

Performance Comparison of 10 Machine Learning Algorithms in Sentiment Classification on Platform X Regarding the Government's Priority Program: *Makanan Bergizi Gratis* (MBG)

Deni Utama^{1*}, Rauhil Fahmi², Muhammad Ridho Kurniawan Pratama³

^{1,2,3}Information Systems and Technology, Faculty of Engineering, Universitas Negeri Jakarta, DKI Jakarta, Indonesia

E-mail: ^{1*}deniutama@unj.ac.id, ²rauhilfahmi@unj.ac.id, ³muhammadridho@unj.ac.id

(Received: 27 Sep 2025, revised: 15 Oct 2025, accepted: 16 Oct 2025)

Abstract

This study evaluates public sentiment on Platform X regarding the Government's Priority Free Nutritious Food Program. A total of 2031 user comments were analyzed using 10 machine learning algorithms: Naive Bayes, Gradient Boosting, AdaBoost, Random Forest, Extra Trees, Logistic Regression, Linear SVM, SGD Classifier, Ridge Classifier, and Bagging. The dataset underwent preprocessing including lowercasing, stopword removal, stemming, and tokenization, followed by TF-IDF vectorization with 5000 features. Models were evaluated using accuracy, precision, recall, weighted F1-score, and 5-fold cross-validation. Bagging achieved the highest accuracy (81%) and weighted F1-score (81%), followed by Gradient Boosting (81%) and Random Forest (77%). Feature analysis revealed negative sentiment indicators such as 'racun', 'stop', 'korupsi', and positive indicators like 'sehat', 'enak', 'bergizi'. These findings provide actionable insights for policy communication and program improvement.

Keywords: Sentiment Analysis, Machine Learning, Ensemble Methods, Naive Bayes, Gradient Boosting, Feature Importance

I. INTRODUCTION

Social media platforms and online news provide important insights into public perception of government programs, including the Free Nutritious Food Program (MBG) aimed at improving the nutrition of school children. Analyzing user comments and public opinions is essential to understand how the program is received, including strengths, weaknesses, and potential risks. Optimal nutrition at school age has a direct impact on physical growth, health, and cognitive development, making public evaluation a critical step before further interventions are implemented.

The National Nutrition Agency (BGN), a non-ministerial institution established through Presidential Regulation No. 83 Year 2024, is responsible for managing the MBG program and ensuring that every child receives adequate nutrition through school kitchens and hygienic menus [1]. However, recent issues indicate challenges in program implementation. For instance, KOMPAS.com reported that only 34 MBG kitchens have hygiene certificates [2], while Liputan6.com reported thousands of school children being poisoned due to MBG menus [3]. These cases highlight the importance of nutrition

and hygiene monitoring, as well as the need for sentiment analysis to identify public perception and potential health risks.

The MBG program targets students from primary to secondary education, including general schools, vocational schools, religious schools, and special education. BGN's strategy includes monitoring child nutrition status, reducing morbidity and mortality, preventing health problems caused by poor nutrition or unhygienic food, and supporting long-term economic growth through a healthy and productive generation. Public opinion analysis via sentiment allows stakeholders to adjust menus, improve kitchen procedures, and enhance service quality to ensure children's health and safety. Additionally, BGN provides geospatial data related to the Nutrition Fulfillment Program through the official portal gina.bgn.go.id to support national nutrition equity [2]. By integrating field data and sentiment analysis from school children and the public, the MBG program can be evaluated comprehensively, offering evidence-based feedback and building a healthy, smart, and resilient future generation while minimizing future health risks.



II. LITERATURE REVIEW

Sentiment analysis employs natural language processing (NLP) and machine learning techniques for text classification. These techniques aim to automatically classify text into categories such as positive, negative, or neutral. Preprocessing steps including TF-IDF vectorization, stopword removal, and stemming are essential to reduce noise and improve classification accuracy [4], [5], [6]. Feature interpretation is also important to identify key words that influence model predictions, providing insights for policy communication and program improvement [7], [8], [9], [10]. Early studies demonstrated the utility of basic probabilistic models like Naive Bayes for text classification, particularly on high-dimensional sparse datasets, despite limitations in capturing contextual word dependencies [4], [5], [11].

Ensemble methods, including Random Forest and Gradient Boosting, have been shown to achieve higher accuracy than single classifiers, particularly in sentiment analysis on Twitter and mobile application reviews [4], [5], [12]. AdaBoost has also proven effective in handling imbalanced datasets by adaptively reweighting misclassified instances [6], [13]. Meanwhile, linear models such as Logistic Regression and SVM remain widely used due to their interpretability and competitive performance in multi-class settings [7], [8], [14]. These models provide a baseline for understanding sentiment distributions while supporting feature analysis for actionable insights.

Recent research highlights the importance of integrating advanced preprocessing strategies with model selection. Techniques such as domain-specific stopword removal, stemming, and multi-word token handling improve both classification performance and interpretability [7], [8], [9], [10]. Additionally, studies have shown that combining TF-IDF with ensemble learning enhances the robustness of sentiment predictions, especially in noisy social media data [15], [16].

Social media platforms and online user reviews have become crucial sources for sentiment analysis. Studies using Twitter datasets, product reviews, and public health feedback have revealed that sentiment classification models benefit from large-scale and diverse datasets, which capture nuanced opinions and emergent trends [12], [13], [14], [17]. These studies also emphasize the challenges of sarcasm, informal language, and multilingual texts, which require careful preprocessing and feature engineering to ensure model reliability [17], [18].

Recent advancements in large language models (LLMs) have expanded sentiment analysis capabilities. Pre-trained LLMs can capture contextual dependencies and semantic nuances that traditional machine learning models might miss, leading to improved performance on both short-text and long-text datasets [17], [18], [19], [20]. Research shows that integrating LLM embeddings with classical classifiers or ensemble methods can further enhance accuracy and interpretability, providing richer insights for policy and program evaluation [18], [19], [21], [22].

Overall, the literature indicates a trend toward combining robust preprocessing, ensemble learning, and modern NLP techniques to improve sentiment classification performance and feature interpretability [7], [8], [9], [10], [19], [20], [21], [22]. These methods allow policymakers, program managers, and researchers to extract actionable insights from textual data efficiently, particularly in evaluating government programs like *Makanan Bergizi Gratis* (MBG).

III. RESEARCH GAP ANALYSIS

Recent studies in sentiment analysis have widely applied traditional or limited machine learning models, yet few have conducted comprehensive multi-algorithm comparisons within the context of Indonesian public programs. Several research gaps can be identified as follows:

1. **Limited Model Diversity in Previous Studies**
Prior works on sentiment analysis such as [5], [11], [23] mostly utilized one to three models, commonly Naive Bayes, Logistic Regression, or SVM, without exploring ensemble approaches like Bagging, Boosting, or Random Forest.
As a result, the comparative understanding of how different algorithms perform under identical conditions was incomplete.
2. **Lack of Contextual Application in Indonesian Public Policy**
Most sentiment analysis studies in Indonesia focused on mobile application reviews or commercial product feedback, e.g., JKN mobile app analysis [7], rather than large-scale government initiatives.
This study introduces novelty by applying sentiment classification to the "*Makanan Bergizi Gratis* (MBG)" program, which directly relates to public health and education policy, offering localized societal relevance.
3. **Minimal Discussion of Model Tuning and Computational Aspects**
Previous studies often reported accuracy scores without analyzing parameter optimization, computational complexity, or overfitting risks.
In contrast, this study provides a systematic evaluation of tuning processes (e.g., learning rate, tree depth, $n_estimators$) and discusses trade-offs between performance and training time for each model.
4. **Insufficient Feature Interpretability**
Earlier works rarely discussed which textual features (keywords) most influenced sentiment predictions.
The present study complements accuracy evaluation with feature importance analysis, identifying domain-specific positive and negative terms (e.g., *sehat*, *gizi*, *korupsi*, *racun*), thereby improving model transparency and practical interpretability.
5. **Limited Validation Approaches**
Prior research often relied solely on single train-test splits, which could bias results.

This paper implements 5-fold cross-validation for a more robust and unbiased performance comparison across ten algorithms.

Table 1. Summary of Identified Gaps

Aspect	Previous Studies	This Study (Proposed)
Number of Models	(NB, SVM, LR) [4], [11], [23]	10 ML models including ensemble methods
Application Domain	App/product reviews [5], [7]	Government policy sentiment (MBG)
Model Tuning	Rarely optimized	Grid Search + 5-Fold Cross Validation
Computational Discussion	Not analyzed	Included time–accuracy trade-off
Feature Interpretability	Minimal [5], [7], [8], [24]	Feature importance via TF-IDF and model coefficients
Validation Approach	Single split	Stratified 5-Fold cross-validation

This study bridges the existing research gap (Table 1) by:

- Conducting a broad comparative analysis across ten machine learning algorithms under uniform preprocessing and feature extraction conditions.
- Introducing Indonesian public program sentiment analysis as a novel and socially impactful domain.
- Providing both technical evaluation (accuracy, tuning, robustness) and semantic insights (feature interpretation) to enhance data-driven policy improvement.

IV. METHODOLOGY

4.1 Data Collection

Data collection in this study refers to the process of gathering user comments from Platform X related to the Makanan Bergizi Gratis (MBG) program. The search keyword “MBG” was used in the Indonesian language, covering the period from September 1 to September 20, 2025, which represented a time of rising public concern over MBG-related food poisoning incidents. The peak occurred when a mass poisoning event took place, prompting the government to declare an extraordinary situation in several regions. This period generated extensive media coverage and public discussion regarding the evaluation of the MBG program. The data were collected using web scraping methods implemented in Node.js to obtain textual data for sentiment analysis. Collecting data from online platforms enables access to large-scale user opinions that are valuable for sentiment classification [7], [8]. Effective sentiment analysis depends not only on the choice of model but also on the quality and diversity of the data collected from social platforms [24]. A total of 2,031 comments were gathered and labeled as positive, negative, or neutral using a bag-of-words approach.

positif words = ['bagus', 'mantap', 'enak', 'lezat', 'lancar',

'puas', 'terbaik', 'keren', 'support', 'terbantu', 'mudah', 'cepat', 'mantul', 'sukses', 'apresiasi', 'transparan', 'favorit', 'senang',...]
negatif_words = ['gagal', 'susah', 'pahit', 'racun', 'basi', 'stop', 'kadaluarsa', 'ribet', 'lelet', 'buruk', 'jelek', 'kecewa', 'tidak enak', 'tidak enak', 'mengecewakan', 'korupsi', 'sampah',...]

After being labeled by the bag of words, the labeler is then validated again manually so that it is more precise if there are words that are sarcastic. The resulting sentiment distribution consisted of 43% neutral, 33% negative, and 24% positive posts out of the total 2,031 comments.

4.2 Preprocessing

Text preprocessing transforms raw text into a clean and structured format, which is crucial for improving model performance in sentiment analysis [6]. Preprocessing steps like stop word removal, lowercasing, and stemming significantly affect classification accuracy in Twitter sentiment analysis [12]. Text preprocessing included the following steps in Figure 1:

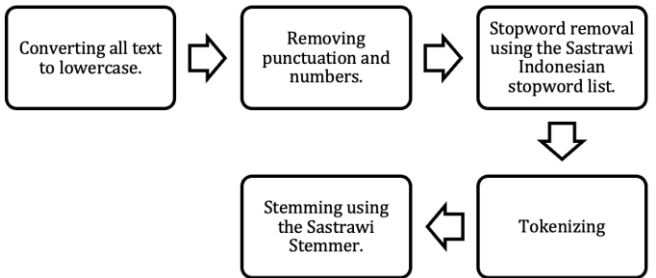


Figure 1. Flow Preprocessing

1. Converting all text to lowercase – standardizes tokens for consistency
2. Removing punctuation and numbers – reduces noise and dimensionality
3. Stopword removal (Sastrawi) – omits high-frequency, uninformative words
4. Stemming (Sastrawi Stemmer) – reduces words to canonical forms while preserving semantics
5. Tokenizing – splits text into discrete words for mapping into TF-IDF or embeddings

The resulting clean text was then vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) method to generate weighted features.

4.3 Algorithms

Naive Bayes is a probabilistic classifier based on Bayes’ theorem, assuming feature independence. It is commonly used for high-dimensional text data because of its simplicity and efficiency [4]. The probability of a class C given a text document d is calculated as shown in Equation 1:

$$P(C | d) = \frac{P(C) \prod_i P(x_i | C)}{P(d)}$$
 (1)



where X_i represents the features (words) in the document, $P(C)$ is the prior probability of class C , and $P(X_i | C)$ is the likelihood of word X_i given class C .

Random Forest is an ensemble method that combines multiple decision trees to improve predictive performance and reduce overfitting [23]. Each tree is trained on a bootstrap sample of the data, and feature subsets are randomly selected at each split. The final prediction is made by majority voting as shown in Equation 2:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (2)$$

where $h_t(x)$ is the prediction of the t -th tree, and T is the total number of trees.

Gradient Boosting builds an ensemble of weak learners sequentially, minimizing residual errors of previous models [13]. The prediction is updated iteratively as shown in Equation 3:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (3)$$

where $F_m(x)$ is the ensemble prediction at iteration m , $h_m(x)$ is the new weak learner, and γ_m is the learning rate.

AdaBoost combines weak classifiers to form a strong classifier by weighting misclassified instances more heavily in successive rounds [14]. The final classifier is shown in Equation 4:

$$H(x) = \text{sign}\left(\sum_t \alpha_t h_t(x)\right) \quad (4)$$

where α_t is the weight of the t -th weak classifier $h_t(x)$, determined by its classification error.

Logistic Regression is a linear model used for binary or multiclass classification. The probability of a class C is modeled using the sigmoid function as shown in Equation 5:

$$P(C = 1 | x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}} \quad (5)$$

where β_i are the model coefficients and X_i are the input features.

Linear SVM finds the optimal hyperplane that separates classes with maximum margin. The decision function is shown in equation 6:

$$f(x) = w^T x + b \quad (6)$$

where w is the weight vector, b is the bias, and the class prediction is based on the sign of $f(x)$.

SGD is an iterative optimization algorithm used for large-scale linear models. Weight updates for each sample i is shown in Equation 7:

$$w_{t+1} = w_t - \eta \nabla L(w_t; x_i, y_i) \quad (7)$$

where η is the learning rate, and ∇L is the gradient of the loss function with respect to w .

Ridge Classifier is a linear model with L2 regularization to prevent overfitting. The objective function minimized is shown in Equation 8:

$$\min_w \|Xw - y\|_2^2 + \alpha \|w\|_2^2 \quad (8)$$

where α is the regularization parameter, X the feature matrix, and y the target labels.

Bagging reduces variance by averaging predictions from multiple base models trained on bootstrap samples [23]. For classification is calculated in Equation 9:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_B(x)\} \quad (9)$$

where B is the number of base learners and $h_b(x)$ their predictions.

4.4 Model Evaluation

Model evaluation is the process of measuring the performance of a machine learning model using metrics such as accuracy, precision, recall, and F1-score. This evaluation is essential for determining how well the model correctly classifies data. Model evaluation metrics such as accuracy and F1-score provide insights into model performance, enabling comparison between classifiers [12]. Accuracy alone is not sufficient for evaluating sentiment analysis models, especially in the case of imbalanced datasets; precision and recall must be considered [13].

The models were evaluated using the following metrics in Equation 10-13:

$$\bullet \text{ Accuracy : } \left[\frac{TP+TN}{TP+TN+FP+FN} \right] \quad (10)$$

$$\bullet \text{ Precision : } \left[\frac{TP}{TP+FP} \right] \quad (11)$$

$$\bullet \text{ Recall : } \frac{TP}{TP+FN} \quad (12)$$

$$\bullet \text{ F1-Score : } \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

Training used 70% of data; 30% was reserved for testing. Models were evaluated using accuracy, precision, recall, weighted F1-score, and 5-fold cross-validation. Feature importance was extracted using `feature_importances_`, `coef_`, or `log probabilities`.

V. RESULTS AND DISCUSSION

5.1 Model Performance

Performance metrics are summarized in Table 2. The performance of ten machine learning algorithms — Naive Bayes, Gradient Boosting, AdaBoost, Random Forest, Extra Trees, Logistic Regression, Linear SVM, SGD Classifier, Ridge Classifier, and Bagging — was evaluated using Accuracy, Precision, Recall, F1-score (weighted), and 5-Fold Cross-Validation Accuracy.

Table 2. Performance Comparison of 10 Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	CV Accuracy
Naive Bayes	62%	65%	53%	56%	60.22%
Gradient Boosting	81%	82%	77%	78%	79.47%
AdaBoost	72%	76%	64%	64%	69.43%
Random Forest	77%	78%	72%	73%	77.20%
Extra Trees	77%	77%	73%	74%	76.42%
Logistic Regression	72%	73%	65%	70%	71.25%
Linear SVM	72%	70%	68%	72%	72.77%
SGD Classifier	69%	67%	67%	69%	71.00%
Ridge Classifier	71%	70%	67%	70%	71.64%
Bagging	81%	81%	80%	81%	80.06%

Among the ten evaluated models, Bagging achieved the highest accuracy (81%) and weighed F1-score (81%), indicating strong overall performance and stability across classes. In contrast, Naive Bayes underperformed in detecting positive sentiment, likely due to its assumption of feature independence, which may not fully capture contextual word dependencies in user comments. Other ensemble methods such as Gradient Boosting and AdaBoost showed competitive performance, benefiting from sequential learning and adaptive weighting of misclassified instances.

The superior performance of ensemble-based models can be attributed to their ability to combine multiple weak learners to reduce variance and bias, thereby improving generalization. By aggregating predictions from diverse models, ensembles are better able to capture complex patterns in the data and are less sensitive to noisy or imbalanced datasets. Bagging, for example, leverages bootstrap sampling and aggregation to stabilize predictions across different subsets of the training data, while boosting methods iteratively focus on correcting errors made by previous learners, enhancing the model’s ability to handle difficult instances.

Overall, these results indicate that ensemble-based algorithms generally outperform simple probabilistic or linear models in multi-class sentiment classification, especially when handling imbalanced datasets. Furthermore, the inclusion of 5-fold cross-validation ensures that the reported metrics are robust and not biased by a particular train-test split, indicating that the models do not suffer from overfitting.

5.2 Feature Analysis

Feature importance and class-specific indicators were analyzed using the ten trained machine learning models as shown in Table 3. Naive Bayes highlighted *racun*, *evaluasi*, *stop*, *korupsi*, *masalah* as top features for negative sentiment, reflecting frequent complaints and dissatisfaction, while positive sentiment was indicated by *makanbergizigratis* and

enak, capturing satisfaction and perceived program usefulness. However, Naive Bayes showed low F1-score for positive sentiment (0.14) and a 5-fold cross-validation accuracy of 0.6022, likely due to its assumption of feature independence.

Table 3. Top Feature Analysis of 10 Machine Learning Models

Model	Top Negative Features	Top Positive Features	5-Fold CV Accuracy
Naive Bayes	<i>racun, evaluasi, stop, korupsi, masalah</i>	<i>Makan bergizi, e</i>	0.6022
Gradient Boosting	<i>racun, stop, baik, evaluasi, korupsi, masalah, basi</i>	<i>makan, sehat, gizi</i>	0.7947
AdaBoost	<i>racun, korupsi, evaluasi, masalah, stop, basi, jelek, sampah</i>	<i>baik, gizi</i>	0.6943
Random Forest	<i>racun, stop, evaluasi, anak, kasus</i>	<i>makan, mbg, baik, sehat, gratis</i>	0.7720
Extra Trees	<i>racun, stop, evaluasi, kasus, korupsi</i>	<i>makan, baik, mbg, sehat, gizi</i>	0.7642
Logistic Regression	<i>racun, stop, evaluasi, korupsi, basi, masalah</i>	<i>sehat, enak, gizi, bermanfaat</i>	0.7125
Linear SVM	<i>racun, stop, evaluasi, masalah, basi, korupsi</i>	<i>sehat, enak, bersih</i>	0.7277
SGD Classifier	<i>racun, evaluasi, masalah, stop, basi, korupsi</i>	<i>sehat, baik, bersih, kenyang</i>	0.7100
Ridge Classifier	<i>racun, stop, evaluasi, basi, masalah, korupsi</i>	<i>sehat, baik, enak, bersih</i>	0.7164
Bagging	<i>racun, masalah</i>	<i>sehat, enak, gizi, makan bergizi</i>	0.8006

Gradient Boosting emphasized both complaint and appreciation terms, with top features including *racun*, *stop*, *baik*, *evaluasi*, *korupsi*, *makan*, *sehat*, *gizi*, *masalah*, and *basi*, achieving a 5-fold cross-validation accuracy of 0.7947. AdaBoost identified *racun*, *korupsi*, *evaluasi*, *masalah*, *stop*, *basi*, *baik*, *jelek*, *sampah*, and *gizi* as influential, highlighting technical issues alongside positive feedback (CV accuracy 0.6943). Random Forest and Extra Trees consistently marked *racun*, *stop*, *makan*, *mbg*, *evaluasi*, *baik*, *sehat*, *anak*, *kasus*, and *gratis* (RF CV 0.7720; ET CV 0.7642), indicating strong importance of both negative and positive sentiment features.

Linear models, including Logistic Regression, Linear SVM, SGD Classifier, and Ridge Classifier, also captured similar patterns. Negative sentiment was associated with *racun*, *stop*, *evaluasi*, *korupsi*, *basi*, and *masalah*, while positive sentiment featured *sehat*, *enak*, *gizi*, *bermanfaat*, and *bersih*. Their 5-fold cross-validation accuracies ranged from 0.7100 to 0.7277, demonstrating stable, though slightly lower, predictive power compared to ensemble methods.

Bagging showed the best overall performance, with top features including *racun*, *sehat*, *enak*, *masalah*, *gizi*, and *makanbergizigratis*, capturing both negative and positive



sentiment effectively, and achieving a 5-fold cross-validation accuracy of 0.8006. Across all models, negative sentiment was consistently driven by words denoting complaints or problematic aspects, while positive sentiment was influenced by words reflecting enjoyment, health benefits, and program usefulness. These findings indicate that ensemble-based models more robustly identify key sentiment-driving features, and suggest that enhanced preprocessing, such as multi-word token handling and domain-specific stopwords removal, could further improve feature discrimination and model interpretability.

To visually explore the most frequent terms in user comments regarding the *Makanan Bergizi Gratis* (MBG) program, a Word Cloud was generated from the preprocessed dataset. The Word Cloud highlights words that appear more frequently by displaying them in larger font sizes, allowing for an intuitive understanding of dominant topics and sentiment indicators.

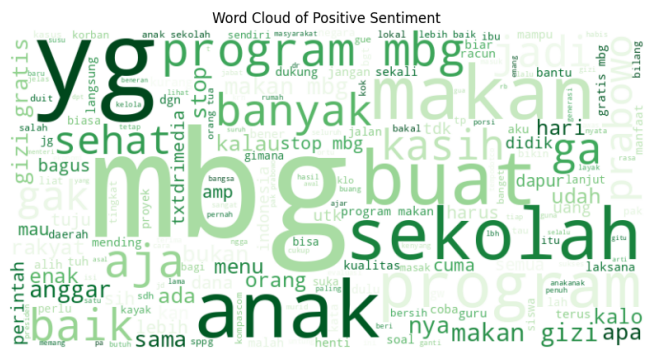


Figure 2. Word Cloud of Positive Sentiment

The positive sentiment Word Cloud, dominated by green hues, clearly highlights public acceptance and high hopes for the MBG Program (Figure 2). The largest and most frequent terms, such as MBG, *Program*, *Makanan* (Food), *Sekolah* (School), and *Anak* (Child), collectively show that positive discussion centers squarely on the program's core purpose: providing food for schoolchildren. Supporting words like *Banyak* (A Lot), *Buat* (For/Make), and *Kasih* (Give) emphasize the desirable act of provision and distribution. Furthermore, medium-sized words like *Sehat* (Healthy), *Gizi* (Nutrition), *Bagus* (Good), *Baik* (Well), and *Dukung* (Support) confirm that the public's appreciation focuses on the program's intended outcomes—namely, health benefits and high quality. The appearance of *Langsung* (Directly) and *Kelola* (Manage) also suggests a positive focus on efficient management and transparent delivery mechanisms. Overall, the positive sentiment strongly expresses support and approval for the MBG program, emphasizing its potential for successful implementation and its valuable contribution to children's health.

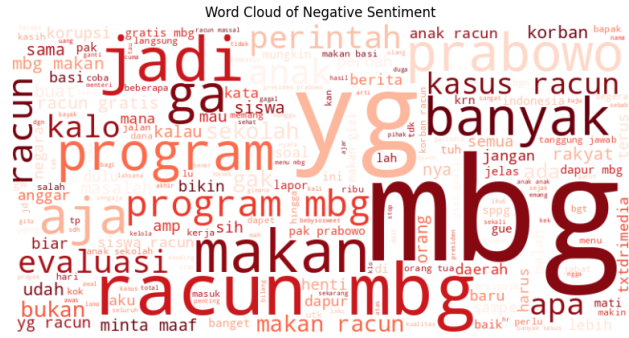


Figure 3. Word Cloud of Positive Sentiment

The negative sentiment Word Cloud, dominated by deep red hues, spotlights the serious concerns and sharp criticism voiced by the public regarding the MBG Program (Figure 3). The most prominent keywords alongside MBG are *Racun* (Poison) and *Basi* (Stale/Expired), which directly reflect the primary fears concerning food safety and quality and the risk of contamination. This criticism is also highly politicized, evidenced by the appearance of Prabowo and objections to the implementation *Perintah* (Order/Command). Supporting terms such as *Korupsi* (Corruption), *Kasus* (Case), and *Anggaran* (Budget) point to significant issues regarding budgetary integrity and transparency, reinforcing allegations of misconduct. Furthermore, the demand for *Evaluasi* (Evaluation) and the keyword *Gagal* (Fail) signal deep dissatisfaction with the program's execution. Overall, the negative sentiment is overwhelmingly driven by acute anxiety over health risks for School Children and program failure resulting from quality issues and potential corruption.

The Word Cloud analysis confirms a clear polarization of sentiment:

- Positive Appreciation focuses on the expected benefits (healthy, nutrition, support) if the program runs as planned.
- Negative Criticism focuses on concrete risks and implementation failures (poison, stale, corruption), which require immediate attention.

This finding provides clear action items for the government: to improve positive sentiment, the focus must be on guaranteeing food quality (eliminating the risk of "poison" and "stale" food) and budget transparency (addressing "corruption" issues). The program's success will heavily rely on the ability to assure the public that the food provided is truly nutritious and safe.

5.3 Sentiment Distribution

The sentiment distribution of the dataset was analyzed to understand the overall pattern of user opinions. Out of the total data, 891 comments (43.87%) were labeled as neutral, indicating that a significant portion of users provided descriptive or balanced feedback without strong positive or negative opinions. Negative sentiment accounted for 676 comments (33.28%), reflecting frequent dissatisfaction, complaints, or concerns related to the program. In contrast, positive sentiment comprised 464 comments (22.85%), showing that fewer users expressed satisfaction or appreciation.

This distribution suggests an imbalanced dataset, with neutral and negative sentiments dominating the responses. Such imbalance can affect model training, as classifiers may be biased toward majority classes. Therefore, careful consideration of evaluation metrics, such as weighted F1-score and cross-validation, is important to ensure that models accurately capture sentiment across all classes, including the minority positive sentiment.

VI. CONCLUSION

Based on the evaluation of ten machine learning algorithms for sentiment classification of user comments on the *Makan Bergizi Gratis* (MBG) program, Bagging and Gradient Boosting emerged as the best-performing models, achieving the highest accuracy and weighted F1-scores, as well as robust 5-fold cross-validation results. Naive Bayes, while less effective in detecting positive sentiment, served as a lightweight baseline and provided interpretable class log-probabilities that helped identify key negative and positive features.

Feature analysis across all models revealed consistent sentiment-driving words. Negative sentiment was strongly associated with terms such as *racun*, *stop*, *korupsi*, *basi*, and *masalah*, while positive sentiment was driven by words like *sehat*, *enak*, *makan bergizi*, *gizi*, and *bermanfaat*. These findings highlight the reliability of ensemble-based models in identifying key sentiment indicators and suggest that preprocessing strategies, such as handling multi-word tokens and domain-specific stopword removal, could further enhance model performance and interpretability.

Overall, the study demonstrates that ensemble methods not only achieve superior predictive performance but also provide meaningful insights into user sentiment, which can inform policy decisions, program improvements, and communication strategies for public nutrition initiatives.

REFERENCES

- [1] "Presidential Regulation of the Republic of Indonesia No. 83 Year 2024 concerning the Establishment of the National Nutrition Agency," 2024, *Government of the Republic of Indonesia*. [Online]. Available: <https://peraturan.go.id>
- [2] "BGN on only 34 MBG kitchens having hygiene certificates: Where does the data come from?," *KOMPAS.com*, 2025. [Online]. Available: <https://nasional.kompas.com/read/2025/09/27/08231551>
- [3] "List of MBG menus that poisoned thousands of school children," *Liputan6.com*, 2025. [Online]. Available: <https://www.liputan6.com/news/read/6169922>
- [4] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012. [Online]. Available: <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>
- [5] A. Fahrezi and S. Solichin, "Analisis Sentimen Ulasan Aplikasi Mobile JKN pada Play Store Menggunakan Metode Multinomial Naive Bayes," in *Proceedings of SENAFTI*, 2024. [Online]. Available: <https://senafiti.budiluhur.ac.id/senafiti/article/view/1302>
- [6] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J Comput Syst Sci*, vol. 55, no. 1, pp. 119–139, 1997. [Online]. Available: <https://doi.org/10.1006/jcss.1997.1504>
- [7] Y. J. F. Harta, "Aspect-Based Sentiment Analysis of JKN Mobile App," *Jurnal Ilmu Komputer Udayana*, 2024. [Online]. Available: <https://ojs.unud.ac.id/index.php/JIKU>
- [8] D. Utama, "Sentiment Analysis of JKN Mobile Application in Health Technology Using Naive Bayes, Gradient Boosting, and AdaBoost," in *Sinergy: Inspire Conference*, 2025.
- [9] M. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intell Syst*, 2016. [Online]. Available: <https://doi.org/10.1109/MIS.2016.31>
- [10] B. Medhat, A. Hassan, and H. Korashy, "Sentiment Analysis Algorithms and Applications: A Survey," *Ain Shams Engineering Journal*, 2014. [Online]. Available: <https://doi.org/10.1016/j.asej.2014.04.011>
- [11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002, pp. 79–86. [Online]. Available: <https://aclanthology.org/W02-1011>
- [12] P. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," in *Proceedings of ICWSM*, 2011. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/14185>
- [13] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *Proceedings of LREC*, 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf
- [14] S. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," Stanford University, 2009. [Online]. Available: <https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf>
- [15] H. Jung, H.-A. Park, and T.-M. Song, "Ontology-Based Approach to Social Data Sentiment Analysis: Detection of Adolescent Depression Signals," *J Med Internet Res*, vol. 19, no. 7, p. e259, 2017. [Online]. Available: <https://www.jmir.org/2017/7/e259/>
- [16] S. Ravi and K. Ravi, "A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications," *Knowl Based Syst*, 2015. [Online]. Available: <https://doi.org/10.1016/j.knosys.2015.06.015>
- [17] P. S. Gatora, S. E. Hosseini, S. Pervez, M. J. Iqbal, and N. Shaukat, "Sentiment Analysis of Product Reviews using Machine Learning and Pre-trained LLM," *Big Data and Cognitive Computing*, vol. 8, no. 12, p. 199,



- 2024, [Online]. Available: <https://doi.org/10.3390/bdcc8120199>
- [18] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment Analysis in the Era of Large Language Models: A Reality Check," *arXiv preprint*, 2023, [Online]. Available: <https://arxiv.org/abs/2305.15005>
- [19] J. R. Jim, "Recent Advancements and Challenges of NLP-Based Sentiment Analysis," *Knowl Based Syst*, 2024, [Online]. Available: <https://doi.org/10.1016/j.knosys.2024.111019>
- [20] Y. Mao, "Sentiment Analysis Methods, Applications, and Challenges," *Information Fusion*, 2024, [Online]. Available: <https://doi.org/10.1016/j.inffus.2024.102287>
- [21] N. A. Sharma, "A Review of Sentiment Analysis: Tasks, Applications, and Challenges," *Journal of AI & Data Mining*, 2025, [Online]. Available: https://jad.shahroodut.ac.ir/article_2475.html
- [22] P. Nandwani, "A Review on Sentiment Analysis and Emotion Detection from Text," *Front Comput Sci*, 2021, [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8870731/>
- [23] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [24] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann Stat*, vol. 29, no. 5, pp. 1189–1232, 2001, [Online]. Available: <https://doi.org/10.1214/aos/1013203451>