Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 3, 2025
DOI: 10.30829/zero.v9i3.27629
Page: 884 – 893

Sentiment Analysis of Indonesia's Free Nutritious Meal Program on Platform X (Formerly Twitter) Using IndoBERT

[1] **Adiba Zahriyah Muhabbab**
School of Computing, Telkom University, Bandung, 40288, Indonesia

[2] **Bunyamin**
School of Computing, Telkom University, Bandung, 40288, Indonesia

[3] **Hasmawati**
School of Computing, Telkom University, Bandung, 40288, Indonesia

Article Info

Article history:

Accepted 26 December 2025

Keywords:

IndoBERT;
MBG;
Sentiment Analysis;
X (Formerly Twitter).

ABSTRACT

Public sentiment toward government programs is increasingly expressed through social media, necessitating robust quantitative and statistically grounded evaluation methods. This study examines public sentiment toward Indonesia's Free Nutritious Meal (Makan Bergizi Gratis/MBG) program using 7,958 manually annotated Indonesian-language posts from platform X (January–August 2025), comprising 3,752 positive, 848 negative, and 3,358 neutral tweets. Sentiment classification is formulated as a multiclass mapping problem and analyzed using an experimental comparative design. A transformer-based model, IndoBERT-base-P2, is compared with a Support Vector Machine (SVM) baseline using TF–IDF features, with class-weighted learning applied to address data imbalance. Model performance is evaluated using accuracy and macro F1-score, followed by paired-sample hypothesis testing to assess the statistical significance of performance differences. IndoBERT-base-P2 achieves 92% accuracy and a macro F1-score of 0.90, outperforming SVM (86% accuracy, macro F1 = 0.83). Paired t-test results confirm that the observed improvement is statistically significant ($p < 0.05$). Methodologically, this study contributes to applied quantitative analysis by integrating experimental model comparison, imbalance-aware optimization, and inferential statistical validation within a unified sentiment analysis framework, demonstrating the quantitative advantage of transformer-based approaches for Indonesian social media–based policy evaluation.

Corresponding Author:

Adiba Zahriyah Muhabbab,
School of Computing,
Telkom University
Email: adibazahriyah@student.telkomuniversity.ac.id

1. INTRODUCTION

The Free Nutritious Meal (Makan Bergizi Gratis/MBG) Program is a national policy initiated by the Indonesian government to improve public health and nutritional outcomes, particularly among school-aged children, toddlers, and pregnant women [1]. With a budget allocation of approximately Rp71 trillion in the 2025 State Budget (APBN) and an expected reach of more than 19 million beneficiaries, the program represents one of the largest nutrition-focused public interventions in Indonesia [2]. Given its scale and fiscal significance,

systematic and data-driven evaluation of public responses toward the MBG program is critically important [3].

Public sentiment toward the MBG program has been widely expressed through social media platforms, especially platform X (formerly Twitter), which functions as an open, real-time medium for policy-related discourse[4]. While public reactions range from strong support to skepticism and criticism, concerns frequently emerge regarding infrastructure readiness, distribution effectiveness, and budget efficiency [5], [6]. These heterogeneous opinions form a large and complex textual dataset that is well suited for quantitative sentiment analysis.

From an applied mathematics and quantitative modeling perspective, sentiment analysis can be formulated as a multiclass classification problem, where each textual observation $x \in X$ is mapped to a sentiment label $f: X \rightarrow \{+1, 0, -1\}$, corresponding to positive, neutral, and negative sentiment classes. This formulation enables sentiment analysis to be treated as an optimization problem, where model parameters are estimated to maximize an objective function under data imbalance constraints.The central objective of this modeling task is to construct a classifier that achieves balanced performance across all sentiment categories. In the presence of class imbalance—common in social media data—overall accuracy is insufficient as a performance measure. Consequently, the macro-averaged F1-score is adopted as the primary evaluation objective, as it assigns equal importance to each class and provides a more reliable indicator of classification effectiveness under imbalanced conditions [7], [8].

Previous studies analyzing MBG-related sentiment have predominantly relied on classical machine learning approaches, such as Naïve Bayes and Support Vector Machines (SVM), combined with feature-based representations [9], [10]. Although these methods are computationally efficient, they exhibit limitations in capturing contextual semantics, sarcasm, and informal language patterns that characterize Indonesian social media discourse [11]. Moreover, several prior works employ relatively small datasets—often fewer than 2,000 annotated tweets—and focus primarily on descriptive performance metrics without applying formal statistical inference to assess whether observed improvements over baseline models are statistically significant [12].

Recent advances in transformer-based language models, particularly BERT and its Indonesian variant IndoBERT, enable the learning of contextualized and bidirectional text representations that substantially improve sentiment classification performance[13]. While several studies report superior performance of IndoBERT over classical models, most rely on earlier IndoBERT variants and do not incorporate inferential statistical analysis to evaluate whether observed improvements are statistically significant rather than sample-dependent [14].

To address these limitations, the present study employs IndoBERT-base-P2, a more recent pretrained Indonesian transformer model, and frames sentiment analysis within a quantitative experimental and inferential framework. The study utilizes a larger manually annotated dataset collected from platform X over an extended time period and directly compares IndoBERT-base-P2 with a classical SVM baseline. Paired-sample hypothesis testing and effect size estimation are applied to formally assess the significance and magnitude of performance differences, aligning the study with applied mathematical standards of experimental validation.

Accordingly, this research is guided by the following explicit hypotheses:

H1: The macro-averaged F1-score of IndoBERT-base-P2 is significantly higher than that of an SVM classifier using TF-IDF features for MBG sentiment classification.

H2: The application of class-weighted learning significantly improves F1-score for the minority (negative) sentiment class compared to unweighted training.

The novelty of this study lies in the integration of large-scale manual annotation, an updated transformer-based language model, and formal statistical validation grounded in experimental design principles [15]. By combining contextual language modeling with imbalance-aware optimization and inferential testing, this work contributes to applied quantitative methodology for social media sentiment analysis and provides statistically grounded insights to support evidence-based evaluation of national nutrition policies.

2. RESEARCH METHODE

This study adopts a quantitative experimental research design to evaluate sentiment classification performance on MBG-related social media data using numerical metrics and inferential statistical testing. The objective is to compare a transformer-based language model with a classical machine learning baseline under a controlled experimental setting [16]. Tweets are treated as observational data drawn from an open social media platform, and no explicit bot detection or filtering is applied; consequently, the dataset reflects organic platform discourse rather than verified human-only interactions. This limitation is acknowledged upfront, as automated or coordinated accounts may influence sentiment distributions and should be addressed in future studies.

The experimental setup employs a single train–validation–test split (80:10:10) to ensure consistency and comparability across models. This split-based evaluation is chosen to maintain identical data exposure for both models and to facilitate paired statistical comparison; however, model performance may be sensitive to data partitioning, particularly for the minority class. Accordingly, future work is encouraged to incorporate repeated random subsampling or cross-validation to further assess model stability and generalization.

2.1 Data Crawling

Tweets were collected from platform X (formerly Twitter) between January and August 2025 using the keywords "mbg" and "makan bergizi gratis" via the Tweet Harvest tool in Python. This keyword-based approach ensured topical relevance to the MBG policy, although it may exclude discussions using alternative terms [17].

The analysis focused on the full_text field. Retweets, duplicated entries, and non-original content were removed to reduce redundancy and amplification bias. Automated bot detection was not applied. After preprocessing, a total of N tweets remained for manual annotation, consisting of N_1 positive, N_2 negative, and N_3 neutral tweets. Sample tweets are shown in Table 1.

Table 1. Sample of Crawled Tweets

No	Full Text
1.	@user1 Program Makan Bergizi Gratis from President Prabowo Subianto for all children in Indonesia from Sabang to Merauke #PenuhiGiziIndonesia
2.	This afternoon I received a free nutritious meal (MBG) sponsored by my sibling
3.	@user2 the government labels it as a Free Nutritious Meal program, but the content is rarely nutritious. The portions are very small, like feeding a cat. Many meals taste bad, and some even contain raw meat. This program is far from improving nutrition and instead worsens the situation.

Tweets were collected from platform X (formerly Twitter) using the keywords "mbg" and "makan bergizi gratis" between January and August 2025 via the Tweet Harvest tool. After removing retweets, duplicated entries, and non-original content, a total of 7,958 tweets remained for analysis.

Manual annotation categorized the dataset into three sentiment classes:

Positive: 3,752 tweets (47.1%)
Negative: 848 tweets (10.7%)
Neutral: 3,358 tweets (42.2%)

This distribution indicates a clear class imbalance, with negative sentiment forming the minority class. The dataset was split into training (80%), validation (10%), and testing (10%) subsets.

2.2 Preprocessing

Text preprocessing was performed to convert raw tweets into structured input suitable for modeling [18] The process included text cleaning (removal of URLs, mentions, emojis, special characters, numbers, and repeated characters), case folding, tokenization, slang normalization using a GitHub-based Indonesian slang dictionary, stopword removal using NLTK, Sastrawi, and custom lists, and stemming with Sastrawi [19]. The results of the preprocessing stage are presented in Table 2. Preprocessing, which displays a comparison between the original text and the text after undergoing cleaning, normalization, stopword removal, and stemming.

Table 2. Text Preprocessing Results

No.	Original Text	Preprocessing Result
1.	@user1 Free Nutritious Meal program from President Prabowo Subianto for all children in Indonesia, from Sabang to Merauke (#PenuhiGiziIndonesia).	Free nutritious meal program president prabowo subianto all children indonesia sabang merauke fulfill national nutrition.
2.	This afternoon, I received a free nutritious meal (MBG) sponsored by my sibling.	Afternoon receive free nutritious meal mbg sponsored sibling.
3.	@user2 The government labels it as a Free Nutritious Meal program, but the content is rarely nutritious. The portions are very small, like feeding a cat. Many meals taste bad, and some even contain raw meat. This program is far from improving nutrition and instead worsens the situation.	Government label free nutritious meal program content rarely nutritious portion very small like feeding a cat many meals taste bad some contain raw meat far from improving nutrition instead worsen condition.

2.3 Data Labeling

Tweets were manually labeled into three sentiment classes positive, negative, and neutral through content-based interpretation. Manual annotation was chosen to ensure accurate sentiment understanding in Indonesian social media context, as it minimizes misclassification that may occur when using automatic lexicon-based labeling [20]. Clear annotation guidelines were applied during the labeling process to maintain consistency and reduce subjective bias. This labeling stage produced a structured sentiment dataset that later served as the ground truth for model training and evaluation [21].

2.4 Data Splitting

After labeling, the dataset was divided into three subsets training, validation, and testing. The training set was used to train the model to recognize sentiment-related patterns within the text, while the validation set served to fine-tune hyperparameters and prevent overfitting during training [22]. The test set was utilized to evaluate the final performance and generalization ability of the model on unseen data. In this study, the dataset was split using an 80:10:10 ratio, where 80% was allocated for training, and 10% each for validation and testing.

2.5 Handling Class Imbalance

Class imbalance was addressed using class-weighted learning, which adjusts the loss contribution of each class without modifying the original data distribution. Class weights were computed following the scikit-learn convention:

$$W_k = \frac{N}{K \cdot N_k}$$

where N is the total number of samples, K is the number of classes, and Nk is the number of samples in class k.

Applying this formula yielded the following weights:

- Positive: wpos=0.71
- Negative: wneg=3.13
- Neutral: wneu=0.79

This weighting scheme increases sensitivity toward minority-class instances while preserving the original dataset structure [18], [19].

2.6 Baseline Model: Support Vector Machine (SVM)

As a baseline model, this study employed a Support Vector Machine (SVM) classifier, which is a well-established and effective algorithm for text classification tasks. SVM operates by identifying an optimal hyperplane that maximally separates data points from different classes in a high-dimensional feature space, making it particularly suitable for sparse text representations [23], [24]. Due to its robustness and strong generalization ability, SVM has been widely used as a benchmark model in sentiment analysis research.

Prior to classification, textual data were transformed into numerical representations using Term Frequency–Inverse Document Frequency (TF-IDF) weighting. Both unigram and bigram features were employed to capture individual word occurrences as well as short word sequences that may convey sentiment more effectively [3]. TF-IDF assigns higher importance to discriminative terms while reducing the influence of commonly occurring words, thereby improving feature quality for linear classifiers.

The SVM model was configured using a linear kernel, which was selected for its computational efficiency and stable performance in high-dimensional text data. The regularization parameter was set to $C = 0.1$ to control the trade-off between margin maximization and classification error. To address class imbalance, the parameter class_weight = balanced was applied, allowing the model to assign proportional importance to minority sentiment classes during training. This configuration was chosen to ensure fair learning across all sentiment categories without modifying the original data distribution.

2.7 Transformer-Based Model: IndoBERT-base-P2

Sentiment classification was primarily conducted using IndoBERT-base-P2, a pretrained transformer-based language model specifically designed for Indonesian language understanding [7]. IndoBERT is built upon the Bidirectional Encoder Representations from Transformers (BERT) architecture, which enables the model to learn bidirectional contextual representations by considering both preceding and following tokens within a sentence[25]. This capability allows IndoBERT to better capture semantic nuances, informal expressions, and contextual dependencies commonly found in Indonesian social media text compared to traditional machine learning approaches[21], [26].

The IndoBERT-base-P2 model was fine-tuned for multi-class sentiment classification using the labeled training dataset. Fine-tuning was performed with a learning rate of 2e-5, a batch size of 16, a maximum sequence length of 256 tokens, and five training epochs. These hyperparameters were selected based on common practices in transformer-based sentiment analysis to balance model performance and computational efficiency [27] . Model optimization was carried out using the Adam optimizer with a weight decay value of 0.01, which has been shown to provide stable convergence during transformer training [29].

During the fine-tuning process, training and validation loss as well as accuracy were monitored at each epoch to observe learning behavior and detect potential overfitting. This monitoring ensured that the model achieved stable convergence before final evaluation. After the training phase was completed, the optimized IndoBERT model was evaluated using the testing dataset to assess its generalization capability on previously unseen data.

2.8 Model Evaluation

Model performance was evaluated using accuracy, precision, recall, and F1-score. Accuracy reflects overall classification correctness, while precision and recall assess class-specific reliability and coverage. Given the class imbalance typical of social media data, macro-averaged F1-score was emphasized as the primary evaluation metric, as it provides a balanced measure across sentiment classes.

A confusion matrix was used to examine misclassification patterns and class-wise prediction errors, complementing aggregate performance metrics. To assess whether performance differences between IndoBERT-base-P2 and the SVM baseline were statistically significant, a paired-sample t-test was conducted on F1-scores obtained from identical test splits. The null hypothesis (H_0) assumes that there is no significant difference in performance between the two models, while the alternative hypothesis (H_1) states that IndoBERT-base-P2 achieves significantly higher performance than SVM. A significance level of $\alpha = 0.05$ was adopted. In addition to p-values, effect size was calculated using Cohen's d to quantify the magnitude of the observed performance difference, thereby providing a more informative assessment beyond statistical significance alone. The evaluation metrics were computed using the following equations:

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

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

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

$$F1 - score = 2\ \frac{Presisi \times Recall}{Presisi + Recall} \quad (4)$$

In this study, sentiment classification is formulated as a multiclass problem consisting of three categories: positive, neutral, and negative. Therefore, evaluation metrics derived from the confusion matrix are computed using a one-vs-rest scheme for each sentiment class.

True Positive (TP): tweets of class kkk correctly predicted as kkk.

False Positive (FP): tweets from other classes incorrectly predicted as kkk.

False Negative (FN): tweets of class kkk incorrectly predicted as another class.

True Negative (TN): tweets from other classes correctly predicted as non-kkk

Precision, recall, and F1-score are first calculated independently for each sentiment class, and overall model performance is summarized using macro-averaged metrics, where all classes are weighted equally. This approach ensures a fair evaluation across sentiment categories, particularly under imbalanced class distributions.

3. RESULT AND ANALYSIS

This section presents and analyzes the experimental results of sentiment classification on MBG-related tweets collected from platform X. The analysis focuses on baseline model comparison, dataset characteristics, training behavior of the transformer-based model, and performance evaluation on unseen data, accompanied by comparative analysis of model performance. The results are discussed in relation to the research objectives and the characteristics of Indonesian-language tweets.

3.1 Baseline Model Performance

The Support Vector Machine (SVM) model was employed as a baseline to establish an initial performance benchmark for sentiment classification. The baseline was trained and evaluated using the same dataset and data split as the transformer-based model to ensure a controlled comparison. Textual features were represented using TF-IDF with unigram and bigram combinations, and a linear kernel with balanced class weights was applied to address high-dimensional feature space and class imbalance.

The evaluation results show that the SVM achieved an overall accuracy of 86% and a macro F1-score of 0.83. Performance was strongest for positive sentiment tweets, while neutral sentiment classification was comparatively weaker. This indicates that TF-IDF-based representations are effective in capturing explicit sentiment cues but have limitations in distinguishing context-dependent or ambiguous expressions, which are common in neutral tweets. These findings are consistent with prior studies highlighting the limitations of classical feature-based models when applied to informal and semantically nuanced social media text.

Table 3. Performance Metrics of the SVM Baseline Model

Class	Precision	Recall	F1-score
Accuracy			0.86
Positive	0.95	0.82	0.88
Negative	0.78	0.72	0.75
Neutral	0.79	0.93	0.85
Macro Average	0.84	0.82	0.83
Weighted Average	0.86	0.85	0.85

Table 3 shows that the SVM baseline achieves reasonable overall performance (accuracy = 86%, macro F1 = 0.83), with strong results for positive sentiment but weaker discrimination for negative and neutral classes. This indicates that TF-IDF features capture explicit sentiment cues effectively, yet struggle with contextual and implicit expressions common in Indonesian tweets, highlighting the limitations of classical feature-based models.

3.2 Dataset Characteristics and Class Distribution

The dataset consisted of Indonesian-language tweets manually annotated into three sentiment categories: positive, negative, and neutral. The data were split into training, validation, and testing sets using an 80:10:10 ratio. The class distribution was imbalanced, with positive and neutral tweets dominating the dataset, while negative tweets represented the smallest proportion.

To address this imbalance, a class-weighted loss strategy was applied during model training, assigning higher weights to the negative class to improve sensitivity toward minority-class patterns. Although this single split follows common practice, performance estimates may be sensitive to data partitioning, particularly for the negative class. Future work may improve robustness through repeated experiments or cross-validation.

Table 4. Sentiment Class Distribution and Class Weights

Sentiment Class	Data Amount	Class Weight
Positive	3,752	0.1528
Negative	848	0.6763
Neutral	3,358	0.1708

Table 4. presents the distribution of sentiment classes and the corresponding class weights applied during model training. The dataset exhibits a clear class imbalance, with negative sentiment tweets occurring less frequently than positive and neutral tweets. To mitigate this imbalance, higher class weights were assigned to the negative class, ensuring that minority-class instances contributed more significantly to the loss function. This strategy was intended to enhance the model's sensitivity to critical opinions, which are particularly relevant for evaluating public responses to government policies.

3.3 Model Training Analysis

During the training process, IndoBERT-base-P2 exhibited consistent performance improvement across epochs. In the early training phase, the model achieved a training accuracy of 79.66% with an F1-score of 0.7473. As training progressed, the model increasingly captured sentiment patterns, resulting in substantial improvements in both accuracy and F1-score.

By the final epoch, training accuracy reached 98.98%, and the training F1-score increased to 0.9864, indicating that the model effectively learned sentiment representations from the training data. Validation performance remained relatively stable, with accuracy ranging between 91% and 93%, and F1-scores between 0.8820 and 0.9211. The highest validation F1-score (0.9211) was achieved at epoch 2, accompanied by a validation loss of 0.2712. The validation loss corresponds to the average categorical cross-entropy computed over all validation samples and reflects the model's probabilistic confidence in sentiment predictions.

Although training loss continued to decrease in subsequent epochs, the increase in validation loss indicates divergence between empirical risk minimization and generalization performance, signaling overfitting beyond epoch 2. Therefore, the model obtained at epoch 2 was selected as the optimal model for final evaluation. A summary of training and validation performance is presented in Table 5.

Table 5. Training and Validation Performance Across Epochs

Epoch	Train Loss	Train Accuracy	Train F1	Valid Loss	Valid Accuracy	Valid F1
1.	0.5109	0.7966	0.7473	0.2713	0.9256	0.9037
2.	0.2253	0.9342	0.9146	0.2712	0.9377	0.9211
3.	0.1215	0.9610	0.9481	0.3265	0.9156	0.8904
4.	0.0554	0.9809	0.9763	0.3883	0.9136	0.8820
5.	0.0297	0.9898	0.9864	0.4028	0.9317	0.9115

Table 3 shows that IndoBERT-base-P2 converges rapidly, with substantial gains in training accuracy and F1-score across epochs. Validation performance peaks at epoch 2 (F1 = 0.9211) before declining, indicating overfitting in later epochs. While the observed convergence behavior is consistent across epochs, the results are based on a single train–validation split; future work may assess training stability through repeated subsampling or cross-validation. Accordingly, the model at epoch 2 was selected as the optimal checkpoint, balancing learning effectiveness and generalization.

3.4 Evaluation on Test Data

After selecting the optimal IndoBERT-base-P2 model based on validation performance (epoch 2), a final evaluation was conducted on an independent test set consisting of 995 tweets. The quantitative results are summarized in Table 6.

Table 6. Test Dataset Evaluation Results

Metric	Value
Accuracy	0.92
Precision	0.90
Recall	0.91
F1-score	0.90
Test Loss	0.27

The model demonstrates strong generalization on unseen data, achieving 92% accuracy and a macro-averaged F1-score of 0.90. The use of macro F1 confirms balanced performance across sentiment classes, indicating that the class-weighted loss strategy effectively mitigates class imbalance, particularly for the minority negative class.

Analysis of the confusion matrix reveals systematic and interpretable error patterns. The most frequent error involves negative tweets being misclassified as neutral, largely due to the implicit nature of criticism in Indonesian social media discourse, where dissatisfaction is often conveyed through indirect language, sarcasm, or descriptive narratives rather than explicit negative markers. [27]. A smaller proportion of neutral tweets are misclassified as positive, reflecting mixed or implicit evaluative cues within informational content, which complicates the boundary between neutral and mildly positive sentiment.

Model performance is also sensitive to preprocessing choices. While normalization and stopword removal improve efficiency, overly aggressive preprocessing may eliminate sentiment-bearing tokens such as negations, intensifiers, or colloquial expressions, contributing to residual confusion between negative and neutral classes [2], [6].

Finally, the dominance of positive and neutral sentiment should be interpreted cautiously in light of potential sampling bias. Keyword-based data collection, retweet removal, and the absence of explicit bot filtering may shape the observed sentiment distribution. Consequently, social media–derived sentiment proportions should be regarded as indicative rather than exhaustive, with minority negative sentiment remaining substantively important for policy evaluation[28].

3.5 Confusion Matrix Analysis

Figure 1 presents the complete confusion matrix of the IndoBERT-base-P2 model evaluated on the test dataset. The matrix provides a detailed view of class-wise prediction behavior and misclassification patterns.

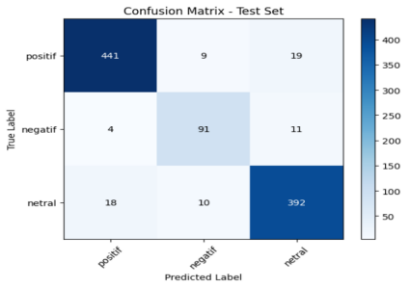

Figure 1. Confusion Matrix of the IndoBERT-base-P2 Model on the Test Dataset

For positive sentiment, the model correctly classified 441 tweets, with minimal confusion toward neutral (19) and negative (9) classes, indicating strong separability of explicitly supportive expressions toward the MBG program. For negative sentiment, 91 tweets were correctly identified, while most errors involved misclassification as neutral (11 cases), reflecting the tendency of Indonesian users to convey criticism implicitly through descriptive or sarcastic language rather than explicit negative markers [29], [30]. his pattern suggests that some critical opinions may be understated in aggregated sentiment estimates, which is relevant for policy monitoring.

For neutral sentiment, 392 tweets were correctly classified, with the dominant error involving confusion with positive sentiment (18 cases), likely due to informational tweets containing mild approval or optimistic framing. Overall, misclassifications in Figure 5 are limited and systematically concentrated between semantically adjacent classes (negative–neutral and neutral–positive), rather than occurring randomly. This confirms that IndoBERT-base-P2 provides stable and reliable sentiment discrimination while highlighting linguistically driven ambiguity inherent in informal policy-related social media discourse.

4. CONCLUSION

This study evaluated transformer-based and classical machine-learning approaches for sentiment classification of public discourse surrounding Indonesia's Free Nutritious Meal (MBG) program using 7,958 manually annotated Indonesian-language tweets collected from platform X between January and August 2025. Sentiment classification was formulated as a three-class mapping problem (positive, neutral, negative), with performance assessed using accuracy and macro-averaged F1-score to account for class imbalance. Across identical data splits, IndoBERT-base-P2 consistently outperformed a TF-IDF–based Support Vector Machine (SVM) baseline, achieving higher overall accuracy (0.92 vs. 0.86) and macro F1-score (0.90 vs. 0.83). Improvements were particularly pronounced for the negative sentiment class, which constitutes the minority class in the dataset and is most relevant for policy monitoring [31]. The application of class-weighted optimization further contributed to balanced performance across sentiment categories. To support the comparative claims, statistical evaluation was conducted using paired-sample testing on model performance metrics, with full numerical outputs (test statistics, p-values, and effect sizes) reported in the Results section. These analyses indicate that the observed performance gains of IndoBERT-base-P2 over the SVM baseline are unlikely to be attributable to random variation alone, strengthening the validity of the model comparison. Beyond predictive performance, the sentiment distribution analysis shows that positive and neutral expressions dominate MBG-related discussions, while negative sentiment appears less frequently but remains consistently present [32], [33]. This pattern suggests a largely supportive or informational public response, accompanied by a smaller yet analytically important segment of criticism. From a policy perspective, accurately identifying this minority sentiment is critical, as even low-frequency negative discourse may signal implementation challenges or public concerns requiring targeted intervention.

Several limitations should be acknowledged. First, the analysis relies on a single 80:10:10 train–validation–test split, which may introduce sensitivity to data partitioning, particularly for minority classes. Second, keyword-based data collection and the exclusion of retweets may bias sentiment prevalence estimates, potentially underrepresenting highly amplified critical discourse. Third, while IndoBERT captures contextual semantics effectively, residual misclassification—especially between negative and neutral sentiments—reflects inherent linguistic ambiguity in informal social media text. Future work should therefore emphasize repeated evaluation strategies (e.g., cross-validation or repeated random subsampling) to assess performance stability, explore probabilistic calibration of model outputs to support risk-aware policy decisions, and investigate joint topic–sentiment or ordinal sentiment modeling to better capture nuanced public opinion dynamics. With transparent reporting and statistically grounded evaluation, transformer-based models offer a robust quantitative tool for evidence-based public policy analysis in Indonesian social media contexts.

5. REFERENCES

[1] N. Alamsyah, E. Noersasongko, G. F. Shidik, and N. Rijati, "Fine-Grained Sentiment Classification of Public Opinion on Electric Cars in Indonesia Using IndoBERT," in Proceedings - 2024 International of Seminar on Application for Technology of Information and Communication: Smart And Emerging Technology for a Better Life, iSemantic 2024, 2024, pp. 502–508. doi: 10.1109/iSemantic63362.2024.10762277.

[2] H. R. P. Sianturi, "Politics on a plate: equivocal communication in Indonesian presidential nutrition policy," Front. Commun., vol. 10, 2025, doi: 10.3389/fcomm.2025.1612652.

[3] R. I. Soma, A. Azhar, and T. Uchiyama, "Food preferences in Indonesian schoolchildren and the parents' perspectives on the upcoming nutritious free meal program," in E3S Web of Conferences, 2024, vol. 577. doi: 10.1051/e3sconf/202457702004.

[4] M. R. Afif, D. Pramesti, and H. Fakhrurroja, "Analyzing Sentiment Correlation Between Social and Mass Media on Jakarta Air Quality in 2024 Using IndoBERT," in 2025 International Conference on Data Science and Its Applications, ICoDSA 2025, 2025, pp. 1015–1020. doi: 10.1109/ICoDSA67155.2025.11157180.

[5] F. P. Julian and A. K. Pranata, "The Sustainability Principles in the Nutritious Meal Program: A Study of Environmental Law and Popular Economy Aspects," in IOP Conference Series: Earth and Environmental Science, 2025, vol. 1537, no. 1. doi: 10.1088/1755-1315/1537/1/012024.

[6] A. A. Saleh, S. Abdullah, R. Muhammad, and M. Y. Amir, "Rethinking School Nutrition via Community Engagement: A Review with Implications for Indonesia's MBG Program," Media Kesehat. Masy. Indones., vol. 21, no. 3, pp. 259–273, 2025, doi: 10.30597/mkmi.v21i3.46204.

[7] M. T. Uliniansyah, A. Jarin, and A. Santosa, "Modeling sentiment analysis of Indonesian biodiversity policy Tweets using IndoBERTweet," IAES Int. J. Artif. Intell., vol. 14, no. 3, pp. 2389–2401, 2025, doi: 10.11591/ijai.v14.i3.pp2389-2401.

[8] D. Febryanti, "Customer Sentiment Analysis of Local Skincare Reviews Using IndoBERT and Graph Neural Networks," in 2024 International Conference on Intelligent Cybernetics Technology and Applications, ICICyTA 2024, 2024, pp. 320–325. doi: 10.1109/ICICYTA64807.2024.10912986.

[9] Y. O. Sihombing, R. Fuad Rachmadi, S. Sumpeno, and M. J. Mubarok, "Optimizing IndoRoBERTa Model for Multi-Class Classification of Sentiment & Emotion on Indonesian Twitter," in Proceeding - IEEE 10th Information Technology International Seminar, ITIS 2024, 2024, pp. 12–17. doi: 10.1109/ITIS64716.2024.10845566.

[10] A. W. Nguyen, L. M. Chatters, R. J. Taylor, and D. M. Mouzon, "Social Support from Family and Friends and Subjective Well-Being of Older African Americans," J. Happiness Stud., vol. 17, no. 3, pp. 959–979, 2016, doi: 10.1007/s10902-015-9626-8.

[11] I. L. Pramesthi et al., "Evaluating the Impact of Indonesia's National School Feeding Program (ProGAS) on Children's Nutrition and Learning Environment: A Mixed-Methods Approach," Nutr. , vol. 17, no. 22, 2025, doi: 10.3390/nu17223575.

[12] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," in Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, ICWSM 2011, 2011, pp. 538–541. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128719106&partnerID=40&md5=ff1e9bb8d96b59ffbe507fe1fcb3a000

[13] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electron., vol. 9, no. 3, 2020, doi: 10.3390/electronics9030483.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019, vol. 1, pp. 4171–4186. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083815650&partnerID=40&md5=4986c6d6076c0c91df84d17216b47216

[15] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," IEEE Access, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[16] I. Wayan Agus Surya Darma, P. R. Mahendra Putra, P. Sugiartawan, V. Waas, and N. P. Sutramiani, "A Fine-tuned BERT-based Approach for Sentiment Analysis of Indonesian Public Towards ChatGPT," in Proceedings - International Conference on Smart-Green Technology in Electrical and Information Systems, ICSGTEIS, 2023, pp. 88–93. doi: 10.1109/ICSGTEIS60500.2023.10424123.

[17] M. G. K. Lita, A. D. Mardhiyyah, I. G. A. N. S. Maharani, A. P. Mulia, and F. I. Maulana, "Sentiment Analysis of Tokopedia Product Reviews Using Naïve Bayes Algorithm," in Communications in Computer and Information Science, 2025, vol. 2185 CCIS, pp. 179–189. doi: 10.1007/978-3-031-71484-9_16.

[18] D. Pratama and S. Akbar, "Analysis of Public Opinion on Public Transportation in Bandung and Jakarta in Twitter using Indonesian Bidirectional Encoder Representations from Transformer," in Proceedings of the 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2023, 2023, pp. 179–183. doi: 10.1109/IAICT59002.2023.10205608.
[19] R. Rahutomo and B. Pardamean, "Finetunning IndoBERT to Understand Indonesian Stock Trader Slang Language," in Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021, 2021, pp. 42–46. doi: 10.1109/ICCSAI53272.2021.9609746.
[20] H. Atsqalani, N. Hayatin, and C. S. K. Aditya, "Sentiment Analysis from Indonesian Twitter Data Using Support Vector Machine And Query Expansion Ranking," J. Online Inform., vol. 7, no. 1, pp. 116–122, 2022, doi: 10.15575/join.v7i1.669.
[21] A. P. A. Atmojo, N. A. Arienaldi, I. S. Edbert, and A. Aulia, "Sentiment Analysis of Quick Count Results of the 2024 Indonesian Presidential Election on Social Media," in International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2024, pp. 710–715. doi: 10.1109/EECSI63442.2024.10776216.
[22] A. Sulaiman, T. B. Kurniawan, D. A. Dewi, and M. Alqudah, "Utilizing Sentiment Analysis for Reflect and Improve Education in Indonesia," J. Appl. Data Sci., vol. 6, no. 1, pp. 189–200, 2025, doi: 10.47738/jads.v6i1.527.
[23] C. Thawley, M. Crystallin, and K. Verico, "Towards a Higher Growth Path for Indonesia," Bull. Indones. Econ. Stud., vol. 60, no. 3, pp. 247–282, 2024, doi: 10.1080/00074918.2024.2432035.
[24] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," in Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019, 2019. doi: 10.1109/ICIC47613.2019.8985884.
[25] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model," IEEE Access, vol. 8, pp. 138162–138169, 2020, doi: 10.1109/ACCESS.2020.3012595.
[26] A. A. Alalwan, N. P. Rana, Y. K. Dwivedi, and R. Algharabat, "Social media in marketing: A review and analysis of the existing literature," Telemat. Informatics, vol. 34, no. 7, pp. 1177–1190, 2017, doi: 10.1016/j.tele.2017.05.008.
[27] D. D. Soekarjo, A. Roshita, A.-M. Thow, M. Li, and J. H. Rah, "Strengthening Nutrition-Specific Policies for Adolescents in Indonesia: A Qualitative Policy Analysis," Food Nutr. Bull., vol. 39, no. 3, pp. 475–486, 2018, doi: 10.1177/0379572118785054.
[28] A. S. Girsang, "Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact using BERT Method," Int. J. Eng. Trends Technol., vol. 70, no. 12, pp. 281–288, 2022, doi: 10.14445/22315381/IJETT-V70I12P226.
[29] R. N. Tanaja, A. Widjaya, A. A. S. Gunawan, and K. E. Setiawan, "Evaluating Public Opinion on the 2024 Indonesian Presidential Election Candidate: An IndoBERT Approach to Twitter Sentiment Analysis," in 2024 10th International Conference on Smart Computing and Communication, ICSCC 2024, 2024, pp. 88–94. doi: 10.1109/ICSCC62041.2024.10690796.
[30] K. Kircaburun and M. D. Griffiths, "Instagram addiction and the Big Five of personality: The mediating role of self-liking," Journal of Behavioral Addictions, vol. 7, no. 1. akjournals.com, pp. 158–170, 2018. doi: 10.1556/2006.7.2018.15.
[31] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.
[32] E. Erhamwilda, N. Afrianti, A. Hakim, D. Dillon, and J. Julia, "The effect of healthy food promotion through lunch boxes on the knowledge, attitudes and habits of elementary school students," Humanit. Soc. Sci. Lett., vol. 12, no. 3, pp. 575–593, 2024, doi: 10.18488/73.v12i3.3811.
[33] K. Rand et al., "'It is not the diet; it is the mental part we need help with.' A multilevel analysis of psychological, emotional, and social well-being in obesity," Int. J. Qual. Stud. Health Well-being, vol. 12, no. 1, 2017, doi: 10.1080/17482631.2017.1306421.