International Journal of Human Computing Studies

e-ISSN: 2615-8159
p-ISSN: 2615-1898

Volume: 8 Issue: 2 | August 2026
https://journals.researchparks.org/index.php/IJHCS

Article

Classification of Sentiment in Reddit Forum Comments using Fine-Tuned IndoBERT (Case Study: Free Nutritious Meal Program)

Bram Aji Saka Putra [1], Suprianto [2*]

1. Muhammadiyah University of Sidoarjo, Indonesia
2. Muhammadiyah University of Sidoarjo, Indonesia

* Correspondence: suprianto@umsida.ac.id

Abstract: This study analyzes public opinion on the Free Nutritious Meals Program (MBG) policy on the Reddit platform using the IndoBERT model. The study uses a dataset of 6,295 Reddit comments collected from 2024 to 2025. The preprocessing stage implemented a comprehensive suite of textual refinement procedures, encompassing normalization of colloquial language, emoji conversion, and translation into Indonesian to establish linguistic consistency across the corpus. Manual annotation was performed with meticulous care on a balanced dataset of 3,000 comments, evenly allocated among positive, negative, and neutral classes. An 80:20 train–test split was applied to enhance the reliability of model training and evaluation. A fine-tuned IndoBERT model exhibited outstanding performance, attaining 99.00% for accuracy, F1-score, and precision. Applied to the full dataset, the model predicted neutral sentiment as predominant (59.44%), followed by positive (34.11%) and negative (6.45%) sentiments, a distribution that suggests a measured and reflective public discourse on the topic. Model reliability was further supported by a mean confidence score of 0.8502 and a processing throughput of 137.93 samples per second, indicating strong potential for deployment as an effective tool for near real-time sentiment monitoring in public policy contexts.

Citation: Putra, B. A. S & Suprianto. Classification of Sentiment in Reddit Forum Comments using Fine-Tuned IndoBERT (Case Study: Free Nutritious Meal Program). International Journal of Human Computing Studies 2026, 8(2), 54-72

Received: 10th Apr 2026
Revised: 21st May 2026
Accepted: 8th June 2026
Published: 15th July 2026

Keywords: IndoBERT; sentiment analysis; Reddit; Free Nutritious Meal Program; Natural Language Processing

1. Introduction

The rapid development of internet technology has fundamentally changed how people communicate and obtain information. In recent years, many users have shifted from traditional news media to digital platforms, with social media becoming a primary source of information [1]. Social media not only provides rapid and open access to news but also enables users to connect, exchange opinions, and participate in public discussions [2].

Reddit was selected as the data source because it supports substantially longer textual content—up to 40,000 characters for posts and 10,000 characters for comments—than platforms such as X/Twitter, which historically limited posts to 280 characters. This capacity encourages more detailed discussion and produces linguistic complexity and sentiment nuance that are valuable for natural language processing research [3], [4]. In addition, Reddit's topic-based subreddit structure, relative anonymity, and application programming interface make it suitable for large-scale extraction of public opinion concerning the Free Nutritious Meal Program [5].

The Free Nutritious Meal Program was selected because it is a large-scale and long-term public policy intended to support children, students, and pregnant women while addressing malnutrition, stunting, and food insecurity. Its implementation may influence human-resource quality and provide a reference for similar social-welfare policies in the future [6]. The study is also temporally relevant because the data were collected from 2024 to 2025, covering the policy-announcement period, early implementation, and subsequent public evaluation. The findings therefore provide both a current description of public opinion and a baseline for monitoring changes in sentiment over time.

Sentiment analysis was adopted to examine public responses to the Free Nutritious Meal Program. Through natural language processing, public comments can be classified into positive, negative, and neutral categories [7], while their emotional tone and evaluative orientation can also be investigated. Because Indonesian social-media discourse includes formal language, slang, code-mixing, regional expressions, and context-dependent meanings, IndoBERT was selected for its ability to represent contextual relationships across diverse writing styles.

This study aims to evaluate the accuracy and computational performance of a fine-tuned IndoBERT model in classifying Reddit comments concerning the Free Nutritious Meal Program. It also examines the predicted sentiment distribution, confidence scores, dominant lexical patterns, and differences in text length among sentiment classes. Through these objectives, the study seeks to provide a more detailed understanding of public expectations, support, concerns, and criticism related to the implementation of a child-welfare policy.

A review of previous studies identified several substantive research gaps that motivated the present investigation.

Nurjoko and Rahardi applied IndoBERT to identify verbal-violence sentiment on Twitter and reported an accuracy of 72%; however, their study did not systematically optimize hyperparameters and the optimizer to maximize the model's ability to capture emotionally nuanced expressions [8].

Nur, Wibisono, and Megasari combined fine-tuned IndoBERT with BERTopic for sentiment analysis of technology brands on platform X. However, the absence of a separate comparison between the pretrained and fine-tuned models made it difficult to isolate the effect of model optimization [9].

Munir analyzed sentiment toward the Free Nutritious Meal Program using support vector machines and a generic BERT model on an imbalanced platform-X dataset. The absence of an optimized Indonesian-specific model and the exclusion of Reddit's distinctive linguistic characteristics leave an important methodological gap [7].

Pham et al. compared VADER, TextBlob, and generic BERT models for Reddit sentiment analysis concerning artificial intelligence. Their use of a generic BERT model rather than an Indonesian-specific model, together with a topic unrelated to Indonesian public policy, demonstrates the need for language- and domain-adapted modeling [4].

Wagay and Jahiruddin demonstrated the effectiveness of an SBERT–BiLSTM architecture for classifying mental-health content in Reddit posts. Nevertheless, their hybrid architecture did not evaluate a standalone fine-tuned IndoBERT model or focus on three-class sentiment distribution in the public-policy domain [10].

Based on these gaps, this study optimizes IndoBERT through carefully selected hyperparameters and the AdamW optimizer for Indonesian public-policy sentiment analysis on Reddit. It further investigates the distribution and linguistic characteristics of public responses to the Free Nutritious Meal Program while accounting for colloquial language, regional variation, and contextual noise in social-media data.

2. Materials and Methods

This study employed a quantitative deep-learning-based sentiment-analysis approach to examine public opinion systematically and objectively. It adopted a positivist experimental paradigm in which model performance was evaluated through measurable quantitative metrics [11].

2.1 Research Workflow

The research workflow consisted of six sequential phases: (1) collecting Reddit data through the PRAW API using the keywords "MBG" and "Makan Bergizi Gratis"; (2) preprocessing through slang normalization, emoji conversion, and translation into Indonesian; (3) preparing the data through partial manual annotation and an 80:20 train–test split; (4) fine-tuning IndoBERT with hyperparameter optimization and the AdamW optimizer; (5) evaluating the model using accuracy, precision, recall, and F1-score; and (6) applying the model to the complete dataset to identify sentiment distributions and characteristics. Figure 1 summarizes the methodological framework and shows the feedback mechanism used when model performance did not satisfy the evaluation criteria.

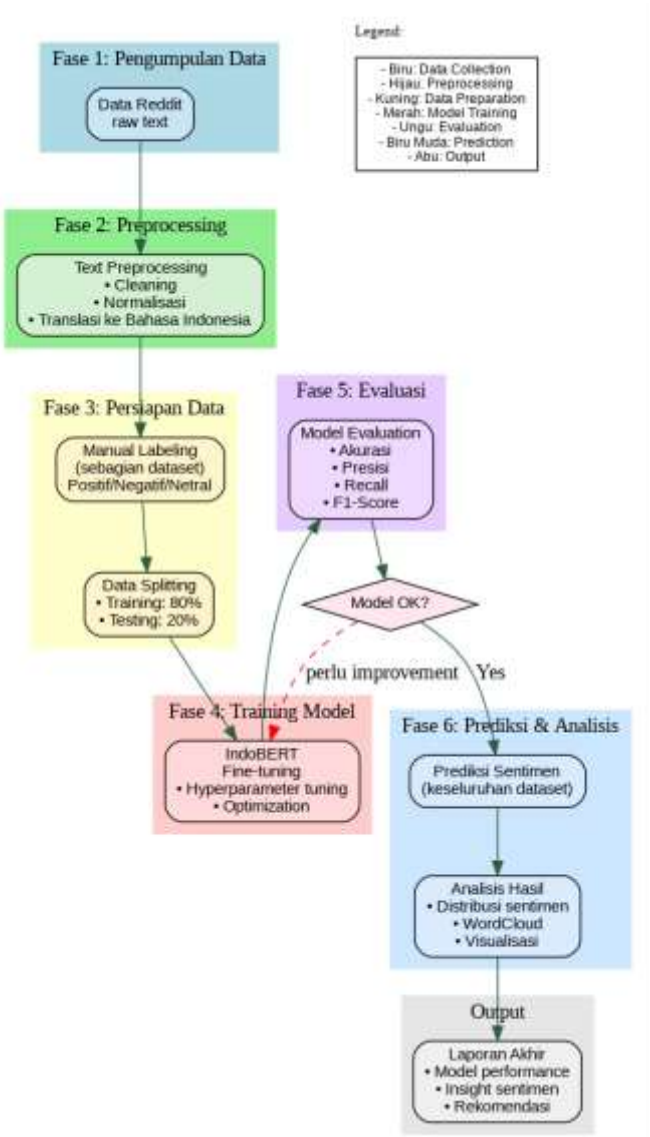

Figure 1. Research Workflow

Figure 1 presents the six methodological phases. Phase 1 acquires raw Reddit data through the API. Phase 2 cleans and normalizes the text. Phase 3 performs manual labeling and divides the data into training and test sets. Phase 4 fine-tunes IndoBERT through hyperparameter optimization. Phase 5 evaluates the model using accuracy, precision, recall, and F1-score. Phase 6 applies the trained model to the complete dataset for sentiment-distribution analysis. The dashed feedback loop indicates that model configuration may be revised when the evaluation results do not meet the required criteria.

2.2 Data Collection

The data were obtained from Reddit, particularly the r/Indonesia subreddit. Reddit supports extended, topic-centered discussion and relative anonymity, making it a valuable source for studying public opinion [12]. Comments were collected through the Reddit API using the Python Reddit API Wrapper (PRAW) [5]. Figure 2 presents the data-collection process.

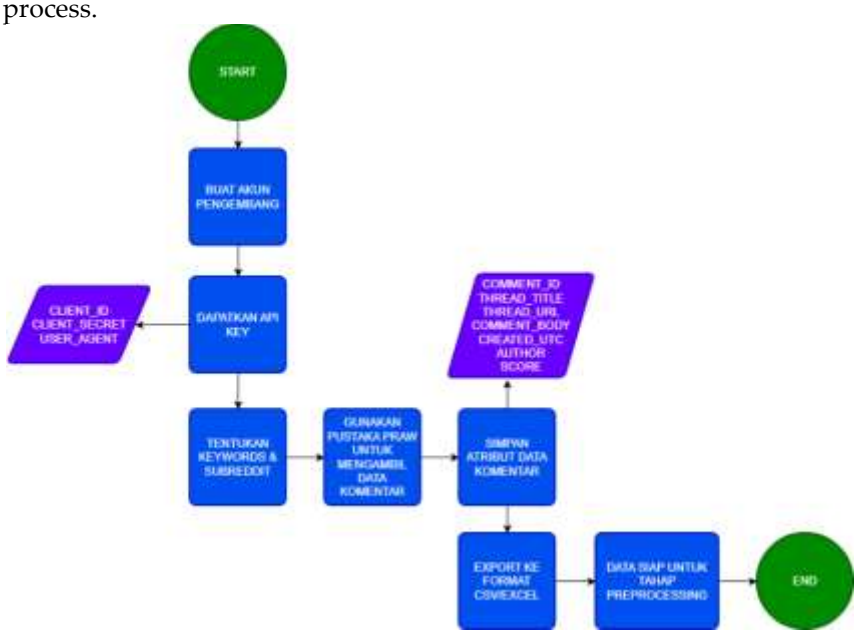

Figure 2. Data Collection Flowchart

Figure 2 illustrates the systematic data-collection procedure. The process began with the creation of a Reddit developer account to obtain an API key consisting of a client ID, client secret, and user agent. The keywords "MBG" and "Makan Bergizi Gratis" and the target subreddit were then defined. PRAW extracted comment data and stored the attributes comment_id, thread_title, thread_url, comment_body, created_utc, author, and score. The extracted data were exported to CSV and Excel formats for preprocessing. This procedure emphasized authenticated API access and comprehensive metadata retention for subsequent analysis.

Data collection used the keywords "MBG" and "Makan Bergizi Gratis" for the period from January 2024 to August 2025 and initially produced 16,268 comments. After strict selection and cleaning, the final dataset contained 6,295 comments. Table 1 summarizes the extracted attributes and provides an overview of the information retained for each Reddit comment.

Table 1. Reddit Comment Attributes

Attribute	Details
comment_id	Unique comment identifier.
Keyword_pencarian	Keyword used during data collection.
Subreddit	Reddit discussion community.
thread_title	Topic or thread title.
thread_url	Topic or thread URL.
comment_body	Comment text.
created_utc	Comment creation time.
author	Username.
score	Number of upvotes or downvotes.

Table 1 lists the principal attributes extracted from each Reddit comment. The comment_id field serves as a unique identifier and supports duplicate detection; keyword_pencarian records the search term used to retrieve the comment; subreddit identifies the discussion community; and thread_title and thread_url preserve

conversational context. The comment_body field contains the text analyzed by the model, created_utc enables temporal analysis, and author and score provide user and engagement metadata.

2.3 Data Preprocessing

Data preprocessing is a critical component of an NLP pipeline because it cleans and normalizes unstructured social-media text [13]. Table 2 provides an example of the transformation from a raw Reddit comment to its normalized form.

Table 2. Example of Preprocessing Results

Raw Comment	Preprocessing
Woyy program MBG keren bgt nih!! 🔥🔥 Program makan bergizi gratis 100% bakal ngebantu bgt buat anak2 sekolah di daerah terpencil 😭😭 Link lengkapnya di sini: https://reddit.com/r/indonesia/comments/xyz123 Gue sih setuju 100% sama program ini, tapi ada yg bilang kualitasnya kurang bagus... Ya gimanaaa lagi ya kalo budgetnya terbatas 😅😅 Btw, menurut kalian gimana? Worth it gak sih? 🤔 #MBG #MakanBergiziGratis #IndonesiaEmas2024	hei program makan bergizi gratis hebat banget ini fire fire program makan bergizi gratis seratus persen akan membantu banget untuk anak anak sekolah di daerah terpencil crying face crying face saya sih setuju seratus persen sama program ini tapi ada yang bilang kualitasnya kurang bagus ya bagaimana lagi ya kalau budgetnya terbatas smiling face with tears smiling face with tears ngomongo ngomong menurut kalian bagaimana layak tidak sih thinking face

Table 2 demonstrates the transformation of a Reddit comment containing slang, abbreviations, emojis, a URL, repeated characters, and mixed Indonesian–English expressions. The preprocessing result standardizes colloquial terms, expands abbreviations, converts emojis into textual descriptions [14], removes the URL, and normalizes the language. This transformation improves corpus consistency while retaining information that may contribute to sentiment classification.

Systematic cleaning and adaptive normalization are necessary to transform raw social-media data into coherent and structured representations [15].

Figure 3. Preprocessing Sequence

Figure 3 presents a ten-stage preprocessing pipeline. The sequence begins with lowercasing, followed by normalization of slang and abbreviations. Special symbols and number-based word repetition are standardized, emojis are converted into text, URLs and non-alphanumeric noise are removed, excessive character repetition is reduced, and redundant spaces are deleted. Translation into Indonesian is performed at the final stage

to obtain a more linguistically homogeneous corpus. The sequence was designed to improve input quality without unnecessarily eliminating affective information.

2.4 Data Labeling and Splitting

Manual labeling was prioritized because it provides high-quality ground truth for supervised sentiment classification [9]. In addition to the three primary sentiment labels, annotators recorded affective-context notes to support consistent interpretation [16], [17]. A total of 3,000 comments were manually labeled with a balanced distribution of 1,000 positive, 1,000 negative, and 1,000 neutral comments. Table 3 summarizes the labeling scheme.

Table 3. Labeling Scheme

Label	Numerical Code	Description	Criteria
Positive	1	A comment that supports, praises, or expresses optimism regarding the program.	Praise, positive expectations, explicit support, or favorable testimony.
Negative	0	A comment that criticizes, rejects, or expresses pessimism regarding the program.	Direct criticism, disappointment, sarcasm or cynicism, or explicit rejection.
Neutral	2	An informative or interrogative comment without a clear evaluative position.	Description, factual question, information without opinion, or balanced positive and negative content.

Table 3 presents the three-class sentiment framework developed for the public-policy context. Each class is defined through an affective orientation, a numerical code for computational processing, and observable operational criteria. The criteria include lexical signals, sentence context, emotional intensity, explicit support or rejection, and the absence of a clear evaluative position. This structure was intended to improve annotation consistency and reproducibility.

The labeled data were divided using an 80:20 stratified split to preserve the same class proportions in the training and test sets. Figure 4 illustrates the resulting distribution and shows that each sentiment class was represented equally in both subsets.

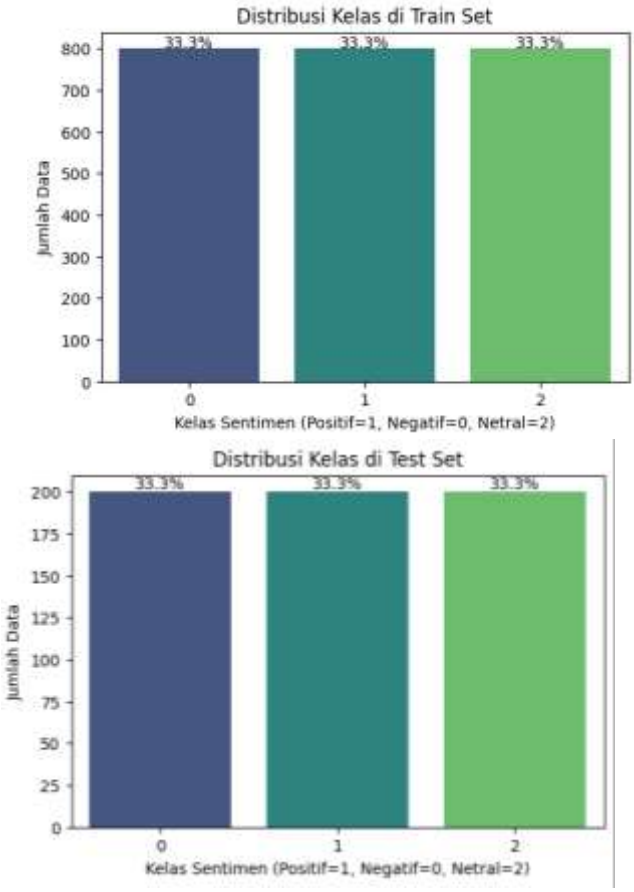

Figure 4. Class Distribution in the Training and Test Sets

Figure 4 shows identical class proportions in the training and test sets. Positive, neutral, and negative comments each account for 33.3% of both subsets, confirming the use of stratified splitting [18]. The training set contains 2,400 comments, with 800 comments per class, while the test set contains 600 comments, with 200 comments per class. A balanced distribution reduces the likelihood that the model will favor a dominant class [19].

2.5 Model Configuration and Training

The study used IndoBERT-base (indobenchmark/indobert-base-p1), which consists of 12 transformer layers, a hidden size of 768, and 12 attention heads [3]. IndoBERT has demonstrated strong performance in Indonesian NLP tasks, including the detection of potential stress sources in media text [20]. Table 4 presents the hyperparameter configuration used in this study.

Table 4. Hyperparameter Configuration

Parameter	Value	Justification
Epochs	5	Sufficient for convergence without excessive overfitting.
Batch Size (Train)	16	Compatible with GPU capacity and gradient stability.
Batch Size (Eval)	32	A larger batch is feasible because evaluation uses only forward passes.
Learning Rate	2×10^{-5}	Commonly effective for BERT fine-tuning.
Warmup Steps	500	Stabilizes optimization during the early training stage.
Weight Decay	0.01	L2 regularization to reduce overfitting.

Parameter	Value	Justification
Gradient Accumulation	2 steps	Simulates a larger effective batch size.
Mixed Precision	FP16	Accelerates computation and reduces memory use.
Evaluation Strategy	Per epoch	Monitors performance at regular intervals.
Save Strategy	Best F1-score	Retains the model with the strongest F1-score.
Early Stopping	Patience 2	Stops training when performance does not improve.
Random Seed	42	Supports reproducibility.

Table 4 lists the 12 principal hyperparameters used to fine-tune IndoBERT. The first group consists of basic training parameters, including the number of epochs, training and evaluation batch sizes, and learning rate. The second group comprises optimization strategies such as warmup steps, weight decay, gradient accumulation, and mixed-precision training. The third group includes evaluation, model-saving, early-stopping, and reproducibility settings. Together, these parameters were selected to balance convergence, computational efficiency, regularization, and model-selection reliability.

The hyperparameter values were selected from prior IndoBERT fine-tuning practices and adjusted to the characteristics of the balanced dataset. The configuration was intended to obtain stable convergence, reduce overfitting, and improve the model's sensitivity to contextual and affective variation in Indonesian public-policy discourse.

3. Results

3.1 Data Collection Results

The collection process retrieved 16,268 Reddit comments before systematic filtering. Figure 5 summarizes the reduction of the dataset to the final analytical sample.

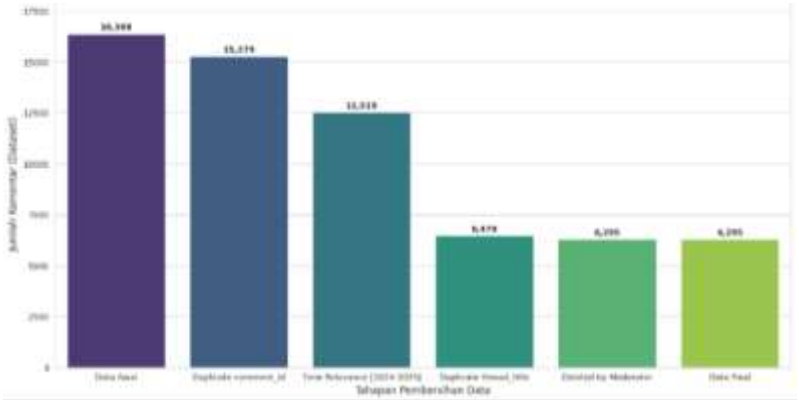

Figure 5. Dataset Filtering Visualization

Figure 5 shows the reduction from 16,268 initial comments to 6,295 final comments. Duplicate removal based on comment_id eliminated 989 comments, the 2024–2025 temporal filter removed 2,760 comments, duplicate thread-title removal eliminated 6,040 comments, and the removal of moderator-deleted comments eliminated 184 records. The final dataset represented 38.7% of the initially collected comments.

The raw scraping output was stored in a structured CSV file. Figure 6 shows the dataset opened in Microsoft Excel before automated preprocessing.

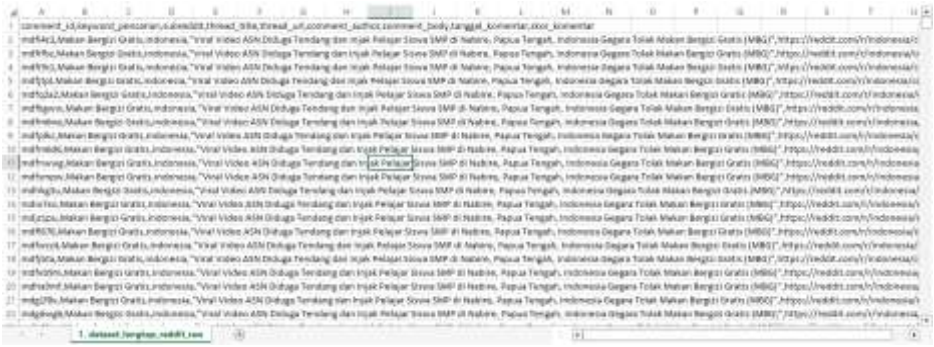

Figure 6. Raw Dataset in Microsoft Excel

Figure 6 displays the raw dataset with the columns comment_id, keyword_pencarian, subreddit, thread_title, thread_url, comment_author, comment_body, tanggal_komentar, and skor_komentar. The file contained 16,268 Reddit-comment records and was used for manual inspection and curation before preprocessing.

3.2 Data Preprocessing Results

After preprocessing with Python and Pandas, the normalized data were exported to Excel for documentation and validation. Figure 7 shows the resulting dataset.

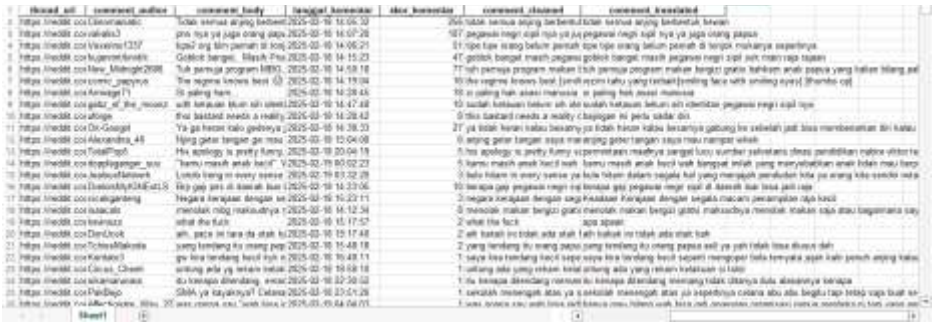

Figure 7. Preprocessed Dataset

Figure 7 presents Reddit comments after cleaning, normalization, emoji conversion, and translation. The processed text was then used as the input for manual labeling and sentiment-model training.

3.3 Data Labeling Results

The labeling stage assigned positive, negative, or neutral sentiment to each selected comment. A sentiment-notes field was also included to record contextual cues and assist annotators in making consistent labeling decisions.

comment_cleaned	comment_translated	sentiment_manual	sentiment_notes
tidak semua anjing berbent	tidak semua anjing berbentuk	negatif	sindiran
pegawai negri sipil nya ya j	pegawai negri sipil nya ya juga	netral	informasi
tipe tipe orang belum pernia	tipe tipe orang belum pernah d	negatif	ofensif
goblok banget masih pega	goblok banget masih pegawai	negatif	sindiran
tuh pemuja program makar	tuh pemuja program makan b	netral	opini
the regime knows best [sm	rezim tahu yang terbaik smile	netral	ambigu
si paling hak asasi manusi	si paling hak asasi manusia	negatif	sindiran
sudah ketauan belum sih ic	sudah ketauan belum sih iden	netral	pertanyaan
this bastard needs a reality	bajingan ini perlu sadar diri	negatif	sindiran
ya tidak heran kalau besan	ya tidak heran kalau besarnya	netral	opini
anjing geter tangan saya m	anjing geter tangan saya mau	negatif	ofensif
his apology is pretty funny	permintaan maafnya sangat lu	negatif	sindiran
kamu masih anak kecil wal	kamu masih anak kecil wah b	negatif	sindiran
bule hitam in every sense y	bule hitam dalam segala hal y	negatif	ejekan
berapa gaji pegawai negri s	berapa gaji pegawai negri sipil	negatif	sindiran
negara kerajaan dengan se	negara kerajaan dengan segal	negatif	sindiran
menolak makan bergizi gra	menolak makan bergizi gratis	netral	opini
what the fuck	apa apaan	netral	ambigu
ah kakak ini tidak ada otak	ah kakak ini tidak ada otak k	negatif	sindiran
yang tendang itu orang pap	yang tendang itu orang papua	netral	informasi
saya kira tendang kecil seg	saya kira tendang kecil seperti	negatif	kata kasar
untung ada yang rekam kel	untung ada yang rekam kelak	negatif	ejekan
itu kenapa ditendang memi	itu kenapa ditendang memang	netral	pertanyaan
sekolah menengah atas ya	sekolah menengah atas ya se	netral	pertanyaan
was gonna say wah bisa ja	hanya mau bilang wah bisa ja	netral	opini

Figure 8. Labeled Dataset

Figure 8 shows the manually labeled dataset after preprocessing. The file contains cleaned and translated comments, sentiment labels, and contextual notes and was prepared for supervised model training.

3.4 Model Training Results

The model achieved its highest evaluation performance at the fifth epoch. Figure 9 presents the training loss, validation loss, accuracy, F1-score, precision, and recall recorded at each epoch.

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	1.080000	0.648165	0.880000	0.880287	0.882936	0.880000
2	0.104800	0.062956	0.980000	0.980054	0.980462	0.980000
3	0.041700	0.057665	0.986667	0.986661	0.986831	0.986667
4	0.029400	0.070039	0.983333	0.983331	0.983636	0.983333
5	0.026500	0.059211	0.990000	0.990000	0.990024	0.990000

Figure 9. IndoBERT Training Results by Epoch

Figure 9 shows that accuracy increased from 88.00% in the first epoch to 98.00% in the second epoch and reached 99.00% in the fifth epoch. Training loss declined from 1.0800 to 0.0265, while validation loss declined from 0.6482 to 0.0592. The fourth epoch produced a temporary increase in validation loss to approximately 0.0700 before the metric improved in the final epoch.

3.5 Model Evaluation

The final model was evaluated on a test set of 600 comments. Table 5 summarizes the evaluation metrics and inference speed.

Table 5. Model Evaluation on the Test Set

Metric	Value
Evaluation Loss	0.0592
Accuracy	0.9900 (99.00%)
F1-Score	0.9900 (99.00%)
Precision	0.9900 (99.00%)
Recall	0.9900 (99.00%)
Evaluation Runtime	4.35 seconds
Samples per Second	137.93

Table 5 shows an evaluation loss of 0.0592 and accuracy, precision, recall, and F1-score of 99.00%. This result corresponds to six misclassified comments among the 600 test examples. The evaluation runtime was 4.35 seconds, equivalent to 137.93 samples per second [21].

```
=== Classification Report ===
              precision    recall  f1-score   support

     positif       0.99      0.99      0.99       200
     negatif       0.99      0.99      0.99       200
      netral       0.99      0.98      0.99       200

    accuracy                           0.99       600
   macro avg       0.99      0.99      0.99       600
weighted avg       0.99      0.99      0.99       600
```

Figure 10. Classification Report

Figure 10 presents precision, recall, F1-score, and support for each class. Positive and negative sentiment obtained precision, recall, and F1-score values of approximately 0.99. Neutral sentiment obtained precision and F1-score values of approximately 0.99 and recall of 0.98. Both the macro and weighted averages were 0.99.

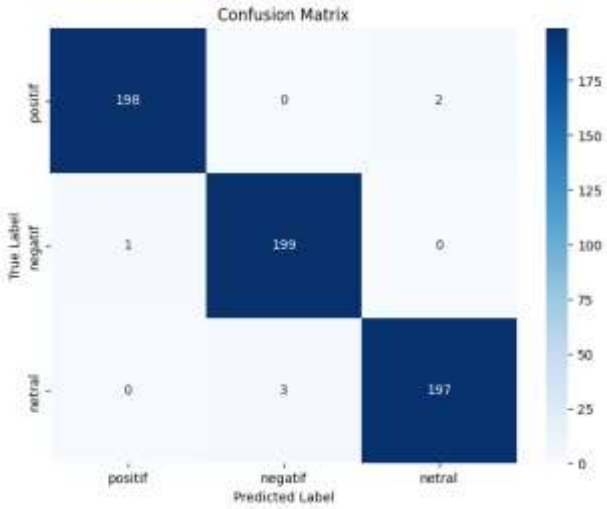

Figure 11. Confusion Matrix

Figure 11 shows 198 correctly classified positive comments, 199 correctly classified negative comments, and 197 correctly classified neutral comments. The six errors consisted of two positive comments predicted as neutral, one negative comment predicted as positive, and three neutral comments predicted as negative.

3.6 Full-Dataset Prediction Results

The trained model was applied to all 6,295 comments. Figure 12 presents the predicted sentiment distribution.

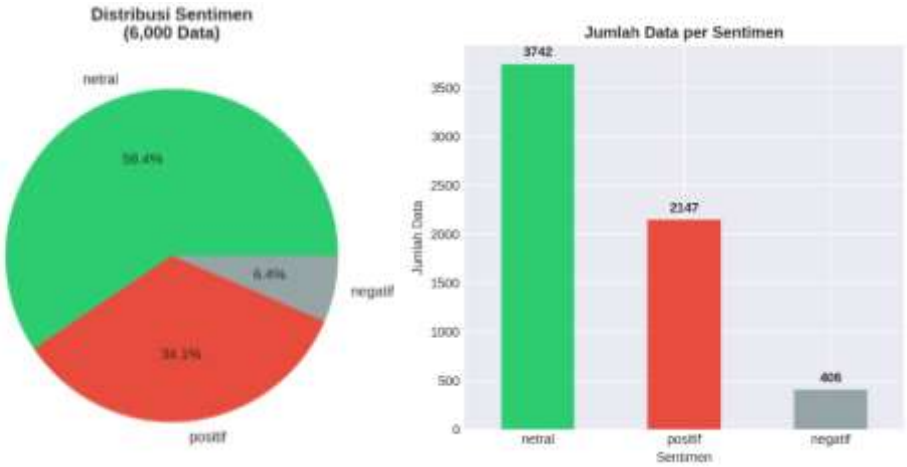

Figure 12. Predicted Sentiment Distribution

Figure 12 shows 3,742 neutral comments (59.44%), 2,147 positive comments (34.11%), and 406 negative comments (6.45%). Neutral sentiment therefore constituted the largest predicted category, followed by positive and negative sentiment.

comment_cleaned	comment_translated	predicted_sentiment	confidence_score	prob_positif	prob_negatif	prob_netral	prob_diff
tidak semua anjing berti	tidak semua anjing berti	netral	0.996797979	0.001332977	0.001869099	0.996797979	0.99492888
pegawai negri sipil nya	pegawai negri sipil nya i	netral	0.999615827	0.000238161	0.000146015	0.999615827	0.999377666
tipe tipe orang belum p	tipe tipe orang belum p	negatif	0.702251554	0.025141621	0.702251554	0.272606879	0.429644674
goblok banget masih pe	goblok banget masih pe	negatif	0.964005709	0.004417186	0.964005709	0.031577043	0.932428665
tuh pemuja program nutuh pemuja program nu		netral	0.551222324	0.435541451	0.013236193	0.551222324	0.115680873
the regime knows best	rezim tahu yang terbaik	netral	0.991795003	0.00317472	0.005030277	0.991795003	0.986764727
si paling hak asasi manu	si paling hak asasi manu	netral	0.994898975	0.002944419	0.002156655	0.994898875	0.992954556
sudah ketauan belum si	sudah ketauan belum si	netral	0.997010946	0.001757625	0.001231429	0.997010946	0.995253321
this bastard needs a rea	bajingan ini perlu sadar	netral	0.834585249	0.103731878	0.06168285	0.834585249	0.730853371
ya tidak heran kalau ber	ya tidak heran kalau ber	positif	0.9380216	0.9380216	0.009630748	0.052347615	0.885673985
anjing geter tangan say	anjing geter tangan say	negatif	0.970774472	0.007832997	0.970774472	0.021392487	0.949163003
his apology is pretty fun	permintaan maafnya sa	positif	0.476003269	0.476003269	0.426868704	0.09710791	0.049134523
kamu masih anak kecil s	kamu masih anak kecil s	netral	0.911578238	0.020361504	0.068060242	0.911578238	0.843517996
bule hitam in every sen	bule hitam dalam segal	netral	0.489782453	0.228814706	0.281402787	0.489782453	0.208379688
berapa gaji pegawai ne	berapa gaji pegawai ne	netral	0.995951056	0.003022069	0.001024917	0.995951056	0.992928139
negara kerajaan dengar	kesdaan kerajaan deng	netral	0.995735407	0.002201657	0.00206287	0.995735407	0.99353375
menolak makan bergizi	menolak makan bergizi	negatif	0.965479573	0.017822923	0.965479573	0.016700443	0.947653849
what the fuck	apa apaan	negatif	0.939988971	0.027709411	0.939988971	0.032301661	0.90768731
aih kakak ini tidak ada o	aih kakak ini tidak ada o	negatif	0.784316599	0.123396601	0.784316599	0.09208677	0.660719998
yang tendang itu orang	yang tendang itu orang j	netral	0.4827438	0.339879185	0.177378986	0.4827438	0.142864615
saya kira tendang kecil r	saya kira tendang kecil r	netral	0.995604396	0.002887208	0.001508477	0.995604396	0.992717188
untung ada yang rekam	untung ada yang rekam	negatif	0.790652772	0.018405253	0.790652772	0.19154203	0.598610742
itu kenapa ditendang m	itu kenapa ditendang m	netral	0.993892312	0.004090228	0.002017446	0.993892312	0.989802084

Figure 13. Prediction-Stage Output

Figure 13 presents the final inference output for the complete dataset. Each record contains the processed comment, the model's predicted sentiment, and the associated confidence score.

3.7 Confidence Score Analysis

Confidence scores indicate the probability assigned by the model to its selected prediction. Table 6 summarizes the confidence-score distribution for the complete dataset.

Table 6. Confidence Score Statistics for the Complete Dataset

Metric	Value
Mean	0.8502
Median	0.9332
Standard Deviation	0.1777
Minimum	0.0000*
Maximum	0.9985

*Note: The minimum value of 0.0000 indicates cases in which the model assigned nearly equal probabilities to multiple classes.

Table 6 shows a mean confidence score of 0.8502, a median of 0.9332, a standard deviation of 0.1777, and a maximum of 0.9985. A total of 57.27% of predictions had confidence scores of at least 0.90, 22.47% were between 0.70 and 0.90, and 20.25% were below 0.70.

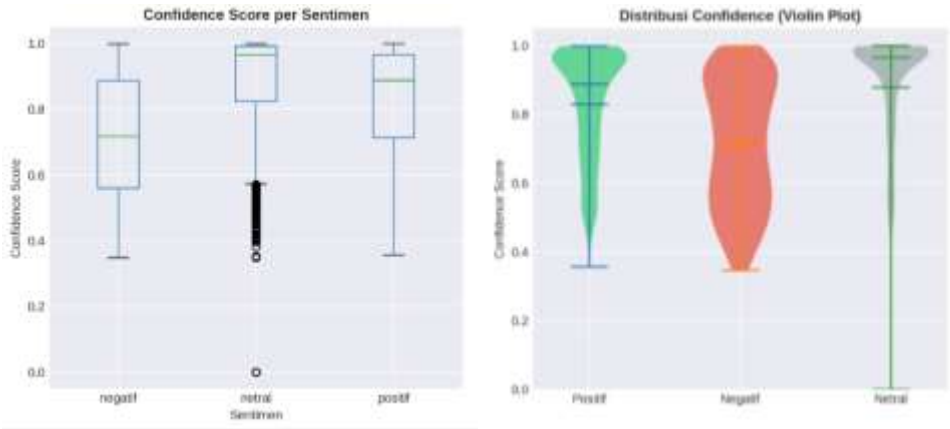

Figure 14. Confidence Scores by Sentiment Category

Figure 14 compares confidence-score distributions across sentiment classes. Neutral sentiment had the highest median confidence, approximately 0.97, and the narrowest interquartile range. Positive sentiment had a median of approximately 0.89 and moderate variability, while negative sentiment had a median of approximately 0.72 and the widest distribution.

3.8 Keyword and Linguistic Pattern Analysis

Keyword analysis was conducted to identify frequent lexical items in each sentiment category. Table 7 lists the ten most frequent terms by class.

Table 7. Top 10 Keywords by Sentiment Category

Rank	Positive Sentiment (Freq.)	Negative Sentiment (Freq.)	Neutral Sentiment (Freq.)
1	makan (866)	makan (54)	makan (1,132)
2	buat (587)	orang (45)	gratis (768)
3	gratis (538)	face (37)	sekolah (693)
4	sekolah (535)	anak (34)	saja (647)
5	tapi (534)	kamu (33)	bergizi (559)
6	saja (532)	saja (31)	jadi (548)
7	orang (505)	buat (31)	makanan (524)
8	jadi (468)	sama (30)	tapi (494)
9	anak (467)	jadi (27)	anak (492)
10	program (376)	anjing (25)	sama (487)

Table 7 shows that the term "makan" appeared most frequently in all three categories, with 1,132 occurrences in neutral comments, 866 in positive comments, and 54 in negative comments. Positive comments frequently contained operational and evaluative terms such as "buat," "jadi," and "program," whereas negative comments included confrontational or vulgar expressions. Neutral comments contained many program-related and descriptive terms, including "gratis," "sekolah," "bergizi," and "makanan." Computational linguistic analysis was used to identify recurrent lexical and discursive patterns [22].

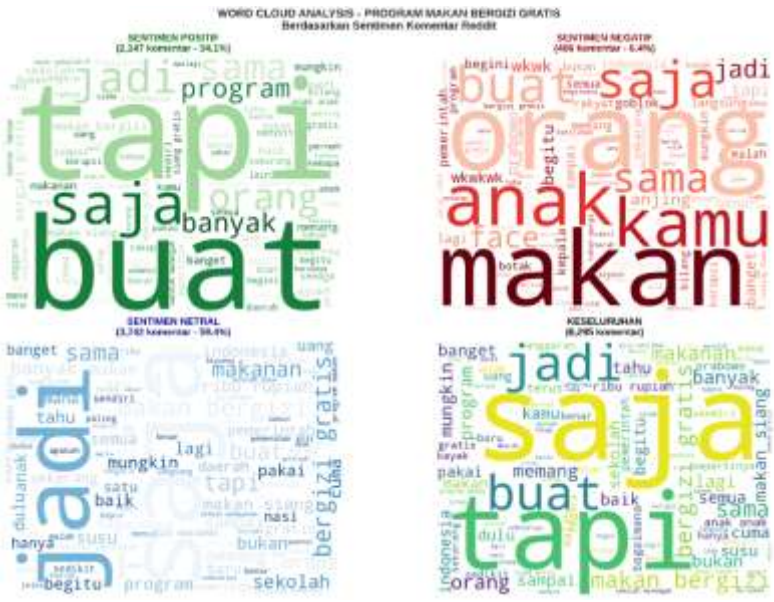

Figure 15. Word-Cloud Analysis of the Free Nutritious Meal Program

Figure 15 presents separate word clouds for positive, negative, neutral, and complete-dataset comments. Font size corresponds to term frequency. The visualizations confirm the dominance of program-related words and common conversational connectors across the dataset.

Table 8. Text-Length Statistics by Sentiment

Sentiment	Mean (words)	Median (words)	Minimum (words)	Maximum (words)
Positive	31.4	18	1	355

Sentiment	Mean (words)	Median (words)	Minimum (words)	Maximum (words)
Negative	12.6	7	1	275
Neutral	25.4	15	1	763

Table 8 shows that positive comments had the greatest mean length at 31.4 words and a median of 18 words. Neutral comments averaged 25.4 words with a median of 15, while negative comments were the shortest, averaging 12.6 words with a median of 7. The longest neutral comment contained 763 words.

4. Discussion

The findings demonstrate that a carefully curated and balanced dataset, combined with language-specific transformer fine-tuning, can produce highly accurate three-class sentiment classification for Indonesian public-policy discourse. The 99.00% accuracy, precision, recall, and F1-score indicate that the model learned highly separable representations of positive, negative, and neutral comments in the labeled test set. This performance is substantially higher than the 72% accuracy reported in earlier IndoBERT research on verbal-violence sentiment [8] and is consistent with the broader observation that domain-specific fine-tuning improves Indonesian-language representations [3], [9], [20]. The result should not be understood only as evidence of IndoBERT's architectural capability. It also reflects the effect of balanced labels, stratified splitting, systematic preprocessing, and a controlled annotation framework. Accordingly, the model's performance is best interpreted as the combined outcome of model selection and data-engineering decisions rather than as an attribute of the pretrained model alone.

The rapid increase in accuracy between the first and second epochs suggests that the pretrained IndoBERT parameters transferred efficiently to the MBG policy domain. Accuracy rose from 88.00% to 98.00% while both training and validation loss declined sharply, indicating that the model quickly adapted to recurrent linguistic patterns in the annotated comments. The temporary increase in validation loss during the fourth epoch may indicate mild instability or the beginning of overfitting, although the decline in the fifth epoch prevented a sustained divergence between training and validation performance. The use of weight decay, warmup steps, gradient accumulation, mixed precision, and early stopping helped maintain training stability. Nevertheless, performance across only five epochs does not establish that the selected configuration is globally optimal. Repeated experiments with multiple random seeds, cross-validation, and systematic ablation studies would be required to determine how strongly each hyperparameter contributed to the final score.

The balanced annotation design was an important strength because each class contributed 1,000 labeled comments before the stratified split. This structure reduced the class-dominance problem that often affects sentiment analysis and enabled macro and weighted averages to remain identical. Previous work has shown that imbalanced data can distort classifier behavior and suppress minority-class recall [18], [19]. In the present study, the confusion matrix contained only six errors, with no evidence that the classifier systematically ignored one class. However, the exceptionally high test performance also requires careful methodological scrutiny. Duplicate removal by comment identifier and thread title reduced obvious repetition, but semantic near-duplicates, quoted text, or highly similar comments could still appear across the training and test sets. Future validation should therefore include similarity-based deduplication, grouped splitting by discussion thread, and an external test set collected from a later period or another platform to rule out hidden information leakage.

The error pattern provides a more informative view of model behavior than the aggregate accuracy alone. Three neutral comments were classified as negative, two positive comments were classified as neutral, and one negative comment was classified as positive. These errors occurred near evaluative boundaries where a comment may combine information, uncertainty, irony, criticism, and conditional support. Neutrality in

public-policy discourse is not always the absence of sentiment; it can include balanced argumentation, indirect evaluation, or factual wording that carries pragmatic implications. This explains why neutral recall was slightly lower than the corresponding values for positive and negative sentiment. Similar ambiguity has motivated hybrid and context-sensitive architectures in Reddit classification [10] and more specialized approaches for code-mixed and imbalanced sentiment data [18]. Qualitative examination of the six misclassified comments would strengthen the analysis by identifying whether the errors arose from annotation uncertainty, sarcasm, translation artifacts, or genuinely ambiguous stance.

The dominance of neutral sentiment in the complete dataset is one of the study's most important substantive findings. A total of 59.44% of comments were classified as neutral, indicating that Reddit discussions about the MBG program frequently involved questions, information sharing, explanation, comparison, or mixed evaluation rather than explicit approval or rejection. This result is plausible because Reddit supports longer and more topic-centered discussion than many short-form platforms [12]. Nevertheless, neutral dominance should not automatically be interpreted as public indifference or objective consensus. A neutral label may conceal implicit criticism, cautious support, uncertainty, or a request for additional evidence. The classification outcome therefore describes the linguistic stance identified by the model, not a direct measurement of policy approval. A complementary stance-detection or aspect-based sentiment framework would be useful for separating factual discussion from ambivalent, conditional, or strategically indirect positions.

Positive sentiment accounted for 34.11% of the predictions, which indicates substantial support for the program while remaining clearly below the neutral proportion. Positive comments were also longer on average than negative comments, suggesting that support was often expressed through explanation, expectations, or arguments concerning program benefits. This finding is relevant to the policy's stated social-welfare objectives, including nutritional support and the reduction of inequality [6]. From a communication perspective, positive discussion may reflect the perceived value of government programs when their intended beneficiaries and social purpose are clearly understood [2]. However, the presence of positive sentiment should not be equated with satisfaction with every implementation component. Users may support the policy's objectives while criticizing menu quality, budget allocation, distribution, targeting, or administrative execution. Aspect-level analysis is therefore needed to distinguish support for the policy principle from evaluations of implementation quality.

Negative sentiment represented only 6.45% of the complete dataset, but it produced the lowest median confidence and the widest confidence distribution. This pattern indicates that negative comments were more linguistically heterogeneous and more difficult for the model to classify consistently. The keyword analysis also showed confrontational expressions and vulgarisms in the negative class, while the shorter average text length suggests that criticism was often delivered through brief, emotionally intense statements. At the same time, negative public discourse frequently uses sarcasm, rhetorical questions, understatement, and implicit comparison rather than direct negative vocabulary. Emoji conversion helps preserve affective cues [14], but it cannot fully reconstruct pragmatic meaning. More advanced preprocessing, context windows, conversation-thread modeling, and sarcasm-specific training data would be necessary to improve performance on such cases [15]–[17].

The confidence-score distribution adds an important layer of reliability assessment. Although the mean confidence was 0.8502 and the median reached 0.9332, approximately one-fifth of predictions had scores below 0.70. The difference between the mean and median indicates that most predictions were highly confident while a smaller group of uncertain cases lowered the average. For practical deployment, the model should not treat all predictions as equally reliable. A human-in-the-loop workflow could automatically accept high-confidence predictions, route medium-confidence cases for periodic sampling, and require manual review for low-confidence comments. Calibration analysis, including reliability diagrams, expected calibration error, and temperature scaling, would

also be necessary because a high softmax probability does not always correspond to a well-calibrated probability of correctness. This is particularly important when model outputs are used to support policy monitoring or decision-making.

The processing rate of 137.93 samples per second demonstrates that the fine-tuned model is computationally suitable for rapid batch analysis and potentially for near-real-time monitoring. At this rate, the complete dataset can be processed in less than one minute under the reported hardware and software conditions. Combined with PRAW-based data collection [5], the model could support a pipeline that retrieves new comments, performs preprocessing, predicts sentiment, and updates a monitoring dashboard. Nevertheless, operational deployment requires more than inference speed. The system would need error handling, data-version control, API-rate-limit management, privacy safeguards, model-drift detection, and scheduled human validation. Processing speed should therefore be presented as evidence of technical feasibility rather than as proof that the system is ready for unsupervised policy use.

The keyword and word-cloud analyses provide a descriptive view of the vocabulary associated with each class. Terms related to eating, schools, children, free provision, and nutrition dominated the dataset, confirming that the collected comments remained topically connected to the program. Positive comments contained terms associated with action and program utility, whereas negative comments included direct interpersonal and emotionally charged expressions. Neutral comments contained many descriptive program terms and discourse connectors. These lexical differences help explain why the classifier achieved strong separation among the classes. However, raw frequency does not reveal whether a term is uniquely associated with one class or simply common throughout the corpus. Statistical measures such as log odds, chi-square association, TF–IDF, contextual embeddings, and keyness analysis would provide a more rigorous account of class-specific language [22].

Differences in text length also reflect the communicative conventions of the platform. Positive comments averaged 31.4 words, neutral comments averaged 25.4 words, and negative comments averaged only 12.6 words. Longer positive and neutral comments may contain justification, policy explanation, personal experience, or comparison, whereas short negative comments may function as direct criticism, ridicule, or emotional reaction. Reddit's support for long-form comments makes these distinctions analytically useful [12]. Even so, text length may become a shortcut feature if the classes differ systematically. The model might partly learn that shorter comments are more likely to be negative rather than relying exclusively on semantic evidence. Future work should evaluate performance across length-controlled subsets and use interpretability techniques to determine whether predictions depend on meaningful contextual tokens or superficial structural cues.

The study also has implications for the interpretation of online public opinion. Reddit users are not a random sample of the Indonesian population, and participation in r/Indonesia is shaped by internet access, platform familiarity, age, education, language preference, and self-selection. Anonymity may encourage candid expression, but it can also increase sarcasm, exaggeration, and adversarial speech. Ethical research on Reddit must therefore consider contextual integrity, user expectations, quotation practices, and the difference between technically public data and ethically unrestricted data [12]. The observed sentiment distribution should be described as discourse within the collected Reddit corpus rather than as a national approval rating for the MBG program. Triangulation with surveys, interviews, news comments, and other social platforms would provide a broader and more representative assessment.

Several methodological strengths support the credibility of the study. The workflow clearly documents data collection, preprocessing, manual annotation, stratified splitting, model configuration, and multiple evaluation metrics. The retention of contextual notes during labeling is particularly useful because sentiment in policy discourse often depends on pragmatic interpretation. The study also reports both predictive performance and computational throughput, followed by complete-dataset inference, confidence analysis, keyword frequencies, and text-length statistics. These components move the analysis

beyond a single accuracy value. At the same time, reproducibility would be improved by reporting the annotation instructions in full, the number and qualifications of annotators, inter-annotator agreement, adjudication procedures, hardware specifications, software versions, preprocessing dictionaries, and the exact code used for training and evaluation.

The principal limitations concern generalizability, annotation reliability, and comparative evaluation. Data were collected from one subreddit and through two keywords, which may exclude relevant discussions that use alternative terms. Translation into Indonesian improves linguistic consistency but may alter sarcasm, cultural references, and code-mixed expressions. The test set was derived from the same curated annotation pool as the training set, and no cross-platform or temporal external validation was conducted. The study also did not report comparisons with simpler baselines, generic multilingual BERT, an unfine-tuned IndoBERT model, or alternative optimizers. Consequently, the 99.00% score demonstrates excellent internal performance but does not yet establish superiority under broader conditions. These limitations should guide the design of subsequent experiments rather than diminish the value of the current findings.

Future development should combine technical refinement with policy-oriented analysis. A stronger system would include grouped and temporal validation, model calibration, explainable predictions, active learning for uncertain cases, and periodic retraining to address language and policy drift. Aspect-based sentiment analysis could distinguish opinions about nutrition, menu variety, distribution, budget, governance, targeting, and implementation. Conversation-aware models could incorporate parent comments and thread context to improve sarcasm and implicit-stance detection. Comparative experiments with IndoBERT variants, multilingual transformers, hybrid neural architectures, and classical baselines would clarify the contribution of each modeling choice. Finally, a responsible monitoring dashboard should communicate uncertainty, display representative themes rather than isolated comments, and prevent automated sentiment scores from being treated as a substitute for inclusive public consultation.

5. Conclusion

This study successfully fine-tuned IndoBERT for three-class Indonesian sentiment classification and obtained accuracy, precision, recall, and F1-score values of 99.00% on a balanced test set. Application of the model to 6,295 Reddit comments showed that neutral sentiment was dominant at 59.44%, followed by positive sentiment at 34.11% and negative sentiment at 6.45%. The model achieved a mean confidence score of 0.8502 and processed 137.93 samples per second, demonstrating its potential for rapid public-policy monitoring. Nevertheless, the lower confidence and greater variability observed in negative comments indicate that sarcasm, implicit criticism, and context-dependent expressions remain challenging. The findings should therefore be interpreted as evidence of model performance on the curated Reddit dataset rather than as a nationally representative measure of public opinion.

5.1 Future Research

Future studies should expand data collection to additional digital platforms and demographic contexts to improve the representativeness of public-opinion analysis. IndoBERT may also be integrated with rule-based or contrastive-learning components to improve sarcasm and implicit-criticism detection. Aspect-based sentiment analysis could identify responses to specific dimensions of the Free Nutritious Meal Program, such as menu quality, distribution, budget efficiency, nutritional adequacy, and program governance. A real-time monitoring dashboard and longitudinal time-series analysis would further support the examination of sentiment changes across the policy cycle. External validation, inter-annotator agreement measurement, baseline-model comparison, and explainability analysis are also required before operational deployment.

5.2 Acknowledgments

The authors express their sincere gratitude to the students who provided moral support and valuable academic discussion throughout this sentiment-classification study. Appreciation is also extended to the lecturers and subject-matter experts who offered

methodological and analytical feedback. The practical assistance of colleagues during the experiments and the continued encouragement of the authors' families contributed substantially to the completion of this research.

REFERENCES

[1] B. Auxier and M. Anderson, "Social Media Use in 2021," Pew Research Center, 2021. [Online]. Available: www.pewresearch.org.
[2] M. A. Afandi and I. Suri, "Effectiveness of Government Communication in Delivering Public Policy on Social Media," Jurnal Komunikasi Pemerintah, vol. 9, no. 1, pp. 23–38, 2025.
[3] W. Wang et al., "IndoBERT: Pre-Trained Model for Bahasa Indonesia," IndoNLP Research Group, 2020. [Online]. Available: https://huggingface.co/indobenchmark.
[4] K. Pham, K. C. Rao Kathala, and S. Palakurthi, "Reddit Sentiment Analysis on the Impact of AI Using VADER, TextBlob, and BERT," Procedia Computer Science, vol. 258, pp. 85–94, 2025.
[5] B. Boe, "PRAW: The Python Reddit API Wrapper," GitHub Documentation, 2023. [Online]. Available: https://praw.readthedocs.io/en/stable/index.html.
[6] A. Kiftiyah et al., "Program Makan Bergizi Gratis (MBG) dalam Perspektif Keadilan Sosial dan Dinamika Sosial-Politik," Pancasila: Jurnal Keindonesiaan, vol. 5, no. 1, pp. 45–62, 2025.
[7] R. A. Munir, "Analisis Sentimen Cuitan di Media Sosial X tentang Program Makan Bergizi Gratis dengan Metode NLP," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 1, pp. 123–135, 2025.
[8] Nurjoko and A. Rahardi, "Model Indo-BERT untuk Identifikasi Sentimen Kekerasan Verbal di Twitter," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 18, no. 2, pp. 156–168, 2024.
[9] M. R. Nur, Y. Wibisono, and R. Megasari, "Analisis Sentimen dan Pemodelan Topik pada Post tentang Merek Teknologi di X Menggunakan Fine-Tuning IndoBERT dan BERTopic," Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI), vol. 4, no. 2, pp. 45–58, 2025.
[10] F. A. Wagay and Jahiruddin, "Classification of Mental Illnesses from Reddit Posts Using Sentence-BERT Embeddings and Neural Networks," Procedia Computer Science, vol. 258, pp. 234–243, 2025.
[11] V. Agustina, A. Herliana, and E. P. Korespondensi, "Analisis Sentimen Publik atas Kebijakan Efisiensi Anggaran 2025 dengan Text Mining dan Natural Language Processing," Jurnal Sistem Informasi dan Teknologi, vol. 7, no. 2, pp. 78–89, 2025.
[12] N. Proferes et al., "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics," Social Media and Society, vol. 7, no. 2, pp. 1–14, 2021.
[13] D. Prasetia et al., "Analisis Sentimen Pengguna Aplikasi MyBluebird dengan Algoritma Naïve Bayes di Playstore," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 2, pp. 145–156, 2025.
[14] Z. Liu et al., "Improving Sentiment Analysis Accuracy with Emoji Embedding," Journal of Safety Science and Resilience, vol. 2, no. 4, pp. 289–298, 2021.
[15] S. S. Berutu et al., "Data Preprocessing Approach for Machine Learning-Based Sentiment Classification," Jurnal Infotel, vol. 15, no. 4, pp. 317–325, 2023.
[16] N. Babanejad, A. Agrawal, A. An, and M. Papagelis, "A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks," 2020. [Online]. Available: https://github.com/NastaranBa/.
[17] H. T. Duong and T. A. Nguyen-Thi, "Preprocessing Techniques and Data Augmentation for Sentiment Analysis," Computational Social Networks, vol. 8, no. 1, pp. 1–15, 2021.
[18] R. Srinivasan and C. N. Subalalitha, "Sentimental Analysis from Imbalanced Code-Mixed Data Using Machine Learning Approaches," Distributed and Parallel Databases, vol. 41, nos. 1–2, pp. 37–52, 2023.
[19] A. Kunaefi, Z. Abidin, and R. Kusumawati, "Klasifikasi Berita Hoaks Bahasa Indonesia Menggunakan IndoBERT Fine-Tuning dengan Pendekatan Focal Loss pada Data Tidak Seimbang," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 10, no. 2, pp. 89–102, 2025.
[20] M. Faza and M. Taufik, "Penerapan Model IndoBERT untuk Deteksi Potensi Sumber Stres dalam Teks Media," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 12, no. 3, pp. 567–578, 2025.
[21] N. Ratnaswari, N. C. Wibowo, and D. S. Y. Kartika, "Analisis Sentimen Menggunakan Metode Lexicon-Based dan Support Vector Machine pada Presiden dan Wakil Presiden Indonesia Periode 2024–2029," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 1, pp. 67–78, 2025.

[22] A. A. P. Wibowo and N. Hidayat, "Eksplorasi Linguistik Komputasional dalam Analisis Bahasa Alami untuk Mengungkap Evolusi Dialek Digital di Era Media Sosial Global," Journal of New Trends in Sciences, vol. 1, no. 3, pp. 45–52, 2023.