Indonesian Journal of Data and Science

Volume 7 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v7i1.381

Research Article

Comparing Sentiment Labeling with RoBERTa and IndoBERTweet on Public Opinion of Program *Makan Bergizi Gratis*

Putri Nur Rezky [1,*]; Dolly Indra [2]; Herdianti Darwis [3]
[1] Universitas Muslim Indonesia, Makassar, 90231, Indonesia, 13120220034@student.umi.ac.id
[2] Universitas Muslim Indonesia, Makassar, 90231, Indonesia, dolly.indra@umi.ac.id
[3] Universitas Muslim Indonesia, Makassar, 90231, Indonesia, herdianti.darwis@umi.ac.id

Correspondence should be addressed to Putri Nur Rezky; 13120220034@student.umi.ac.id

Received 09 November 2025; Accepted 18 Jan 2026; Published 30 March 2026

Abstract:
The Program Makan Bergizi Gratis (MBG) is a flagship program of the Prabowo Subianto administration launched in 2024, triggering diverse public responses on social media. Sentiment analysis using deep learning models offers an effective approach to understanding public opinion at scale. However, selecting the appropriate model for Indonesian social media text remains challenging. This study aims to compare the performance of two pretrained transformer models, RoBERTa Base and IndoBERTweet Base, in conducting automatic sentiment labeling on Indonesian tweets related to the MBG program using a zero-shot labeling approach without human-annotated ground truth. A total of 1,831 tweets were collected from platform X and preprocessed using case folding, normalization, and stopword removal. Both models were applied in parallel to label each tweet with sentiment categories (positive, neutral, negative) along with confidence scores. The comparison was evaluated using agreement rate, Cohen's Kappa, and confidence score analysis. RoBERTa Base exhibits a conservative tendency with 75.20% neutral labels, while IndoBERTweet Base produces a more balanced distribution (68.16% neutral). The comparison shows 77.28% agreement with Cohen's Kappa of 0.490 (Moderate Agreement). RoBERTa Base achieves higher confidence (mean: 0.9559, 83.01% above 0.95) compared to IndoBERTweet Base (mean: 0.9236, 68.65% above 0.95). IndoBERTweet Base is more effective in detecting negative sentiment, identifying nearly twice as many negative tweets (13.54% vs. 7.48%). This study recommends IndoBERTweet Base for exploratory research requiring sensitive sentiment detection and RoBERTa Base for precision-critical applications. An ensemble approach combining both models is recommended for production-critical applications.

Keywords: *Makan Bergizi Gratis* (MBG) Program, IndoBERTweet, Public Opinion, RoBERTa, Sentiment Analysis, Transformer.

1. Introduction

The *Makan Bergizi Gratis* (MBG) Program is one of the strategic policies of the Prabowo Subianto-Gibran Rakabuming Raka administration officially launched in early 2024. This program aims to improve the nutritional quality of Indonesian society, particularly children and pregnant women, through the provision of free nutritious meals. The program target is to reach 15-16.5 million beneficiaries among students as well as pregnant or breastfeeding mothers in the initial implementation phase [1]. As a program involving a large state budget, the MBG Program has received widespread attention from various community groups [2].

Social media, especially Platform X (formerly Twitter), has become one of the main channels for the public to express opinions, criticism, and support for public policies. Platform X has unique characteristics as a dynamic public discussion space, with more than 24 million active users in Indonesia, placing it in the 5th position globally [3]. Public opinion spread on this platform becomes a valuable data source for understanding public perception of the MBG Program.

Sentiment analysis is an effective method for extracting and classifying opinions from large volumes of text data. Sentiment analysis is a computational process to identify and extract subjective information from text, such as opinions, attitudes, and emotions toward a specific topic or entity [4]. In the context of public policy, sentiment analysis helps decision-makers understand public responses quantitatively and qualitatively [5].

The development of Natural Language Processing (NLP) technology, particularly transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), has revolutionized the sentiment analysis approach. Pretrained models such as RoBERTa (Robustly Optimized BERT Pretraining Approach) and IndoBERTweet (Indonesian BERT for Tweet) have shown superior performance in various NLP tasks, including sentiment classification [6], [7].

RoBERTa Base is a development of the BERT architecture optimized with more robust pretraining techniques, using larger datasets and longer training duration [8]. This model has proven effective in various English and multilingual NLP tasks. On the other hand, IndoBERTweet Base is a model specifically trained using Indonesian tweet corpus, thus having better understanding of informal language characteristics, abbreviations, and slang commonly used in Indonesian social media [9].

Several studies have implemented transformer models for Indonesian sentiment analysis. Jayadianti et al. [10] used fine-tuning IndoBERT with RCNN for Indonesian review sentiment analysis, achieving 95.16% accuracy, 94.05% precision, 92.74% recall, and 93.27% f1-score. Mukarramah et al. [11] compared SVM with linear and polynomial kernels for multiclass sentiment analysis on Covid-19 pandemic tweets, showing that SVM with polynomial kernel and trigram features produced the best performance with 51.2% accuracy.

Indra et al. [12] applied the Naïve Bayes classifier with TF-IDF weighting and SMOTE upsampling for sentiment analysis, achieving 80.65% accuracy, demonstrating the effectiveness of traditional machine learning approaches when combined with appropriate data balancing techniques.

Koto et al. [13] developed IndoLEM and IndoBERT as benchmark dataset and pretrained language model for Indonesian NLP, achieving 89.2% accuracy on Twitter sentiment analysis tasks. This research became an important foundation for evaluating Indonesian language models including IndoBERTweet.

Setyawan et al. [14] compared CNN-BiLSTM with and without multi-head attention for sentiment analysis of President Jokowi post-presidency on Platform X using 52,643 tweets. Results showed that CNN-BiLSTM with multi-head attention and SMOTE achieved 98.78% accuracy, demonstrating that the attention mechanism can capture complex contexts such as irony and sarcasm, while SMOTE effectively handles data imbalance.

Sianturi [15] analyzed public response to the MBG Program using 4,041 news reports and 138,301 social media posts from January 6-12, 2025. Results showed sentiment distribution tending toward negative (41.2% negative, 45.7% neutral, 13.1% positive) with problem typology covering planning, implementation, and monitoring aspects.

However, comparative research directly comparing RoBERTa Base and IndoBERTweet Base performance for Indonesian public policy cases, particularly the MBG Program, remains limited. The characteristic differences between multilingual RoBERTa Base and Indonesian tweet-specific IndoBERTweet Base raise an important question about which model is more appropriate for analyzing Indonesian public opinion sentiment on Platform X. Previous studies in this area have primarily focused on classification accuracy using supervised approaches with manually annotated datasets. In contrast, this study employs a zero-shot labeling approach and focuses on agreement analysis and confidence score comparison between models, providing a different perspective on model behavior rather than traditional accuracy metrics. This approach offers practical insights for researchers and practitioners who need to select appropriate models for sentiment analysis without the availability of labeled training data

This research aims to compare sentiment labeling results using RoBERTa Base and IndoBERTweet Base comprehensively including: (1) sentiment distribution; (2) agreement level between models; (3) confidence scores; (4) classification difference patterns; and (5) characteristics of tweets producing disagreement. Three-label classification (positive, negative, neutral) was chosen as many tweets are informative without explicit sentiment [16].

The main contributions are: (1) comprehensive performance comparison of both models in Indonesian public policy cases; (2) in-depth analysis of agreement and disagreement patterns; (3) practical recommendations for model selection; and (4) annotated dataset about MBG Program sentiment for further research.

2. Method

The research methodology section explains the structured stages in conducting sentiment analysis and comparing labeling results between RoBERTa Base and IndoBERTweet Base models. The stages include data collection, preprocessing, sentiment labeling using both models, and comparison evaluation. The research framework is visualized in the flowchart shown in **Figure 1**.

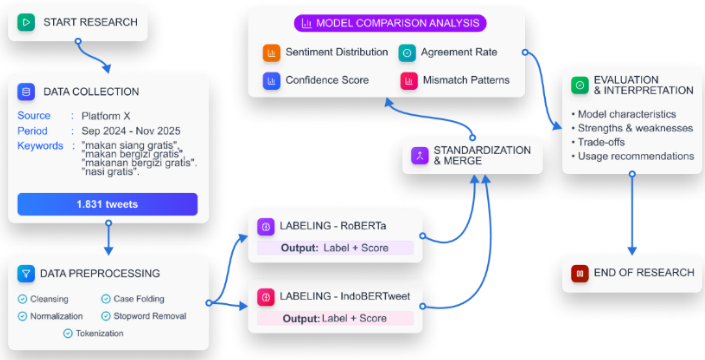

Figure 1: Research Workflow

Dataset:

The dataset used in this research consists of tweet data collected from Platform X using web scraping techniques with Python library (TweetHarvest API). The crawling or scraping method has been commonly used in social media sentiment analysis research [17]. Data collection was conducted from September 2024 to November 2025, with search keywords including: "*makan bergizi gratis*" (free nutritious meal), "*makanan bergizi gratis*," "*makan siang gratis*," "*nasi gratis*," and related combinations.

The total data successfully collected was 1,831 Indonesian tweets. The data was then saved in CSV format with column structure: tweet_id, username, tweet_text, and created_at. **Table 1** displays sample tweet examples used in

Table 1. Sample Dataset of *Makan Bergizi Gratis* Program Tweets

No.	Tweet
1	*Cegah Stunting Mas Pram-Bang Doel janji laksanakan UU Kesejahteraan Ibu dan Anak Di wujudkan dengan : 1. Day Care 2. Ruang untuk laktasi 3. Posyandu Juga ikut menyukseskan program makan bergizi gratis yang digagas Presiden Prabowo. #JakartaMenyala #PrakarUntukJakarta https://t.co/PY4YY3hApf.*
2	*Saya kebetulan dari keluarga yang sederhana untuk itu kami akan mendukung program yang dilakukan oleh pemerintah pusat oleh Presiden Prabowo Subianto yaitu makan siang gratis maka kami akan memberikan sarapan gratis di pagi hari #03JakartaMenyala #PramonoAnung #RanoKarno https://t.co/w3wAt7ZrOE.*
3	*Dari kejadian2 seperti ini program MAKAN BERGIZI GRATIS sudah sangat tepat. Bahkan Pak Prabowo Subianto sudah mengatakan siapa yg ga setuju dengan program Beliau ya GO OUT aja dr kabinet . https://t.co/VlbvL1ucXu.*

Table 1 displays raw tweet samples collected from Platform X. These examples show the diversity of text structures containing hashtags, URLs, numbers, and varying writing styles, which confirms the necessity of the preprocessing stage before the data can be processed by the model.

Pre-processing Data:

Although transformer models can process raw text effectively, preprocessing remains essential for social media data due to several factors. Indonesian tweets often contain excessive noise such as URLs, mentions, hashtags, and non-standard language that can interfere with model attention mechanisms. Additionally, normalization of slang words and informal abbreviations helps standardize the vocabulary and improve tokenization quality, particularly for models not specifically trained on Indonesian social media text.

Collected tweet data is raw, unstructured data containing noise. The preprocessing stage is required to transform raw data into clean and structured data ready for analysis [18]. Preprocessing is a crucial stage that greatly affects sentiment analysis quality [19]. Following are the preprocessing stages conducted:

a. Cleansing

This stage involves removing irrelevant elements from tweets, including: URLs/links, mentions (@username), hashtags (#), emoticons, emojis, punctuation marks, special characters, numbers, and newlines. These components are removed because they are considered not to significantly contribute to sentiment analysis [20].

b. Case Folding

The process of converting all letters in text to lowercase. This step is important to maintain consistency and avoid duplication of the same words with different capitalization [21]. Case folding results are shown in **Table 2**.

Table 2. Example of Case Folding Results

No.	Tweet	After Case Folding
1	*Cegah Stunting Mas Pram-Bang Doel janji laksanakan UU Kesejahteraan Ibu dan Anak Di wujudkan dengan : 1. Day Care 2. Ruang untuk laktasi 3. Posyandu Juga ikut menyukseskan program makan bergizi gratis yang digagas Presiden Prabowo. #JakartaMenyala #PrakarUntukJakarta https://t.co/PY4YY3hApf.*	*cegah stunting mas prambang doel janji laksanakan uu kesejahteraan ibu dan anak di wujudkan dengan day care ruang untuk laktasi posyandu juga ikut menyukseskan program makan bergizi gratis yang digagas presiden prabowo.*
2	*Saya kebetulan dari keluarga yang sederhana untuk itu kami akan mendukung program yang dilakukan oleh pemerintah pusat oleh Presiden Prabowo Subianto yaitu makan siang gratis maka kami akan memberikan sarapan gratis di pagi hari #03JakartaMenyala #PramonoAnung #RanoKarno https://t.co/w3wAt7ZrOE.*	*saya kebetulan dari keluarga yang sederhana untuk itu kami akan mendukung program yang dilakukan oleh pemerintah pusat oleh presiden prabowo subianto yaitu makan siang gratis maka kami akan memberikan sarapan gratis di pagi hari*
3	*Dari kejadian2 seperti ini program MAKAN BERGIZI GRATIS sudah sangat tepat. Bahkan Pak Prabowo Subianto sudah mengatakan siapa yg ga setuju dengan program Beliau ya GO OUT aja dr kabinet . https://t.co/VlbvL1ucXu.*	*dari kejadian seperti ini program makan bergizi gratis sudah sangat tepat bahkan pak prabowo subianto sudah mengatakan siapa yg ga setuju dengan program beliau ya go out aja dr kabinet.*

Based on **Table 2**, the case folding process successfully converts all uppercase letters to lowercase and removes non-textual characters such as numbers in "*kejadian2*" which becomes "*kejadian*," as well as removing URLs and hashtags, making the text more uniform.

c. Normalization

The process of returning words to standard form by converting abbreviations, non-standard words, or slang to standard forms according to the normalization dictionary. Normalization is very important for social media data containing much informal language. Examples: "*gak*" → "*tidak*" (no/not), "*bgus*" → "*bagus*" (good), "*bgt*" → "*banget*" (very) [21].

d. Stopword Removal

Removing words that do not have significant meaning in sentiment analysis [22], such as conjunctions, prepositions, and articles. The Sastrawi library is used as reference for Indonesian stopwords [23]. Stopword removal results are displayed in **Table 3**.

Table 3. Example of Stopword Removal Results

No.	Tweet	After Stopword Removal
1	*cegah stunting mas prambang doel janji laksanakan uu kesejahteraan ibu dan anak di wujudkan dengan day care ruang untuk laktasi posyandu juga ikut menyukseskan*	*cegah stunting prambang doel janji laksanakan undang-undang kesejahteraan anak wujudkan day ruang laktasi posyandu menyukseskan*

No.	Tweet	After Stopword Removal
	program makan bergizi gratis yang digagas presiden prabowo.	*program makan bergizi gratis digagas presiden prabowo.*
2	*saya kebetulan dari keluarga yang sederhana untuk itu kami akan mendukung program yang dilakukan oleh pemerintah pusat oleh presiden prabowo subianto yaitu makan siang gratis maka kami akan memberikan sarapan gratis di pagi hari.*	*keluarga sederhana mendukung program pemerintah pusat presiden prabowo subianto makan siang gratis sarapan gratis pagi.*
3	*dari kejadian seperti ini program makan bergizi gratis sudah sangat tepat bahkan pak prabowo subianto sudah mengatakan siapa yg ga setuju dengan program beliau ya go out aja dr kabinet.*	*kejadian program makan bergizi gratis prabowo subianto setuju program beliau out kabinet.*

Table 3 shows the results after removing common words (stopwords). As seen in the table, words like "*yang*," "*dan*," "*untuk*," and "*di*" have been removed, leaving only keywords that carry the main information content of the tweet.

e. Tokenization

The process of breaking or separating text into individual word units (tokens). Tokenization is a fundamental stage in text processing that affects the quality of subsequent analysis [24]. Sentences in the dataset are converted to token arrays word by word, as shown in **Table 4**.

Table 4. Example of Tokenization Results

No.	Tweet	After Tokenization
1	*cegah stunting prambang doel janji laksanakan undang-undang kesejahteraan anak wujudkan day ruang laktasi posyandu menyukseskan program makan bergizi gratis digagas presiden prabowo.*	*['cegah', 'stunting', 'prambang', 'doel', 'janji', 'laksana', 'undang', 'sejahtera', 'anak', 'wujud', 'day', 'ruang', 'laktasi', 'posyandu', 'sukses', 'program', 'makan', 'gizi', 'gratis', 'gagas', 'presiden', 'prabowo'].*
2	*keluarga sederhana mendukung program pemerintah pusat presiden prabowo subianto makan siang gratis sarapan gratis pagi.*	*['keluarga', 'sederhana', 'dukung', 'program', 'perintah', 'pusat', 'presiden', 'prabowo', 'subianto', 'makan', 'siang', 'gratis', 'sarap', 'gratis', 'pagi']*
3	*kejadian program makan bergizi gratis prabowo subianto setuju program beliau out kabinet.*	*['jadi', 'program', 'makan', 'gizi', 'gratis', 'prabowo', 'subianto', 'tuju', 'program', 'beliau', 'out', 'kabinet'].*

Table 4 illustrates how sentences are broken down into lists of individual words or tokens. This format is the final input required before the data can be vectorized or processed by the model.

Sentiment Labeling Using Pretrained Models

After preprocessing is complete, the next stage is sentiment labeling using two pretrained transformer models: RoBERTa Base and IndoBERTweet Base. The use of pretrained models has proven effective for various NLP tasks including sentiment analysis, with advantages in capturing complex semantic contexts [25].

a. RoBERTa Base for Indonesian Language

The RoBERTa Base model used is w11wo/indonesian-roberta-base-sentiment-classifier which is a RoBERTa model that has been fine-tuned specifically for Indonesian sentiment classification. This model is trained using RoBERTa architecture with 12 transformer layers, 768 hidden units, and 12 attention heads [8]. a standard configuration that has proven effective for various NLP tasks. Fine-tuning transformer-based models for Indonesian sentiment analysis has shown very good performance, with previous research achieving accuracy up to 95.16% on Indonesian review datasets [10].

b. IndoBERTweet Base Model

The IndoBERTweet Base model used is Aardiiiiy/indobertweet-base-Indonesian-sentiment-analysis which is an IndoBERTweet variant that has been fine-tuned specifically for sentiment analysis. This model is built based on IndoBERTweet developed by Koto et al. [13], which was specifically trained with Indonesian tweet corpus totaling 200 million tweets. This model has architecture similar to BERT base (12 layers, 768 hidden units) but with vocabulary specially initialized to handle informal Indonesian Twitter language characteristics such as slang, abbreviations, and emoticons.

c. Automatic Labeling Implementation

Sentiment labeling implementation in this research uses automatic functions that iterate for each data row in the CSV file. This process is conducted separately for both models with consistent steps to ensure result comparability. The labeling procedure is designed to process tweets in batches with progress tracking for monitoring the running process.

The implementation stages begin with loading models and tokenizers, loading appropriate pretrained models and tokenizers from Hugging Face Model Hub using the Transformers library [26]. The data used as input is the result of the final preprocessing stage, namely tweet text that has been cleaned of URLs, mentions, hashtags, emoticons, punctuation marks, and special characters. The text is then normalized and goes through the tokenization process, but is rejoined so it remains in complete sentence form. Using text from recombined tokenization as input allows transformer models to tokenize themselves according to their built-in vocabulary and tokenization mechanisms, thus maintaining more complete semantic context.

Next, data iteration is performed by looping for each tweet in the dataframe with progress indicator every 100 tweets for monitoring. Each final result tweet is directly processed by the model tokenizer which automatically converts text into tokens with max_length of 128 tokens, truncation for long tweets, and padding for input consistency [27]. The inference stage performs forward pass through the model to obtain logits, then applies softmax to obtain probability distribution [28]. In the prediction stage, the class with highest probability is taken as predicted label and confidence score is saved.

The labeling process produces two separate CSV files: final_result(model roberta).csv for RoBERTa Base labeling results and final_result(model tweetbert).csv for IndoBERTweet Base labeling results. Each file contains three main columns: tweet (cleansed text), label (Negative/Neutral/Positive), and accuracy (confidence score with range 0-1).

d. Merging Labeling Results

After both models produce sentiment labels for the entire dataset, the next stage is merging the labeling results from both CSV files based on the same tweet. The merge process is conducted using the tweet column as key, producing one combined dataframe with structure: tweet, label_roberta, label_tweetbert, confidence_roberta, and confidence_tweetbert. The merged data structure is displayed in **Table 5**.

Table 5. Structure of Merged Labeling Data

Column	Description	Data Type
tweet	Cleansed tweet text	String
label_roberta	Sentiment label from RoBERTa Base	Categorical (Negative/Neutral/Positive)
label_indobertweet	Sentiment label from IndoBERTweet Base	Categorical (Negative/Neutral/Positive)

Based on **Table 5**, the merged dataset consists of five main columns that align the prediction results from both models for the same tweet. This structure facilitates direct row-by-row comparison for evaluation metrics such as agreement rate and confusion matrix.

The merge process ensures that each tweet has labeling results from both models on the same row, allowing direct comparison between RoBERTa Base and IndoBERTweet Base predictions. Confidence scores from both models indicate the model's confidence level toward given predictions, where values approaching 1 indicate high confidence and values approaching 0 indicate uncertainty in classification [29]. This merged dataset becomes the basis for all comparison analyses conducted in the evaluation stage.

This study adopts a zero-shot labeling approach where pretrained models are applied directly without fine-tuning on domain-specific annotated data. This approach is chosen for several reasons: (1) it reflects real-world scenarios where labeled training data for specific policy topics may not be available, (2) it enables fair comparison of models' out-of-the-box capabilities, and (3) it reduces potential bias from subjective manual annotation. However, this approach has limitations: the absence of human-annotated ground truth means that performance cannot be evaluated against a gold standard, and the comparison relies on agreement metrics rather than accuracy measures.

Comparison Evaluation Methods

After both models produce sentiment labels for the entire dataset, comprehensive evaluation is conducted to compare their performance. Evaluation is conducted in several aspects to provide comprehensive understanding of each model's characteristics.

a. Agreement Rate

Agreement rate measures the percentage of tweets that receive the same label from both models. This metric provides a general picture of classification consistency between RoBERTa Base and IndoBERTweet Base. The agreement rate calculation formula is shown in equation (1).

$$Agreement\ Rate = \left(\frac{Number\ of\ Tweets\ with\ Same\ Label}{\text{Total Tweets}}\right) \times 100\% \quad (1)$$

Agreement rate around 65-80% is considered good for automatic sentiment classification, while minimum level of 50% is still considered effective for initial assessment [30]. Higher agreement rate values indicate better consistency in sentiment interpretation between the two models [30].

b. Cohen's Kappa Score

Cohen's Kappa Score measures the agreement level between two raters (in this case two models) while accounting for the possibility of agreement occurring by chance (chance agreement) [31]. The Cohen's Kappa Score formula is shown in equation (2).

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (2)$$

Where Pr(a) is observed agreement (proportion of observed agreement) and Pr(e) is expected agreement by chance (proportion of agreement expected by chance). Cohen's Kappa is designed to eliminate the possibility of agreement occurring by chance due to random guessing [32]. Interpretation of Kappa values based on Landis and Koch [32] widely used in classification research is as follows: values less than 0 indicate poor agreement, values 0.00-0.20 indicate slight agreement, values 0.21-0.40 indicate fair agreement, values 0.41-0.60 indicate moderate agreement, values 0.61-0.80 indicate substantial agreement, and values 0.81-1.00 indicate almost perfect agreement [32].

c. Confusion Matrix

Confusion matrix is used for detailed visualization of agreement and disagreement patterns between the two models in multiclass classification tasks. Confusion matrix is an evaluation method used to measure classification model performance, providing complete information about correct and incorrect predictions for each class [33].

d. Confidence Score Analysis

e. Descriptive statistical analysis is conducted on confidence scores from each model to understand prediction confidence levels. Analysis includes mean, median, standard deviation, minimum, maximum, and confidence score distribution in high (>0.95), medium (0.80-0.95), and low (<0.80) categories.

f. Disagreement Pattern Analysis

In-depth analysis is conducted on tweets that receive different labels from both models to understand systematic differences patterns. Analysis includes disagreement pattern frequency, linguistic characteristics, qualitative sample analysis, and identification of strengths and weaknesses of each model.

g. Testing Scenario

This research uses a zero-shot learning approach, where both pretrained models are directly applied to the MBG Program dataset without additional fine-tuning, chosen to evaluate pretrained model generalization ability, save time and computational resources, and assess model readiness for real-world applications where fine-tuning with domain-specific data is not always feasible. All experiments were conducted using Google Colab with Python 3.12, PyTorch 2.8, and Transformers library 4.57 from Hugging Face, utilizing 12.67 GB RAM with CPU as processing device, completing the sentiment labeling process for 1,831 tweets using both models in approximately 45 minutes.

3. Result and Discussion

From 1,831 tweets collected, all data successfully passed the preprocessing stage with 3 data losses. The preprocessing process produced clean dataset ready for sentiment labeling. Labeling using both models required total time of approximately 45 minutes for 1,831 tweets, with average of 1.5 seconds per tweet including tokenization and inference time.

Sentiment Distribution

Sentiment label distribution produced by each model shows different characteristics. Distribution results are displayed in **Table 6**.

Table 6. Sentiment Distribution RoBERTa Base vs IndoBERTweet Base

Label	RoBERTa Base	IndoBERTweet Base
Negative	137 (7.48%)	248 (13.54%)
Neutral	1.377 (75.20%)	1.248 (68.16%)
Positive	317 (17.31%)	335 (18.30%)
Total	1.831 (100%)	1.831 (100%)

Based on **Table 6**, RoBERTa Base shows very conservative tendencies with 75.20% tweets labeled as Neutral. This indicates the model tends to play it safe and only gives Positive or Negative labels to tweets with very explicit sentiment. Conversely, IndoBERTweet Base has more balanced distribution with 68.16% Neutral, showing this model is more sensitive in detecting sentiment nuances. Significant difference is seen in the Negative label, where IndoBERTweet detects nearly twice as many negative sentiments (13.54%) compared to RoBERTa (7.48%). This shows IndoBERTweet is more capable of capturing negative expressions in informal Indonesian language common on social media.

Word Cloud Sentiment Visualization

Word cloud analysis was conducted to visualize dominant themes in each sentiment category (**Figure 2**). The positive sentiment is characterized by words like "prabowo," "*semangat*" (spirit), and "*dukung*" (support), indicating public appreciation. In contrast, negative sentiment displays words such as "*sampah*" (garbage), "*korupsi*" (corruption), and "*bohong*" (lie), reflecting criticism. Meanwhile, neutral sentiment focuses on informative terms like "*anggaran*" (budget), "*pemerintah*" (government), and "program," showing factual discussion without clear bias.

(a) (b) (c)

Figure 2. Word Cloud of Positive (a), Negative (b), and Neutral (c) Sentiments

Figure 2 visualizes these frequent terms where larger font sizes indicate higher occurrence frequency. The distinct vocabulary separation across the three clouds confirms that the classification models successfully captured coherent thematic content, distinguishing clearly between supportive expressions, specific grievances, and objective policy discussions.

Agreement Rate Analysis

Agreement analysis results show that both models have fairly good conformity level but with room for improvement. **Table 7** displays agreement rate results.

Table 7. Agreement Rate Results

Metric	Number	Percentage
Total Tweets	1,831	100%
Agreement	1,415	77.28%
Disagreement	416	22.72%
Cohen's Kappa	0.490	Moderate

Agreement Rate Analysis **Table 7** shows an agreement rate of 77.28%, meaning the models agree on approximately three out of four tweets. The Cohen's Kappa Score of 0.490 indicates Moderate Agreement. This suggests that while there is consistency, systematic differences exist in how each model interprets sentiment, likely due to RoBERTa's general-purpose nature versus IndoBERTweet's domain-specific training. Highest agreement was found in the Neutral label (61.88%), while the Negative label showed the most divergence. Detailed confusion matrix is displayed and visualized in **Figure 3** to provide visualization of agreement and disagreement patterns.

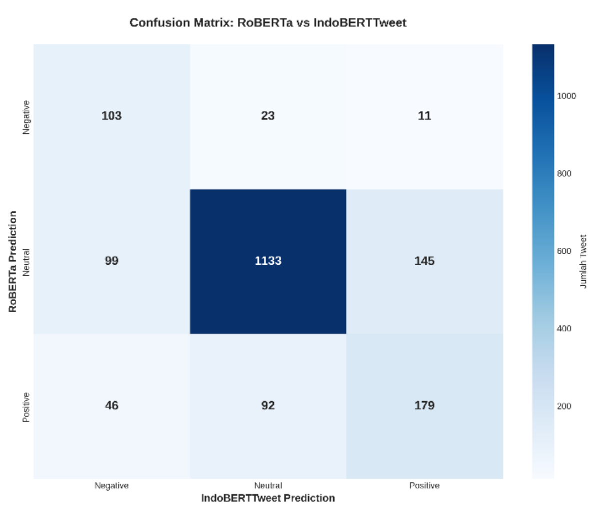

Figure 3. Confusion Matrix: RoBERTa Base vs IndoBERTweet Base.

Confusion Matrix Analysis **Figure 3** illustrates the prediction distribution. The main diagonal (1,415 tweets) represents agreement. The most significant disagreement pattern (7.92%) involves RoBERTa labeling tweets as Neutral while IndoBERTweet labels them Positive. This often occurs in tweets expressing implicit appreciation. The second major pattern (5.41%) sees RoBERTa labeling Neutral where IndoBERTweet detects Negative sentiment, highlighting IndoBERTweet's superior sensitivity to sarcasm and implicit criticism.

Analysis of Tweets with Different Labels

An analysis of 416 discordant tweets reveals key differences in model sensitivity. The largest disagreement pattern involves 145 tweets (7.92%) labeled Neutral by RoBERTa but Positive by IndoBERTweet. For example, the tweet 'presiden prabowo berhasil angkat derajat bangsa...' was labeled Neutral by RoBERTa (conf: 0.64) but Positive by IndoBERTweet (conf: 0.98), demonstrating IndoBERTweet's superior ability to capture implicit appreciation using context clues like 'berhasil' and 'diapresiasi'. Similarly, 99 tweets (5.41%) labeled Neutral by RoBERTa were classified as Negative by IndoBERTweet. A representative case, '*makan bergizi gratis* berubah makan sampah gratis...', was missed by RoBERTa (Neutral, conf: 0.67) but correctly identified as Negative by IndoBERTweet (conf: 0.98), highlighting its advantage in detecting sarcasm and irony typical of Indonesian social media.

Performance Evaluation Based on Confidence Score

Confidence score analysis provides insights into prediction confidence levels of each model. Statistics are displayed in **Table 8**.

Table 8. Confidence Score Statistics

Metric	RoBERTa Base	IndoBERTweet Base
Mean	0.9559	0.9236
Median	0.9976	0.9854
Std Deviation	0.1025	0.1243
Min	0.4686	0.3848
Max	0.9995	0.9991
Confidence > 0.99	1,281 (69.96%)	775 (42.33%)
Confidence > 0.95	1,520 (83.01%)	1,257 (68.65%)
Confidence > 0.90	239 (13.05%)	418 (22.83%)
Confidence > 0.80	162 (8.85%)	250 (13.65%)

Table 8 compares the confidence metrics of both models. RoBERTa Base shows consistently higher confidence scores compared to IndoBERTweet Base. RoBERTa mean confidence (95.59%) is 3.23 percentage points higher than IndoBERTweet (92.36%). This difference is statistically and practically significant. RoBERTa standard deviation (0.1025) is lower than IndoBERTweet (0.1243), indicating RoBERTa predictions are more consistent and stable. This is a common characteristic of more conservative models; when the model decides to give a certain label, it is very confident in its decision.

Confidence score distribution shows 69.96% of RoBERTa predictions have confidence above 99%, while IndoBERTweet only 42.33%. Conversely, IndoBERTweet has more predictions with low confidence (less than 90%): 22.83% versus 13.05% for RoBERTa. This pattern confirms the trade-off between conservatism and sensitivity, where RoBERTa is more confident but less sensitive, while IndoBERTweet is more sensitive but with more varied confidence.

Discussion

Model Characteristics, Research results confirm that RoBERTa Base and IndoBERTweet Base have different characteristics reflecting fundamental trade-offs in sentiment analysis. RoBERTa Base shows high precision and conservative characteristics, where the model excels in precision with low false positive rate, has high and consistent confidence scores, very skewed distribution to Neutral (75.20%), and suitable for applications prioritizing certainty.

This conservative tendency aligns with findings from previous studies showing that general-purpose multilingual models tend to favor majority classes in zero-shot settings [34]. Conversely, IndoBERTweet Base shows high sensitivity and balanced characteristics, where the model excels in recall and sensitivity, capable of detecting implicit sentiment and sarcasm, has more balanced distribution (68.16% Neutral), and better for exploration and maximum coverage This behavior is consistent with prior research indicating that domain-specific models trained on social media data demonstrate higher sensitivity to informal language patterns [13].

Domain Specificity versus Generalization, Performance differences reflect trade-offs between domain specificity and generalization capability. IndoBERTweet trained specifically on Indonesian tweets has advantages in understanding slang and colloquial language, detecting sarcasm and irony, and recognizing Indonesian Twitter-specific communication patterns. However, RoBERTa with multilingual basis and larger training data has more robust semantic representation, better generalization on formal language, and higher consistency. Similar trade-offs have been observed in comparative studies between multilingual and monolingual BERT models for low-resource languages [35].

Implications for Public Policy Sentiment Analysis , In the context of public policy sentiment analysis such as the *Makan Bergizi Gratis* Program, model selection must be adjusted to specific application needs. The dominance of neutral sentiment (68-75%) in our results is consistent with previous Indonesian sentiment analysis studies on political topics, where public opinion often shows mixed or ambivalent responses to government policies [36], [37]. RoBERTa Base is recommended if priority is avoiding false alarms, crisis or sensitive issue monitoring is required, decisions require high confidence, and resources are limited for manual verification. IndoBERTweet Base is recommended if the goal is in-depth public opinion exploration, wanting to capture the widest sentiment spectrum possible, resources are available for manual verification, and analysis is academic or research-oriented. Ensemble approach using both models is recommended if balanced approach is needed, budget and resources are adequate, application is production-critical, and using voting mechanism or stacking.

4. Conclusion

This study compares RoBERTa Base and IndoBERTweet Base on 1,831 tweets regarding the Makan Bergizi Gratis Program. Results show RoBERTa is highly conservative (75.20% Neutral) with superior confidence (mean 0.9559), while IndoBERTweet is more sensitive (68.16% Neutral) in detecting implicit sentiments. The moderate agreement of 77.28% (Cohen’s Kappa 0.490) reveals systematic differences, where IndoBERTweet significantly outperforms RoBERTa in identifying implicit positive sentiment (7.92% disagreement) and sarcasm (5.41% disagreement), making it more effective for capturing nuanced public opinion.

This study has several limitations that should be acknowledged. First, the use of zero-shot labeling without human-annotated ground truth means that the comparison is based on agreement metrics rather than accuracy against a gold standard. Second, the absence of manual validation limits our ability to determine which model produces more "correct" labels when disagreements occur. Third, the dataset is limited to a specific time period and may not capture the full evolution of public opinion on the MBG program.

Therefore, RoBERTa Base is recommended for high-precision monitoring to minimize false positives, whereas IndoBERTweet Base is optimal for research requiring high sensitivity. For balanced performance, an ensemble approach is advised. Future work should focus on fine-tuning with domain-specific datasets, validating results with human-annotated samples to establish ground truth accuracy metrics and exploring advanced ensemble strategies or multimodal analysis to further enhance sentiment detection accuracy in the Indonesian context.

References:

[1] A. A. Saleh, S. Abdullah, R. Muhammad, M. Y. Amir, and S. Sulvinajayanti, “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, Sep. 2025, doi: 10.30597/mkmi.v21i3.46204.

[2] A. Koswara and L. Herlina, “A Collaborative Model for Funding Indonesiaâ€™s MBG Program Through Government and Philanthropy Partnerships,” *J. Islam. Econ. Philanthr.*, vol. 7, no. 4, pp. 266–288, 2025, doi:

10.21111/jiep.v7i4.13960.

[3] E. Wildanu and H. Nauli Fitri Agasya, “Presidential Election Campaign of the Republic of Indonesia via Twitter (X),” *J. Polisci*, vol. 1, no. 3, pp. 132–141, 2024, doi: 10.62885/polisci.v1i3.159.

[4] E. Haddi, X. Liu, and Y. Shi, “The Role of Text Pre-processing in Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[5] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications : A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[6] K. Hilary, K. Fitri, I. Anindaputri, P. Aziza, F. Aji, and I. Anindaputri, “ScienceDirect Performance Performance comparison comparison of of deep deep learning learning approaches approaches for for Indonesian twitter hate speech detection using IndoBERTweet Indonesian twitter hate speech detection using IndoBERTweet embedding e,” *Procedia Comput. Sci.*, vol. 269, pp. 1663–1671, 2025, doi: 10.1016/j.procs.2025.09.109.

[7] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, “Improving sentiment classification using a RoBERTa-based hybrid model,” *Front. Hum. Neurosci.*, vol. 17, no. December, pp. 1–10, 2023, doi: 10.3389/fnhum.2023.1292010.

[8] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” no. 1, Jul. 2019, doi: 10.48550/arXiv.1907.11692.

[9] F. Koto, J. H. Lau, and T. Baldwin, “IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.

[10] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, “Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN,” *Ilk. J. Ilm.*, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[11] R. Mukarramah, D. Atmajaya, and L. B. Ilmawan, “Performance comparison of support vector machine (SVM) with linear kernel and polynomial kernel for multiclass sentiment analysis on twitter,” *Ilk. J. Ilm.*, vol. 13, no. 2, pp. 168–174, 2021, doi: 10.33096/ilkom.v13i2.851.168-174.

[12] D. Indra, R. Ramdaniah, and W. Sukur, “Analysis of Hybrid Learning Sentiment among Information Systems Students using The Naïve Bayes Classifier,” *J. ELTIKOM*, vol. 8, no. 2, pp. 91–99, Dec. 2024, doi: 10.31961/eltikom.v8i2.1144.

[13] 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.

[14] M. R. Setyawan, ; Fajar, R. B. Putra, and A. Ramadhani, “Sentiment Analysis towards Jokowi Post-Presidential Term Using CNN-BiLSTM with Multi-head Attention on Platform X,” *Ilk. J. Ilm.*, vol. 17, no. 2, pp. 150–161, 2025, doi: 10.33096/ilkom.v17i2.2843.150-161.

[15] H. R. P. Sianturi, “Politics on a plate: equivocal communication in Indonesian presidential nutrition policy,” *Front. Commun.*, vol. 10, no. September 2025, pp. 1–7, 2025, doi: 10.3389/fcomm.2025.1612652.

[16] M. K. Anam, T. A. Fitri, A. Agustin, L. Lusiana, M. B. Firdaus, and A. T. Nurhuda, “Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm,” *Ilk. J. Ilm.*, vol. 15, no. 2, pp. 290–302, 2023, doi: 10.33096/ilkom.v15i2.1590.290-302.

[17] Hengky Triyo, Aji Primajaya, and Purwantoro, “Public Sentiment Analysis on the Increase of Value Added Tax in Indonesia Through Tweet-Harvest,” *J. Artif. Intell. Eng. Appl.*, vol. 5, no. 1, pp. 1996–2000, Oct. 2025, doi: 10.59934/jaiea.v5i1.1772.

[18] R. Rahmaddeni, M. K. Anam, Y. Irawan, S. Susanti, and M. Jamaris, “Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination,” *Ilk. J. Ilm.*, vol. 14, no. 1, pp. 32–38, 2022, doi: 10.33096/ilkom.v14i1.1090.32-38.

[19] T. A. Putra, V. Ariandi, and S. Defit, "Enhancing Accuracy by Using Boosting and Stacking Techniques on the Random Forest Algorithm on Data from Social Media X," *Ilk. J. Ilm.*, vol. 16, no. 2, pp. 184–189, 2024, doi: 10.33096/ilkom.v16i2.2058.184-189.

[20] S. Suswadi and M. Erkamim, "Sentiment Analysis of Shopee App Reviews Using Random Forest and Support Vector Machine," *Ilk. J. Ilm.*, vol. 15, no. 3, pp. 427–435, 2023, doi: 10.33096/ilkom.v15i3.1610.427-435.

[21] I. Irawanto, C. Widodo, A. Hasanah, P. A. Dharma Kusumah, K. Kusrini, and K. Kusnawi, "Sentiment Analysis and Classification of Forest Fires in Indonesia," *Ilk. J. Ilm.*, vol. 15, no. 1, pp. 175–185, 2023, doi: 10.33096/ilkom.v15i1.1337.175-185.

[22] L. B. Ilmawan, M. Muladi, and D. D. Prasetya, "Feature Space Augmentation for Negation Handling on Sentiment Analysis," *Ilk. J. Ilm.*, vol. 15, no. 2, pp. 353–357, 2023, doi: 10.33096/ilkom.v15i2.1695.353-357.

[23] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.

[24] I. As'ad, M. A. Asis, H. M. Pakka, R. Mursalim, and Y. binti M. Noor, "K-Nearest Neighbors Analysis for Public Sentiment towards Implementation of Booster Vaccines in Indonesia," *Ilk. J. Ilm.*, vol. 15, no. 2, pp. 365–372, 2023, doi: 10.33096/ilkom.v15i2.1561.365-372.

[25] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024, doi: 10.11591/eei.v13i5.8032.

[26] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, May 2019, doi: 10.48550/arXiv.1810.04805.

[28] E. P. A. Akhmad, "Sentiment Analysis of DLU Ferry Application Reviews on the Google Play Store Using Bidirectional Encoder Representations from Transformers," *J. Apl. PELAYARAN DAN KEPELABUHANAN*, vol. 13, no. 2 SE-Articles, pp. 104–112, Mar. 2023, doi: 10.30649/japk.v13i2.94.

[29] J. Grabinski, P. Gavrikov, J. Keuper, and M. Keuper, "Robust Models are less Over-Confident," *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, Dec. 2022, doi: 10.48550/arXiv.2210.05938.

[30] D. Contreras, S. Wilkinson, E. Alterman, and J. Hervás, "Accuracy of a pre-trained sentiment analysis (SA) classification model on tweets related to emergency response and early recovery assessment: the case of 2019 Albanian earthquake," *Nat. Hazards*, vol. 113, no. 1, pp. 403–421, 2022, doi: 10.1007/s11069-022-05307-w.

[31] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012, doi: 10.11613/BM.2012.031.

[32] M. Albrektsson, O. Wolf, A. Enocson, and M. Sundfeldt, "Validation of the classification of surgically treated acetabular fractures in the Swedish Fracture Register," *Injury*, vol. 53, no. 6, pp. 2145–2149, 2022, doi: 10.1016/j.injury.2022.03.002.

[33] A. Michael and M. Garonga, "Classification model of 'Toraja' arabica coffee fruit ripeness levels using convolution neural network approach," *Ilk. J. Ilm.*, vol. 13, no. 3, pp. 226–234, 2021, doi: 10.33096/ilkom.v13i3.861.226-234.

[34] S. Wu and M. Dredze, "Are All Languages Created Equal in Multilingual BERT?," in *Proceedings of the 5th Workshop on Representation Learning for NLP*, S. Gella, J. Welbl, M. Rei, F. Petroni, P. Lewis, E. Strubell, M. Seo, and H. Hajishirzi, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2020, pp. 120–130. doi: 10.18653/v1/2020.repl4nlp-1.16.

[35] R. Joshi, "L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources," *6th Work. Indian Lang. Data Resour. Eval. WILDRE 2022 - held conjunction with*

Int. Conf. Lang. Resour. Eval. Lr. 2022 - Proc., no. June, pp. 97–101, Mar. 2022.

[36] F. Z. Lubis and R. Kurniawan, "Sentiment analysis towards naturalization of Indonesian National Team Players on social media x using the Naive Bayes method," *J. Mandiri IT*, vol. 14, no. 1, pp. 208–218, 2025, doi: 10.35335/mandiri.v14i1.412.

[37] N. K. Ode, D. Indra, and R. Ramdaniah, "Sentiment Analysis of Public Opinion on Deforestation in Papua on YouTube Platform Using Long Short-Term Memory (LSTM) Method," *G-Tech J. Teknol. Terap.*, vol. 9, no. 4, pp. 1878–1888, Oct. 2025, doi: 10.70609/g-tech.v9i4.7855.