

Application of IndoBERT for Granular Emotion Detection in Public Sentiment toward the Free Nutritious Meals Program (MBG)

Alifiya Seftika Putri¹, Fajri Rakhmat Umbara², Ridwan Ilyas³

^{1,2,3}Universitas Jenderal Achmad Yani, Cimahi, Indonesia

Email: alifiyaseftikaputri@gmail.com

Copyright © 2026 Alifiya Seftika Putri et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. The Free Nutritious Meals Program (Program Makan Bergizi Gratis/MBG) is a government policy that has generated diverse public responses on social media. Previous studies have generally focused on sentiment analysis, which may not adequately capture the fine-grained emotional responses of the public, while the use of data augmentation and misclassification analysis remains limited. This study aims to classify public emotions toward the MBG Program using IndoBERT based on Robert Plutchik's emotion theory, which consists of eight emotion categories: anger, happiness, sadness, trust, fear, disgust, interest, and surprise. Data were collected through web scraping from X (Twitter), TikTok, and Facebook and were subsequently annotated by the author and two expert annotators. The preprocessing steps included case folding, data cleaning, and normalization. Back-translation data augmentation was applied to minority classes in the training set to address class imbalance. The experiments compared baseline and augmented data scenarios using several learning-rate configurations. Model performance was evaluated using accuracy, precision, recall, and F1-score, along with misclassification analysis. The results show that Model M4, which employed data augmentation with a learning rate of 1×10^{-5} , a batch size of 16, and 5 epochs, achieved an accuracy of 85.13% and a macro F1-score of 85.39%. Based on the 733 test records, anger was the dominant emotion (31.79%), followed by happiness (16.78%) and trust (15.96%). Anger was the dominant emotion on Facebook and TikTok, whereas trust was dominant on X (Twitter). The misclassification analysis revealed classification errors among several emotions with semantically similar characteristics. These findings demonstrate the applicability of IndoBERT combined with data augmentation for granular emotion classification of public responses to the MBG Program.

Keywords: *IndoBERT, Emotion Detection, Robert Plutchik, Free Nutritious Meals Program, Data Augmentation, Natural Language Processing.*

A. INTRODUCTION

The Free Nutritious Meals Program (Program Makan Bergizi Gratis/MBG) is a government policy aimed at providing free nutritious meals for students at primary and secondary education levels. The program is intended to address nutritional challenges while contributing to the improvement of Indonesia's human resources. Since its implementation, the MBG Program has attracted considerable public attention on social media, generating both support and criticism, particularly regarding logistics, food quality, and distribution equity. In the digital era, social media has become an important medium for the public to express opinions and responses to public policies. Therefore, analyzing public responses expressed through social media is important for understanding public perceptions and the dynamics surrounding the MBG Program (Laia et al., 2025).

Several previous studies have examined public responses to the MBG Program using sentiment analysis. However, most of these studies have focused on three-class sentiment classification, namely positive, negative, and neutral. The Naïve Bayes algorithm with TF-IDF features achieved an accuracy of 89.93%, although class imbalance remained a challenge (Tsabitah et al., 2025). Support Vector Machine (SVM) and BERT were compared using data collected from X (Twitter), with SVM achieving an accuracy of 92% (Munir et al., n.d.). YouTube comments were analyzed using Naïve Bayes, which obtained an accuracy of 84.69%,

although the resulting AUC indicated limitations in distinguishing sentiment patterns (Rahmatullah et al., 2025). Transformer-based approaches have also been investigated. IndoBERT was applied to Instagram comments and achieved an accuracy of 93.95% (Destryanto et al., 2025). IndoBERT also outperformed Bi-LSTM with an accuracy of 80% on X data, while another study reported that IndoBERT outperformed SVM and Random Forest with an accuracy of 87.60% (Saepul Rohman, 2025; Subarkah et al., 2025).

Although these studies demonstrate promising performance, three-class sentiment classification does not provide sufficient detail to represent the specific emotional responses expressed by the public (Mohammad et al., 2017). Emotion classification can provide a more fine-grained representation of public responses by identifying specific emotional categories rather than simply distinguishing between positive, negative, and neutral sentiment. A previous study employed Paul Ekman's basic emotion framework to examine public perceptions of the MBG Program in Papua (Arfan & Badii, 2025). However, the study remained primarily qualitative and did not implement the emotion framework as a multi-class classification scheme at the data level. In terms of language-model performance, IndoBERT was demonstrated to perform better than mBERT for Indonesian-language text (Dwiyansaputra et al., 2025). IndoBERT was also applied to emotion detection and achieved an accuracy of 73%, but the model still encountered misclassification between semantically similar emotions, such as anger and sadness (Saputra et al., 2025). The study also suggested the exploration of data augmentation and more extensive hyperparameter tuning to improve classification performance.

Another challenge concerns the process of emotion labeling and class imbalance. The use of automatic models such as RoBERTa for emotion labeling can introduce potential bias and has been applied to only five emotion categories (Roberta, 2026). Meanwhile, approaches for handling class imbalance, including class weighting, SMOTE, and Bayesian Optimization, have also been explored (Ezer et al., 2026; Sahron Angelina Ihalauw, 2025). However, these approaches still rely on automatic labeling based on the InSet Lexicon, which may have limitations in capturing sarcasm, implicit meanings, and other contextual expressions in social media data.

Based on the findings of previous studies, several research gaps remain. First, research on public responses to the MBG Program predominantly focuses on three-class sentiment classification rather than granular emotion classification. Second, previous studies are often limited to a single social media platform, which may restrict the diversity of linguistic and contextual characteristics represented in the dataset. Third, automatic emotion labeling may introduce potential bias, while expert-based labeling provides an alternative approach for establishing emotion labels based on psychological interpretation. Fourth, previous studies have explored various approaches to class imbalance, but the application of back-translation specifically to minority emotion classes in the training data remains limited. Finally, misclassification between semantically similar emotions remains a challenge that requires further analysis to understand the limitations of the model.

To address these gaps, this study proposes a granular emotion classification framework using IndoBERT for public responses to the MBG Program. The first contribution is the use of cross-platform social media data collected from X (Twitter), TikTok, and Facebook, allowing the study to capture a broader range of informal language and contextual characteristics. The second contribution is the adoption of Robert Plutchik's Wheel of Emotions as the psychological framework for granular emotion classification, consisting of eight emotion categories: anger, happiness, sadness, trust, fear, disgust, interest, and surprise (Alan Tusa Bagus W, 2022). The third contribution is the use of expert labeling through a joint-labeling process involving two professional psychology annotators to support the consistency and

psychological interpretation of the emotion labels. The fourth contribution is the application of back-translation data augmentation using English and Malay as intermediate languages specifically to minority classes in the training set, thereby addressing class imbalance while avoiding data leakage. Finally, this study performs error analysis to examine the model's misclassification patterns, particularly among emotions with similar linguistic characteristics.

Therefore, this study aims to fine-tune IndoBERT for granular emotion classification of public responses to the MBG Program, investigate the effect of data augmentation on model performance, and analyze emotion distributions and classification errors across different social media platforms. Through this approach, the study is expected to provide a more detailed representation of public emotional responses toward the MBG Program and contribute to the development of Indonesian-language Natural Language Processing (NLP), particularly in the area of granular emotion classification.

B. METHOD

This study employed a quantitative approach to classify public emotions toward the Free Nutritious Meals (MBG) Program using the IndoBERT model. The research workflow consists of several stages, including data collection, exploratory data analysis, granular emotion annotation, text preprocessing, dataset splitting, data augmentation, IndoBERT fine-tuning, model evaluation, and emotion distribution and error analysis. The overall research workflow is presented in Figure 1.

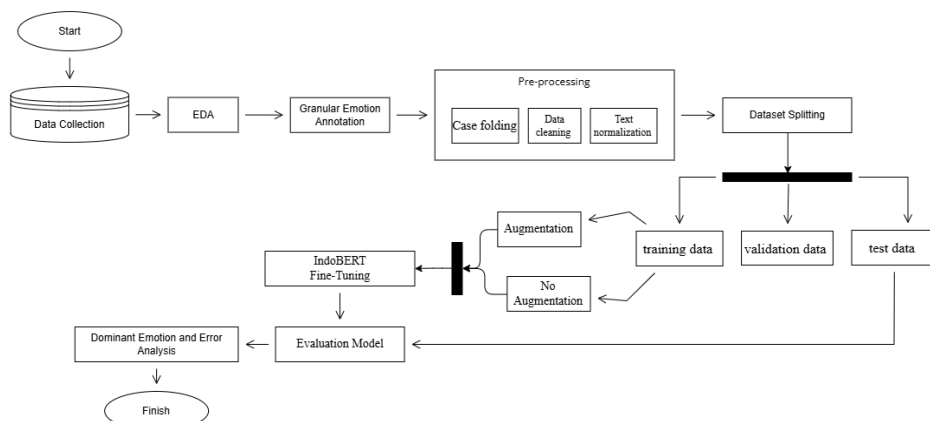


Figure 1. Overview of the Research Method

1. Dataset and Sampling

The dataset was collected through web scraping using Python from X (Twitter), TikTok, and Facebook. Public posts and comments related to the Free Nutritious Meals (MBG) Program were collected using keywords and hashtags, including #MBG, #ProgramMakanGratis, #BadanGiziNasional, and #MakanBergiziGratis. Data were collected from November 1, 2024, to January 17, 2026, resulting in 7,121 raw records: 3,832 from TikTok, 1,918 from X (Twitter), and 1,371 from Facebook.

2. Exploratory Data Analysis

EDA was conducted to assess data quality by identifying missing values, duplicates, non-Indonesian content, and emoji- or symbol-only entries. The analysis identified 1,835 duplicates, 24 records with missing values, 105 non-Indonesian records, and 121 emoji- or symbol-only records. Because some records belonged to multiple categories, the category counts were not summed directly. After removing 1,953 unique records, 5,168 records remained. Manual filtering then removed irrelevant content, resulting in 4,885 final records:

2,510 from TikTok, 1,497 from X (Twitter), and 878 from Facebook. An example of the collected dataset is presented in Table 1.

Table 1. Example of Collected Dataset

id_str	conversati on_id_str	created_at	username	full_text
17573995 11947936 145	17573995 11947936 145	2024-11-11 13:41:09	AirinDatang Lagi	Program Makan Gratis & Pendidikan Gratis itu Hasilnya cuma beda Satu Huruf Doang Klw Makan Gratis Hasilnya jd TAI Klw Pendidikan Gratis Hasilnya jd TAU Jd lu mw pilih mana? Yg Jadi TAI atau yg Jadi TAU..?? Masih ada waktu utk berpikir..!! • #GanjarMahfud2024
17601564 69763420 247	17601564 69763420 247	21/11/2024 04:16:19	manggalaku rnia	Dari sebelum kampanye sampai sekarang dah selesai, tetep gak bakal berubah bahwa ini satu satunya program kampanye yang saya benci banget. Mengatakan di awal program perbaikan gizi anak, kemudian jadi makan siang gratis, yang mana semua anak sekolah dapat nantinya, namun program

3. Granular Emotion Annotation

Each cleaned text was assigned a single dominant emotion based on Robert Plutchik's eight emotion categories: anger, disgust, fear, happiness, sadness, surprise, trust, and interest (Alan Tusa Bagus W, 2022). Annotation was initially performed by the author, followed by a joint-labeling process with two expert annotators from a psychology clinic. Each annotation was reviewed and discussed to determine the most appropriate dominant emotion and reach a final consensus. The resulting human-annotated dataset showed an imbalanced class distribution, with anger as the majority class (31.7%), while surprise and disgust were among the minority classes. These final emotion labels served as the ground-truth labels for model training and evaluation.

4. Text Preprocessing

The annotated dataset was preprocessed to obtain cleaner and more consistent text before model training. The preprocessing steps included case folding, data cleaning, and text normalization. Case folding converted all text to lowercase. Data cleaning removed irrelevant elements, including URLs, user mentions, hashtags, punctuation, special characters, repeated characters, and excessive whitespace. Text normalization standardized non-standard words, abbreviations, and slang commonly used in social media (Khairani et al., 2024). An example of the text preprocessing process is presented in Table 2.

Table 2. Example of Text Preprocessing

Original Text	Case Folding	Data Cleaning	Normalization
Program Makan Gratis & Pendidikan Gratis itu Hasilnya cuma beda Satu Huruf	program makan gratis & pendidikan gratis itu hasilnya cuma beda satu huruf doang klw	program makan gratis pendidikan gratis itu hasilnya cuma beda satu	program makan gratis pendidikan gratis itu hasilnya hanya beda satu

<i>Doang Klw Makan Gratis Hasilnya jd TAI Klw Pendidikan Gratis Hasilnya jd TAU Jd lu mw pilih mana? Yg Jadi TAI atau yg Jadi TAU..?? Masih ada waktu utk berpikir..!! • #GanjarMahfud2024</i>	<i>makan gratis hasilnya jd tai klw pendidikan gratis hasilnya jd tau jd lu mw pilih mana? yg jadi tai atau yg jadi tau..?? masih ada waktu utk berpikir..!! • #ganjarmahfud2024</i>	<i>huruf doang klw makan gratis hasilnya jd tai klw pendidikan gratis hasilnya jd tau jd lu mw pilih mana yg jadi tai atau yg jadi tau masih ada waktu utk berpikir</i>	<i>huruf saja kalau makan gratis hasilnya jadi tahi kalau pendidikan gratis hasilnya jadi tahu jadi kamu mau pilih mana yang jadi tahi atau yang jadi tahu masih ada waktu untuk berpikir</i>
--	--	---	---

5. Dataset Splitting

After preprocessing, the dataset was divided using stratified random sampling into 70% training data (3,419 records), 15% validation data (733 records), and 15% test data (733 records). Stratification was applied to maintain the proportion of each emotion class across the three subsets (Khairani et al., 2024). The training set was used to fine-tune IndoBERT, the validation set was used to monitor training performance and select hyperparameters, and the test set was reserved exclusively for the final model evaluation.

6. Data Augmentation

After dataset splitting, data augmentation was applied to the minority emotion classes in the training set to address class imbalance and increase data diversity (Marchelo et al., 2025; Pornama et al., 2025). Augmentation was performed only on the training set to prevent data leakage. The technique used was back translation, which generates text variations while preserving the original meaning and emotion label (Rahma et al., 2023). Indonesian texts were translated into English and Malay (Malaysia) and then translated back into Indonesian. This process generated variations in sentence structure while maintaining the original semantic meaning and emotion labels. After augmentation and removal of duplicate records, the training set increased from 3,419 to 5,449 records. An example of the back-translation process is presented in Table 3.

Table 3. Example of the Back Translation Process

Original Text (Indonesian)	Translation into English	Back Translation into Indonesian
<i>Program makan gratis pendidikan gratis itu hasilnya hanya beda satu huruf saja kalau makan gratis hasilnya jadi tahi kalau pendidikan gratis hasilnya jadi tahu jadi kamu mau pilih mana yang jadi tahi atau yang jadi tahu masih ada waktu untuk berpikir.</i>	The free meal program and free education program differ by only one letter. If it is free meals, the result is <i>tahi</i> (feces), while if it is free education, the result is <i>tahu</i> (know). So, which one do you want to choose, becoming <i>tahi</i> or becoming <i>tahu</i> ? There is still time to think.	<i>Program makan gratis dan pendidikan gratis hanya berbeda satu huruf. Jika makan gratis, hasilnya menjadi tahi, sedangkan jika pendidikan gratis, hasilnya menjadi tahu. Jadi, kamu mau memilih menjadi tahi atau menjadi tahu? Masih ada waktu untuk berpikir.</i>

7. IndoBERT Fine-Tuning

This study employed IndoBERT (indobenchmark/indobert-base-p1), a pretrained Indonesian language model based on the BERT Transformer encoder architecture, and fine-tuned it for eight-class granular emotion classification. The experiments used a maximum

sequence length of 128, batch size of 16, dropout of 0.1, AdamW optimizer, and five epochs. Three learning rates were evaluated under two conditions, namely without and with data augmentation, resulting in six experimental scenarios. The experimental configurations used for IndoBERT fine-tuning are summarized in Table 4.

Table 4. IndoBERT Fine-Tuning Experimental Scenarios

Model	Augmentation	Learning Rate	Batch Size	Epochs
M1	No	1×10^{-5}	16	5
M2	No	2×10^{-5}	16	5
M3	No	3×10^{-5}	16	5
M4	Yes	1×10^{-5}	16	5
M5	Yes	2×10^{-5}	16	5
M6	Yes	3×10^{-5}	16	5

8. Evaluation

Model performance was evaluated using a confusion matrix and four standard classification metrics: accuracy, precision, recall, and F1-score. The confusion matrix was used to compare the predicted labels with the actual labels and to identify patterns of correct and incorrect classifications. Precision, recall, and F1-score were calculated based on True Positive (TP), False Positive (FP), and False Negative (FN) values for each emotion class (Tsabitah et al., 2025)

Precision measures the proportion of correctly predicted instances among all instances predicted as a particular emotion.

$$Precision = \frac{TP}{TP+FP}$$

Recall measures the proportion of correctly identified instances among all actual instances of a particular emotion.

$$Recall = \frac{TP}{TP+FN}$$

Accuracy measures the proportion of correctly classified instances among all evaluated instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

F1-score provides a balanced measure of precision and recall and is particularly useful for imbalanced class distributions.

$$F1 - Score = 2 \frac{Precision \times Recall}{Precision + Recall}$$

For the eight-class emotion classification task, macro-averaged metrics were used to provide an equal contribution from each emotion class, regardless of the number of instances in each class. Macro precision, macro recall, and macro F1-score were calculated by averaging the corresponding metric values across all eight emotion classes.

Macro Precision is calculated as the arithmetic mean of the precision values across all emotion classes.

$$Macro Precision = \frac{1}{N} \sum_{i=1}^N Precision_i$$

Macro Recall is calculated as the arithmetic mean of the recall values across all emotion classes.

$$Macro Recall = \frac{1}{N} \sum_{i=1}^N Recall_i$$

Macro F1-score is calculated as the arithmetic mean of the F1-score values across all emotion classes

$$Macro\ F1 = \frac{1}{N} \sum_{i=1}^N F1_i$$

9. Dominant Emotion and Error Analysis

After model evaluation, an error analysis was conducted to examine misclassified instances and identify potential causes, such as informal language, slang, ambiguous contexts, and semantic similarities between emotion classes. In addition, the predicted emotion distribution was analyzed across the test dataset and each social media platform (X/Twitter, TikTok, and Facebook) to identify dominant emotions and compare predicted emotion patterns related to the MBG Program across platforms.

C. RESULTS AND DISCUSSION

1. Performance Evaluation of Experimental Scenarios

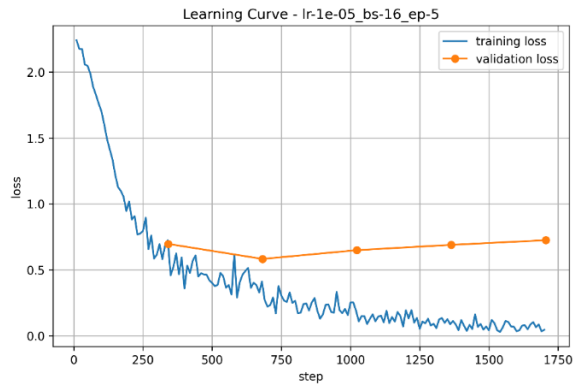
All six experimental scenarios were evaluated using an independent test set of 733 records, with the results presented in Table 5.

Table 5. Performance Comparison of Experimental Scenarios

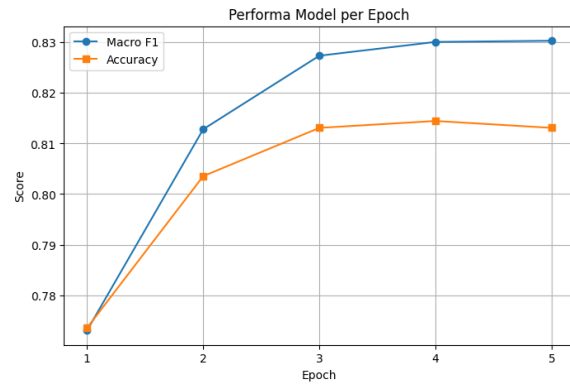
Model	Augmentation	Learning Rate	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
M1	No	1×10^{-5}	84.45%	86.56%	82.86%	84.35%
M2	No	2×10^{-5}	82.40%	85.03%	80.34%	82.34%
M3	No	3×10^{-5}	81.45%	83.19%	80.96%	81.86%
M4	Yes	1×10^{-5}	85.13%	87.14%	84.11%	85.39%
M5	Yes	2×10^{-5}	84.04%	85.33%	82.81%	83.92%
M6	Yes	3×10^{-5}	84.86%	85.77%	84.08%	84.80%

As shown in Table 5, the six experimental scenarios achieved accuracy values ranging from 81.45% to 85.13%. Data augmentation improved the overall classification performance at all three evaluated learning rates compared with their corresponding baseline models. At a learning rate of 1×10^{-5} , for example, accuracy increased from 84.45% (M1) to 85.13% (M4), while the macro F1-score increased from 84.35% to 85.39%. Similar improvements were observed at learning rates of 2×10^{-5} and 3×10^{-5} , suggesting that back-translation data augmentation can improve classification performance across different learning rate configurations.

Among the six experimental configurations, M4 achieved the highest overall performance, with an accuracy of 85.13%, macro-averaged precision of 87.14%, recall of 84.11%, and F1-score of 85.39%. These results indicate that the combination of back-translation data augmentation and a learning rate of 1×10^{-5} produced the best classification performance within the evaluated experimental configurations. Therefore, M4 was selected as the primary model for further analysis, including validation performance tracking and error analysis. Its validation performance across five training epochs is presented in Figure 2.



a. Learning Curve Model M4



b. Accuracy dan Macro F1-Score Model M4

Figure 2. Validation Performance of M4 Across Five Epochs

The validation results indicate rapid early-stage optimization. Evaluation accuracy increased from 77.35% in the first epoch to 80.35% in the second epoch, reaching 81.31% by the third epoch. Correspondingly, the macro F1-score improved from 77.30% to 81.28% and 82.74% over the same initial three epochs. Evaluation loss reached its lowest point (0.5824) at epoch 2 before steadily climbing to 0.7243 by epoch 5. Despite the increasing evaluation loss, evaluation accuracy and macro F1-score plateaued and remained relatively stable (reaching 81.31% and 83.03%, respectively, at epoch 5). This divergence between rising validation loss and stagnant evaluation metrics points toward a mild overfitting trend in the later training epochs. Complete epoch-wise metrics are documented in Table 6.

Table 6. Validation Performance of M4 Across Five Epochs

Epoch	Step	Eval Loss	Eval Accuracy	Eval Macro F1	Eval Macro Precision	Eval Macro Recall
1	341	0.696362	0.773533	0.773033	0.781468	0.773264
2	682	0.582358	0.803547	0.812823	0.817698	0.811975
3	1023	0.648788	0.813097	0.827371	0.834207	0.826217
4	1364	0.689191	0.814461	0.830094	0.851590	0.817731
5	1705	0.724303	0.813097	0.830331	0.843197	0.823757

To evaluate class-level performance, the confusion matrix of model M4 evaluated on the 733 independent test set records is illustrated in Figure 3.

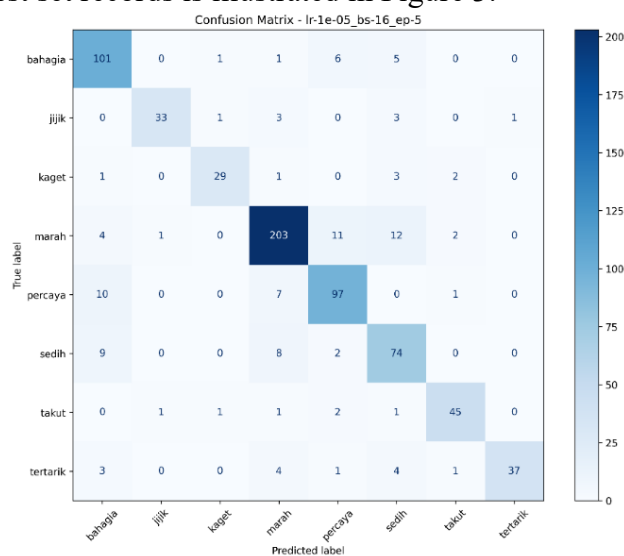


Figure 3. Confusion Matrix of Model M4

The confusion matrix illustrates the classification performance across the eight target emotion categories: happiness, disgust, surprise, anger, trust, sadness, fear, and interest. The diagonal elements represent correctly classified instances, whereas the off-diagonal elements indicate misclassified instances.

The diagonal values show that M4 achieved relatively good classification performance across the major emotion classes, particularly anger with 203 correctly classified instances, followed by happiness with 101 and trust with 97. However, several misclassification patterns were observed between emotion categories with similar semantic or contextual characteristics.

A notable confusion occurred between anger and sadness, with 12 instances of true anger classified as sadness and 8 instances of true sadness classified as anger. This pattern may occur in comments expressing dissatisfaction, disappointment, or negative experiences without explicit indicators of anger. Confusion was also observed between happiness and trust, with 10 instances of true trust classified as happiness and 6 instances of true happiness classified as trust. This overlap may be associated with similar positive expressions, such as appreciation, support, and hope.

In addition, anger acted as a frequent predicted class for instances from several other emotion categories, including sadness, trust, and interest. This pattern may be related to the relatively high representation of anger in the test data. Overall, the confusion patterns indicate that contextual interpretation remains challenging when emotional expressions are implicit, subtle, or contain multiple emotional cues within a single comment.

2. Error Analysis

Error analysis was conducted using the confusion matrix of M4 to identify patterns of misclassification. Several misclassification patterns were observed between emotion classes with similar semantic or contextual characteristics. Confusion between anger and sadness was observed in comments describing dissatisfaction or poor experiences without explicit anger-related expressions. Confusion between trust and happiness occurred in positive comments where support, expectation, appreciation, and gratitude were expressed in similar contexts.

Sarcasm and implicit expressions also contributed to misclassification, particularly when anger was conveyed without explicit emotional words. In addition, removing emoticons during preprocessing may have eliminated important emotional cues in short comments, which could affect the classification of certain emotion classes. Representative examples of misclassified instances produced by M4 are presented in Table 7.

Table 7. Representative Examples of Misclassified Instances by M4

No.	Text	True Emotion	Predicted Emotion	Error Pattern
1	<i>bukti nyata bukan hanya ujaran kebencian yang ditebarkan untuk memecah belah bangsa dengan gagasan dan teori yang belum ada bukti nyata sehat selalu orang baik presiden kabinet merah putih</i>	Trust	Anger	Strong negative lexical cues
2	<i>nah makan bergizi gratis anak aku sekolah padahal sekolah dasar negeri sekolah gratis tapi makan bergizi gratis nya layak tapi anak ku sekolah swasta bayaran per bulan nya tapi masa makan bergizi gratis nya seperti ini</i>	Anger	Sadness	Similar negative context

3	<i>semakin ke sini semakin ke sana sudah seperti jajanan di sekolah saja</i>	Anger	Sadness	Sarcasm and implicit expression
4	<i>bersyukur yang sekolahnya sudah bisa makan gratis daripada kami di pelosok terpencil belum merasakan yang namanya makan gratis anak anak harus rela tahan lapar sampai jam satu siang baru sampai rumah dan bisa makan</i>	Sadness	Happiness	Positive lexical cues vs. negative context
5	<i>jangan di hentikan makan bergizi gratis makan bagi sebagian orang yang berkecukupan tidak butuh makan bergizi gratis tapi bagi orang yang tidak mampu itu sangat berarti sekali buat mereka</i>	Happiness	Trust	Similar positive context
6	<i>bacot sekali anak ibu tidak usah disuruh ambil mana tahu ibu bagaimana rasanya menahan lapar pulang sekolah menunggu sampai malam supaya makan dirapel dengan makan malam dengan adanya makan bergizi gratis banyak lapangan pekerjaan banyak anak anak bahagia</i>	Anger	Happiness	Multiple emotional cues
7	<i>makan bergizi gratis kami</i>	Sadness	Happiness	Loss of emoticon cue during preprocessing

The examples in Table 7 illustrate several factors contributing to misclassification by M4. The first case was labeled as trust because it expressed support and positive expectations, but the model predicted anger, possibly because of strongly negative expressions such as “hate speech” and “dividing the nation.” This indicates that strong lexical cues may influence the prediction even when the overall context conveys a different dominant emotion.

Confusion between anger and sadness was also observed in comments expressing dissatisfaction, disappointment, or unfavorable experiences. In the second and third examples, the texts conveyed criticism without explicit anger-related expressions, causing the model to predict sadness. The third example also demonstrates the difficulty of interpreting sarcasm and implicit emotional expressions. Conversely, in the fourth example, the model predicted happiness because of positive words such as “gratitude” and “free meals,” although the overall context primarily described sadness caused by unequal access to the program.

Confusion among positive emotions was also found. The fifth example was labeled as happiness but predicted as trust because the comment expressed both appreciation and confidence in the benefits of the program. The sixth example demonstrates another challenge, where anger and happiness were expressed within the same comment. The model predicted happiness after encountering the explicit positive expression referring to children being happy, despite the presence of an angry expression at the beginning of the comment.

The seventh example highlights a potential effect of preprocessing. The original data contained an emoticon indicating dissatisfaction, but the emoticon was removed during preprocessing, leaving only a very short textual expression. Without this emotional cue, the model predicted happiness instead of sadness. This case suggests that emoticons may contain

useful emotional information in short social media texts and that their removal can contribute to certain classification errors.

3. Public Emotion Distribution and Platform Comparison

The predictions generated by M4 were further analyzed to examine the distribution of emotions across the test data and social media platforms. Overall, anger was the most frequently predicted emotion, accounting for 31.79% of the test data (233 records), followed by happiness at 16.78% (123 records), trust at 15.96% (117 records), and sadness at 13.64% (100 records). The remaining predictions consisted of fear (6.55%), disgust (5.59%), interest (5.46%), and surprise (4.23%). The dominant predicted emotion for each social media platform is summarized in Table 8.

Table 8. Dominant Emotion by Social Media Platform

Platform	Dominant Emotion	Percentage
Facebook	Anger	54.00%
TikTok	Anger	31.65%
X (Twitter)	Trust	19.81%

The platform-level results show differences in the dominant predicted emotions. Anger was the dominant emotion on Facebook and TikTok, whereas trust was the most frequently predicted emotion on X (Twitter). These results indicate that the distribution of predicted emotional expressions varied across social media platforms.

The results indicate that data augmentation improved the performance of IndoBERT across the tested learning rates. M4, which combined back-translation augmentation with a learning rate of 1×10^{-5} , achieved the highest accuracy (85.13%) and macro F1-score (85.39%). This suggests that increasing the diversity of minority-class training data, combined with an appropriate learning rate, can improve granular emotion classification. The finding is consistent with previous studies reporting the usefulness of data augmentation in addressing class imbalance in text classification (Marchelo et al., 2025; Pornama et al., 2025).

The error analysis shows that misclassification mainly occurred between semantically related emotions, particularly anger and sadness, as well as happiness and trust. Sarcasm, implicit expressions, and multiple emotional cues also presented challenges for the model. These findings are consistent with previous research reporting difficulties in distinguishing closely related emotions in Indonesian text (Saputra et al., 2025).

In addition, the removal of emoticons during preprocessing may have eliminated useful emotional cues in short texts. Overall, the results demonstrate that IndoBERT with back-translation augmentation can effectively support granular emotion classification, although contextual and implicit emotional expressions remain challenging.

D. CONCLUSION

This study applied IndoBERT for granular emotion classification of public responses to the Free Nutritious Meals (MBG) Program using eight emotion categories based on Robert Plutchik's framework: anger, happiness, sadness, trust, fear, interest, disgust, and surprise. The experimental results showed that the M4 configuration, using back-translation data augmentation, a learning rate of 1×10^{-5} , a batch size of 16, and five training epochs, achieved the best performance, with an accuracy of 85.13% and a macro F1-score of 85.39%.

The comparison between baseline and augmented scenarios indicated that back-translation contributed to improved classification performance, particularly at a learning rate of 1×10^{-5} . The analysis of predicted emotions on the test set showed that anger was the dominant emotion, followed by happiness, trust, and sadness. At the platform level, anger was the dominant predicted emotion on Facebook and TikTok, while trust was dominant on X

(Twitter), indicating differences in the distribution of predicted emotional expressions across platforms.

Error analysis showed that misclassifications mainly occurred between emotion categories with similar semantic and contextual characteristics, particularly between anger and sadness and between happiness and trust. Sarcasm, implicit expressions, multiple emotional cues within a single comment, and the removal of emoticons during preprocessing were also identified as potential factors contributing to classification errors. Overall, the findings demonstrate that IndoBERT combined with data augmentation can be applied to granular emotion classification of Indonesian social media texts and provide a more detailed representation of public emotional responses to the MBG Program.

Future research may explore K-Fold Cross Validation to obtain more robust model evaluation, approaches for detecting sarcasm and implicit expressions, and multi-label classification to represent comments containing multiple emotions. Further experiments may also investigate larger transformer-based models and additional data for minority emotion classes to improve recognition across all emotion categories.

ACKNOWLEDGMENT

The authors gratefully acknowledge Rati Badriyati, S.Psi., M.Psi., Psikolog, and Risa Nurhayati, S.Psi., M.Psi., Psikolog, as expert annotators from the field of psychology for their valuable contributions to the joint-labeling process and the development of the final emotion-labeled dataset.

REFERENCES

- Alan, T. B. W. (2022). *Klasifikasi emosi pada teks menggunakan metode deep learning [Skripsi sarjana, Universitas Islam Indonesia]*. DSpace UII. <https://dspace.uui.ac.id/handle/123456789/40586>.
- Arfan, U., & Badii, E. (2025). *Sentiment Analysis of Papuan Public Perceptions Toward the Free Nutritional Meal Program in Indonesia Analisis Sentimen Masyarakat Papua Terhadap Program Makan Gizi Gratis di Indonesia*. 5(October), 1538–1546.
- Destryanto, F., Rizqiyah, P., & Sokibi, P. (2025). *Sentiment Analysis of Public Response to the Free Nutritious Meal Program on Instagram Using IndoBERT*. 5(1), 1–5.
- Dwiyansaputra, R., Murpratiwi, S. I., & Aranta, A. (2025). *Analisis Sentimen Pengguna Platform Media Sosial X pada Topik Pemilihan Presiden 2024 Menggunakan Perbandingan Model*. 9(1), 626–634.
- Ezer, E., Marpaung, O., & Tarigan, M. (2026). *Perbandingan Model Indobert dan Indonesian Roberta Base Dalam Analisis Sentimen Pengguna Twitter terhadap Timnas Indonesia di Kualifikasi Piala Dunia 2026*. 10(1), 30–37.
- Khairani, U., Mutiawani, V., Ahmadian, H., Kuala, U. S., Aceh, B., Ar-raniry, U. I. N., Aceh, B., & Korespondensi, P. (2024). *Pengaruh Tahapan Preprocessing terhadap Model Indobert dan The Influence of Preprocessing Stages on Indobert and Indobertweet Models For Emotion Detection In Instagram News*. 11(4). <https://doi.org/10.25126/jtiik.1148315>
- Laia, M. M., Hasan, F. N., Kuntoro, A. Y., Informatika, B. S., Pusat, J., Kramat, J., No, R., & Pusat, J. (2025). *Analisis Sentimen Program Makan Gratis Pada Platform X Algoritma Naïve Bayes Menggunakan*.
- Marchelo, N., Fatyanosa, T. N., & Perdana, R. S. (2025). *Klasifikasi Sentimen Lintas Domain pada Komentar Media Sosial Menggunakan IndoBERT dengan Fine-Tuning dan Data Augmentation*. 9(12).
- Mohammad, S. M., Bravo-marquez, F., & Zealand, N. (2017). *Emotion Intensities in Tweets*.

- 65–77.
- Munir, R. A., Voutama, A., Karawang, U. S., Timur, T., & Barat, J. (n.d.). *Analisis Sentimen Cuitan di Media Sosial X Tentang Program Makan Bergizi Gratis*. 13(3), 589–595.
- Pornama, A., Fitri, S., & Wati, A. (2025). *Text-Based Emotion Sentiment Analysis Dengan Pendekatan Deep Learning : Akuisisi Tiktok terhadap Tokopedia*. 9(1), 518–526.
- Rahma, I. A., Suadaa, L. H., Timur, J., & Korespondensi, P. (2023). *Penerapan Text Augmentation intuk Mengatasi Data yang Tidak Seimbang Pada Klasifikasi Teks Berbahasa Indonesia Studi Kasus : Deteksi Judul Clickbait dan Komentar Hate Speech Application of Text Augmentation To Handling Imbalanced Data Problem in Indonesian Text Classification Case Study : Clickbait Headline And Hate Speech Comment*. 10(6), 1329–1340. <https://doi.org/10.25126/jtiik.2023107325>
- Rahmatullah, B., Saputra, S. A., Budiono, P., & Wigandi, D. P. (2025). *Sentimen Analisis Makan Bergizi Gratis Menggunakan Algoritma Naive Bayes*. 05(01).
- Roberta, M. (2026). *Klasifikasi Emosi Pada Data Teks Pidato Politik Menggunakan*. 13(1).
- Saepul Rohman, I. W. P. A. (2025). *Dalam Analisis Sentimen Program Makan Bergizi Gratis*. 9(6), 9464–9471.
- Sahron Angelina Ihalauw. (2025). *Analisis Sentimen Kinerja 100 Hari Pemerintahan Prabowo Gibran Pada Media Sosial X Menggunakan INDOBERT*.
- Saputra, A. C., Saragih, A. S., & Ronaldo, D. (2025). *Prediksi Emosi dalam Teks Bahasa Indonesia*. 19(1), 1–15.
- Subarkah, P., Ikhsan, A. N., Anggraeni, E., & Sabaniyah, A. P. (2025). *Sentiment Perspective of Government ' s Free Nutritious Meal Policy on Social Media X using Indo-BERT and Bi-LTSM*. 7(2), 201–210.
- Tsabitah, L., Karima, D. A., Diva, Z., Munaspin, P., Titiana, N. M., Studi, P., Informasi, S., Sriwijaya, U., Ilir, K. O., Makan, P., Gratis, S., & Gratis, S. (2025). *Analisis Sentimen Program Makan Siang Gratis Dalam*. 9(4), 6288–6294.