Cyberspace: Jurnal Pendidikan Teknologi informasi
Volume 10, Nomor 1, April 2026, hal. 26 – 39
ISSN 2598-2079 (print) | ISSN 2597-9671 (online)

PERFORMANCE ANALYSIS OF MACHINE LEARNING AND INDOBERT IN CLASSIFYING SENTIMENTS ON INDONESIA'S FREE NUTRITIOUS MEAL

Maulyanda[1*], Sri Azizah Nazhifah[2], Syafrial Fachri Pane[3], Irvanizam[4]

[1,2,4]Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia
[3]Informatics Engineering, Universitas Logistik Bisnis Internasional, Indonesia
E-mail: [1]maulyanda@usk.ac.id, [2]sriazizah07@usk.ac.id, [3]syafrial.fachri@ulbi.ac.id
Corresponding Author: maulyanda@usk.ac.id

Abstract

The Free Nutritious Meal (MBG) program, as one of the national strategic policies, has generated diverse responses on social media; however, the patterns of public sentiment and opinion toward the program have not been systematically identified, thus requiring a Natural Language Processing (NLP) approach to analyze public attitudes and the dominant issues emerging in relation to its implementation. NLP is a branch of artificial intelligence that is widely used to analyze whether a sentence contains positive, negative, or neutral sentiment, particularly in the context of expressing opinions in the online environment. This study compares several models to identify the most optimal one, namely Naïve Bayes, Support Vector Machine (SVM), XGBoost, and IndoBERT. The dataset used in this research was obtained from Kaggle and consists of 5,644 data points in the neutral class, 2,934 data points in the positive class, and 2,606 data points in the negative class. Prior to model implementation, the dataset underwent a preprocessing stage that included case folding, cleansing, tokenization, slang word conversion, stemming, and stopword removal. Subsequently, the data were trained using the four aforementioned methods. The results indicate that Naïve Bayes achieved an accuracy of 75%, SVM reached 79%, XGBoost obtained 76%, while IndoBERT achieved the highest accuracy at 85%. Therefore, it can be concluded that, using this dataset, IndoBERT performed sentiment [1]ification very effectively.

Keywords: *social media, platform X, NLP, deep learning, classification*

Abstrak

Program Makan Bergizi Gratis (MBG) sebagai salah satu kebijakan nasional strategis masih menimbulkan beragam respons di media sosial, namun pola sentimen dan opini publik terhadap program tersebut belum teridentifikasi secara sistematis sehingga diperlukan pendekatan Natural Language Processing (NLP) untuk menganalisis kecenderungan sikap masyarakat serta isu-isu dominan yang berkembang terkait implementasinya. NLP merupakan salah satu cabang kecerdasan buatan yang banyak digunakan untuk menganalisis sentimen suatu kalimat, apakah mengandung unsur positif, negatif, atau netral, khususnya dalam konteks penyampaian opini di dunia maya. Penelitian ini membandingkan beberapa model untuk memperoleh model yang paling optimal, yaitu Naïve Bayes, Support Vector Machine (SVM), XGBoost, dan IndoBERT. Dataset yang digunakan diperoleh dari Kaggle, terdiri atas 5.644 data pada kelas netral, 2.934 data pada kelas positif, dan 2.606 data pada kelas negatif. Sebelum diterapkan pada model, dataset terlebih

dahulu melalui tahap prapemrosesan yang meliputi case folding, cleansing, tokenization, slang word conversion, stemming, dan stopword removal. Selanjutnya, data dilatih menggunakan keempat metode tersebut. Hasil penelitian menunjukkan bahwa Naïve Bayes mencapai akurasi sebesar 75%, SVM sebesar 79%, XGBoost sebesar 76%, sementara IndoBERT memperoleh akurasi tertinggi, yaitu 85%. Oleh karena itu, dapat disimpulkan bahwa dengan menggunakan dataset ini, IndoBERT mampu melakukan klasifikasi sentimen dengan sangat baik.

Kata Kunci: *social media, platform X, NLP, deep learning, classification*

1. Introduction

Currently, the research related to sentiment analysis is developing rapidly, especially in accordance with the increasing growth of social media users. A survey conducted by the Indonesian Internet Service Providers Association (APJII) in mid-2025 revealed that approximately eight out of ten Indonesians—around 229 million people—are currently connected to the internet, indicating a significant increase compared to the previous six years [1].

Under the leadership of Prabowo Subianto, the Indonesian government has launched a flagship program to provide free nutritious meals (MBG) to students from preschool through high school, reflecting the government's commitment to supporting the growth and development of Indonesian children. Following the implementation of the program, public responses in Indonesia were divided, with both support and opposition clearly reflected in opinions expressed on social media.

A large number of opinions are circulated on social media platforms such as X. To systematically identify whether Indonesian public sentiment is positive (agreement), negative (disagreement), or neutral (neither supportive nor opposing), this approach employs Natural Language Processing (NLP) techniques to analyze and interpret the opinions embedded within the collected data.

In today's digital age, the role of NLP is becoming increasingly significant. This technology enables fast and efficient processing of large-scale text data, supporting various applications in business [2], [3], healthcare [4], education [5], [6], politics [7]–[9], and online video platform [10]–[12]. NLP helps organizations understand user perceptions and opinions through sentiment analysis and improves customer service through ChatBots that provide natural and contextual responses [13].

A previous study conducted by [14] investigated sentiment analysis concerning the development of Indonesia's new capital city (IKN) using Support Vector Machine (SVM) and Decision Tree (DC) methods, based on 11,600 data samples obtained from platform X. The results demonstrated that the SVM method outperformed the DC method in terms of accuracy. In addition, sentiment analysis was conducted on user reviews of the Jamsostek mobile application using data obtained from the Google Play Store. By integrating the IndoBERT and BERTopic models, the study aimed to gain a more comprehensive understanding of user feedback. The findings revealed that the proposed model was able to accurately detect around 75% of the 2,846 reviews that expressed negative sentiment [4].

The Free Nutritious Meal (MBG) program, as a national strategic policy, has generated diverse reactions and debates across social media platforms; however, comprehensive and continuous mapping of public sentiment patterns and opinions remains necessary, thereby motivating the use of a NLP approach to identify prevailing public attitudes and key issues emerging throughout its implementation.

Therefore, this study employs NLP models to analyze sentiment toward the MBG program implemented by the Indonesian government. The data used consists of user comments collected from platform X and obtained through Kaggle. This research aims to identify the most optimal model for sentiment analysis of MBG, namely Naïve Bayes, Support Vector Machine (SVM), XGBoost, and IndoBERT. Subsequently, Latent Dirichlet Allocation (LDA) is applied to identify hidden patterns and relationships within the text dataset, enabling the extraction of the main topics emerging from the existing data. The performance of these four models is evaluated using Accuracy, Precision, Recall, and F1-score metrics.

2. Method

In this study, an analysis of public perception of the Free Nutritious Meal Program in Indonesia was conducted using sentiment analysis based on machine learning and deep learning methods. The proposed method is described in Figure 1. The first step involves data collection, where this study uses text data collected from the Kaggle platform originating from social media X containing user responses related to the implementation of the Free Nutritious Meal Program launched by the government. After the data exploration process, relevant responses discussing the program were filtered and collected, resulting in a total of 11.184 data points used for further analysis.

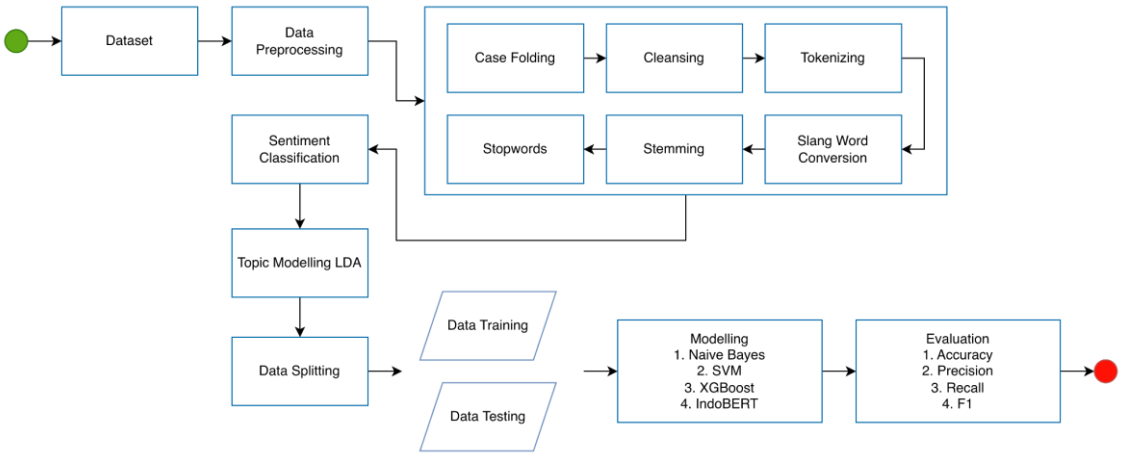

Figure 1. Workflow Diagram Method

The evaluation stage was conducted to test the model's efficacy in identifying patterns in the available dataset. After dividing the data into training and testing subsets, feature extraction was performed using the TF-IDF Vectorizer method to transform the text into a numerical vector representation. This step is crucial so that the data can be processed optimally by the machine learning algorithms used, namely Naïve Bayes, Support Vector Machine (SVM), XGBoost, and IndoBERT. The performance of the resulting model was then validated using a confusion matrix, with precision, recall, and F1-score metrics as the main indicators in measuring the accuracy of the classification results.

A. Dataset

The data used in this study was collected from the Kaggle Platform using Tweet Harvest Twitter (X) [15]. The data collected consists of tweets in Indonesian, covering

the period from January 6, 2025 to October 6, 2025. A total of 11.184 raw data entries were collected and stored in CSV format.

B. Data Processing

Data preprocessing is a stage carried out to clean, organize, and prepare data before analysis. Since this research is in the field of Natural Language Processing, the preprocessing stages applied include case folding, data cleansing, tokenization, slang word conversion, stemming and stopword removal. This preprocessing process aims to ensure that the extracted features are of good quality so that they can be used effectively in the model training and testing processes. Thus, the resulting model is expected to achieve an optimal level of accuracy.

1. Case Folding

The first preprocessing step is case folding, which aims to standardize the text by converting all characters into lowercase. This step is performed to ensure consistency in text representation and to reduce variations caused by differences in letter casing during further natural language processing (NLP) tasks, as seen in Table 1.

TABLE 1. RESULT CASE FOLDING

Raw Text	Result Case Folding
Dampak Program MBG Kepala BPOM Aceh Yudi Noviandi bersama Kepala Dinas Pendidikan Kota Banda Aceh Sulaiman Bakri hadir sebagai narasumber dalam dialog publik RRI Pro 1 Banda Aceh. Saksikan video lengkapnya di kanal YouTube RRI Banda Aceh #BPOMAceh #ppidbpomaceh https://t.co/74RA5IdJEB	dampak program mbg kepala bpom aceh yudi noviandi bersama kepala dinas pendidikan kota banda aceh sulaiman bakri hadir sebagai narasumber dalam dialog publik rri pro 1 banda aceh. saksikan video lengkapnya di kanal youtube rri banda aceh #bpomaceh #ppidbpomaceh https://t.co/74ra5idjeb

2. Cleansing

The second step is data cleaning, which is done to remove irrelevant elements from the text, including usernames, retweet indicators, hashtags, single characters, numerical values, URLs, punctuation marks, and excessive spaces. In addition, emoticons are removed to improve the quality of the text data for further analysis, as seen in Table 2.

TABLE 2. RESULT CLEANSING

Raw Text	Result Cleansing
dampak program mbg kepala bpom aceh yudi noviandi bersama kepala dinas pendidikan kota banda aceh sulaiman bakri hadir sebagai narasumber dalam dialog publik rri pro 1 banda aceh. saksikan video lengkapnya di kanal youtube rri banda aceh #bpomaceh #ppidbpomaceh https://t.co/74ra5idjeb	dampak program mbg kepala bpom aceh yudi noviandi sama kepala dinas didik kota banda aceh sulaiman bakri hadir narasumber dialog publik rri pro banda aceh saksi video lengkap kanal youtube rri banda aceh

3. Tokenizing

Tokenization is performed to break sentences into individual words, which

represent the smallest units of meaning in a text and serve as the basic elements for further text processing, as seen in Table 3.

TABLE 3. RESULT TOKENIZING

Raw Text	Result Tokenizing
dampak program mbg kepala bpom aceh yudi noviandi sama kepala dinas didik kota banda aceh sulaiman bakri hadir narasumber dialog publik rri pro banda aceh saksi video lengkap kanal youtube rri banda aceh	‘dampak’, ‘program’, ‘mbg’, ‘kepala’, ‘bpom’, ‘aceh’, ‘yudi’, ‘noviandi’, ‘sama’, ‘kepala’, ‘dinas’, ‘didik’, ‘kota’, ‘banda’, ‘aceh’, ‘sulaiman’, ‘bakri’, ‘hadir’, ‘narasumber’, ‘dialog’, ‘publik’, ‘rri’, ‘pro’, ‘banda’, ‘aceh’, ‘saksi’, ‘video’, ‘lengkap’, ‘kanal’, ‘youtube’, ‘rri’, ‘banda’, ‘aceh’

4. Slang Word Conversion

   Text standardization is implemented to address linguistic irregularities by converting non-standard vocabulary into formal Indonesian. Furthermore, negation units within the text are grouped to enhance analytical accuracy. For instance, phrases such as 'tidak' and 'bisa' are transformed into 'tidak_bisa', enabling the system to recognize the entity as a single feature that preserves the negative semantic meaning.

5. Stemming

   Stemming is applied to return words to their base form by removing affixes, so that different word variations can be treated as a single term during the analysis process, as seen in Table 4.

TABLE 4. RESULT STEMMING

Raw Text	Result Stemming
dampak program mbg kepala bpom aceh yudi noviandi sama kepala dinas didik kota banda aceh sulaiman bakri hadir narasumber dialog publik rri pro banda aceh saksi video lengkap kanal youtube rri banda aceh	['dampak', 'program', 'mbg', 'bpom', 'aceh', 'yudi', 'noviandi', 'sama', 'dinas', 'duduk', 'kota', 'banda', 'aceh', 'sulaiman', 'bakri', 'hadir', 'dialog', 'publik', 'rri', 'pro', 'banda', 'aceh', 'saksi', 'video', 'lengkap', 'kanal', 'youtube', 'rri', 'banda', 'aceh']

6. Stopwords

   Stopword removal is performed to eliminate common conjunction words, such as and, from, as well as, or, but, and other frequently occurring terms that do not contribute significant meaning to the analysis, as seen in Table 5.

TABLE 5. RESULT STOPWORDS

Raw Text	Result STOPWORDS
['dampak', 'program', 'mbg', 'bpom', 'aceh', 'yudi', 'noviandi', 'sama', 'dinas', 'duduk', 'kota', 'banda', 'aceh', 'sulaiman', 'bakri', 'hadir', 'dialog', 'publik', 'rri', 'pro', 'banda', 'aceh', 'saksi', 'video', 'lengkap', 'kanal', 'youtube', 'rri', 'banda', 'aceh']	['mbg', 'bpom', 'aceh', 'yudi', 'noviandi', 'dinas', 'didik', 'kota', 'banda', 'aceh', 'sulaiman', 'bakri', 'rri', 'banda', 'aceh']

C. Sentiment Classification

After the preprocessing stage, a total of 11.184 clean data samples were obtained and

prepared for further analysis. The next step involved sentiment tagging to provide contextual information for the machine learning classification process. Public perceptions were categorized into sentiment classes, including positive, neutral, and negative. This tagging process enables the model to understand the sentiment expressed in the text data. Sentiment classification in this study was performed using a RoBERTa-based model [16], namely indonesian-roberta-base-sentiment-classifier, which utilizes contextual word representations to effectively classify sentiment in Indonesian text, as seen in Table 6.

TABLE 6. DATA DISTRIBUTION BASED ON SENTIMENT CLASSIFICATION

Sentiment Class	Amount	Percentage (%)
neutral	5.644	50.46%
positive	2.934	26,23%
negative	2.606	23.30%

D. Topic Modeling Latent Dirichlet Allocation (LDA)

Topic modeling using the Latent Dirichlet Allocation (LDA) method [17]. LDA is a text mining technique used to identify hidden patterns and relationships in text data sets, so that the main topics that emerge can be extracted. This allows the main topics discussed in the data to be clearly identified. A visual representation of the LDA architecture is presented in Figure 2.

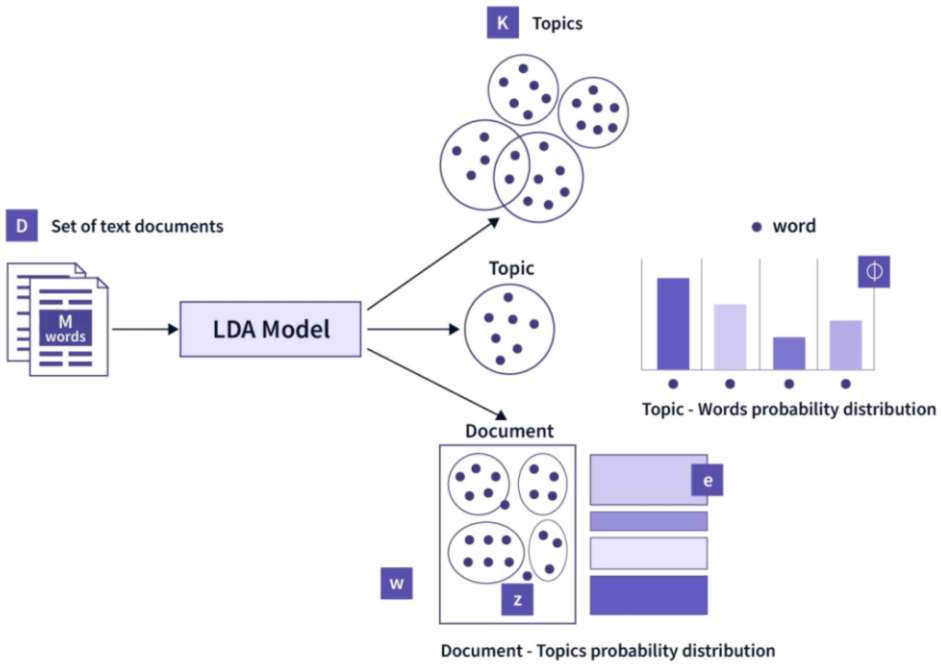

Figure 2. Architecture Latent Dirichlet Allocation (LDA)

As illustrated in Figure 2, LDA treats documents as probabilistic blends of topics, while defining each topic as a distribution across a vocabulary. By analyzing word co-occurrence within a corpus, the model extracts latent themes to generate two primary metrics: the topic word distribution (defining a topic's content) and the document topic distribution (quantifying a document's thematic makeup). These outputs facilitate a structured examination of the underlying themes within a dataset.

E. Data Splitting

Data splitting is performed to evaluate the performance of the machine learning model by dividing the dataset into training and testing subsets. In this study, the data were split using an 80:20 ratio, where 8,948 data samples were used for training and 2,237 data samples were allocated for testing, based on a total of 11,184 data. This ratio was selected to strike a balance between providing sufficient data for the model to capture underlying patterns while maintaining a substantial, independent set for robust performance evaluation. To ensure the representative distribution of classes in both subsets, stratified sampling was employed. The training data were utilized to build the model, while the testing data were used to assess the model's performance on unseen data.

F.Modelling

Before entering the modeling stage, text data undergoes a feature extraction process using TF-IDF Vectorizer to transform unstructured textual information into numerical vector representations. The TF-IDF mechanism integrates two main components: Term Frequency (TF), which measures the density of word occurrence in a document, and Inverse Document Frequency (IDF), which serves to reduce the weight of common words that appear massively throughout the dataset. The result of this stage is a vector representation that reflects the significance of each term in the document.

Classification is performed by implementing various machine learning architectures. This study uses Naïve Bayes, a probabilistic approach that refers to Bayes' theorem with the assumption of feature independence. In addition, Support Vector Machine (SVM) is used, which works by mapping the optimal hyperplane to separate classes through maximum margin. The ensemble learning technique is also applied through XGBoost, which integrates several gradient boosting-based decision trees to improve accuracy. Finally, the transformer-based deep learning model, IndoBERT, is used due to its advantages that have undergone pre-training on the Indonesian language corpus.

G. Evaluation

The performance of the model that has been tested on the training and testing datasets is then analyzed using a confusion matrix. This evaluation matrix is commonly used in classification tasks to present a statistical summary of the accuracy of the model's predictions, both in terms of correctly identified categories and errors that occurred during the classification process. Model performance analysis was conducted by referring to four main variables: True Positive, False Positive, False Negative, and True Negative. These four variables form the basis for calculating precision, recall, and F1-score. In the context of this study, model performance was measured using these three metrics to verify the effectiveness of the classification results.

3. Result and Analysis

Sentiment analysis of the Free Nutritious Food initiative is conducted in this study by utilizing a variety of computational models, including IndoBERT, XGBoost, SVM, and Naïve Bayes, to interpret public perception. Sentiment classification is applied directly to text data using a RoBERTa-based model, and topic modeling is performed using the Latent Dirichlet Allocation (LDA) method to extract dominant themes from the responses. The labeling process yielded the following data distribution: Neutral class with 5.644 data points, Positive class with 2.934 data points and Negative class with 2.606 data points.

The study implements cost-sensitive learning by adjusting class weights and prioritizes the Macro averaged F1-Score as the primary evaluation metric to ensure that the performance on the minority classes (Positive and Negative) is measured with equal importance to the majority class [18]. These strategies ensure that the IndoBERT and RoBERTa-based models maintain high sensitivity across all sentiment categories despite the inherent data disproportion.

The research began with the acquisition of user tweets from the Kaggle platform via the Tweet Harvest tool. This process yielded a dataset of 11,184 reviews, representing public discourse on Indonesia's Free Nutritious Food (MBG) initiative. The primary attributes extracted during this data collection phase are detailed in Table 7.

TABLE 7. DATA ATTRIBUTES

Attributes	Descriptions
created_at	Shows the date and time each post or comment was made
full_text	Contains the original text from the post or comment

The initial dataset obtained from social media X contains raw text related to the Free Nutritious Food Program. Before analysis, this data underwent processing to improve the quality and consistency of the text. The processing stages applied included case folding, cleansing, slang word conversion, stemming and stopwords. The attributes used after the processing stage can be seen in Table 8.

TABLE 8. DATA ATTRIBUTES AFTER PROCESSING

Attributes	Descriptions
created_at	Shows the date and time each post or comment was made
text_clean	Contains text that has undergone a cleaning process
sentiment	Predicted sentiment categories, classified as positive, neutral, and negative.
confidence	The probability value or confidence level of the model regarding sentiment prediction, usually in the range of 0–100%.

This data processing attribute data that converts raw text data into a clean and structured dataset, which is necessary for sentiment analysis. With this data, sentiment predictions can be made with greater accuracy and reliability in interpreting public responses.

In topic modelling was performed using cleaned text data (text_clean). Word frequencies were extracted with CountVectorizer using the parameters max_df=0.95, min_df=2, and a maximum of 1,000 features, thereby forming a representative document-word matrix. The Latent Dirichlet Allocation (LDA) model was then trained on this matrix with five topics, using max_iter=10, learning_method='online', and random_state=42 to ensure consistency of results.

The LDA results identified five main topics that reflect the dominant aspects in the text documents. Topic labels were determined based on the words with the highest contribution to each topic. In addition, semantic embeddings were generated using

SentenceTransformer (all-MiniLM-L6-v2) to support further analysis and measurement of similarity between texts.

The sentiment distribution in the dataset is shown in Figure 3. The analysis shows that neutral sentiment dominates user tweets, with a percentage of 50.5%. Positive and negative sentiments each accounted for 26.2% and 23.3% respectively. These findings indicate that most people respond neutrally to government policies related to free nutritious food.

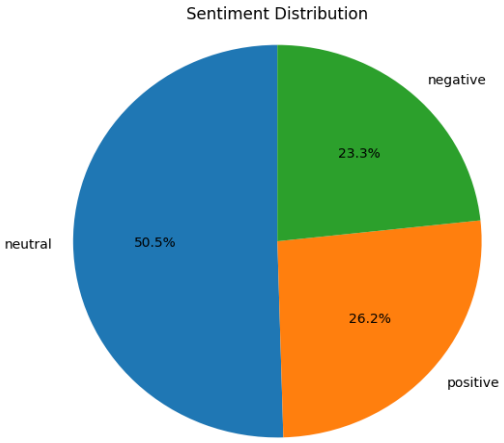

FIGURE 3. SENTIMENT DISTRIBUTION

The semantic landscape of the dataset was comprehensively analyzed using BERTopic, revealing distinct thematic clusters. The Intertopic Distance Map in Figure 4 provides a two-dimensional projection of these topics, illustrating their semantic proximity. While Topic 0, 1, and 2 appear in closer proximity, suggesting shared contextual elements, Topic 4 is positioned more distinctly, indicating that this topic has unique linguistic characteristics and differs significantly from the mainstream discussion in the dataset.

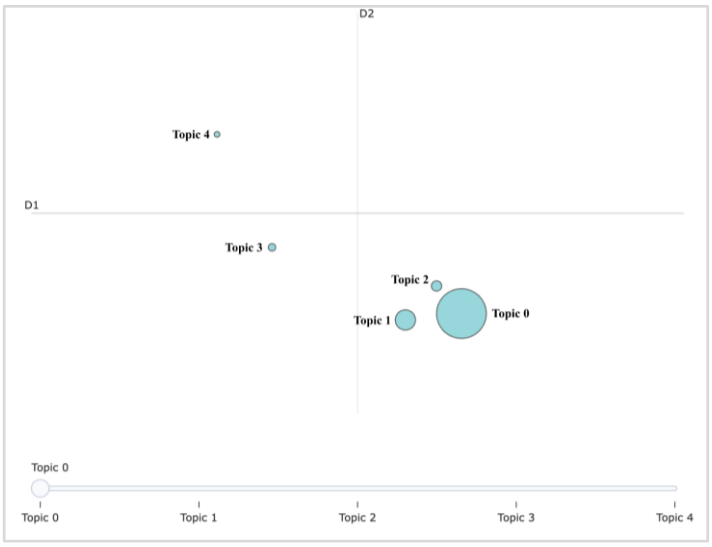

FIGURE 4. INTERTOPIC DISTANCE MAP

The Distribution of Dominant Topics in Figure 5 quantifies the prevalence of each cluster within the dataset. It highlights that Topic 3 is the most dominant theme, encompassing 48.2% (5,390 documents) of the total corpus. This is followed by Topic 4

(29.1%), while Topics 0, 1, and 2 represent smaller, more specialized discussions, contributing 5.4%, 11.3%, and 5.9% respectively.

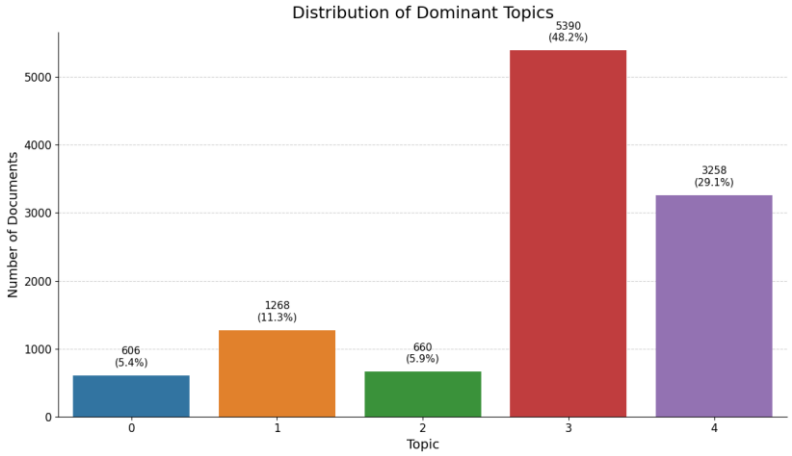

FIGURE 5. DISTRIBUTION OF DOMINANT/PROBABILITY TOPICS

To further interpret the semantic composition of each identified cluster, a keyword contribution analysis was performed. As shown in the Topic Contribution graph in Figure 6, Topic 0 is predominantly characterized by political and administrative figures, with 'Prabowo' (28.6%) and 'Presiden' (25.3%) acting as the primary descriptors. This suggests that the largest topic is heavily framed around government leadership. Interestingly, Topic 1 reveals a more critical or specific discourse, where the term 'racun' (26.4%) appears significantly alongside 'mbg' and 'makan', indicating potential sub-discussions regarding food safety or controversial issues within the program.

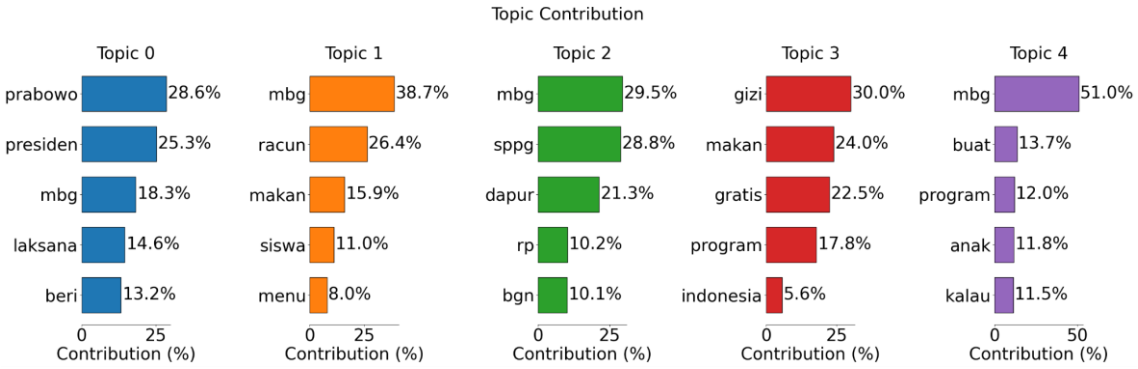

FIGURE 6. TOPIC CONTRIBUTION

The semantic landscape of the identified topics was further elucidated through a word cloud analysis in Figure 7, which highlights the most frequent and significant terms within the corpus. As illustrated in the visualization, the discourse is heavily centralized around the keywords 'makan', 'gizi', and 'gratis', which serve as the core linguistic anchors of the dataset. These terms, alongside associated descriptors such as 'program', 'sekolah', and 'anak', confirm that Topic 0 the largest cluster represents public discourse regarding the national free nutritious meal program. The prominence of terms like 'presiden prabowo' and 'dukung' suggests a strong thematic link between the government's policy initiatives and public sentiment. Furthermore, the presence of localized and evaluative

terms such as 'masa depan' and 'nasional' indicates that the discussion transcends mere logistical aspects, touching upon broader socio-economic aspirations. This textual distribution aligns with the spatial proximity observed in the *Intertopic Distance Map*, where the dominance of the primary cluster is driven by high-frequency, interrelated keywords that define the central narrative of the study.

FIGURE 7. WORLD CLOUD

After analyzing the dataset, the next step is the modeling process to build a model that can perform sentiment analysis on the processed dataset. In the modeling process, the model is trained using training data so that it can learn and recognize patterns in the data. After the model has successfully learned these patterns, testing is carried out using test data. The sentiment results for each identified class are then evaluated and presented in Table 9.

TABLE 9. PERFORMANCE MODEL BASED ON SENTIMENTS

Model	Class	Precision	Recall	F1-Score
Naïve Bayes	Negative	0.6745	0.7678	0.7181
	Neutral	0.7938	0.8459	0.8190
	Positive	0.7778	0.5843	0.6673
Accuracy NB				**0.7591**
SVM	Negative	0.7192	0.7620	0.7400
	Neutral	0.8365	0.8521	0.8442
	Positive	0.7776	0.7087	0.7415
Accuracy SVM				**0.7935**
XGBoost	Negative	0.69	0.71	0.70
	Neutral	0.78	0.85	0.82
	Positive	0.78	0.63	0.69
Accuracy XGBoost				**0.76**
IndoBERT	Negative	0.9080	0.7198	0.8030
	Neutral	0.8743	0.9114	0.8925
	Positive	0.7821	0.8620	0.8201
Accuracy IndoBERT				**0.8538**

Based on the performance evaluation in Table 9, the IndoBERT model shows superior performance compared to traditional machine learning models (Naïve Bayes, SVM, and XGBoost) in classifying Indonesian text sentiment. This model achieved an overall accuracy of 0.854 or 85%, confirming the transformer's ability to understand linguistic context and capture semantic nuances in unstructured text data. This advantage is particularly evident in the model's ability to handle unbalanced sentiment distributions and identify subtle patterns in public opinion expressions.

Analysis by sentiment class shows more in-depth results. In the neutral class, IndoBERT achieved a precision of 0.874 or 87%, a recall of 0.911 or 91%, and an F1-score of 0.893 or 89%, indicating that this model is not only capable of accurately classifying neutral texts, but also has high sensitivity in capturing truly neutral texts, thereby minimizing false negatives. For the positive class, an F1-score of 0.820 or 82% indicates that the model is able to consistently recognize expressions of support or positive perceptions, even though positive texts often have a wide variety of vocabulary. In the negative class, an F1-score of 0.803 or 80% shows that this model is still able to detect critical opinions or negative sentiments, even though they are usually less represented than the neutral and positive classes.

When compared to other models, it is evident that Naïve Bayes and XGBoost have significant weaknesses in the positive class F1-score of 0.667 or 67% and 0.690 or 69%, indicating their limitations in capturing vocabulary variation and semantic context. SVM performs relatively better than these two traditional models, but still lags behind IndoBERT on all metrics. The superiority of IndoBERT shows that transformer-based contextual representations are capable of extracting complex language patterns, considering word interactions in sentence context, and distinguishing subtle nuances of sentiment.

Overall, these results confirm that the implementation of IndoBERT not only significantly improves classification accuracy but also produces more balanced predictions across all sentiment classes compared to traditional models. The performance of IndoBERT is primarily attributed to its Transformer based architecture, which utilizes a self attention mechanism to generate contextualized word representations. Unlike traditional machine learning models such as Naïve Bayes, SVM, and XGBoost.

The performance gap observed, particularly in the lower F1-scores of Naïve Bayes and XGBoost for the positive and negative classes, stems from their reliance on static frequency-based features. These traditional models treat words as independent entities, failing to account for word order and semantic shifts. Consequently, they struggle to generalize when faced with the diverse vocabulary and informal structures common in public opinion data. In contrast, IndoBERT’s pre-training on a massive Indonesian corpus enables it to extract complex linguistic patterns that traditional algorithms overlook. With this capability, Indonesian language-based transformer models can be reliably used for high-stakes public opinion analysis, data-driven decision making, and government policy evaluation, especially when the analyzed data is unstructured and contains highly diverse expressions of sentiment.

4. Conclusion

Based on the results of this study, a total of 11.184 data points obtained from Kaggle were utilized and deemed suitable for analysis. Prior to the training process, the dataset underwent a comprehensive preprocessing stage, which included case folding to convert all text to lowercase, cleansing to remove irrelevant characters or words in the

comments such as usernames, tokenization to decompose sentences into meaningful lexical units, slang word conversion to change informal terms into standard formal forms, stemming to eliminate affixes from words, and stopword removal to filter out conjunctions and other non-informative terms.

Following the preprocessing stage, the dataset was partitioned into training and testing sets. The training process was then conducted using four classification models, namely Naïve Bayes, Support Vector Machine (SVM), XGBoost, and IndoBERT. The performance of each model was subsequently evaluated using precision, recall, and F1-score as evaluation metrics. The experimental results indicate that among the four models, IndoBERT achieved the most optimal performance for sentiment analysis, with an F1-score of 0.85. Nevertheless, this performance can potentially be further improved by increasing the size of the training dataset, which may lead to a higher F1-score approaching 0.9.

For future research, it is recommended to address the issue of class imbalance within the dataset and to explore transformer-based architectures alongside hybrid deep learning models, while optimizing hyperparameters, applying advanced imbalance handling techniques, experimenting with alternative architectures, and evaluating the models on larger and more diverse datasets in order to enhance model performance, generalizability, and overall robustness.

References

[1] Indonesian Internet Service Providers Association (APJII), "Lanskap Pengguna Internet Indonesia 2025_ Koneksi Meluas, Literasi Masih Lemas – internetsehat." 2025, [Online]. Available: https://internetsehat.id/.

[2] F. Sudirjo, K. Diantoro, J. A. Al-gasawneh, H. K. Azzaakiyyah, and M. A. Ausat, "Application of ChatGPT in Improving Customer Sentiment Analysis for Businesses," vol. 5, no. 3, pp. 283–288, 2023, [Online]. Available: https://doi.org/10.47233/jteksis.v5i3.871 .

[3] H. Taherdoost and M. Mitra, "Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research." pp. 1–15, 2023.

[4] S. W. Nisa Akbarilah Putri, Anindya Apriliyanti Pravitasari, "Topic-based sentiment analysis of Jamsostek mobile application reviews using the BERTopic method," *Commun. Math. Biol. Neurosci*, pp. 1–20, 2026, [Online]. Available: https://doi.org/10.28919/cmbn/9682.

[5] R. Hajrizi and K. P. Nuçi, "Aspect-Based Sentiment Analysis in Education Domain," 2020.

[6] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Nat. Lang. Process. J.*, vol. 2, no. December 2022, p. 100003, 2023, doi: 10.1016/j.nlp.2022.100003.

[7] S. Adi, S. Mola, Y. T. Polly, and Y. C. Atok, "Analisis Sentimen Terhadap Data Komentar Publik Mengenai Isu UU Pilkada 2024 Menggunakan Metode Naïve Bayes dan K-Nearest Neighbor," vol. 5, no. 3, pp. 337–343, 2025, doi: 10.47065/jimat.v5i3.514.

[8] A. Haris, K. Sandi, E. Haerani, L. Oktavia, and F. Kurnia, "Klasifikasi Sentimen Masyarakat Terhadap Revisi Undang-Undang Tentara Nasional Indonesia Menggunakan Naïve Bayes Classifier," vol. 5, no. 4, pp. 594–603, 2025, doi: 10.47065/bulletincsr.v5i4.615.

[9] A. A. Nurhasananda and M. Akbar, "Analisis Sentimen Masyarakat Terhadap Kebijakan Undang-Undang Tentara Nasional Indonesia (UU TNI) Menggunakan Support Vector Machine," vol. 5, no. 1, pp. 1–14, 2025, [Online]. Available: https://doi.org/10.53697/jkomitek.v5i1.2603.

[10] J. Li, Z. Li, X. Ma, Q. Zhao, C. Zhang, and G. Yu, "Sentiment Analysis on Online Videos by Time-Sync Comments," *entropy*, vol. 25, no. 1016, pp. 1–18, 2023, [Online]. Available: https://doi.org/10.3390/%0Ae25071016.

[11] D. A. Musleh, I. Alkhwaja, A. Alkhwaja, and M. Alghamdi, "Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation," *Big Data Cogn. Comput.*, vol. 7, no. 127, pp. 1–16, 2023, [Online]. Available: https://doi.org/10.3390/bdcc7030127%0A.

[12] R. Singh and A. Tiwari, "Youtube Comments Sentiment Analysis," *Int. J. Sci. Res. Eng. Manag.*, vol. 05, no. 05, pp. 1–11, 2021.

[13] R. Danar, D. M. Kom, and M. M. Kom, "Dasar Dasar Natural Language Processing (NLP)."

[14] W. Nurharjadmo, F. Ansoriyah, and M. A. Khadija, "Analyzing Public Perception using Aspect Based Sentiment Analysis : Case Study of Capital Relocation Planning of Indonesia," *2024 7th Int. Conf. Informatics Comput. Sci.*, pp. 191–196, 2024, doi: 10.1109/ICICoS62600.2024.10636903.

[15] Kaggle"MBG 06 Januari - 06 Oktober 2025." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/ravhihz/mbg-06-januari-06-oktober-2025.

[16] A. P. Maretta and A. Meiriza, "Aspect-Based Sentiment Analysis of Hospital Service Reviews Using Fine- Tuned IndoBERT," vol. 9, no. 5, 2025.

[17] E. Puspita, D. F. Shiddieq, and F. F. Roji, "Topic Modeling on Online News Media Using Latent Diriclet Allocation (Case Study Somethinc Brand) Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc)," vol. 4, no. April, pp. 481–489, 2024.

[18] V. Cerqueira, L. Torgo, P. Branco, and C. Bellinger, "Automated imbalanced classification via layered learning," Machine Learning, vol. 112, no. 6, pp. 2083–2104, 2023, doi: 10.1007/s10994-022-06282-w.