Published online on the journal's webpage: **http://journals.telkomuniversity.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 6 No. 1 (2026) 18 - 24 E-ISSN: 2809-8064

Implementation of IndoBERT for Text-Based Emotion Classification on TikTok Application Reviews in Google Play Store

Syahrudin Athyabunnaja^{1*}, Teguh Tamrin^{2}, Harminto Mulyo^{3}

*Informatics Engineering, Nahdlatul Ulama Islamic University of Jepara

^{1*}211240001126@unisnu.ac.id, ^{2}teguh@unisnu.ac.id, ^{3}minto@unisnu.ac.id

Abstract

This study aims to perform emotion classification on user reviews of the TikTok application obtained from the Google Play Store using the IndoBERT model. The emotion categories used in this research are based on Paul Ekman's six basic emotions, namely happiness, anger, sadness, fear, surprise, and disgust. The dataset was collected using a web scraping technique and processed through several preprocessing stages, including text cleaning to remove irrelevant elements, normalization of informal language, duplication removal, and lexicon-based automatic labeling. The labeled data were then divided into training, validation, and testing datasets using a stratified split technique to maintain proportional class distribution and ensure a more representative training process. The IndoBERT model was subsequently fine-tuned to adapt to the linguistic characteristics of TikTok user reviews. To minimize the impact of class imbalance, the evaluation process did not rely solely on accuracy but also employed macro-F1 and weighted-F1 as the main evaluation metrics. The experimental results show that IndoBERT achieved an accuracy of 97.49%, a macro-F1 score of 0.90, and a weighted-F1 score of 0.97. Class-level evaluation indicates very high performance in the happiness class, while minority classes such as surprise and disgust were still identifiable despite having a limited number of samples. Overall, these findings demonstrate that IndoBERT is effective for Indonesian text-based emotion classification on TikTok application reviews and has strong potential to be applied in the development of automated emotion analysis systems across various digital platforms.

Keywords: *emotion analysis, indobert, text classification, tiktok, application reviews.*

1. Introduction

The rapid development of social media has brought significant changes to the way people interact and express themselves in the digital world. Platforms such as TikTok have now become major spaces for users to share opinions, criticisms, and experiences related to the features and services they use. In this context, user reviews on the Google Play Store do not only serve as a medium for evaluating application quality but also reflect the emotional responses of users toward their application usage experience [1].

Analyzing emotions expressed in review texts is crucial for application developers and relevant stakeholders. Compared with sentiment analysis, which commonly classifies opinions into positive, negative, and neutral categories, emotion classification offers a more specific perspective. Emotions can be categorized into several classes such as happiness, anger, sadness, fear, surprise, and disgust, as proposed in Paul Ekman's theory. However, this task is not simple, as it involves natural language complexity, semantic ambiguity, and limited Indonesian language resources [2].

Several previous studies have attempted to develop Indonesian text-based emotion classification methods using various approaches. Some employed traditional machine learning algorithms such as Logistic Regression and Naïve Bayes, as well as lexicon-based approaches which are considered effective but still limited in capturing complex semantic contexts [3]. Other studies utilized deep learning methods such as Convolutional Neural Networks (CNN), which offer better feature extraction capabilities than conventional methods [4].

The advancement of deep learning further introduced transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), which can understand bidirectional sentence context. Several studies have shown that BERT and its variants, including

Received: 21-11-2025 | Revised: 27-01-2026 | Published: 24-02-2026

RoBERTa and DistilBERT, achieve superior performance compared to conventional models in text classification tasks. For the Indonesian language context, IndoBERT was developed as an adaptation of BERT specifically designed to process Indonesian text and has demonstrated strong performance in various NLP tasks such as sentiment analysis and text classification [5].

Despite the growing number of studies conducted, most research still focuses on platforms such as Twitter or general datasets, while studies specifically utilizing user application reviews from the Google Play Store remain limited. In fact, reviews on this platform possess unique characteristics, including being short, informal, and often containing spontaneous emotional expressions. Moreover, TikTok, as a platform with high user engagement, potentially generates diverse and complex emotional expressions within its reviews [6].

As one of the largest social media platforms with high user activity, TikTok user reviews provide a rich source of information. However, many previous studies remain limited to simple sentiment analysis, which only differentiates opinions into positive, negative, and neutral categories, thus failing to fully represent users' emotional conditions. Therefore, an approach capable of identifying more specific emotional types through emotion-based text analysis is required.

Based on this background, this study focuses on implementing the IndoBERT model for emotion classification in TikTok user reviews on the Google Play Store. This study aims to identify dominant emotion categories and evaluate model performance using precision, recall, and F1-score metrics. The findings are expected to contribute to the development of Indonesian-language automatic emotion analysis systems and provide practical benefits for industry stakeholders in understanding user perceptions and experiences more comprehensively [7].

2. Research Methods

This research begins with the collection of TikTok application user reviews from the Google Play Store using a web scraping technique. The dataset obtained consists of 4,414 Indonesian-language user reviews, including review text, rating, upload date, and application version. These reviews serve as the primary source for emotion analysis because they reflect users' subjective opinions and experiences toward the TikTok application.

After the data were collected, a preprocessing stage was carried out to ensure data quality and consistency. This stage included cleaning the text from symbols, irrelevant numbers, spam, and certain emojis; converting all text to lowercase; normalizing informal language and abbreviations; removing duplicate reviews; and selecting reviews with a minimum length of three words to ensure that the data were sufficiently informative for analysis.

The cleaned data were then automatically assigned emotional labels using a lexicon-based approach grounded in Paul Ekman's emotion theory, consisting of six categories: happiness, anger, sadness, fear, surprise, and disgust. Reviews that did not contain emotional representation were removed, so only labeled data were used. The dataset was then divided into training data (1,587 reviews), validation data (198 reviews), and testing data (199 reviews) using a stratified split technique to maintain the proportion of each emotion class, although the distribution remained imbalanced, with happiness as the dominant class.

The IndoBERT model was subsequently fine-tuned using the research dataset with 5 epochs, a batch size of 16, and the AdamW optimizer. Model performance was evaluated using accuracy, precision, recall, F1-score, macro-F1, weighted-F1, and confusion matrix. This evaluation aims to assess the model's ability to accurately recognize and classify user emotions, thereby providing a deeper understanding of user perceptions toward the TikTok application. Overall, the research workflow is illustrated in Figure 1.

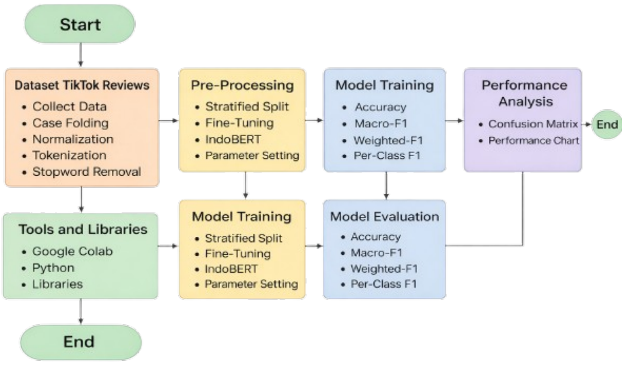

Figure 1. Research methods

2.1. *Data Collection*

The data were collected from the Google Play Store using a web scraping technique. The obtained dataset consists of 4,414 Indonesian-language user reviews, including review text, rating, upload date, and application version. These reviews were selected because they contain users' subjective expressions that can be analyzed for emotion classification.

2.2 *Data Preprocessing*

The preprocessing stage aims to clean and prepare the text data so that it matches the input format required by the IndoBERT model [8]. The preprocessing steps include the following:

2.2.1 *Data Cleaning*

This stage removes non-linguistic elements such as symbols, irrelevant numbers, spam, URLs, hashtags, mentions, and certain emojis. In addition, duplicate reviews were deleted to avoid bias during model training.

2.2.2 *Case Folding*

All text was converted into lowercase letters. For example, the words “Bagus” and “bagus” are treated the same by the model, thereby reducing duplication in the model’s token dictionary.

2.2.3 *Language Normalization*

Slang words, abbreviations, and non-standard language expressions were normalized into their standard forms so that the model could consistently understand the context.

2.2.4 *Review Length Selection*

Only reviews containing a minimum of three words were retained to ensure that the text was sufficiently informative for emotion analysis.

2.2.5 *Emotion Labeling*

Each review was assigned an emotional label based on Paul Ekman’s theory, consisting of six categories: happiness, sadness, anger, fear, surprise, and disgust. Labeling was performed automatically using a lexicon-based approach, and reviews without emotional representation were removed.

2.3 *Dataset Splitting*

The dataset was divided into three subsets using the stratified split method to maintain proportional distribution across emotion classes, namely training data consisting of 1,587 reviews, validation data consisting of 198 reviews, and testing data consisting of 199 reviews. Furthermore, the class distribution shows imbalance, where the happiness class is dominant, while surprise and disgust contain very few samples.

2.4 *Model Training (Fine-Tuning IndoBERT)*

This stage is the core of the research, where the IndoBERT model was trained to recognize linguistic patterns representing emotions in text.

2.4.1 Fine-Tuning IndoBERT

The IndoBERT-base model was used as the pretrained model. A dense layer was added for emotion classification, and the final layer employed a Softmax activation function to generate probability outputs for each emotion class. The model was then fine-tuned using the TikTok review dataset [9].

2.4.2. *Training Parameters*

Training parameters were defined to maintain learning stability and efficiency, including batch size of 16, epochs of 5, optimizer AdamW, and loss function Cross-Entropy Loss.

2.4.3 *Training process*

The model was trained iteratively by calculating the loss value at each epoch. If the validation loss began to increase, the training process was stopped to prevent overfitting.

2.5 *Evaluasi Model*

The performance of the model was evaluated using several commonly used evaluation metrics in text classification:

2.5.1 *Accuracy*

Accuracy represents the percentage of correct predictions compared to the total number of predictions made by the classification model. This metric provides a general overview of how often the model successfully predicts the correct label, although it may be misleading when applied to imbalanced datasets [10].

2.5.2 *Precision*

Precision measures how many of the predicted positive instances are truly relevant to the targeted class, calculated as the proportion between True Positive and the total predicted positive instances (True Positive + False Positive) [11].

2.5.3 *Recall*

Recall measures the model’s ability to correctly identify all actual data belonging to a particular class. It is calculated by dividing the number of True Positive by the total actual positive samples (True Positive + False Negative) [12].

2.5.4 *F1-Score*

F1-score represents the balance between precision and recall using the harmonic mean. This metric is particularly important for imbalanced datasets because it considers both precision and recall simultaneously.

The formulas used are as follows [13]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

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

$$\text{F1-Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3)$$

Description: TP (True Positive) refers to the number of data that are truly positive and correctly predicted as positive by the model, FP (False Positive) refers to the number of data that are actually negative but predicted as positive by the model, FN (False Negative) refers to the number of data that are actually positive but predicted as negative by the model, and TN (True Negative) refers to the number of data that are truly negative and correctly predicted as negative by the model.

These values are obtained from the confusion matrix, which illustrates how well the model recognizes each emotion category.

2.5.5 *Macro-F1 and Weighted-F1*

Macro-F1 and Weighted-F1 are aggregation techniques of F1-score for multi-class classification [14], where Macro-F1 assigns equal weight to all classes regardless of their support, while Weighted-F1 assigns weight based on the number of samples in each class, making it more representative for imbalanced datasets.

2.5.6 *Confusion Matrix*

A confusion matrix is an evaluation table that shows the distribution of predictions against actual labels, enabling the calculation of True Positive, True Negative, False Positive, and False Negative. It also serves as the basis for deriving evaluation metrics such as accuracy, precision, recall, and F1-score [15].

2.6. *Tools dan Pustaka*

This research utilizes several supporting tools and libraries used in data processing, model training, and system performance evaluation.

2.6.1 *Google Colab*

Google Colab was used as a cloud-based computational environment for training the IndoBERT model. This platform was selected because it provides GPU access, supports integration with various NLP libraries, and allows interactive and efficient notebook execution.

2.6.2 *Python*

Python was used as the primary programming language due to its rich ecosystem, flexibility, and wide adoption in machine learning and natural language processing research.

2.6.3 Supporting *Libraries*

Transformers (Hugging Face) were used as supporting libraries, while Pandas and NumPy were used during data processing, text manipulation, and dataset structuring. In addition, Scikit-learn was used for dataset splitting (training, validation, testing) and calculating evaluation metrics such as accuracy, precision, recall, F1-score, Macro-F1, and Weighted-F1. Furthermore, Matplotlib and Seaborn were used for visualization, including confusion matrix plots and model performance graphs across emotion classes.

2.7. *Visualization*

The visualization stage was conducted to present model evaluation results in graphical form, making them easier to analyze comprehensively. The visualizations used in this study include the confusion matrix, emotion distribution graph, and comparison graphs of precision, recall, and F1-score for each class.

The confusion matrix is used to observe the distribution of model predictions against actual labels, enabling identification of the model's success rate in recognizing each emotion class and the classification errors that occur. The emotion distribution graph displays the proportion of data in each class, highlighting dataset imbalance. Meanwhile, performance comparison graphs help illustrate the model's precision, recall, and F1-score across emotion categories, and identify the easiest and most challenging classes to recognize [16].

Through these visualization processes, the evaluation results can be analyzed more thoroughly, providing a clear depiction of IndoBERT's effectiveness in performing emotion classification on TikTok user reviews.

3. Results and Discussion

3.1. *Dataset Description*

The dataset used in this research consists of user reviews of the TikTok application collected from Google Play Store using web scraping techniques. A total of 4,414 Indonesian-language reviews were obtained. However, not all collected data were directly used, as they first underwent a selection and automatic emotion-labeling process. Based on the lexicon-based labeling approach referring to Paul Ekman's six basic emotions, 1,984 reviews were successfully labeled into six emotion categories, namely happiness, anger, sadness, fear, surprise, and disgust. Meanwhile, 2,430 reviews could not be assigned any emotion label due to the absence of clear emotional expressions; therefore, they were excluded from the dataset. The distribution of labeled data is shown in Table 1.

The labeled dataset was further divided using a stratified split technique into three subsets: 80% training data (1,587 reviews), 10% validation data (198 reviews), and 10% testing data (199 reviews). This division aimed to maintain the proportion of each emotion class across subsets and ensure objective model training and evaluation while minimizing the risk of overfitting.

Table 1. Distribution of Labeled Review Data

Emotion	Count
Happiness	1,595
Anger	160
Sadness	121
Fear	77
Surprise	21
Disgust	10
Noun	1,984

The labeled dataset was further divided using a stratified split technique into three subsets: 80% training data (1,587 reviews), 10% validation data (198 reviews), and 10% testing data (199 reviews). This division aimed to maintain the proportion of each emotion class across subsets and ensure objective model training and evaluation while minimizing the risk of overfitting.

3.2. *Pre-Processing Results*

Pre-processing was conducted to ensure that the review texts were clean and compatible with the IndoBERT input format. The stages included:

3.2.1. *Data Cleaning*, removing unnecessary characters such as punctuation, emojis, URLs, mentions, and hashtags.

3.2.2. *Case Folding*, converting all letters to lowercase to standardize word representation.

3.2.3. *Tokenization*, using IndoBERT's built-in tokenizer to match the transformer model structure.

3.2.4. *Stopword Removal*, eliminating common words that do not significantly affect emotional context.

3.2.5. *Emotion Labeling*, based on the six primary emotion categories.

These steps were proven effective in improving text quality and helping the model better understand emotional context, particularly in distinguishing subtle positive and negative expressions in short user reviews commonly found in Google Play Store.

3.3. *IndoBERT Model Training Results*

The IndoBERT-base model was employed as the pretrained model and then fine-tuned using the TikTok review dataset. Training was conducted using key parameters including a batch size of 16, learning rate of 5×10^{-5}, 5 epochs, AdamW optimizer, and Cross-Entropy Loss function.

During the training process, the model demonstrated a stable decrease in loss values across epochs for both training and validation datasets. This indicates that the model successfully adapted to the data without experiencing significant overfitting. Therefore, the applied training configuration can be considered optimal for supporting IndoBERT's performance in emotion classification of TikTok user reviews.

3.4. *Model Performance Evaluation*

Model performance was evaluated using accuracy, precision, recall, F1-score, Macro-F1, and Weighted-F1 metrics. The testing results show that IndoBERT achieved excellent performance with the following values, namely Accuracy = 97,49%, Macro F1 = 0,90, and Weighted F1 = 0,97.

A detailed performance report per class is presented in Table 2.

Table 2. IndoBERT Emotion Classification Performance Report

Emotion Class	Precision	Recall	F1-Score	Support
Happiness	0.9876	0.9938	0.9907	160
Anger	0.8889	1.0000	0.9412	16
Sadness	0.9091	0.8333	0.8696	12
Fear	1.0000	0.8750	0.9333	8
Surprise	1.0000	0.5000	0.6667	2
Disgust	1.0000	1.00001	0.9412	16
Accuracy			0.96	199
Macro Avg	0.9643	0.8670	0.9002	199
Weighted Avg	0.9756	0.9749	0.9739	199

The Macro-F1 score of 0.90 indicates that the model did not rely solely on majority classes, while the Weighted-F1 score of 0.97 reflects strong overall performance despite class imbalance. The confusion matrix presented in Figure 2 further illustrates the prediction distribution of each emotion class.

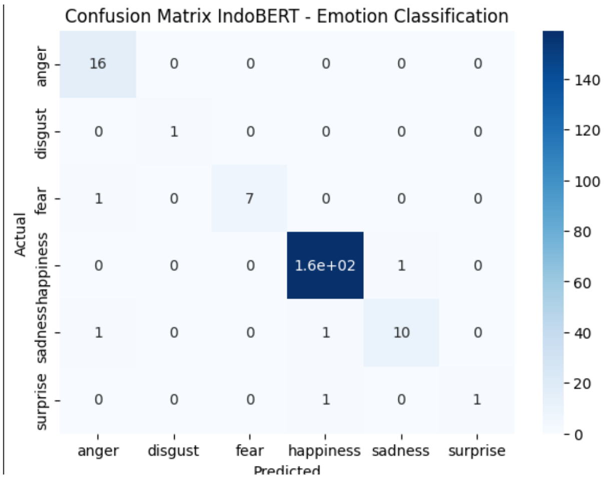

Figure 2. IndoBERT Emotion Classification Confusion Matrix Results

Based on Figure 2, the happiness class demonstrates the most stable predictions with very low misclassification. Meanwhile, the highest uncertainty appears in the surprise and sadness classes, where several samples were misclassified into other emotions.

3.5. *Result Analysis and Visualization*

In addition to the confusion matrix, performance was also visualized through precision, recall, and F1-score comparison charts for each class, as shown in Figure 3.

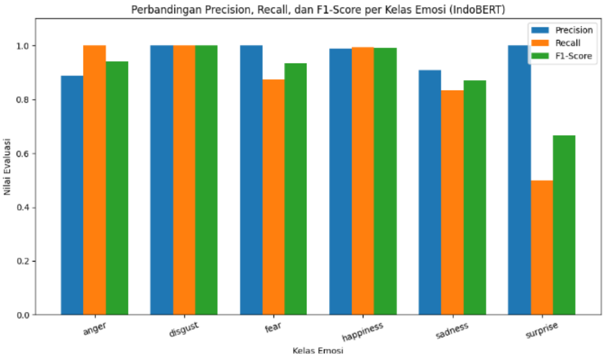

Figure 3. Comparison of Precision, Recall, and F1-Score per Emotion Class (IndoBERT)

The visualization shows that the happiness class achieved the most consistent performance with precision, recall, and F1-score values approaching 1.0. The anger and fear classes also achieved strong performance. However, the surprise class presents the largest gap between precision and recall, indicating that the model still struggles to fully detect samples belonging to this class. Although the disgust class achieved near-perfect performance, this result is not fully representative due to its very limited sample size.

Overall, IndoBERT demonstrated high capability in classifying emotional expressions in TikTok user reviews. Nevertheless, class imbalance, particularly in the surprise and disgust classes, remains a challenge affecting model performance.

4. Conclusion

This study implemented the IndoBERT model to perform text-based emotion classification on user reviews of the TikTok application obtained from Google Play Store. Through a series of stages including data collection, preprocessing, emotion labeling based on Paul Ekman's theory, dataset splitting, as well as model training and evaluation, the results demonstrate that IndoBERT can achieve excellent performance in recognizing emotional expressions. The developed model successfully achieved an accuracy of 97.49%, a Macro F1-score of 0.90, and a Weighted F1-score of 0.97, indicating strong generalization capability on unseen test data.

At the class level, the best performance was obtained in the happiness category, with an F1-score approaching 1.0, reflecting the model's stability in identifying dominant positive emotions within user reviews. The anger and fear classes also demonstrated satisfactory performance. However, performance slightly decreased in the sadness class, while surprise recorded the lowest F1-score. Meanwhile, the disgust class achieved a perfect F1-score; however, this result cannot be considered fully representative due to the extremely limited number of samples in this category.

Overall, the findings confirm that IndoBERT is highly effective for Indonesian-language emotion analysis in social media application reviews. The model is able to produce accurate classification results and has strong potential to be applied in automatic emotion analysis systems to help developers better understand users' emotional responses.

For future research, it is recommended to address class imbalance by increasing the quantity of minority class data, applying data augmentation techniques, or utilizing more optimal class weighting strategies. Furthermore, exploring other transformer-based models such as IndoBERTweet or implementing ensemble-based approaches may further enhance classification performance, particularly for emotion classes with limited data representation.

References

[1] A. S. Agil Rafsanjani, D. L. Fithri, and S. Supriyono, “Sentiment Analysis of User Reviews of the KitaLulus Application on Google Play Store using the Support Vector Machine (SVM) Algorithm,” *Sistemasi*, vol. 14, no. 5, p. 2519, 2025, doi: 10.32520/stmsi.v14i5.5519.

[2] P. Ekman, *Emotions revealed*, vol. 328, no. Suppl S5. 2004. doi: 10.1136/sbmj.0405184.

[3] N. Hilmiaji, K. M. Lhaksmana, and M. D. Purbolaksono, “Identifying Emotion on Indonesian Tweets using Convolutional Neural Networks,” *J. RESTI*, vol. 5, no. 3, pp. 584–593, 2021, doi: 10.29207/resti.v5i3.3137.

[4] G. S. Rasyad and W. Maharani, “Logistic Regression and Naïve Bayes Comparison in Classifying Emotions on Indonesian X Social Media,” *Edumatic J. Pendidik. Inform.*, vol. 9, no. 1, pp. 31–40, 2025, doi: 10.29408/edumatic.v9i1.29120.

[5] F. Basbeth, “Klasifikasi Emosi Pada Data Text Bahasa Indonesia Menggunakan,” vol. 8, no. April, pp. 1160–1170, 2024.

[6] A. N. Azhar, “Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis,” *2023 10th Int. Conf. Adv. Informatics Concept, Theory Appl. ICAICTA 2023*, 2023.

[7] W. Agastya and Aripin, “Pemetaan Emosi Dominan pada Kalimat Majemuk Bahasa Indonesia Menggunakan Multinomial Naïve Bayes (Mapping Dominant Emotion in Indonesian Compound Sentences Using Multinomial Naïve Bayes),” *J. Nas. Tek. Elektro dan Teknol. Inf.* |, vol. 9, no. 2, pp. 171–179, 2020.

[8] R. Setiabudi, N. M. S. Iswari, and A. Rusli, “Enhancing text classification performance by preprocessing misspelled words in Indonesian language,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 19, no. 4, pp. 1234–1241, 2021, doi: 10.12928/TELKOMNIKA.v19i4.20369.

[9] H. F. Karim and A. P. Wibowo, “Kinerja Metode Fine-Tuning IndoBERT untuk Klasifikasi Emosi Multi-Kelas pada Teks Informal Bahasa Indonesia,” vol. 6, no. 1, pp. 63–74, 2025, doi: 10.47065/bulletincsr.v6i1.850.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning.* 2023. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

[11] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[12] P. Fränti and R. Mariescu-istodor, “Pattern Recognition Letters,” vol. 167, pp. 115–121, 2023, doi: 10.1016/j.patrec.2023.02.005.

[13] G. Zeng, “Invariance Properties and Evaluation Metrics Derived from the Confusion Matrix in Multiclass Classification,” *Mathematics*, vol. 13, no. 16, 2025, doi: 10.3390/math13162609.

[14] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, “Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores,” pp. 4961–4972, 2022.

[15] S. Sathyanarayanan and B. R. Tantri, “Confusion Matrix-Based Performance Evaluation Metrics,” vol. 27, no. 4, 2024.

[16] J. Erbani, P.-édouard Portier, E. Egyed-zsigmond, and D. Nurbakova, “Confusion Matrices : A Unified Theory,” vol. 12, no. November, 2024, doi: 10.1109/ACCESS.2024.3507199.

.