Vol. 8, No. 1, February 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Sentiment Analysis of the Free Nutritious Meal Program on Twitter Using Naive Bayes and IndoBERT-Based Labeling

Mikel Frewinta Manao[1], Sri Mujiyono[2]

[1,2]Informatics Engineering Department, Universitas Ngudi Waluyo, Indonesia

Received:
November 11, 2025
Revised:
December 1, 2025
Accepted:
February 13, 2026
Published:
March 7, 2026

Corresponding Author:
Author Name*:
Mikel Frewinta Manao
Email*:
mikelfremanao@gmail.com

DOI:
10.63158/journalisi.v8i1.1345

Abstract. The Free Nutritious Meal Program is a government initiative aimed at improving the nutritional status of primary school children in Indonesia. However, its implementation has generated diverse public reactions on the X/Twitter platform, making systematic sentiment analysis essential for policy evaluation. This study analyzes public sentiment using two labeling approaches—translation-based TextBlob and IndoBERT contextual labeling—combined with Naïve Bayes and Linear SVC classifiers. A total of 2,903 Indonesian-language tweets were collected, preprocessed, and classified to compare the performance impact of each labeling method. The evaluation was conducted using accuracy, precision, recall, and macro F1-score. Sentiment distribution under IndoBERT indicates a predominance of negative and neutral opinions, particularly related to budget concerns, implementation quality, and food distribution issues. This study is subject to several limitations. The dataset size (2,903 tweets) and restricted temporal window may limit the generalizability of findings to long-term public discourse. The analysis also relies on a single social media platform (X/Twitter), excluding perspectives from other platforms such as Instagram or TikTok. Moreover, although IndoBERT improves contextual understanding, transformer-based labeling still may not fully capture sarcasm or highly colloquial expressions. Despite these limitations, the study demonstrates the effectiveness of combining Indonesian transformer models with conventional classifiers to support data-driven policy evaluation.

Keywords: Free Nutritious Meal Program (MBG), Sentiment Analysis, IndoBERT, Naïve Bayes, Linear SVC

1. INTRODUCTION

The Free Nutritious Meal Program has become one of Indonesia's most prominent strategic policies for improving the nutritional status of primary school children and strengthening long-term human capital development. Despite its policy importance and potential benefits for health and education, the program has also generated intense public debate. On social media, especially X (formerly Twitter), discussions about the program reveal not only support but also criticism, skepticism, concerns over budget allocation, food quality, safety, distribution, and implementation effectiveness. This growing polarization shows that public perception is not a peripheral issue, but a central factor that can influence policy acceptance, legitimacy, and sustainability. Because X functions as an open and dynamic space for policy discourse, the large volume of real-time conversations provides a valuable opportunity to capture emerging public reactions and concerns in a way that conventional surveys may not fully achieve [1], [2].

Although public debate around the program is highly visible online, systematic empirical evidence that maps sentiment toward the Free Nutritious Meal Program using digital trace data remains limited. This creates an important research problem. Without a structured and data-driven understanding of how people respond to the program, policymakers may struggle to distinguish between isolated opinions and broader patterns of acceptance or resistance. The urgency of this issue is amplified by the nature of social media itself, where millions of short posts are produced every day, often using informal vocabulary, abbreviations, sarcasm, and context-dependent expressions. These linguistic characteristics make manual interpretation difficult and may reduce the reliability of policy evaluation if public reactions are not analyzed using computational methods specifically designed for noisy social media text [6], [7]. In this context, sentiment analysis offers a practical way to transform large-scale online discussions into interpretable evidence about public opinion.

Sentiment analysis has been widely recognized as an effective approach for extracting opinion patterns from social media data by classifying textual content into positive, negative, and neutral categories. In Indonesia, previous studies have shown that machine-learning algorithms such as Naïve Bayes remain relevant because they are simple, efficient, and reliable for large-scale text classification tasks [3], [5]. Other studies have

also demonstrated the usefulness of sentiment analysis for examining societal responses to public issues. For example, Mubarok [1] analyzed netizens' opinions regarding large-scale social restrictions, while Rasiban and Riyadi [3] compared Naïve Bayes and SVM for identifying public reactions to healthcare service applications. These studies confirm that sentiment analysis can support the interpretation of public discourse and that traditional machine-learning methods still provide meaningful performance, especially when computational efficiency and model transparency are important.

However, the existing literature still leaves a significant methodological gap. Prior Indonesian studies on sentiment analysis have applied methods such as Naïve Bayes, SVM, BiLSTM, and transformer-based models across domains including policy, healthcare, and digital services, yet most of them rely on only one labeling mechanism, typically either a lexicon-based approach or a contextual transformer-based approach, without examining how the choice of labeling strategy affects downstream classification performance. This limitation is important because the quality of labels directly influences the quality of the sentiment model. In particular, translation-based pseudo-labeling may introduce semantic distortion or polarity shifts when Indonesian expressions are translated before sentiment scoring, whereas contextual models such as IndoBERT are better positioned to capture local linguistic nuance, slang, and context in Indonesian social media language. Even though Python and its ecosystem—supported by tools such as NLTK, spaCy, and scikit-learn—have made this type of analysis increasingly accessible in academic and applied settings [4], [5], the comparative impact of different labeling approaches remains underexplored.

Based on this gap, this study offers a clear novel contribution. To the best of current knowledge, no previous Indonesian study has compared translation-based pseudo-labeling and IndoBERT contextual labeling in the context of sentiment analysis on national policy discussions related to the Free Nutritious Meal Program. This study therefore does not only examine public sentiment toward an important government program, but also tests how two different labeling strategies shape the effectiveness of sentiment classification. By combining TextBlob translation-based pseudo-labeling and IndoBERT contextual labeling with Naïve Bayes and Linear SVC classifiers, this research provides a comparative framework for identifying which approach is more suitable for Indonesian policy discourse on social media. This comparison is valuable both

methodologically and substantively: methodologically, it contributes to the development of more robust Indonesian-language sentiment analysis pipelines; substantively, it provides evidence-based insight into how the public perceives a major nutrition policy in real time.

Accordingly, the objective of this study is twofold: first, to identify patterns of public sentiment toward the Free Nutritious Meal Program on X; and second, to evaluate the comparative performance of two labeling approaches and two classification algorithms in order to produce a more comprehensive and accurate sentiment map. Through this contribution, the study seeks to strengthen the use of computational social media analytics for policy evaluation and to expand the empirical foundation of Indonesian-language sentiment analysis research [1]–[7].

2. METHODS

This study employed a quantitative research design using a text mining approach within an ex post facto framework, in which researchers analyzed naturally occurring data without manipulating any variables. The study is computational in nature because it applies natural language processing (NLP) and machine learning algorithms to classify public sentiment toward the Free Nutritious Meal Program based on user-generated content on social media. This design was selected because it enables large-scale, systematic, and objective analysis of public opinion expressed in digital environments, in line with common procedures in text data management and text mining research [8].

2.1. Data Source and Research Subject

The data source of this study consisted of public posts uploaded on the X/Twitter platform discussing the Free Nutritious Meal Program. In this context, the research subjects were not human participants in the conventional survey sense, but rather tweets produced by users who voluntarily expressed opinions about the policy online. Accordingly, the unit of analysis in this study was the individual tweet. The sampling technique used was purposive keyword-based sampling, designed to collect only posts directly relevant to the research topic. Data were retrieved using the keywords “makan bergizi gratis” and “makanbergizigratis”. To maintain contextual relevance, only tweets written in Indonesian (lang:id) and posted within a predetermined time window were

included. This restriction was intended to ensure that the dataset reflected discussion within a specific policy context rather than unrelated or multilingual discourse.

2.2. Research Instruments

The research instruments consisted of both data collection tools and analytical tools. Data collection was conducted using a Python-based scraping script executed in Google Colab, with the Tweet-Harvest module serving as the primary scraping utility. The collected data were stored in CSV format for subsequent processing and analysis. The analytical environment was built in Python and supported by several libraries. Sastrawi was used for Indonesian stemming, scikit-learn for feature extraction and machine learning implementation, TextBlob for translation-based pseudo-labeling, wordcloud and visualization libraries for result presentation, and the w11wo/indonesian-roberta-base-sentiment-classifier model for contextual sentiment labeling.

2.3. Research Procedure

The research procedure consisted of several sequential stages, beginning with problem identification, continuing through data collection, noise filtering, preprocessing, feature extraction, sentiment labeling, model training and testing, and ending with result visualization and interpretation. The overall workflow follows a standard social media sentiment analysis pipeline in which raw textual data are transformed into structured representations for classification and evaluation [9].

Figure 1. Research Flowchart

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2.4. Data Collection and Research Period

Data were collected from the X/Twitter platform using the Tweet-Harvest scraping module executed in Google Colab. The collection process employed a purposive keyword-based sampling technique using the keywords "makan bergizi gratis" and "makanbergizigratis" to capture tweets specifically related to the Free Nutritious Meal Program. To maintain contextual relevance, only tweets written in Indonesian (lang:id) were included in the dataset.

The data collection process was conducted within the overall research period from December 2024 to March 2025, while the tweet scraping and dataset preparation were restricted to a predetermined observation window within that period in order to reflect public discourse in a specific policy context. This restriction was important to ensure that the collected tweets represented reactions that were temporally relevant to the implementation and discussion of the program, rather than unrelated or outdated conversations. After the scraping stage, the collected tweets underwent cleaning and validation procedures to remove irrelevant and low-quality entries. Following these procedures, the final dataset consisted of 2,903 unique Indonesian-language tweets that were considered suitable for sentiment analysis. The resulting corpus represents public discussion on the Free Nutritious Meal Program during the selected data collection period.

2.5. Noise Filtering and Dataset Integrity

To ensure the validity of the corpus, several filtering procedures were applied to remove content that could distort sentiment distribution, including retweets, duplicate posts, spam-like content, and bot-generated tweets. This stage was essential because raw social media data often contain repeated, automated, or non-substantive content that can bias machine learning models.

First, retweets and quote retweets were handled to reduce repetition. Standard retweets in the form of RT @username were excluded because they merely duplicated existing content and could artificially inflate the visibility of particular opinions. Quote retweets were retained only when the added text exceeded a defined character threshold, indicating that the user had contributed a substantial original comment rather than simply resharing a message.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Second, duplicate detection was performed to identify repeated posts generated either manually or automatically. Exact duplicates were removed using text hashing, while near-duplicate tweets were detected through normalized text similarity using a cosine similarity threshold of 0.90 on TF-IDF vectors. Third, spam filtering was applied to remove tweets containing excessive promotional links, unrelated hashtags, random character strings, or promotional keywords such as “promo,” “link di bio,” and “giveaway.” In addition, accounts posting at unusually high frequency within short time intervals were flagged as potentially spam-like.

Fourth, a heuristic bot filtering strategy was implemented. Although external tools such as Botometer were not used due to platform and access constraints, several indicators were employed to identify bot-like accounts, including extremely high posting rates, usernames with long numeric sequences, default profile images, low follower-to-following ratios, and highly repetitive posting patterns. Tweets originating from accounts meeting multiple bot-like criteria were excluded from the analysis. Finally, a manual integrity check was conducted through random sampling of the cleaned dataset to verify that irrelevant content had been removed and that the remaining corpus still reflected diverse and organically generated public opinion.

2.6. Text Preprocessing

Before modeling, all tweets underwent a structured preprocessing pipeline to standardize the text and reduce noise. The preprocessing steps included the removal of URLs, mentions, hashtags, punctuation, numbers, and non-relevant symbols, followed by the elimination of empty or non-informative entries. Because Indonesian social media text frequently contains slang, abbreviations, and informal orthography, slang normalization was performed using a custom normalization dictionary. The cleaned text then underwent stemming using the Sastrawi library, followed by stopword removal using an Indonesian stopword list adapted from NLTK and enriched with locally relevant stopwords. These preprocessing steps were intended to improve textual consistency and support more reliable downstream sentiment classification [8].

2.7. Feature Extraction

Two feature extraction strategies were used to match the characteristics of the selected classifiers. For the Naïve Bayes model, tweets were transformed into TF-IDF vectors, which represent the relative importance of words within the corpus. To improve vocabulary quality and reduce sparsity, the TF-IDF vectorizer was configured with min_df = 3 and max_features = 10,000. After filtering, the effective vocabulary size for this representation was approximately 7,200 features. For the Linear SVC model, a hybrid representation combining word n-grams and character n-grams was used. Specifically, word n-grams (1–2) were employed to capture lexical and short phrase patterns, while character n-grams (4–5) were used to improve robustness against spelling variation, expressive elongation, and informal writing patterns frequently observed in social media text. This feature space produced approximately 28,000–32,000 sparse features, depending on the labeling scheme used. The use of Naïve Bayes and Support Vector Machine/Linear SVC was motivated by prior studies showing that both algorithms remain effective for Twitter sentiment classification, with SVM generally offering stronger discriminative performance and Naïve Bayes providing simplicity and computational efficiency [10].

2.8. Sentiment Labeling

This study compared two sentiment labeling approaches in order to evaluate how different annotation strategies affect classification performance. The first approach was translation-based pseudo-labeling using TextBlob. In this method, Indonesian tweets were translated into English before sentiment polarity was estimated using TextBlob's polarity scoring mechanism. The resulting scores were then mapped into positive, neutral, and negative categories. This approach is practical and computationally lightweight, but it is vulnerable to semantic drift because sentiment is inferred after translation rather than directly from the source language. The use of lexicon-based or pseudo-labeling approaches remains relevant in sentiment analysis research, particularly when manually labeled Indonesian datasets are limited [11]. The second approach used contextual labeling with the Indonesian transformer model w11wo/indonesian-roberta-base-sentiment-classifier. This model was selected because it is specifically trained for Indonesian sentiment classification and is better suited to capturing contextual meaning, slang, idiomatic expressions, and social-media-specific language patterns. At the time of experimentation, the model version used was v1.0.

Several examples observed during experimentation illustrate the limitations of translation-based polarity scoring. For instance, the phrase "tidak buruk" may be translated as "not bad," which TextBlob often classifies as positive even though in Indonesian discourse it may function as neutral or mildly negative depending on context. Likewise, "kurang bagus" may be interpreted as neutral despite conveying dissatisfaction, while sarcastic statements such as "Mantap, anggaran bocor lagi" may be misread as positive due to the literal interpretation of the word "mantap."

2.9. Model Training and Testing

Two traditional machine learning classifiers were used in this study: Complement Naïve Bayes and Linear Support Vector Classifier (Linear SVC). Each classifier was trained under both labeling schemes, resulting in four experimental configurations: TextBlob + Naïve Bayes, TextBlob + Linear SVC, IndoBERT + Naïve Bayes, and IndoBERT + Linear SVC. The dataset was divided using an 80:20 train-test split, where 80% of the data were used for model training and 20% for testing. Model performance was assessed using accuracy, precision, recall, and macro-averaged F1-score, because these metrics provide a more balanced assessment across sentiment classes, especially in the presence of class imbalance. The use of multiple evaluation metrics is important in tweet sentiment classification because overall accuracy alone may not adequately reflect performance across all sentiment categories [12]. For Linear SVC, hyperparameter optimization was performed using Grid Search with 5-fold cross-validation. The parameter space included C = {0.1, 0.25, 0.5, 1.0}, class weight = {None, balanced}, loss = {hinge, squared_hinge}, and penalty = l2. For Complement Naïve Bayes, tuning was limited to the smoothing parameter alpha, with trial values of 0.1, 0.5, 1.0, and 2.0. Because variations in alpha yielded negligible improvements, alpha = 1.0 was retained in the final model.

2.10. Evaluation and Error Analysis

To enhance transparency, confusion matrices were generated for both the best-performing and worst-performing models. These matrices made it possible to identify how each model handled positive, neutral, and negative classes and to observe where misclassification patterns were concentrated. The worst-performing model, TextBlob + Naïve Bayes, tended to over-predict positive and neutral sentiments while showing weak separation of the negative class. By contrast, the best-performing model, IndoBERT +

Linear SVC, showed stronger separation of negative sentiment, improved recall for positive sentiment, and more balanced performance across all three classes.

2.11. Visualization

To support interpretation, the study employed several visualization techniques. Word clouds were generated to display the most frequent words associated with positive, neutral, and negative sentiment classes under each labeling scheme. In addition, sentiment distribution charts were produced to compare the proportion of positive, neutral, and negative tweets under the two labeling approaches.

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation

This study examines public sentiment toward the Free Nutritious Meal Program (MBG) on the X/Twitter platform using two sentiment-labeling schemes: TextBlob-based pseudo-labeling, which depends on Indonesian-to-English translation, and IndoBERT-based contextual labeling, which processes Indonesian text directly [14], [16]. Based on these two labeling sources, two conventional classifiers were trained and evaluated, namely Naïve Bayes with TF-IDF features and Linear SVC with word- and character-level features. This comparative design makes it possible to evaluate not only classification performance but also the extent to which the labeling strategy influences the distribution of sentiment classes and the interpretability of the results. Such a framework is consistent with recent sentiment analysis studies that compare lexicon-based or pseudo-labeling approaches with contextual transformer-based methods [13], [17].

Before discussing model performance, it is important to examine the sentiment distribution generated by each labeling scheme because the quality and balance of labels strongly affect downstream classification. The first pattern can be observed in the TextBlob-based label distribution, which is presented in Figure 2. As shown in the figure, the Positive and Neutral classes dominate the dataset, while the Negative class appears substantially smaller.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

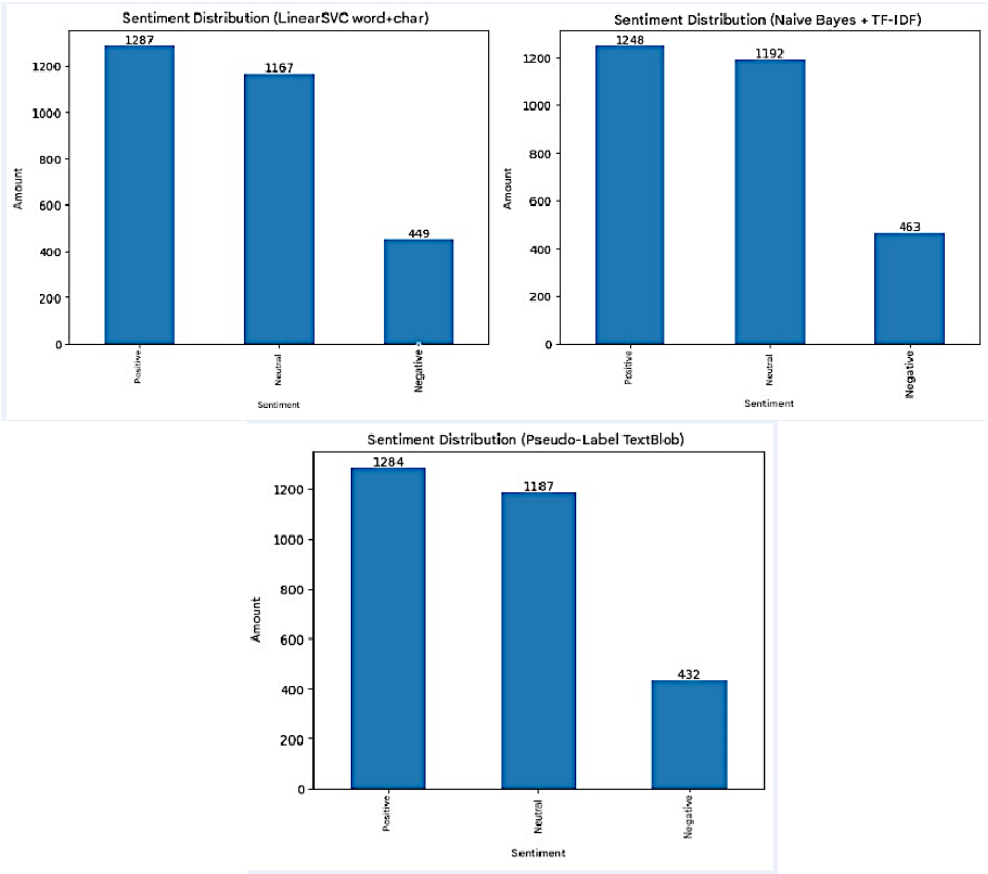

Figure 2. TextBlob Distribution

This distribution indicates a positive bias resulting from the use of TextBlob, which relies on an English polarity lexicon. Because TextBlob requires Indonesian tweets to be translated into English before scoring, many expressions that are negative or critical in Indonesian can be transformed into more neutral or even positive constructions in English. A frequently cited example is “tidak buruk”, which becomes “not bad” and often receives a positive polarity score [13], [14]. This translation step reduces sensitivity to negative cues that are linguistically and culturally specific to Indonesian discourse. Recent studies also show that lexicon-based approaches have important limitations in capturing idiomatic meaning, contextual negation, and culturally embedded expressions in local-language social media data [15], [16].

The consequence of this bias is not merely descriptive; it also affects model learning. When TextBlob-generated pseudo-labels are used as supervisory signals for models such as Naïve Bayes and Linear SVC, the resulting classifiers inherit the skewed label distribution and become less capable of identifying negative sentiment. This explains why

the proportion of Negative labels under the TextBlob scheme is much lower than under the contextual scheme. Prior studies similarly report that lexicon-based pseudo-labeling can create minority-class underrepresentation and weaken the model's ability to capture negative sentiment reliably [16], [17].

In contrast to the TextBlob results, the distribution produced by the contextual Indonesian model reveals a markedly different pattern. Figure 3 presents the class distribution generated by IndoBERT, and the figure shows that Negative sentiment becomes the dominant class, followed by Neutral and Positive. This contrast is important because it suggests that the overall picture of public opinion changes substantially depending on the labeling strategy used.

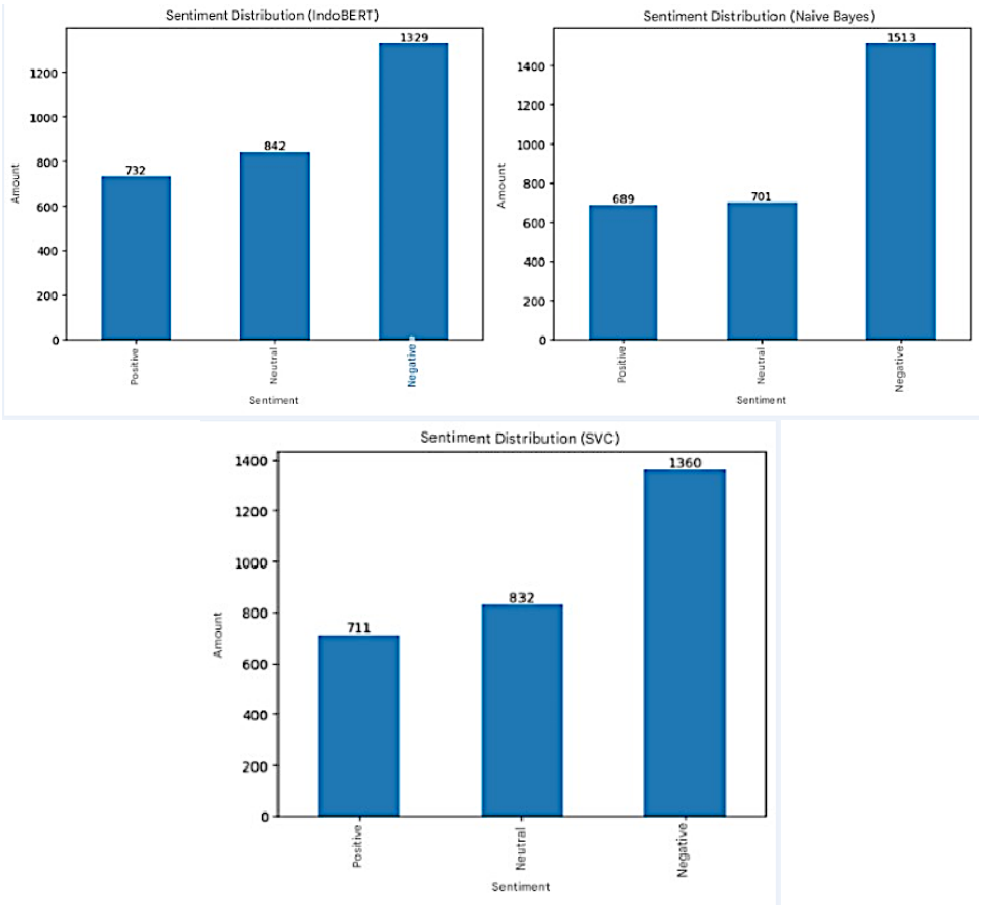

Figure 3. Distribution of IndoBERT

The distribution in Figure 3 suggests that public conversations surrounding the MBG program are more strongly oriented toward complaints, criticism, and dissatisfaction than

toward praise or explicit support [16]. Unlike TextBlob, IndoBERT captures Indonesian linguistic context directly, including negation patterns such as "tidak bagus" and "kurang maksimal", as well as sarcasm and informal social media expressions that frequently appear in online policy debates [14], [17]. This ability allows the model to detect sentiment cues that are often lost in translation-based methods. As a result, the labels generated by IndoBERT are more realistic and more closely aligned with the contextual meaning of Indonesian user-generated text.

From a methodological perspective, the IndoBERT-based distribution also indicates that contextual transformer labeling provides a more credible representation of public perception, especially for identifying negative sentiment. Recent studies confirm that Indonesian transformer-based models substantially improve the accuracy and robustness of sentiment classification compared with traditional lexicon-driven approaches, particularly in noisy social media environments [16], [18]. Therefore, the contrast between Figure 2 and Figure 3 is not only a technical difference; it also has substantive implications for how public reactions to the MBG program should be interpreted.

After establishing the differences in label distribution, the next step is to compare how those labels affect downstream classifier performance. The overall evaluation results for all experimental configurations are summarized in Table 1. This table presents the results of four model combinations: TextBlob → Naïve Bayes, TextBlob → Linear SVC, IndoBERT → Naïve Bayes, and IndoBERT → Linear SVC.

Table 1. Evaluation Metric

Label & model	Accuracy	Macro Precision	Macro Recall	Macro F1
TextBlob → Naive Bayes (TF-IDF)	0.573	0.526	0.530	0.528
TextBlob → SVC (word+char)	0.633	0.587	0.593	0.589
IndoBERT → Naive Bayes (TF-IDF)	0.716	0.715	0.684	0.692
IndoBERT → SVC (balanced, optimized)	0.742	0.734	0.719	0.725

As shown in Table 1, both accuracy and macro F1-score increase substantially when the models are trained using IndoBERT-derived labels. Under the Naïve Bayes setting,

accuracy rises from 0.573 with TextBlob labels to 0.716 with IndoBERT labels, while macro F1 improves from 0.528 to 0.692. The best overall result is achieved by Linear SVC trained on IndoBERT labels, which reaches an accuracy of 0.742 and a macro F1-score of 0.725 after class weighting and hyperparameter optimization. These results indicate that label quality has a strong impact on downstream classification performance and that contextual Indonesian labeling produces a more learnable and discriminative supervisory signal than translation-based pseudo-labeling.

Another important pattern in Table 1 is that Linear SVC consistently outperforms Naïve Bayes under both labeling schemes. This difference can be explained by the characteristics of the feature space. In high-dimensional and sparse textual representations, especially those built from TF-IDF, word n-grams, and character n-grams, SVC can learn a more effective decision boundary through margin maximization. By comparison, Naïve Bayes relies on feature independence assumptions that become restrictive when many correlated lexical and subword patterns are present. This issue is particularly relevant in Indonesian social media text, where negation and sentiment are often conveyed through short but variable forms such as "nggak," "gak," "ga," "ngga," "tdk," "tidak," and "kurang." Character-level n-grams help Linear SVC capture these morphological variations more effectively, enabling it to distinguish subtle negative expressions that are diluted under a simple bag-of-words assumption. This explains why the IndoBERT + Linear SVC combination provides the strongest and most stable classification results.

To understand more clearly how each labeling scheme structures the sentiment classes, the class distributions are presented separately in Table 2 and Table 3. Table 2 shows the distribution under the TextBlob-oriented scheme, while Table 3 presents the distribution under the IndoBERT-oriented scheme. These tables are important because they reveal whether the trained models preserve or distort the label tendencies of their respective annotation sources.

Table 2. Class Distribution with TextBlob

Scenarios	Positive	Neutral	Negative
TextBlob	1284	1187	432
TextBlob → SVC (word+char)	1248	1192	463
IndoBERT → Naive Bayes (TF-IDF)	1287	1167	449

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 3. Class Distribution with IndoBERT

Scenarios	Positive	Neutral	Negative
TextBlob	732	842	1329
TextBlob → SVC (word+char)	689	701	1513
IndoBERT → Naive Bayes (TF-IDF)	704	838	1361

The class distributions in Table 2 and Table 3 reveal a striking contrast between the two labeling schemes. Under the TextBlob scheme, the corpus is dominated by Positive and Neutral classes, with the Negative class appearing as a relatively small minority (1284 / 1187 / 432). Under the IndoBERT scheme, however, the pattern reverses: Negative sentiment becomes dominant (1329) and clearly exceeds both Positive (732) and Neutral (842) [13], [16]. This difference strongly suggests that translation-based labeling suppresses negative cues, whereas contextual Indonesian labeling captures them more effectively.

The prediction patterns of the trained models broadly follow the tendencies of their label sources. Under the contextual scheme, the classifier outputs remain closer to the original class proportions, indicating better alignment between the learned decision boundary and the sentiment structure of the data. This is especially important because public discussions about the MBG program appear to be driven less by celebratory discourse and more by concerns regarding budget transparency, implementation quality, food safety, and operational effectiveness. In that sense, the class distribution generated by IndoBERT is not merely more negative; it is also more substantively plausible in relation to the topics commonly debated in public online conversations.

A closer examination of the IndoBERT-labeled distribution also shows that the dataset is moderately imbalanced, with the Negative class representing the majority. This imbalance reflects the real dynamics of policy-related discourse on X/Twitter, where users are often more motivated to post complaints, skepticism, or criticism than routine approval. From a modeling perspective, such imbalance creates at least three challenges. First, classifiers may become biased toward the dominant negative class, resulting in apparently high overall accuracy but weaker sensitivity to minority classes. Second, negative sentiment in Indonesian social media tends to be lexically diverse, often expressed through sarcasm, negation, idioms, and colloquial criticism, which broadens the feature space for negative

examples. Third, because IndoBERT more accurately recognizes these nuanced negative cues, it also makes the underlying imbalance more visible than TextBlob does. This is why class-weight balancing in the Linear SVC model becomes methodologically important. The improved recall and macro F1-score under the balanced SVC configuration indicate that handling class imbalance is essential for fairer and more reliable sentiment classification outcomes.

Beyond overall metrics and class counts, the interpretive layer of the analysis can be strengthened through lexical visualization. To examine the terms that most frequently occur within each sentiment category, word cloud visualizations were generated for the Positive, Neutral, and Negative classes. These visualizations help identify the most salient words associated with each class and provide an intuitive overview of the semantic focus of public discussion [19], [21]. The resulting word cloud is presented in Figure 4.

Figure 4. WordCloud

The lexical patterns shown in Figure 4 add substantive depth to the quantitative findings. In the Positive class, dominant words such as anak, program, makan, gratis, sehat, and dukung indicate that supportive discourse tends to emphasize the perceived social and nutritional benefits of the program, especially for children [16]. These terms suggest that positive sentiment is closely linked to the normative appeal of improving welfare and access to nutrition. In the Neutral class, commonly occurring words such as pelaksanaan, pemerintah, gizi, sekolah, and Prabowo reflect a more descriptive or informational style of communication. These tweets are less evaluative and more likely to contain updates, references to implementation, or statements about the policy itself.

By contrast, the Negative class is characterized by terms such as anggaran, uang, kurang, masalah, racun, and ampas, which collectively indicate dissatisfaction, distrust, and criticism of the MBG program [17], [18]. These lexical patterns point to public concern not only about the concept of the program but, more importantly, about its execution, cost,

and perceived risks. The word cloud therefore reinforces the broader conclusion that online discourse surrounding MBG is shaped largely by practical and political concerns rather than by purely symbolic support. This finding is consistent with prior studies emphasizing the value of visualization techniques for interpreting large-scale public sentiment on social media platforms [19], [20].

The comparison between labeling approaches also reinforces a broader methodological point: language compatibility matters greatly in sentiment analysis [22]. TextBlob depends on English lexical polarity and therefore requires Indonesian-to-English translation, a process that is inherently vulnerable to semantic distortion [23]. Negative Indonesian constructions are often translated into softer or more ambiguous English forms, which weakens the negative signal before classification even begins. This weakness is clearly reflected in the label distribution, where the Negative class becomes underrepresented, and in the relatively modest performance range of the TextBlob-based models (accuracy 0.573–0.633) [15]. Prior studies have similarly shown that translation errors and lexicon mismatch can obscure negative sentiment and degrade downstream performance [16], [17].

In contrast, IndoBERT captures the contextual, morphological, and idiomatic properties of Indonesian text directly. Its labels provide stronger and more reliable supervision for conventional classifiers, enabling Naïve Bayes to improve substantially and Linear SVC to achieve the best overall result after balancing and optimization. The strong improvement in the recall of the Negative class indicates that contextual representations are particularly effective for distinguishing genuine criticism from neutral reporting. This is especially important in Indonesian social media, where meaning often depends on phrase-level context, sarcasm, or informal constructions rather than on isolated sentiment words [17], [18], [20].

To further clarify the weaknesses of translation-based pseudo-labeling, several tweets were examined manually. Cases such as “Programnya nggak buruk, tapi pelaksanaannya ampas banget” show how TextBlob can be misled by the phrase “not bad”, producing a positive polarity despite the strongly negative evaluation conveyed by “ampas banget.” Similarly, “Mantap, anggarannya bocor lagi” can be read as positive under literal lexical scoring because of the word “mantap,” even though the intended meaning is sarcastic

criticism. Another example, "Kurang masuk akal kalau biaya segini tapi kualitas makanannya begitu," may be weakened into a neutral interpretation by translation, even though it clearly expresses dissatisfaction. These examples highlight why contextual transformer models are better suited to Indonesian social media discourse: they can capture sarcasm, negation, idiomatic criticism, and implicit evaluative meaning that lexicon-based translated systems often miss.

From a substantive perspective, the results indicate that MBG-related discourse on X/Twitter is dominated by negative sentiment, followed by neutral discussion, while explicitly positive support occupies a smaller portion of the conversation. This does not necessarily mean that the program is broadly rejected; rather, it suggests that online discussion is driven largely by contested issues and implementation concerns. Social media users appear more inclined to post when they perceive problems, risks, or inconsistencies than when they simply agree with the policy. At the same time, the best-performing model, IndoBERT + Linear SVC (balanced), maintains adequate predictive sensitivity for the minority Positive class, preventing the analysis from collapsing into an oversimplified "all negative" interpretation.

Overall, for Indonesian-language sentiment analysis, the IndoBERT + Linear SVC (balanced) configuration provides the most effective combination of accuracy, class stability, and contextual sensitivity. TextBlob with translation may still function as a quick baseline, but it is insufficient for capturing the rich negative nuances, informal expressions, and sarcasm that characterize Indonesian social media discourse [13], [14]. Taken together, the findings of this study align closely with recent developments in Indonesian sentiment analysis research and further confirm that contextual language-specific labeling is essential for obtaining reliable evidence from public policy discussions online [13]–[20].

3.2. Discussion

The present study provides a comprehensive analysis of public sentiment toward the Free Nutritious Meal Program (MBG) by combining two labeling schemes—TextBlob pseudo-labeling and IndoBERT contextual labeling—with two traditional machine-learning classifiers, Naïve Bayes and Linear SVC. Overall, the findings highlight several methodological and substantive insights that deepen the understanding of Indonesian-

language sentiment analysis and the nature of public discourse related to national policies.

The first major finding of this research is the significant performance disparity between the TextBlob-based labeling and IndoBERT-based labeling. TextBlob, which classifies polarity based on an English lexicon after translation, produced labels with a disproportionately small share of negative sentiment. This is largely due to the structural differences between Indonesian and English expressions, especially regarding negation and idiomatic forms. The translation step tends to soften or distort negative expressions—particularly those involving sarcasm, informal phrasing, or implicit criticism—leading to misclassifications such as "tidak buruk" → "not bad" being labeled as positive. This phenomenon is reflected in the lower classifier performance under the TextBlob scheme, with Naïve Bayes achieving only 0.573 accuracy.

In contrast, IndoBERT-based labeling demonstrates superior sensitivity to Indonesian linguistic structures. IndoBERT captures contextual features such as negation, morphological variations, slang, and sarcasm, enabling the generation of labeling signals that align more naturally with Indonesian social media discourse. The dominance of negative sentiment in the IndoBERT-labeled dataset suggests that public discussions around MBG emphasize concerns about budget allocation, implementation inefficiencies, and doubts about sustainability. This sentiment pattern is consistent with issue-driven discourse commonly found in policy-related conversations on social platforms.

A second important finding is that Linear SVC consistently outperforms Naïve Bayes across both labeling schemes. This aligns with well-established evidence in text classification research showing that SVC performs robustly in high-dimensional, sparse feature spaces—especially those involving hybrid word–character n-grams. Character n-grams (4–5 grams), in particular, capture Indonesian negation patterns ("tdk," "tidaaak," "ga," "nggak"), informal spellings, and expressive elongations that are extremely common on Twitter. Meanwhile, Naïve Bayes' conditional independence assumption restricts its ability to model such subword nuances effectively. Consequently, SVC's margin-maximization behavior allows it to distinguish negative sentiment more reliably, especially when paired with IndoBERT labels and class-weight balancing.

The third major finding concerns class imbalance. IndoBERT labeling reveals an approximate 45–50% dominance of negative sentiment, with positive sentiment making up the smallest class. This imbalance is both a reflection of real public discourse and a modeling challenge. Without class weighting, classifiers would naturally skew toward the majority class, reducing precision and recall for minority positive sentiment. The adoption of class_weight = "balanced" in the Linear SVC model mitigates this issue by enforcing equal learning emphasis across classes.

Mubarok analyzed sentiment toward COVID-19 PSBB policies using Naïve Bayes and found that Indonesian public discourse tends to cluster around negative sentiment due to policy dissatisfaction. Similar to the present study, negative sentiment was dominant, suggesting that discussions of government policies in Indonesia commonly lean toward criticism rather than support. However, Mubarok relied solely on lexicon-based methods, which can oversimplify sentiment interpretation. The current study advances this by integrating transformer-based labeling (IndoBERT), resulting in more accurate context-sensitive sentiment extraction [1]. Rasiban and Riyadi compared Naïve Bayes and SVM on healthcare service applications and concluded that SVM generally outperforms Naïve Bayes, especially when handling high-dimensional data. Their findings align strongly with the results of this study: Linear SVC consistently provided higher accuracy and F1-scores across both TextBlob and IndoBERT labeling schemes. This reinforces the generalizability of SVC's superior performance in Indonesian social media text classification, regardless of domain (healthcare vs. national policy) [3]. Imaduddin et al. used IndoBERT to analyze sentiment in Indonesian healthcare applications and reported significant improvements compared to traditional machine-learning methods. IndoBERT's ability to understand contextual and morphological features—especially negation—was highlighted as the key performance driver. The present study corroborates these findings: IndoBERT labeling enhanced both Naïve Bayes and SVC performance. Moreover, IndoBERT captured sentiment more realistically than TextBlob, particularly for tweets with sarcasm and informal language—characteristics also frequently discussed in Imaduddin et al.'s study [14]. Pramana et al. compared BiLSTM, BERT, and ensemble methods, showing that transformer-based models produce more accurate sentiment classification for Indonesian text. The authors emphasize that Indonesian-language nuances—slang, cultural expressions, local idioms—require deep contextual understanding that lexicon-based or simple machine-learning models cannot fully capture. This directly supports the

results obtained in this study: IndoBERT contextual labeling significantly improves classification metrics, and character-level n-grams further enhance sensitivity to Indonesian linguistic variation [16]. In an aspect-based sentiment analysis study of hospital reviews, the use of Fine-Tuned IndoBERT yielded high accuracy and fine-grained sentiment differentiation. The authors highlight the importance of domain-specific contextual embeddings. Although the present study focuses on general sentiment (not aspect-based), IndoBERT's strengths remain consistent: it effectively captures domain-specific expressions related to public policy, such as concerns about "anggaran," "pelaksanaan," and "kualitas makanan." This demonstrates IndoBERT's adaptability across domains, supporting its use as a labeling mechanism in various Indonesian NLP tasks [18]. Beyond the five specific studies, several broader patterns emerge:

Lexicon-based methods like TextBlob are consistently weaker for Indonesian sentiment analysis due to translation errors and the absence of morphological understanding. Studies such as Ahmadian et al. (2023) [15] and Mustofa & Prasetiyo (2021) [11] note similar limitations. Transformer-based Indonesian models (IndoBERT, IndoRoBERTa, idT5) consistently outperform traditional models, especially in datasets containing slang, sarcasm, and contextual negation [15], [17], [19], [20], High-dimensional n-gram features (word + character) enhance SVC performance—a finding supported by Styawati (2023) [10] and other SVM-focused studies, Negative sentiment tends to dominate policy-related discourse on Indonesian social media, consistent with findings in studies on public policy, healthcare, and education [1], [7], [16]. The alignment between this research and the broader literature confirms the methodological validity of integrating IndoBERT with Linear SVC for Indonesian sentiment analysis.

Substantively, the dominance of negative sentiment indicates potential areas of concern for policymakers—especially regarding budget transparency, food distribution mechanisms, and perceived quality of meal provision. Public skepticism suggests that the government needs stronger communication strategies and clearer implementation guidelines to improve trust. Methodologically, the results demonstrate the effectiveness of combining transformer-based labeling with classical classifiers, offering a hybrid approach that is accurate yet computationally efficient. This may serve as a useful framework for public policy analytics in Indonesia, particularly for institutions with limited GPU resources.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

A deeper statistical interpretation of the confusion matrices reveals several critical insights regarding model errors and the underlying distribution of sentiment in the MBG dataset. In the TextBlob → Naïve Bayes configuration, the rate of false positives (FP) for the Positive class is notably high—more than 380 Neutral tweets and nearly 100 Negative tweets are incorrectly classified as Positive. Statistically, this indicates that the posterior probabilities estimated by Naïve Bayes are disproportionately influenced by lexically positive English terms introduced during translation. When Indonesian negative expressions such as "ampas," "kurang masuk akal," or "nggak beres" are translated into English, these terms do not carry sufficiently negative polarity within TextBlob's lexicon. As a consequence, the likelihood of the Negative class is significantly weakened, resulting in systematic misclassification.

Conversely, the largest false negative (FN) burden also appears in the Negative class, where more than 400 originally negative tweets are misclassified as Positive or Neutral. In contrast, the IndoBERT → Linear SVC model demonstrates a substantially higher true positive rate (TPR) for the Negative class, correctly identifying 728 out of 1,044 instances and achieving a sensitivity close to 0.70—compared to less than 0.40 under the TextBlob scheme. Statistically, this reflects the stronger decision boundary produced by the margin-maximizing SVC, which can effectively separate high-dimensional n-gram feature vectors characteristic of criticism, complaints, and negative assessments.

Furthermore, the precision of the Positive class improves markedly under IndoBERT → SVC due to a significant reduction in false positives. This suggests that linguistic patterns associated with supportive or appreciative statements—such as endorsements of the program's benefits, health value, or social impact—are captured more consistently. Overall, the statistical evidence indicates that performance gains are driven not merely by improvements in average accuracy but also by more effective error control across classes, particularly in the context of imbalanced sentiment distribution. These findings reinforce that transformer-based contextual labeling provides a more stable class structure and increases the reliability of sentiment mapping in Indonesian social media discourse.

4. CONCLUSION

This study demonstrates that the use of IndoBERT contextual labeling combined with Linear SVC produces the highest sentiment classification performance for Indonesian social media text. Quantitatively, the IndoBERT + SVC configuration achieved an accuracy of 0.742, representing an improvement of approximately 17% over the TextBlob + Naïve Bayes model, which obtained 0.573 accuracy. The findings also reveal a strong dominance of negative sentiment in public discussions of the Free Nutritious Meal Program, indicating that discourse surrounding the policy is shaped by concerns about budget transparency, implementation quality, and logistical feasibility. From a policy perspective, these results offer practical insights for government agencies seeking to gauge real-time public sentiment toward large-scale social programs. By adopting context-aware NLP approaches such as IndoBERT, policymakers can monitor emerging public concerns more accurately, identify critical issues that require intervention, and adjust program communication strategies accordingly. Sentiment trends extracted from social media can be integrated into feedback loops for policy refinement, early detection of dissatisfaction, and more targeted delivery of information. Overall, this research provides both a methodological contribution to Indonesian sentiment analysis and a substantive foundation for evidence-based policy evaluation.

Despite these contributions, the study has several limitations. First, the dataset size (2,903 tweets) and restricted temporal window may limit the representativeness of long-term sentiment fluctuations. Second, the analysis relies solely on Twitter/X, excluding public perspectives from other platforms such as Instagram, TikTok, or Facebook, which may contain different demographic and linguistic characteristics. Third, although IndoBERT offers strong contextual understanding, transformer-based models still face challenges in fully capturing sarcasm, hyperbole, or highly localized slang. Future research may address these limitations by expanding cross-platform datasets, incorporating temporal modeling, or leveraging fine-tuned domain-specific transformer architectures. From a policy perspective, the findings demonstrate that social media sentiment can serve as a real-time diagnostic tool for evaluating public reactions to national programs. Negative sentiment patterns identified in this study provide early signals of public dissatisfaction that may guide improvements in communication strategies, resource allocation, and program delivery mechanisms. By integrating sentiment monitoring into policy feedback

loops, government agencies can detect emerging concerns more rapidly and refine implementation practices to enhance public trust and program effectiveness.

REFERENCES

[1] F. Al isfahani and R. Mubarok, "Analisis sentimen pengguna Twitter terhadap kebijakan pemberlakuan pembatasan sosial berskala besar (PSBB) dengan metode Naïve Bayes," *J. Siliwangi Seri Sains dan Teknol.*, vol. 7, no. 1, pp. 19–24, 2021.

[2] I. J. T. Gurning, P. P. Adikara, dan R. S. Perdana, "Analisis Sentimen Dokumen Twitter menggunakan Metode Naïve Bayes dengan Seleksi Fitur GU Metric," *J. Pengemb. Teknol. Inf. Ilmu Komput.*, vol. 7, no. 5, pp. 2169–2177, 2023.

[3] F. M. Sarimole dan Kudrat, "Analisis Sentimen terhadap Aplikasi Satu Sehat pada Twitter Menggunakan Algoritma Naïve Bayes dan Support Vector Machine," *J. Sains Teknol.*, vol. 5, no. 3, pp. 783–790, 2024.

[4] R. Hidayat, M. Fikry, Yusra, F. Yanto, dan E. P. Cynthia, "Penerapan Naïve Bayes Classifier dalam Klasifikasi Sentimen Publik di Twitter terhadap Puan Maharani," *JUKI J. Komput. Inform.*, vol. 6, no. 1, pp. 100–108, 2024, doi: 10.53842/juki.v6i1.479.

[5] N. Desiani dan A. Syafiq, "Efektivitas Program Makan Gratis pada Status Gizi Siswa Sekolah Dasar: Tinjauan Sistematis," *Malahayati Nurs. J.*, vol. 7, no. 1, pp. 27–48, 2025, doi: 10.33024/mnj.v7i1.17497.

[6] U. Agustini dan S. Mulyani, "Efektivitas dan Tantangan Kebijakan Program Makan Bergizi Gratis sebagai Intervensi Pendidikan di Indonesia," *J. Kiprah Pendidik.*, vol. 4, no. 3, pp. 362–368, 2025, doi: 10.33578/kpd.v4i3.p362-368.

[7] R. Widaryanti, Casnuri, dan Metty, "Penurunan Masalah Gizi Pada Anak Usia Dini Melalui Edukasi PMT-AS," *Dinamisia*, vol. 6, no. 5, pp. 1168–1173, 2022, doi: 10.31849/dinamisia.v6i5.10762.

[8] K. L. Tan, C. P. Lee, dan K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," *Appl. Sci.*, vol. 13, no. 7, Art. no. 4550, 2023, doi: 10.3390/app13074550.

[9] M. AminiMotlagh, H. Shahhoseini, dan N. Fatehi, "A reliable sentiment analysis for classification of tweets in social networks," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, Art. no. 7, 2023, doi: 10.1007/s13278-022-00998-2.

[10] S. Styawati, A. R. Isnain, N. Hendrastuty, dan L. Andraini, "Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis," *J. Inform. J. Pengemb. IT*, vol. 6, no. 1, pp. 56–60, 2021, doi: 10.30591/jpit.v6i1.3245.

[11] R. L. Mustofa dan B. Prasetiyo, "Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter," *J. Phys. Conf. Ser.*, vol. 1918, no. 4, Art. no. 042155, 2021, doi: 10.1088/1742-6596/1918/4/042155.

[12] S. K. Assayed, K. Shaalan, M. Alkhatib, dan S. Maghaydah, "Machine Learning ChatBot for Sentiment Analysis of COVID-19 Tweets," in *Comput. Sci. Inf. Technol.*, vol. 13, 2023, pp. 41–55, doi: 10.5121/csit.2023.130404.

[13] C.-H. Lin dan U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *J. Big Data*, vol. 10, Art. no. 88, 2023, doi: 10.1186/s40537-023-00782-9.

[14] H. Imaduddin, F. Y. A'la, dan Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, pp. 113–117, 2023, doi: 10.14569/IJACSA.2023.0140813.

[15] H. Ahmadian, T. F. Abidin, H. Riza, dan K. Muchtar, "Transformer-Based Indonesian Language Model for Emotion Classification and Sentiment Analysis," in *Proc. 2023 IEEE Int. Conf. Inf. Technol. Comput. (ICITCOM)*, 2023, pp. 209–214, doi: 10.1109/ICITCOM60176.2023.10442970.

[16] R. Pramana, M. Jonathan, H. S. Yani, and R. Sutoyo, "A comparison of BiLSTM, BERT, and ensemble method for emotion recognition on Indonesian product reviews," *Procedia Comput. Sci.*, vol. 245, pp. 399–408, 2024, doi: 10.1016/j.procs.2024.10.266.

[17] U. K. Das, R. S. Ani, N. Datta, I. Fahad, J. Sikder, dan U. Sara, "Enhancing sentiment analysis accuracy on social media comments using a tuned BERT model," *Discov. Comput.*, vol. 28, Art. no. 198, 2025, doi: 10.1007/s10791-025-09599-x.

[18] A. Maretta dan A. Meiriza, "Aspect-Based Sentiment Analysis of Hospital Service Reviews Using Fine-Tuned IndoBERT," *J. Appl. Inform. Comput.*, vol. 9, no. 5, pp. 2541–2551, 2025, doi: 10.30871/jaic.v9i5.10765.

[19] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardaneswari, dan S. Nurrohmah, "BERT-Based Combination of Convolutional and Recurrent Neural Network for Indonesian Sentiment Analysis," *Appl. Soft Comput.*, vol. 151, Art. no. 111112, 2024, doi: 10.1016/j.asoc.2023.111112.

[20] M. Fuadi, A. D. Wibawa, dan S. Sumpeno, "idT5: Indonesian Version of Multilingual T5 Transformer," arXiv:2302.00856, 2023, doi: 10.48550/arXiv.2302.00856.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[21] K. Kucher, C. Paradis, dan A. Kerren, "The state of the art in sentiment visualization," *Comput. Graph. Forum*, vol. 37, no. 1, pp. 71–96, 2018, doi: 10.1111/cgf.13217.

[22] J. Tao dan X. Fang, "Toward multi-label sentiment analysis: A transfer learning based approach," *J. Big Data*, vol. 7, Art. no. 1, 2020, doi: 10.1186/s40537-019-0278-0.

[23] S. Ahmad, S. M. Saqib, A. H. Syed, N. Alromema, dan A. Kararay, "Exploring the best fit: A comparative analysis of AFINN, TextBlob, VADER, and Pattern on Arabic reviews for optimal dictionary extraction," *Mehran Univ. Res. J. Eng. Technol*, vol. 44, no. 2, pp. 197–216, 2025, doi: 10.22581/muet1982.3449.