Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2115~2127
e-ISSN: 2548-6861

An Applied Data Science Approach for Detecting Depression Symptoms in Indonesian Social Media Text Using Transformer Models

Winson [1]*, Puguh Hiskiawan [2]*
Data Science, Faculty of Technology and Design, Bunda Mulia University
s36220016@student.ubm.ac.id [1], phiskiawan@bundamulia.ac.id [2]

Article Info

Article history:

Received 2026-03-07
Revised 2026-04-18
Accepted 2026-04-30

Keyword:

Applied Data Science,
Depression Detection,
IndoBERT,
XLM-RoBERTa,
Natural Language Processing.

ABSTRACT

Depression is a mental health disorder that often remains undetected due to limited access to mental health services and persistent social stigma. Social media platforms provide an alternative source for identifying depressive symptoms through linguistic expressions shared by users in textual posts. This study proposes an applied data science approach for detecting depression symptoms in Indonesian social media text using Transformer-based models. The dataset was constructed by combining the DEPTWEET dataset with social media posts collected through keyword-based scraping guided by PHQ-9 indicators. The proposed framework consists of dataset construction, text preprocessing, Transformer-based modeling, and performance evaluation. Two pre-trained language models, IndoBERT and XLM-RoBERTa, were evaluated under two preprocessing configurations, namely normal preprocessing and light preprocessing. Experimental results show that preprocessing strategies significantly influence classification performance. Light preprocessing consistently improves contextual representation and leads to better results compared with normal preprocessing. XLM-RoBERTa combined with light preprocessing achieves the best overall performance with a test accuracy of 0.77 and an F1-score of 0.77. Additional robustness analysis and pairwise model agreement evaluation further indicate that both models maintain relatively stable predictions when processing noisy social media text. Findings from this study demonstrate the effectiveness of Transformer-based models for multi-class depression detection in Indonesian social media environments. The proposed framework provides insights into how applied data science techniques can support large-scale analysis of mental health signals in online platforms and contribute to the development of data-driven approaches for early detection of depression symptoms.

I. INTRODUCTION

Mental health disorders represent a major public health concern due to their long-term impact on emotional stability, social interaction, and overall quality of life. Depression is widely recognized as one of the most prevalent psychological conditions affecting individuals across different age groups and social backgrounds [1]. Many cases remain undiagnosed or untreated during the early stages, which may lead to worsening psychological conditions and reduced well-being. Limited access to mental health services and persistent social stigma often discourage individuals from seeking professional assistance. Early identification of depression symptoms therefore becomes an important step for improving mental health awareness and enabling preventive intervention [2].

Online social media platforms have become an important medium for communication and personal expression. Users frequently share their experiences, emotions, and daily thoughts through short textual posts that reflect their psychological conditions [3], [4]. Linguistic patterns contained in these posts often reveal emotional signals such as sadness, loneliness, anxiety, or feelings of worthlessness [5]. Large volumes of user-generated content produced in

social media environments create new opportunities for analyzing emotional and behavioral indicators at scale. Computational approaches capable of extracting meaningful information from such textual data can therefore support the identification of potential mental health risks [6].

Advances in Natural Language Processing (NLP) have enabled automated analysis of textual data for various applications including sentiment analysis, emotion recognition, and psychological state classification [7], [8]. Early studies in text classification often relied on conventional machine learning algorithms combined with manually engineered statistical features [9], [10]. Feature-based approaches typically represent text using bag-of-words or term-frequency representations that fail to capture deeper contextual relationships between words. Informal language structures, slang expressions, spelling variations, and noisy text frequently found in social media further complicate the effectiveness of these traditional approaches [11].

Recent developments in deep learning have introduced Transformer-based architectures that significantly improve contextual language representation [12]. Transformer models utilize self-attention mechanisms capable of modeling long-range dependencies within sentences and capturing semantic relationships more effectively. Pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) [13], [14] and Robustly Optimized BERT Approach (RoBERTa) [15], [16] have demonstrated strong performance across numerous NLP tasks including sentiment classification and mental health detection. Multilingual adaptations and language-specific variants of these models provide promising opportunities for analyzing social media text written in Indonesian [17].

Existing research on depression detection using textual data has predominantly focused on English-language datasets and controlled experimental environments. Studies involving Indonesian-language social media remain relatively limited despite the rapid growth of digital communication in the region [18]. Several challenges remain in handling linguistic diversity, informal expressions, and noise commonly found in Indonesian social media text. Comparative investigations that examine the effectiveness of different Transformer architectures for Indonesian depression detection tasks are still scarce, particularly those that analyze the influence of preprocessing strategies and model robustness in real-world social media data [19], [20].

This study proposes an applied data science approach for detecting depression symptoms in Indonesian social media text using Transformer-based models. The research investigates the performance of two pre-trained language models, IndoBERT and XLM-RoBERTa, within a multi-class classification framework representing different levels of depressive symptoms [21]. Experimental analysis evaluates the impact of preprocessing strategies on classification accuracy and predictive stability. Multiple evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC are used for performance assessment [22]. Findings from this study provide empirical insights into the effectiveness of Transformer models for supporting automated early detection of depression indicators in Indonesian social media environments [23].

The present study provides several contributions to the development of computational approaches for mental health analysis in Indonesian social media environments. A structured applied data science framework is introduced for detecting depression symptoms from user-generated textual data collected from social media platforms [24]. Integration of social media scraping guided by PHQ-9 indicators with the publicly available DEPTWEET dataset enables the construction of a dataset that reflects real-world linguistic expressions related to depression in Indonesian online communication [25]. Comparative evaluation between two Transformer-based language models, IndoBERT and XLM-RoBERTa, is conducted within a multi-class classification setting that represents different levels of depressive symptoms [26]. Examination of preprocessing strategies is also performed in order to analyze how different levels of text normalization influence contextual representation and classification performance. Robustness analysis and pairwise comparison using Cohen's Kappa further provide additional insights into model stability and agreement when processing noisy social media text [27]. Findings from this study contribute empirical evidence regarding the effectiveness of Transformer-based models for Indonesian depression detection tasks and demonstrate the potential of applied data science methodologies for supporting early identification of mental health signals in large-scale social media data [28].

II. METHOD

A. Dataset Construction

Social media text provides a rich source of linguistic signals that reflect emotional expressions and psychological conditions shared by users in online environments [29]. Textual posts written in Indonesian were collected from the social media platform X through a keyword-based scraping process, where keywords were derived from indicators in the Patient Health Questionnaire-9 (PHQ-9), a widely used instrument for assessing depression symptoms. Linguistic expressions associated with each PHQ-9 indicator were identified based on common phrases frequently used in Indonesian social media conversations [30]. The DEPTWEET dataset was incorporated to enrich the diversity of depression-related textual samples and improve dataset variability [25], [31]. Data from both sources were integrated through a structured process involving normalization, removal of duplicate entries, filtering of incomplete or irrelevant posts, and alignment of textual formats to ensure consistency prior to model training.

Labeling was performed using a rule-based approach guided by PHQ-9 indicators [34]. Each textual post was assigned to a depression severity category based on predefined keyword patterns corresponding to specific PHQ-9 symptoms. Representative keywords were mapped to each indicator to capture linguistic expressions commonly associated with emotional distress in Indonesian social media communication [35]. This approach enables scalable labeling for large datasets; however, the contextual nature of language may introduce ambiguity, as similar expressions can reflect different emotional intensities depending on usage. PHQ-9 indicators were further used as a conceptual framework for organizing linguistic expressions into multiple levels of depression severity, as summarized in Table 1.

TABLE 1.
PHQ-9-BASED KEYWORD CATEGORIES FOR SOCIAL MEDIA DATA COLLECTION

Depression Level	PHQ-9 Indicator	Example Keywords (Indonesian)
Mild	0	kurang gembira, kehilangan keinginan, tidak ada kegembiraan, tidak ada tujuan, tidak ada motivasi
	1	kesepian, hari buruk, menangis, kecewa, merasa sedih
	2	kurang tidur, terjaga, tidur berlebih, aktif malam hari
Moderate	3	lelah, tidak melakukan apa-apa, malas, lemah
	4	nafsu makan buruk, diet, merasa gemuk
	5	tidak berguna, diabaikan, merasa malu
Severe	6	linglung, tidak fokus, overthinking
	7	cemas, marah, panik, gelisah
	8	bunuh diri, menyakiti diri, lebih baik mati

Keyword-based data collection introduces potential sampling bias, as posts containing explicit depressive expressions are more likely to be retrieved compared with subtle or implicit emotional signals. This limitation may affect dataset representativeness and the generalizability of the model in capturing nuanced depression-related language in social media environments. Class imbalance observed in the combined dataset was addressed using an undersampling strategy applied to the majority class in order to produce a more balanced distribution across depression categories [[32], [33]. This step reduces model bias toward dominant classes and supports more reliable learning across multiple levels of depression severity.

B. Text Pre-processing

Text data collected from social media platforms often contain informal language structures, abbreviations, repeated characters, and various forms of textual noise. Raw social media posts therefore require preprocessing in order to improve data consistency and reduce irrelevant elements before being processed by machine learning models [10], [36]. Preprocessing also helps standardize textual inputs so that linguistic patterns can be more effectively captured during model training.

Several preprocessing operations were applied to the collected dataset. Lowercasing was performed to convert all characters into a consistent format. URLs, user mentions, and hashtags were removed because these elements do not contribute meaningful semantic information for depression detection. Slang normalization was applied to convert informal expressions commonly used in Indonesian social media into their standard forms [37]. Repeated characters were reduced in order to normalize exaggerated expressions frequently found in online conversations. Stopword removal and stemming were also applied to reduce lexical variation and simplify word representations [38].

Two preprocessing configurations were examined in this study in order to analyze their influence on model performance [39]. Normal preprocessing included the full sequence of operations consisting of lowercasing, noise removal, slang normalization, stopword removal, and stemming. Light preprocessing applied only minimal normalization steps such as lowercasing and basic noise removal while preserving most of the original linguistic structure. This configuration was designed to retain contextual information that may carry emotional signals relevant for depression detection in Transformer-based models [40].

C. System Architecture

The proposed system follows an applied data science pipeline designed for detecting depression symptoms from Indonesian social media text. The overall architecture consists of several stages including data collection, text preprocessing, Transformer-based modeling, and prediction generation. Each stage plays an important role in transforming raw textual data into structured representations that can be processed by machine learning models.

Social media posts collected from the dataset construction stage are first processed through the text preprocessing module. This stage removes irrelevant elements and standardizes the textual format so that meaningful linguistic patterns can be preserved. Preprocessed text is then converted into tokenized representations that serve as input for the Transformer-based language models. Contextual embeddings generated by the Transformer encoder capture semantic relationships between words and phrases within each sentence.

Feature representations produced by the Transformer model are subsequently forwarded to a classification layer that predicts the corresponding depression category. The output layer produces a multi-class prediction representing different levels of depression severity derived from the PHQ-9 indicators. Prediction results generated by the model can be further analyzed using evaluation metrics and additional robustness assessments. Figure 1 illustrates the overall architecture of the proposed depression detection framework.

D. Transformer Models

Transformer architectures have become a dominant approach in Natural Language Processing due to their ability to capture contextual relationships within textual data. Self-attention mechanisms enable Transformer models to represent semantic dependencies between words across long textual sequences [41]. Pre-trained language models built on this architecture provide contextual embeddings that significantly improve performance in various text classification tasks, including sentiment analysis and mental health detection [26], [42].

Two Transformer-based language models were employed in this study, namely IndoBERT and XLM-RoBERTa. IndoBERT is a pre-trained language model specifically developed for the Indonesian language and trained on large-scale Indonesian textual corpora. Linguistic representations produced by IndoBERT are therefore well adapted to the vocabulary, grammar, and contextual patterns commonly found in Indonesian text. XLM-RoBERTa is a multilingual Transformer model trained on a large multilingual corpus covering numerous languages. Multilingual training enables the model to capture cross-lingual semantic patterns while still maintaining strong contextual understanding for individual languages [43].

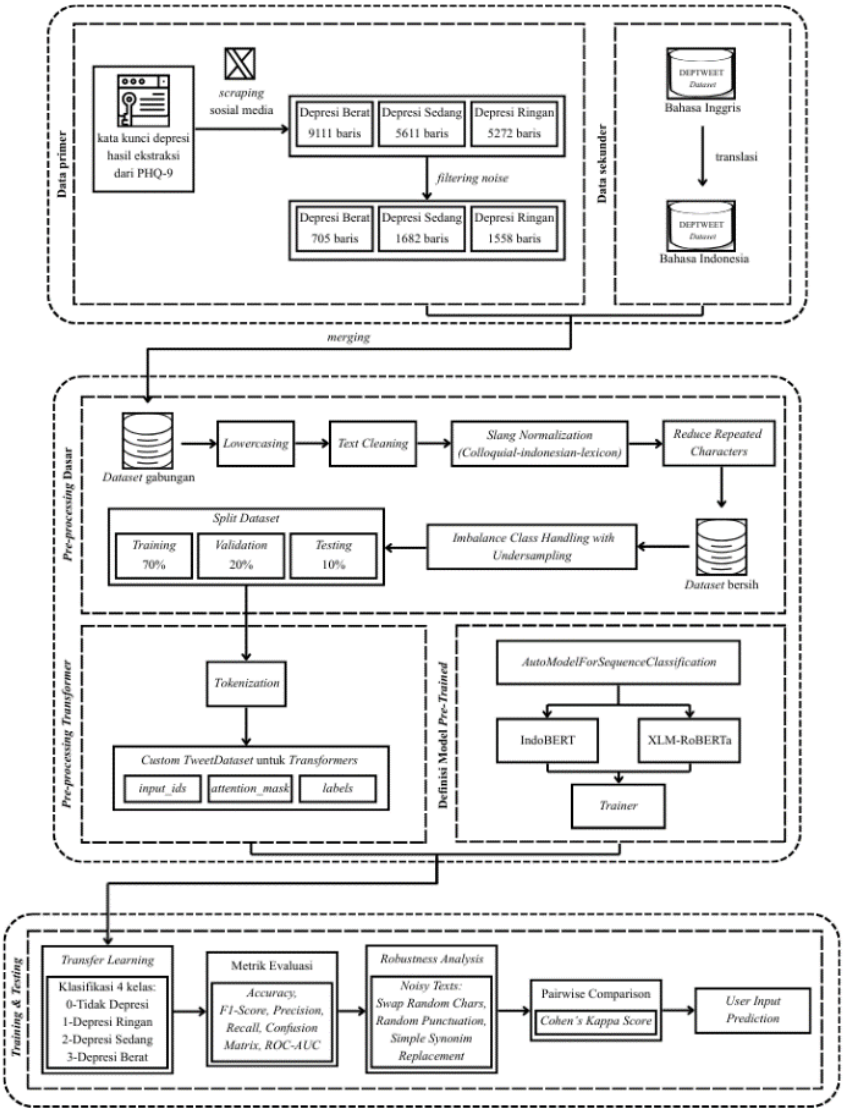

Figure 1. Overall Architecture of the Proposed Depression Detection Framework

Fine-tuning was performed by adapting each pre-trained Transformer model to a multi-class classification task representing different levels of depression severity. Tokenized input text was processed through the Transformer encoder to generate contextual feature embeddings [44], [45]. A classification layer was added on top of the encoder output in order to map these representations into depression categories derived from PHQ-9 indicators. Model parameters were updated during training so that the Transformer representations could be optimized for detecting depression-related linguistic signals in Indonesian social media text [46].

E. Model Training Procedure

Model training was performed through a fine-tuning process using the labeled dataset generated during the dataset construction stage. Pre-trained Transformer models were adapted to the depression classification task by updating model parameters based on the training data [47], [48]. Fine-tuning enables the contextual representations learned during large-scale pretraining to be adjusted for identifying linguistic patterns related to depression symptoms in Indonesian social media text [49]. The dataset was partitioned into training, validation, and testing subsets to ensure reliable model evaluation, where the training set was used for parameter learning, the validation set for model tuning, and the testing set for final performance assessment.

Hyperparameter screening was conducted to determine the optimal training configuration for each model. Several combinations of batch size, learning rate, and weight decay were evaluated during the screening stage [50]. Each configuration was trained on the training dataset and evaluated using the validation dataset in order to identify parameter settings that produced the best classification performance. This approach ensures that model selection is not biased toward the test data and helps prevent overfitting during the training process.

Final model training was conducted using the best hyperparameter configuration obtained from the screening stage. The models were trained for a fixed number of epochs using the training dataset, while performance was continuously monitored using validation metrics. The final evaluation was performed on the unseen test dataset to provide an unbiased estimate of model performance. This training procedure enables the models to learn contextual representations that capture linguistic signals associated with different levels of depression severity in Indonesian social media text [51], [52].

F. Evaluation Metrics

Performance evaluation was conducted to assess the effectiveness of the Transformer-based models in detecting depression symptoms from Indonesian social media text [53]. Several quantitative metrics commonly used in text classification tasks were employed to evaluate the predictive performance of the models. These metrics include accuracy, precision, recall, and F1-score, which provide complementary perspectives on classification capability across multiple classes [54].

Accuracy measures the proportion of correctly predicted instances among the total number of predictions produced by the model [55]. The accuracy score is calculated as:

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

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Precision reflects the proportion of correctly predicted positive instances among all instances predicted as positive. Recall measures the ability of the model to correctly identify relevant instances belonging to a particular class [56]. These metrics are defined as:

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

The F1-score represents the harmonic mean of precision and recall and is commonly used to evaluate classification performance when class distribution is imbalanced.

$$F1 - score = \frac{2 \times precision \times recall}{precision + recal} \quad (3)$$

Additional evaluation was conducted using confusion matrix analysis and Receiver Operating Characteristic Area Under Curve (ROC-AUC). Confusion matrix analysis provides detailed information about classification errors across different depression categories, while ROC-AUC measures the discriminative capability of the model across classification thresholds [57].

G. Robustness Analysis

Robustness analysis was conducted to evaluate the stability of the Transformer-based models when processing noisy or modified textual inputs [10]. Social media text frequently contains spelling variations, punctuation irregularities, abbreviations, and informal linguistic expressions that may influence model predictions. Evaluation of model robustness therefore provides insights into the ability of the models to maintain consistent classification performance when the input text undergoes minor perturbations [58].

Several perturbation strategies were applied to simulate common forms of textual noise observed in social media communication [59]. Character-level modification was introduced through random character swapping within selected words in order to imitate typographical errors frequently produced in informal online writing. Punctuation perturbation was also applied by inserting random punctuation symbols into the text to reflect stylistic variations often present in social media posts [28].

Lexical variation was further simulated through synonym replacement, where selected words were substituted with semantically similar expressions [29]. This modification preserves the overall meaning of the sentence while altering its surface representation. Performance of the models under these perturbation conditions was evaluated using the same classification metrics employed in the primary experiments. Robustness analysis therefore provides additional evidence regarding the reliability of the Transformer models when handling noisy Indonesian social media text [30], [60].

H. Pairwise Model Agreement

Pairwise model agreement analysis was conducted to examine the consistency between predictions produced by different Transformer models when processing the same textual inputs [61], [62]. Agreement analysis provides additional insight beyond standard performance metrics by evaluating how similarly the models classify each instance in the dataset. Consistency between model predictions indicates that both models capture similar linguistic patterns associated with depression-related expressions[63].

Cohen's Kappa coefficient was employed to measure the level of agreement between the predictions generated by IndoBERT and XLM-RoBERTa [26], [64]. This statistical measure accounts for the possibility of agreement occurring by chance, making it more reliable than simple percentage agreement. The coefficient is calculated using the following formulation:

$$\kappa = \frac{P_o - P_e}{1 + P_e} \quad (4)$$

where P_o represents the observed agreement between the two models, while P_e represents the expected agreement that may occur randomly. Higher values of the Kappa coefficient indicate stronger agreement between the classification results produced by the models [65].

TABLE 2.
INTERPRETATION OF COHEN'S KAPPA VALUES

Kappa Value	Interpretation
0.00 – 0.20	None
0.21 – 0.39	Minimal
0.40 – 0.59	Weak
0.60 – 0.79	Moderate
0.80 – 0.90	Strong
> 0.90	Almost Perfect

Interpretation of the Kappa coefficient follows commonly accepted agreement categories used in classification analysis [66]. These categories provide qualitative descriptions of agreement strength ranging from no agreement to almost perfect agreement. Table 2 presents the interpretation ranges used for evaluating the agreement between the two models.

III. RESULTS AND DISCUSSIONS

A. Dataset Distribution

Dataset distribution analysis was conducted to examine the class composition of the collected textual data before and after preprocessing and class balancing procedures. Understanding the distribution of depression categories is important because highly imbalanced datasets may lead to biased model predictions and reduced classification performance. In multi-class classification tasks, imbalance can cause the model to overfit toward dominant classes while underrepresenting minority classes, particularly those associated with moderate and severe depression levels. Visualization of class distribution therefore provides critical insights into how the dataset was constructed and how it may influence model learning behavior.

Figure 2 illustrates the initial class distribution of the DEPTWEET dataset, where the non-depression class dominates the dataset while other categories contain significantly fewer instances. This imbalance reflects real-world conditions in which explicit expressions of severe depression are relatively rare compared with neutral or non-depressive content. However, such distribution poses challenges for machine learning models, as the model may develop a bias toward predicting the majority class, leading to reduced sensitivity in detecting minority classes. Figure 3 presents the class distribution obtained from the keyword-based scraping process after filtering procedures were applied. The integration of scraped data with the DEPTWEET dataset increases the diversity of linguistic expressions, particularly for depression-related categories, although it may still retain certain biases toward explicitly expressed emotional content.

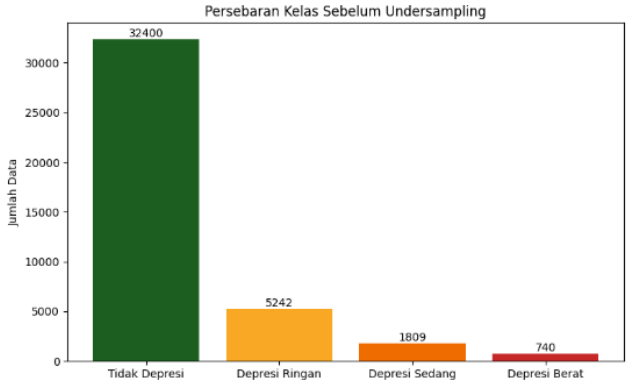

Figure 2. Class Distribution of the DEPTWEET Dataset Before Undersampling

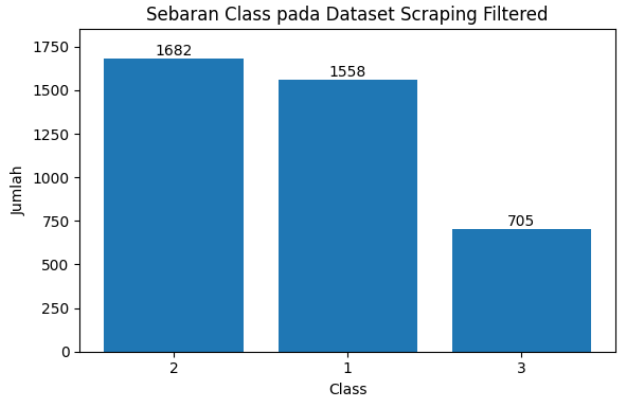

Figure 3. Class Distribution of Scraped Social Media Data

Class balancing was subsequently performed using an undersampling strategy applied to the majority class, as illustrated in Figure 4. This approach reduces the dominance of the majority class and produces a more uniform distribution across depression categories. A more balanced dataset enables the Transformer models to learn decision boundaries more effectively across all classes, particularly for intermediate categories that exhibit higher linguistic ambiguity. However, undersampling may also reduce the overall volume of training data, potentially limiting the richness of contextual patterns learned by the model. Despite this trade-off, the balancing process contributes to improved model fairness and more reliable classification performance across different levels of depression severity.

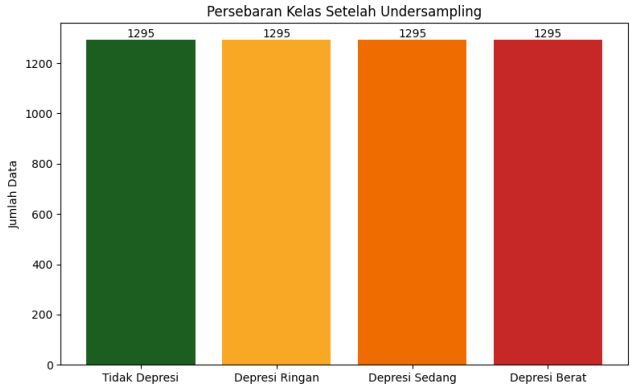

Figure 4. Final Dataset Distribution After Undersampling

B. *Model Training Results*

Model training experiments were conducted to evaluate the performance of IndoBERT and XLM-RoBERTa under different preprocessing configurations. Hyperparameter screening was performed to determine the optimal combination of batch size, learning rate, and weight decay for each model. The screening results are summarized in Table 3 and Table 4 for normal preprocessing and light preprocessing respectively. The results indicate that different Transformer architectures require distinct parameter configurations to achieve optimal performance, reflecting differences in model capacity and training dynamics.

IndoBERT achieved its best performance using a smaller batch size and relatively higher learning rate, suggesting that the model benefits from more frequent parameter updates when learning language-specific representations. In contrast, XLM-RoBERTa performed better with a larger batch size and non-zero weight decay, indicating that regularization plays a more important role in stabilizing training for multilingual representations. Light preprocessing consistently produced higher F1-score and accuracy values compared with normal preprocessing. This improvement can be attributed to the preservation of contextual and expressive linguistic features in social media text, such as informal structure, repetition, and emotionally rich phrases. Transformer-based models rely heavily on contextual embeddings; therefore, excessive normalization through stemming and stopword removal may distort semantic relationships and reduce the model's ability to capture nuanced emotional signals

TABLE 3
HYPERPARAMETER SCREENING RESULTS USING NORMAL PREPROCESSING

Model	Batch Size	Learning Rate	Weight Decay	F1-score	Acc
IndoBERT	8	3	0.00	0.7561	0.7000
XLM-RoBERTa	16	2×10^{-5}	0.01	0.7295	0.7307

TABLE 4
HYPERPARAMETER SCREENING RESULTS USING LIGHT PREPROCESSING

Model	Batch Size	Learning Rate	Weight Decay	F1-score	Acc
IndoBERT	8	2×10^{-5}	0.0	0.7988	0.7982
XLM-RoBERTa	16	3×10^{-5}	0.01	0.7804	0.7809

TABLE 5
FINAL MODEL PERFORMANCE USING NORMAL PREPROCESSING

Model	Test Accuracy	Test F1	Validation Accuracy	Validation F1	F1-Class 0	F1-Class 1	F1-Class 2	F1-Class 3
IndoBERT	0.76	0.76	0.75	0.75	0.75	0.64	0.73	0.88
XLM-RoBERTa	0.70	0.70	0.76	0.76	0.76	0.67	0.74	0.88

TABLE 6
FINAL MODEL PERFORMANCE USING LIGHT PREPROCESSING

Model	Test Accuracy	Test F1	Validation Accuracy	Validation F1	F1-Class 0	F1-Class 1	F1-Class 2	F1-Class 3
IndoBERT	0.76	0.76	0.78	0.78	0.80	0.70	0.74	0.88
XLM-RoBERTa	0.77	0.77	0.79	0.79	0.78	0.71	0.77	0.90

Final model training was subsequently conducted using the optimal hyperparameter configurations obtained from the screening stage. Table 5 and Table 6 present the final performance results for normal preprocessing and light preprocessing respectively. Performance comparison shows that XLM-RoBERTa combined with light preprocessing achieved the best overall results with a test accuracy of 0.77 and an F1-score of 0.77. This result suggests that multilingual pretraining enables the model to capture more diverse semantic patterns, which is advantageous when processing noisy and informal social media text. Class-wise F1-score analysis further indicates that both models perform better in detecting extreme depression categories, particularly Class 0 and Class 3, where linguistic signals tend to be more distinctive. In contrast, intermediate classes remain more challenging due to overlapping semantic characteristics and ambiguity in emotional expression, which makes fine-grained classification more difficult.

C. Classification Performance Analysis

Classification performance analysis was conducted to examine how effectively the Transformer models distinguish between different levels of depression severity in Indonesian social media text. Confusion matrix visualization and ROC-AUC analysis were used to provide detailed insights into prediction patterns and the discriminative capability of the models across different classes. These evaluation approaches enable a deeper understanding of how the models handle class imbalance and semantic ambiguity in multi-class depression classification.

Confusion matrix results for both models using normal preprocessing are presented in Figure 5 and Figure 6. The matrices indicate that both IndoBERT and XLM-RoBERTa perform well in predicting extreme classes, particularly Class 0 and Class 3, which represent non-depression and severe depression categories. Linguistic signals associated with these classes tend to be more distinctive and explicit, allowing the models to identify them more accurately. In contrast, misclassification occurs more frequently between intermediate classes, especially Class 1 and Class 2, where textual expressions often share overlapping emotional characteristics. This ambiguity makes it more difficult for the models to establish clear decision boundaries, leading to increased classification errors in these categories.

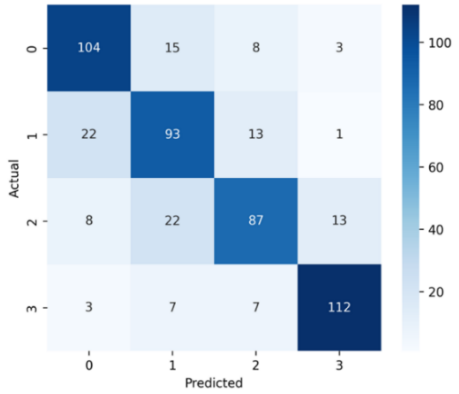

Figure 5. Confusion Matrix of IndoBERT with Normal Preprocessing

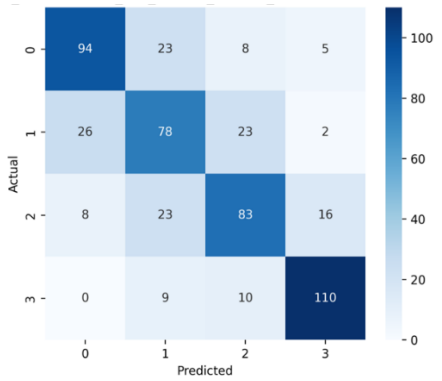

Figure 6. Confusion Matrix of XLM-RoBERTa with Normal Preprocessing

Figures 7 and 8 present the confusion matrices obtained using light preprocessing. Performance improvements can be observed in several classes, particularly in the classification of Class 1 and Class 2. Preservation of contextual linguistic structures during light preprocessing appears to help the Transformer models capture subtle emotional signals embedded in social media text. Informal expressions, repetition, and contextual cues that are retained in light preprocessing contribute to richer semantic representations, which are essential for distinguishing nuanced differences between moderate depression levels. XLM-RoBERTa demonstrates stronger performance in detecting higher depression levels, suggesting that multilingual contextual representations enhance sensitivity to more complex emotional patterns, while IndoBERT shows relatively better performance in identifying lower severity categories due to its stronger adaptation to Indonesian linguistic structures.

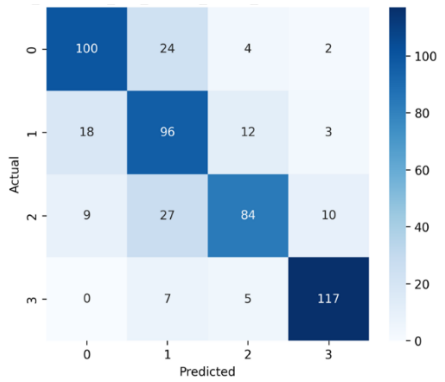

Figure 7. Confusion Matrix of IndoBERT with Light Preprocessing

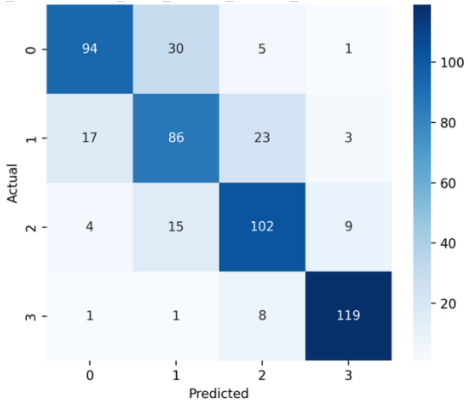

Figure 8. Confusion Matrix of XLM-RoBERTa with Light Preprocessing

Additional evaluation using ROC-AUC analysis further highlights the discriminative capability of the models. Figures 9 and 10 illustrate the ROC-AUC curves for IndoBERT and XLM-RoBERTa under normal preprocessing conditions. Both models achieve high AUC values for extreme classes, particularly the severe depression category, indicating strong capability in distinguishing clearly defined emotional states. However, relatively lower AUC values for intermediate classes suggest that these categories remain more challenging due to semantic overlap and variability in expression.

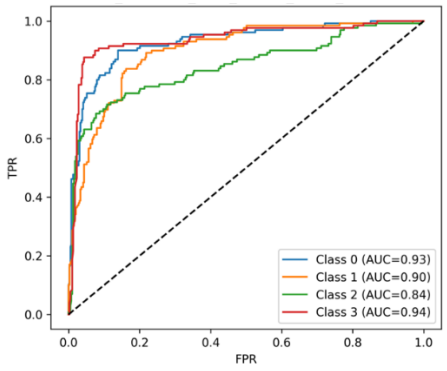

Figure 9. ROC-AUC Curve of IndoBERT with Normal Preprocessing

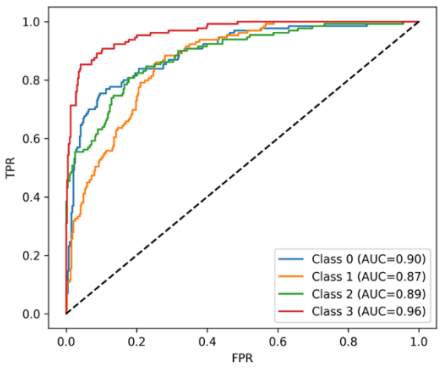

Figure 10. ROC-AUC Curve of XLM-RoBERTa with Normal Preprocessing

Figures 11 and 12 present the ROC-AUC results obtained with light preprocessing. Overall AUC scores increase compared with normal preprocessing, reinforcing the importance of preserving contextual information in Transformer-based classification. XLM-RoBERTa with light preprocessing demonstrates the most stable classification performance across all depression categories, indicating that the multilingual Transformer architecture can effectively capture diverse semantic patterns present in Indonesian social media text, even under conditions of linguistic variability and noise.

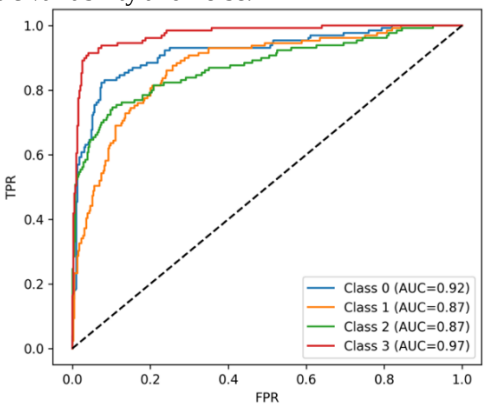

Figure 11. ROC-AUC Curve of IndoBERT with Light Preprocessing

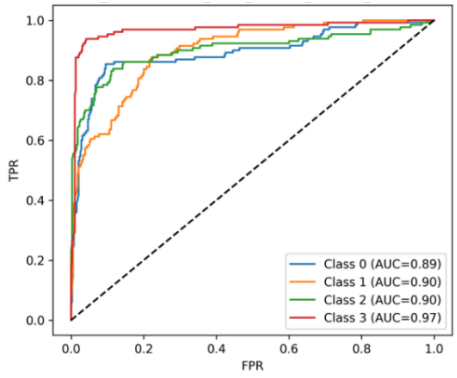

Figure 12. ROC-AUC Curve of XLM-RoBERTa with Light Preprocessing

D. Robustness and Model Agreement Analysis

Robustness evaluation was conducted to examine the stability of the Transformer models when processing noisy or perturbed textual inputs. Social media text frequently contains typographical errors, punctuation variations, and lexical substitutions that may affect model predictions. Robustness analysis therefore provides insights into how well the models maintain classification performance under such perturbation conditions and how resilient their contextual representations are when facing variations commonly found in real-world social media data.

TABLE 7
ROBUSTNESS ANALYSIS RESULTS USING NORMAL PREPROCESSING

Model	Perturbation Type	Noisy Accuracy	Noisy F1-score
IndoBERT	Character swap	0.722	0.722
	Random punctuation	0.733	0.734
	Synonym replacement	0.743	0.743
XLM-RoBERTa	Character swap	0.679	0.678
	Random punctuation	0.710	0.710
	Synonym replacement	0.702	0.702

Table 7 presents the robustness results obtained using normal preprocessing. IndoBERT demonstrates relatively stronger stability across all perturbation types compared with XLM-RoBERTa. Performance degradation is more visible in the XLM-RoBERTa model when character-level noise is introduced, suggesting that multilingual tokenization may be more sensitive to disruptions in word structure. Character swapping alters subword segmentation, which can significantly affect embedding generation in Transformer models.

TABLE 8
ROBUSTNESS ANALYSIS RESULTS USING LIGHT PREPROCESSING

Model	Perturbation Type	Noisy Accuracy	Noisy F1-score
IndoBERT	Character swap	0.749	0.749
	Random punctuation	0.770	0.770
	Synonym replacement	0.766	0.766
XLM-RoBERTa	Character swap	0.766	0.767
	Random punctuation	0.779	0.780
	Synonym replacement	0.760	0.760

Synonym replacement produces the highest noisy performance among the tested perturbations, indicating that semantic preservation within the sentence allows the models to maintain prediction accuracy even when lexical variations occur. This finding highlights that Transformer models are more robust to semantic-level variations than to structural noise that disrupts token-level representations.

Results obtained from light preprocessing are summarized in Table 8. Both models exhibit improved robustness compared with the normal preprocessing configuration. Preservation of contextual linguistic structures in light preprocessing contributes to more stable semantic representations, allowing the models to better tolerate perturbations. XLM-RoBERTa achieves the highest noisy accuracy and F1-score under punctuation perturbation, suggesting that its multilingual training enhances robustness to stylistic variations commonly found in informal text. IndoBERT also maintains consistent performance across perturbation variants, indicating strong adaptation to Indonesian linguistic patterns, although its overall performance remains slightly lower than XLM-RoBERTa in this configuration.

TABLE 9
PAIRWISE MODEL AGREEMENT BASED ON COHEN'S KAPPA

Preprocessing Type	Cohen's Kappa	Interpretation
Normal Preprocessing	0.7219	Moderate
Light Preprocessing	0.7453	Moderate

Pairwise agreement analysis between the two models was further conducted using Cohen's Kappa coefficient. Table 9 presents the agreement scores obtained under both preprocessing settings. The Kappa values indicate moderate agreement between IndoBERT and XLM-RoBERTa predictions, suggesting that while both models capture similar high-level patterns in depression-related language, they differ in handling ambiguous or context-dependent cases. Light preprocessing produces a slightly higher agreement score compared with normal preprocessing, indicating that preserving contextual information leads to more consistent classification behavior across models. This result further supports the importance of maintaining linguistic context in Transformer-based approaches for detecting nuanced emotional signals in social media text.

IV. CONCLUSION

Findings from this study demonstrate that Indonesian social media text can serve as a valuable data source for identifying depression-related linguistic signals through computational analysis. Integration of the DEPTWEET dataset with social media data collected using PHQ-9–based keywords enables the construction of a dataset suitable for multi-class depression detection. Experimental evaluation shows that preprocessing strategies significantly influence model performance. Light preprocessing consistently produces better results compared with normal preprocessing because contextual linguistic structures remain preserved for

Transformer-based models. Performance comparison further indicates that XLM-RoBERTa combined with light preprocessing achieves the best overall classification performance, while IndoBERT shows relatively strong capability in identifying lower severity depression categories.

Several limitations should be acknowledged in this study. Dataset size and class distribution remain constrained by the availability of depression-related expressions in Indonesian social media posts, which may influence model generalization capability. Informal language variations and contextual ambiguity in social media text also create challenges in accurately distinguishing intermediate depression categories. Future research may explore larger and more diverse datasets, investigate advanced architectures or ensemble approaches, and incorporate additional contextual information such as temporal posting patterns or user interaction signals. Practical implementation of the proposed framework may support early-stage monitoring of mental health signals within large-scale social media environments, enabling researchers and mental health professionals to identify potential emotional trends and develop data-driven strategies for preventive mental health interventions.

REFERENCES

[1] T. Wang, "Major Depressive Disorder: A General Overview," *SHS Web of Conferences*, vol. 193, p. 03007, 2024, doi: 10.1051/shsconf/202419303007.

[2] X. Hong, Y. Li, and Z. Xue, "A Review of Studies on Major Depressive Disorder," *Advances in Social Science, Education and Humanities Research*, 2022.

[3] T. Joseph, "Natural Language Processing (NLP) for Sentiment Analysis in Social Media," 2024. [Online]. Available: www.carijournals.orgwww.carijournals.org

[4] L. Lina, A. Chris, R. Ranny, and P. Hiskiawan, "Monitoring Crowd Behavior For Campus Surveillance In Indonesia Using Convolutional Neural Network," *International Journal of Innovative Computing, Information and Control*, vol. 22, no. 1, pp. 95–107, Feb. 2026, doi: 10.24507/ijicic.22.01.95.

[5] C. A. Arango-Dávila and H. G. Rincón-Hoyos, "Depressive Disorder, Anxiety Disorder and Chronic Pain: Multiple Manifestations of a Common Clinical and Pathophysiological Core," Jan. 01, 2018, *Elsevier Doyma*. doi: 10.1016/j.rcp.2016.10.007.

[6] A. Albladi, M. Islam, and C. Seals, "Sentiment Analysis of Twitter Data Using NLP Models: A Comprehensive Review," 2025, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2025.3541494.

[7] Y. Saputri, F. Syaki, and N. Hadinata, "Sentiment Analysis of Trending Topics on Social Media X Using Natural Language Processing and LSTM," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[8] S. M. Padmaja *et al.*, "Depression Detection in Social Media Using NLP and Hybrid Deep Learning Models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, p. 2025, 2025, [Online]. Available: www.ijacsa.thesai.org

[9] E. Wallace, M. Gardner, and S. Singh, "Interpreting Predictions of NLP Models," in *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Tutorial Abstracts*, Association for Computational Linguistics (ACL), 2020, pp. 20–23. doi: 10.18653/v1/P17.

[10] P. Hiskiawan, J. William, L. Feliepe, and T. Jansel, "A Hybrid Data Science Framework for Forecasting Bitcoin Prices using Traditional and AI Models," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2089–2101, 2025.

[11] S. Fathoniah and C. Rozikin, "Analisis Sentimen Masyarakat terhadap Teroris dalam Media Sosial Twitter menggunakan NLP," *Jurnal Ilmiah Wahana Pendidikan*, vol. 2022, no. 13, pp. 412–419, 2022, doi: 10.5281/zenodo.6962682.

[12] N. Alyaa Anindyaputri and A. Suganda Girsang, "A Comparative Study of Deep Learning Models for Detecting Depressive Disorder in Tweets," *Journal of System and Management Sciences*, vol. 14, no. 3, Feb. 2024, doi: 10.33168/jsms.2024.0318.

[13] G. F. Situmorang and R. Purba, "Deteksi Potensi Depresi dari Unggahan Media Sosial X Menggunakan IndoBERT," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, pp. 649–661, Sep. 2024, doi: 10.47065/bits.v6i2.5496.

[14] I. R. Hidayat and W. Maharani, "General Depression Detection Analysis Using IndoBERT Method," *International Journal on Information and Communication Technology (IJoICT)*, vol. 8, no. 1, pp. 41–51, Aug. 2022, doi: 10.21108/ijoict.v8i1.634.

[15] V. Sharma, G. Sikka, and A. K. Sharma, "A Time-Aware Multilingual Multimodal Framework for Depression Detection on Social Media," *Reserch Square*, Nov. 2025, doi: 10.21203/rs.3.rs-8067865/v1.

[16] S. Islam *et al.*, "Ensemble Transformer with Post-hoc Explanations for Depression Emotion and Severity Detection," *iScience*, p. 114605, Feb. 2026, doi: 10.1016/j.isci.2025.114605.

[17] P. Triawan, I. Tahyudin, and P. Purwadi, "Impact of NLP Algorithms on Sentiment Analysis Efficiency and Accuracy," *Journal of Information Systems and Informatics*, vol. 7, no. 3, pp. 2684–2709, Sep. 2025, doi: 10.51519/journalisi.v7i3.1222.

[18] J. Y. M. Nip and B. Berthelier, "Social Media Sentiment Analysis," *Encyclopedia*, vol. 4, no. 4, pp. 1590–1598, Dec. 2024, doi: 10.3390/encyclopedia4040104.

[19] B. A. Mustofa, W. Laksito, and Y. Saptomo, "Use of Natural Language Processing in Social Media Text Analysis," *Journal of Artificial Intelligence and Engineering Applications*, vol. 4, no. 2, pp. 2808–4519, 2025, [Online]. Available: https://ioinformatic.org/

[20] Rakibul Hasan Chowdhury, "Sentiment analysis and social media analytics in brand management: Techniques, trends, and implications," *World Journal of Advanced Research and Reviews*, vol. 23, no. 2, pp. 287–296, Aug. 2024, doi: 10.30574/wjarr.2024.23.2.2369.

[21] N. Hussain *et al.*, "Multi-Level Depression Severity Detection with Deep Transformers and Enhanced Machine Learning Techniques," *AI (Switzerland)*, vol. 6, no. 7, Jul. 2025, doi: 10.3390/ai6070157.

[22] N. Tötsch and D. Hoffmann, "Classifier uncertainty: evidence, potential impact, and probabilistic treatment," *PeerJ Comput. Sci.*, vol. 7, 2021, doi: 10.7717/peerj-cs.398.

[23] A. Tiwari, Y. Gaidhani, G. Katare, K. Mehta, and M. M. Raghuwanshi, "Sentiment Analysis for Social Media Using NLP," 2021. [Online]. Available: www.ijcrt.org

[24] S. U. Rahaman, R. Kokku, and S. Suddala, "Sentiment Analysis Revolution: Using NLP to Uncover Social Media's Hidden Marketing Power," *International Journal of Novel Research and development (IJNRD)*, 2022, [Online]. Available: www.ijnrd.org

[25] M. Kabir *et al.*, "DEPTWEET: A Typology for Social Media Texts to Detect Depression Severities," *arXiv:221005372v1*, Oct. 2022, doi: 10.1016/j.chb.2022.107503.

[26] F. I. Kurniadi, N. L. P. S. P. Paramita, E. F. A. Sihotang, M. S. Anggreainy, and R. Zhang, "BERT and RoBERTa Models for Enhanced Detection of Depression in Social Media Text," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 202–209. doi: 10.1016/j.procs.2024.10.244.

[27] S. Padmalal, I. Edwin Dayanand, G. S. Rao, and S. Gore, "Enhancing Sentiment Analysis in Social Media Texts Using

Transformer-Based NLP Models," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 11, no. 8, pp. 208–216, Aug. 2024, doi: 10.14445/23488379/IJEEE-V11I8P118.

[28] M. S. Alam, M. S. H. Mrida, and M. A. Rahman, "Sentiment Analysis In Social Media: How Data Science Impacts Public Opinion Knowledge Integrates Natural Language Processing (NLP) With Artificial Intelligence (AI)," *American Journal of Scholarly Research and Innovation*, vol. 4, no. 1, pp. 63–100, Jan. 2025, doi: 10.63125/r3sq6p80.

[29] J. Lappeman, R. Clark, J. Evans, L. Sierra-Rubia, and P. Gordon, "Studying social media sentiment using human validated analysis," *MethodsX*, vol. 7, Jan. 2020, doi: 10.1016/j.mex.2020.100867.

[30] M. Rana, R. Khokale, and S. Sall, "Exploring Sentiment Analysis in Social Media: A Natural Language Processing Case Study," *International Journal on Recent and Innovation Trends in Computing and Communication*, p. 12, 2023, [Online]. Available: http://www.ijritcc.org

[31] A. Qasim, G. Mehak, N. Hussain, A. Gelbukh, and G. Sidorov, "Detection of Depression Severity in Social Media Text Using Transformer-Based Models," *Information (Switzerland)*, vol. 16, no. 2, Feb. 2025, doi: 10.3390/info16020114.

[32] S. S. Dhawale, R. Ponnusamy, P. K. Kumaresan, S. Thavareesan, S. Rajiakodi, and B. R. Chakravarthi, "RACHNA: Racial hoax code mixed Hindi–English with novel language augmentation," *Natural Language Processing Journal*, vol. 13, p. 100183, Dec. 2025, doi: 10.1016/j.nlp.2025.100183.

[33] F. Mahardhika, M. L. Haryanti, and P. Hiskiawan, "Performance Evaluation of Speech Emotion Recognition Using Hybrid Feature Selection and Machine Learning," in *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)*, 2025, pp. 1–7. doi: 10.1109/ICCIT65724.2025.11166879.

[34] B. Shelia M., "Implementing the Patient Health Questionnaire-9 (PHQ-9) to Identify and Refer Adults with Depression," *International Journal of Depression and Anxiety*, vol. 6, no. 1, Dec. 2023, doi: 10.23937/2643-4059/1710040.

[35] E. Fonseca-Pedrero, A. Díez-Gómez, A. Pérez-Albéniz, S. Al-Halabí, B. Lucas-Molina, and M. Debbané, "Youth screening depression: Validation of the Patient Health Questionnaire-9 (PHQ-9) in a representative sample of adolescents," *Psychiatry Res.*, vol. 328, Oct. 2023, doi: 10.1016/j.psychres.2023.115486.

[36] G. S. B. Simanullang and J. A. The, "Roles of Natural Language Processing in New Product Development Process: Literature Review," *Jurnal Rekayasa Sistem Industri*, vol. 13, no. 1, pp. 117–130, Apr. 2024, doi: 10.26593/jrsi.v13i1.6790.117-130.

[37] S. Joshi, G. J. Laxmi Priya, U. G. Student, and B. Durga Bhavani, "Social Media Sentiment Analysis using NLP and AI Concepts," *Industrial Engineering Journal*, 2023.

[38] B. Kaldarova, A. Tursynbayev, G. Zhakypbekova, G. Beissenova, L. Zhaidakbayeva, and S. Aldeshov, "Applying artificial intelligence to detect depressive disorders in adolescents via social network generated contents," *Int. J. Health Sci. (Qassim).*, pp. 1706–1724, Aug. 2022, doi: 10.53730/ijhs.v6ns8.12287.

[39] J. You, S. Wang, X. Gong, and X. Wan, "M3L: A Multi-Modal and Multi-Lingual Depression Detection Framework," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2025, pp. 5268–5272. doi: 10.21437/Interspeech.2025-329.

[40] F. Ayu, D. Aryanti, A. Luthfiarta, D. Adiwinata, and I. Soeroso, "Aspect-Based Sentiment Analysis with LDA and IndoBERT Algorithm on Mental Health App: Riliv," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[41] E. Lim, M. Jhon, J. W. Kim, S. H. Kim, S. Kim, and H. J. Yang, "A lightweight approach based on cross-modality for depression detection," *Comput. Biol. Med.*, vol. 186, Mar. 2025, doi: 10.1016/j.compbiomed.2024.109618.

[42] A. E. Putra and W. Maharani, "Depression Levels Detection Through Twitter Tweets Using RoBERTa Method," *Journal of Information System Research (JOSH)*, vol. 3, no. 4, pp. 453–459, Jul. 2022, doi: 10.47065/josh.v3i4.1872.

[43] J. Al Abrar, M. Bin, K. M. R. Chowdhury, and M. A. Bahari, "A Hybrid Transformer-Sequential Model for Depression Detection in Bangla-English Code-Mixed Text," in *Proceedings ofthe Second Workshop on Bangla Language Processing (BLP-2025)*, 2025, p. 2025.

[44] D. William and D. Suhartono, "Text-based Depression Detection on Social Media Posts: A Systematic Literature Review," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 582–589. doi: 10.1016/j.procs.2021.01.043.

[45] A. S. Rizky and E. Y. Hidayat, "Emotion Classification in Indonesian Text Using IndoBERT," *Computer Engineering and Applications*, 2024.

[46] F. Nuraini, M. Najamuddin, D. Miharja, and A. H. Anshor, "Mental Health Chatbot for Detecting Depression Symptoms Using Natural Language Processing and DASS-21," *Jurnal Teknologi Universitas Muhammadiyah Jakarta*, 2025, doi: 10.24853/jurtek.17.2.133-142.

[47] N. Ahmed, A. K. Saha, Md. A. Al Noman, J. R. Jim, M. F. Mridha, and M. M. Kabir, "Deep learning-based natural language processing in human–agent interaction: Applications, advancements and challenges," *Natural Language Processing Journal*, vol. 9, p. 100112, Dec. 2024, doi: 10.1016/j.nlp.2024.100112.

[48] P. Hiskiawan, C. Chih, C. Zheng, and K. Ye, "Processing of electrical resistivity tomography data using convolutional neural network in ERT - NET architectures," *Arabian Journal of Geosciences*, pp. 1–14, 2023, doi: 10.1007/s12517-023-11690-w.

[49] Murat Başal, "Natural Language Processing for Sentiment Analysis in Social Media Marketing," *Economics World*, vol. 12, no. 1, Mar. 2025, doi: 10.17265/2328-7144/2025.01.004.

[50] M. Claesen and B. De Moor, "Hyperparameter Search in Machine Learning," *CoRR*, vol. abs/1502.0, 2015.

[51] M. Iqbal, Hendri Mahmud Nawawi, M. R. Ramadhan Saelan, M. Sony Maulana, Yudhistira, and A. Mustopa, "Optimasi Hyperparameter Multilayer Perceptron Untuk Prediksi Daya Beli Mobil," *Jurnal Manajemen Informatika dan Sistem Informasi*, vol. 6, no. 1, pp. 73–81, 2023, doi: 10.36595/misi.v6i1.739.

[52] N. A. Rahmi, S. Defit, and Okfalisa, "The Use of Hyperparameter Tuning in Model Classification: A Scientific Work Area Identification," *International Journal on Informatics Visualization*, vol. 8, no. 4, pp. 2181–2188, 2024, doi: 10.62527/joiv.8.4.3092.

[53] Z. Zhang, J. Zhu, Z. Guo, Y. Zhang, Z. Li, and B. Hu, "Natural Language Processing for Depression Prediction on Sina Weibo: Method Study and Analysis," *JMIR Ment. Health*, vol. 11, 2024, doi: 10.2196/58259.

[54] Y. Zhou, "Depression Prediction Model based on NLP," *Applied and Computational Engineering*, vol. 109, no. 1, pp. 109–112, Nov. 2024, doi: 10.54254/2755-2721/109/20241284.

[55] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, no. Nips, pp. 4905–4913, 2016.

[56] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2008.05756

[57] M. Studer, G. Ritschard, A. Gabadinho, and N. S. Muller, "Discrepancy Analysis of State Sequences," *Sociol. Methods Res.*, vol. 40, no. 3, pp. 471–510, 2011, doi: 10.1177/0049124111415372.

[58] P. Hiskiawan, E. Stephanie, H. Heryanto, and S. A. Feri, "Trustworthy Data Science Framework for Non-Invasive Nutritional Screening Using Computer Vision," in *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2025, pp. 1744–1748.

[59] M. Ranjan, S. Tiwari, A. M. Sattar, and N. S. Tatkar, "A New Approach for Carrying Out Sentiment Analysis of Social Media Comments Using Natural Language Processing," *Engineering Proceedings*, vol. 59, no. 1, 2023, doi: 10.3390/engproc2023059181.

[60] F. Joanda Kaunang, A. Pramana Thenata, B. Hakim, D. Fernando Nainggolan, P. Hiskiawan, and Ranny, "Sound Engine Based In-Situ Environment Leveraging Neural Network Classification Algorithm," in *2025 IEEE International Conference on Artificial Intelligence for Learning and Optimization (ICoAILO)*, 2025, pp. 352–358. doi: 10.1109/ICoAILO66760.2025.11156048.

[61] and L. F. T. J. P. Hiskiawan, J. William, "A Hybrid Data Science Framework for Forecasting Bitcoin Prices using Traditional and AI Models," *JAIC*, vol. 9, no. 5, pp. 2089–2101, 2025, doi: https://doi.org/10.30871/jaic.v9i5.10631.

[62] P. Hiskiawan, S. A. Yasodhara, and D. Alexander, "Mel-Frequency Cepstral Coefficients and Neural Networks for Indonesian Traditional Music Recognition," in *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2025, pp. 1707–1712.

[63] M. Widmann, "Cohen's kappa: Learn it, use it, judge it," *www.knime.com*, 2024.

[64] I. Ramadhani and W. Maharani, "Predicting Depressive Disorder Based on DASS-42 on Twitter Using XLNet's Pretrained Model Classification Text," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 379–385, Sep. 2022, doi: 10.47065/josyc.v3i4.2157.

[65] H. Mao and Q. Han, "Enhancing TextGCN for depression detection on social media with emotion representation," *Front. Psychol.*, vol. 16, 2025, doi: 10.3389/fpsyg.2025.1612769.

[66] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond Accuracy: Behavioral Testing of NLP Models with Checklist (Extended Abstract)," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021. [Online]. Available: https://github.com/marcotcr/checklist.