from Textual Data. *Electronics*, *11*(5), 676. https://doi.org/10.3390/electronics11050676

Balasaranya, K., & Ezhumalai, P. (2026). Text Optimized XLNet Based Sentimental Analysis for Opinion Mining Towards Products From Twitter Data. *International Journal of Computational Intelligence Systems*, *19*(1), 165. https://doi.org/10.1007/s44196-026-01265-4

Baydili, İ., Tasci, B., & Tasci, G. (2025). Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection. *Behavioral Sciences*, *15*(3), 352. https://doi.org/10.3390/bs15030352

Bendebane, L., Laboudi, Z., Saighi, A., Al-Tarawneh, H., Ouannas, A., & Grassi, G. (2023). A Multi-Class Deep Learning Approach for Early Detection of Depressive and Anxiety Disorders Using Twitter Data. *Algorithms*, *16*(12), 543. https://doi.org/10.3390/a16120543

Bokolo, B. G., & Liu, Q. (2023). Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques. *Electronics*, *12*(21), 4396. https://doi.org/10.3390/electronics12214396

Budiman, A. (2021). DatasetIndikasiDepresi. GitHub Repository. https://github.com/andrebudiman/DatasetIndikasiDepresi.

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., … Fung, P. (2021). IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8875–8898. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.699

Darmawan, F., Joe, M., Kurniawan, Y. I., & Afuan, L. (2023). Analisis Sentimen Kemungkinan Depresi dan Kecemasan pada Twitter Menggunakan Support Vector Machine. *Jurnal Eksplora Informatika*, *13*(1), 24–36. https://doi.org/10.30864/eksplora.v13i1.854

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Efron, Bradley., & Tibshirani, Robert. (1998). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

Hidayat, I. R., & Maharani, W. (2022). General Depression Detection Analysis Using IndoBERT Method. *International Journal on Information and Communication Technology (IJoICT)*, *8*(1), 41–51. https://doi.org/10.21108/ijoict.v8i1.634

Jude Chukwura Obi. (2023). A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences*, *8*(1), 308–314. https://doi.org/10.30574/wjaets.2023.8.1.0054

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10660–10668. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.833

Liu, Q., He, H., Yang, J., Feng, X., Zhao, F., & Lyu, J. (2020). Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study. *Journal of Psychiatric Research*, *126*, 134–140. https://doi.org/10.1016/j.jpsychires.2019.08.002

McNemar, Q. (1947). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

*name of corresponding author

Porwal, A., Saritha, S. K., & Ahirwal, M. K. (2024). *A Hybrid Approach for Depression Classification Using BERT and SVM*. https://doi.org/10.1007/978-981-97-3180-0_30

Poświata, R., & Perełkiewicz, M. (2022). Detecting Signs of Depression from Social Media Text using RoBERTa Pre-trained Language Models. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 276–282. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.ltedi-1.40

Qasim, R., Bangyal, W. H., Alqarni, M. A., & Ali Almazroi, A. (2022). A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *Journal of Healthcare Engineering*, *2022*, 1–17. https://doi.org/10.1155/2022/3498123

Rafi Jonathan Siger, Muhammad Naufal, & Farrikh Alzami. (2026). Analisis Performa Class Weight Dan Focal Loss Pada Model Indobert Untuk Klasifikasi Teks Depresi Berbahasa Indonesia. *JURIKOM (Jurnal Riset Komputer)*, *13*(2), 558–568. https://doi.org/10.30865/jurikom.v13i2.9620

Rahayu, K., Fitria, V., Septhya, D., Rahmaddeni, R., & Efrizoni, L. (2023). Klasifikasi Teks untuk Mendeteksi Depresi dan Kecemasan pada Pengguna Twitter Berbasis Machine Learning. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 108–114. https://doi.org/10.57152/malcom.v3i2.780

Ridha, M., Nurjanah, D., & Rakha, M. (2024). Multilabel Classification Abusive Language and Hate Speech on Indonesian Twitter Using Transformer Model: IndoBERTweet & IndoRoBERTa. *2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 48–54. IEEE. https://doi.org/10.1109/ICICYTA64807.2024.10912874

Saadah, S., Kaenova Mahendra Auditama, Ananda Affan Fattahila, Fendi Irfan Amorokhman, Annisa Aditsania, & Aniq Atiqi Rohmawati. (2022). Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(4), 648–655. https://doi.org/10.29207/resti.v6i4.4215

Saxena, A., & Santhanavijayan, A. (2026). A layer-wise survey on internal modifications in BERT and its variants: techniques, applications, and performance trade-offs. *International Journal of Data Science and Analytics*, *22*(1), 1. https://doi.org/10.1007/s41060-025-00986-7

Segev, E. (2023). Sharing Feelings and User Engagement on Twitter: It's All About Me and You. *Social Media + Society*, *9*(2). https://doi.org/10.1177/20563051231183430

Sulak, S. A., & Koklu, N. (2024). Analysis of Depression, Anxiety, Stress Scale (DASS-42) With Methods of Data Mining. *European Journal of Education*, *59*(4). https://doi.org/10.1111/ejed.12778

Tejaswini, V., Sathya Babu, K., & Sahoo, B. (2024). Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(1), 1–20. https://doi.org/10.1145/3569580

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., … Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 843–857. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.aacl-main.85

*name of corresponding author