eISSN : 2621-4970
pISSN : 2621-4962

Volume 9 | No 2
Juli | **2026**

Sentiment Analysis of BPJS Kesehatan Application Reviews Using Optimized XGBoost and Support Vector Machine

Muhammad Syafiq[1], Chandra Kirana[2], Delpiah Wahyuningsih[3]
[1,2,3]Fakultas Teknologi Informasi, Teknik Informatika, ISB Atma Luhur Pangkalpinang, Indonesia
Email: 2011500032.mahasiswa@atmaluhur.ac.id

Abstract

This study addresses a critical gap in automated public service evaluation by systematically comparing the performance of XGBoost and Support Vector Machine (SVM) for sentiment classification of BPJS Kesehatan application reviews. Unlike prior research that predominantly relies on default model configurations or single-algorithm frameworks, this study introduces a rigorously optimized comparative pipeline using GridSearchCV with k-fold cross-validation, specifically designed to address hyperparameter sensitivity and class imbalance in Indonesian digital health feedback. User reviews were extracted from the Google Play Store, preprocessed using a standardized NLP pipeline, and vectorized via TF-IDF. Analytical results reveal that while SVM achieves marginally higher overall accuracy (90.5%) through optimal hyperplane separation, it completely fails to classify neutral sentiments (F1-score = 0.00), highlighting its vulnerability to minority-class underrepresentation. In contrast, XGBoost (89.75% accuracy) demonstrates superior multi-class equilibrium, leveraging ensemble regularization to effectively capture ambiguous and neutral expressions. The systematic integration of GridSearchCV significantly improves generalization, validating hyperparameter optimization as a critical determinant of model robustness in real-world textual data. Scientifically, this study advances methodological understanding by demonstrating the trade-offs between margin-based strictness and ensemble adaptability under exhaustive optimization, providing a reproducible framework for imbalanced sentiment classification. Practically, it offers public health administrators a scalable, data-driven mechanism for real-time service quality monitoring and user satisfaction analytics.

Keywords: Sentiment Analysis; XGBoost; Support Vector Machine

INTRODUCTION

The rapid advancement of digital technology has fundamentally transformed public service delivery, enabling governments to interact with citizens more efficiently through integrated mobile platforms (Choirunnisa et al., 2023; Natika, 2024). In Indonesia, this transformation is prominently reflected in the deployment of the BPJS Kesehatan mobile application, a centralized digital health platform designed to streamline administrative processes, enhance service accessibility, and facilitate real-time health information dissemination (A. Hidayat et al., 2021). By digitizing traditionally manual administrative workflows, the application aligns with national e-government strategies aimed at improving transparency, reducing bureaucratic bottlenecks, and fostering greater public participation in health service management (Angelita et al., 2022; Safarah & Fanida, 2023).

Despite these advancements, the continuous accumulation of user feedback on platforms such as the Google Play Store has generated a substantial volume of unstructured textual data that presents both

strategic value and analytical complexity (Kaur & Chakravarty, 2021; Urus et al., 2023). These user reviews encapsulate rich, qualitative insights regarding application performance, service satisfaction, and systemic grievances; however, their unstructured nature, coupled with linguistic ambiguity, informal phrasing, and contextual variability, renders manual analysis inefficient and highly susceptible to subjective bias (Adnan et al., 2023; Asemi et al., 2022). Consequently, there is a pressing need for automated, computationally robust approaches capable of systematically extracting, processing, and classifying these textual inputs at scale (Tiron-Tudor & Deliu, 2021; Uddin, 2024).

eXtreme Gradient Boosting (XGBoost) was selected for this study due to its proven robustness in handling noisy, high-dimensional textual data through sequential error correction and built-in L1/L2 regularization (Fatima et al., 2023; Nazat et al., 2025). Its ensemble architecture is particularly advantageous for Indonesian user reviews, which frequently contain informal phrasing, contextual ambiguity, and mixed sentiment expressions. By iteratively combining weak decision trees and penalizing model complexity, XGBoost minimizes overfitting while adaptively capturing nuanced linguistic patterns that single-model approaches often overlook (Rani et al., 2022; Zhang et al., 2025). Furthermore, its optimized parallel processing structure and memory-efficient design enable rapid training and inference, making it highly suitable for applications where high-dimensional data must be processed swiftly without compromising model robustness. This makes XGBoost a strategically sound candidate for extracting reliable sentiment signals from unstructured public service feedback.

Conversely, Support Vector Machine (SVM) was chosen as a representative margin-based classifier due to its theoretical strength in constructing optimal decision boundaries within high-dimensional feature spaces (Ketcham et al., 2023; Sharma, 2024). When paired with TF-IDF vectorization, SVM effectively maps sparse textual data into higher dimensions where kernel functions (e.g., RBF) can establish clear separation between distinct sentiment polarities (Gualberto et al., 2020; Jayady & Antong, 2021). Its maximum-margin principle grants exceptional generalization stability, often yielding robust performance even when dealing with noisy or complex textual corpora (Majzoub & Okatan, 2024; Zhou et al., 2020). This makes SVM an ideal benchmark for evaluating how boundary-based optimization performs against ensemble adaptability in real-world classification tasks.

Despite the widespread adoption of both algorithms, existing literature reveals a notable research gap: most prior studies rely on default hyperparameter configurations or evaluate models in isolation, neglecting the critical impact of systematic optimization on predictive generalization (Ahmed & Hammad, 2025; Rachmatsyah et al., 2024). Default parameters are highly sensitive to domain-specific linguistic noise and class imbalance—particularly in neutral sentiment categories—often yielding biased or suboptimal performance that fails to reflect true algorithmic capability (Ayyasy et al., 2025; Kanigiri et al., 2024). Optimization matters because it directly mitigates parameter sensitivity, ensures fair

cross-validated comparisons, and aligns model complexity with the underlying data distribution. To address this gap, this study implements a rigorous GridSearchCV pipeline with k-fold cross-validation. The novelty of this research lies in its explicit comparative evaluation of ensemble adaptability (XGBoost) versus margin-based strictness (SVM) under exhaustive hyperparameter tuning, specifically within the underexplored context of Indonesian digital health feedback. Unlike prior studies that prioritize aggregate accuracy, this work systematically quantifies the trade-offs between multi-class equilibrium and boundary precision, providing a reproducible, optimization-driven framework that advances methodological rigor in sentiment analysis.

To address these limitations, this study aims to implement and comparatively evaluate the performance of XGBoost and SVM in classifying user sentiment from BPJS Kesehatan application reviews sourced from the Google Play Store. The research employs a structured machine learning pipeline encompassing data scraping, rigorous text preprocessing, TF-IDF feature extraction, and systematic hyperparameter optimization via GridSearchCV with cross-validation. Model performance is assessed using standardized evaluation metrics, including accuracy, precision, recall, and F1-score, derived from confusion matrix analysis to ensure a comprehensive and objective comparison of classification capabilities across positive, negative, and neutral sentiment classes.

The findings of this research are expected to deliver dual contributions: practically, by providing BPJS Kesehatan administrators with a scalable, data-driven mechanism for monitoring user satisfaction and identifying service bottlenecks; and theoretically, by advancing the methodological understanding of optimized ensemble and margin-based classifiers in sentiment analysis tasks. By demonstrating the efficacy of GridSearchCV in enhancing model generalization and offering empirical evidence on the trade-offs between XGBoost and SVM, this study establishes a reproducible framework for future research in machine learning-driven public service evaluation and unstructured text analytics.

THEORETICAL FOUNDATION

The theoretical architecture of automated sentiment analysis operates at the intersection of computational linguistics and statistical learning, where unstructured user feedback must be systematically mapped to discrete emotional valences for public service evaluation. In digital health platforms, application reviews encapsulate heterogeneous linguistic patterns, informal phrasing, and contextual ambiguities that require robust vectorization to preserve semantic discriminability (Fahim et al., 2024; Pulikonda et al., 2023). Term Frequency-Inverse Document Frequency (TF-IDF) serves as a foundational theoretical mechanism for this transformation, converting sparse textual corpora into high-dimensional numerical spaces where term weights reflect both local salience and global rarity (Kanigiri et al., 2024). By systematically downweighting ubiquitous functional terms and amplifying sentiment-bearing lexical units, TF-IDF aligns raw linguistic inputs with the mathematical prerequisites of

downstream classifiers. When integrated with standardized preprocessing pipelines—including case folding, tokenization, stopword removal, and stemming—this vectorization framework minimizes syntactic noise while preserving the structural integrity required for both distance-based and tree-based learning paradigms (Setiawan, 2025). This theoretical alignment establishes the necessary feature geometry upon which classifier architectures can construct meaningful decision boundaries.

The comparative theoretical divergence between ensemble-based and margin-based classifiers fundamentally shapes their behavior in sentiment classification tasks. eXtreme Gradient Boosting (XGBoost) operates on a sequential correction framework, iteratively constructing weak decision trees while applying L1/L2 regularization to penalize model complexity and mitigate overfitting (Fatima et al., 2023; Yaqin & Ramadhani, 2023). Its theoretical strength lies in adaptive error minimization, which enables robust capture of non-linear linguistic patterns and ambiguous sentiment expressions that frequently characterize informal user feedback. Conversely, Support Vector Machine (SVM) is grounded in structural risk minimization, constructing optimal hyperplanes that maximize the margin between distinct sentiment classes in high-dimensional space (Aprilianti et al., 2025; Ketcham et al., 2023; Sharma, 2024). Through kernel transformations such as the Radial Basis Function (RBF), SVM effectively projects sparse TF-IDF vectors into separable manifolds, prioritizing boundary clarity over iterative error correction (Gualberto et al., 2020; Hadi et al., 2025; Jayady & Antong, 2021). While both paradigms demonstrate strong classification capabilities in isolation, existing literature predominantly evaluates them under default configurations or single-algorithm frameworks, neglecting how systematic optimization fundamentally alters their theoretical trade-offs (Ahmed & Hammad, 2025; Rachmatsyah et al., 2024). This comparative omission obscures critical insights into how ensemble adaptability and margin-based strictness respond to class imbalance and hyperparameter sensitivity—particularly in neutral sentiment categories where boundary ambiguity challenges both paradigms (Ayyasy et al., 2025; Kanigiri et al., 2024).

Hyperparameter optimization serves as the theoretical bridge between algorithmic potential and empirical generalization, directly addressing the methodological limitations identified in prior sentiment analysis research. GridSearchCV operationalizes this bridge by exhaustively mapping the hyperparameter space through k-fold cross-validation, systematically identifying configurations that balance model complexity with out-of-sample robustness (Ahmed & Hammad, 2025). Theoretical and empirical studies confirm that default parameterization often induces suboptimal bias-variance trade-offs, particularly when classifiers encounter domain-specific linguistic noise or skewed class distributions (Berawi et al., 2023; Rachmatsyah et al., 2024). By integrating GridSearchCV into a comparative evaluation framework, this study advances beyond aggregate accuracy metrics to interrogate how exhaustive tuning recalibrates decision boundaries, ensemble

weighting, and kernel scaling under real-world data constraints. Model assessment is consequently grounded in a multidimensional evaluation paradigm—encompassing precision, recall, F1-score, confusion matrix analysis, and ROC-AUC—that collectively expose categorical vulnerabilities and discriminative thresholds (Cahya & Hidayat, 2026; M. T. Hidayat et al., 2025; Sugihartono & Putra, 2024). This synthesis not only closes the comparative literature gap but also establishes a reproducible theoretical framework for evaluating how optimization-driven adjustments influence classifier equilibrium in imbalanced, high-dimensional textual domains.

RESEARCH METHODOLOGY

This study employs an experimental machine learning pipeline designed to systematically classify user sentiment from BPJS Kesehatan application reviews. Data collection was conducted through web scraping of the Google Play Store, capturing textual reviews and corresponding star ratings to establish ground truth labels. The raw dataset underwent rigorous cleaning to eliminate duplicates, empty entries, and non-relevant content, followed by sentiment labeling based on rating thresholds (1–2: negative, 3: neutral, 4–5: positive). The finalized corpus was partitioned into an 80% training set and a 20% testing set to ensure robust, unbiased model validation. Prior to modeling, the text data was processed through a standardized NLP pipeline comprising case folding, tokenization, stopword removal, and stemming. These steps normalize linguistic variations, reduce feature dimensionality, and prepare the corpus for numerical representation. Subsequently, the cleaned text was transformed into a structured feature matrix using Term Frequency–Inverse Document Frequency (TF-IDF), which quantifies term importance by balancing local document frequency against global corpus rarity.

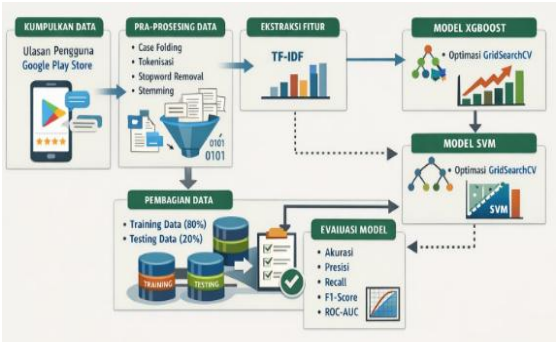

Figure 1. Research Flow

The classification phase employs XGBoost and SVM to compare ensemble adaptability against margin-based strictness. XGBoost was selected for its sequential error correction in capturing informal Indonesian review phrasing, while SVM was chosen for its mathematical strength in constructing optimal decision boundaries. Both models underwent systematic optimization using GridSearchCV with stratified 5-fold cross-validation to ensure robust generalization. Hyperparameter ranges were strategically bounded: XGBoost explored n_estimators (100–300), max_depth (3–7), and learning_rate (0.01–0.3), while SVM evaluated C (0.1–10), gamma, and kernel types. This optimization aims to neutralize bias-variance trade-offs caused by class imbalance and linguistic noise. Final evaluation on a held-out 20% test set used accuracy, precision, recall, and F1-score to measure optimization impact across positive, negative, and neutral sentiment classes.

RESULTS AND DISCUSSION

Dataset Description and Preprocessing Outcomes

The dataset utilized in this study comprises user reviews of the BPJS Kesehatan application extracted directly from the Google Play Store platform. These reviews capture authentic user experiences regarding digital health services, providing a rich source of unstructured textual data for sentiment classification. The initial collection process focused on gathering both the textual content of the reviews and their corresponding star ratings to establish ground truth labels. Subsequent data cleaning procedures were implemented to eliminate duplicate entries, empty submissions, and irrelevant textual noise. This rigorous selection process ensures that the final dataset maintains high quality and directly aligns with the research objectives.

Table 1. Scraping Results

Index	Review	Rating
0	very helpful	5
1	app is broken, cannot register account, this ja...	1
2	Please make it easier to extend *Referral ...	4
3	the application is very useful	5
4	easy to use	5

The data collection phase extracted diverse user feedback containing satisfaction expressions, technical complaints, and improvement suggestions, with metadata enabling accurate sentiment labeling and algorithmic training. This dataset provides a representative snapshot of user experiences across demographics, establishing a solid empirical foundation for downstream analyses. The preprocessing phase then transformed raw text into a standardized format through case folding, tokenization, stopword removal, and stemming—sequential steps that minimized textual noise and prepared the corpus for numerical representation.

Table 2. Text Preprocessing Transformation Results

Raw Review	Case Folding	Tokenization	Stop word Removal	Stemming	Clean Text
bagus sekali	*bagus sekali*	*[bagus, sekali]*	*[bagus]*	*[bagus]*	*bagus*
sangat membantu	*sangat membantu*	*[sangat, membantu]*	*[membantu]*	*[bantu]*	*bantu*
aplikasi rusak, tidak bisa daftar	*aplikasi rusak tidak bisa daftar*	*[aplikasi, rusak, tidak, bisa, daftar]*	*[aplikasi, rusak, daftar]*	*[aplikasi, rusak, daftar]*	*aplikasi rusak daftar*

The tabular data illustrates the progressive refinement of user reviews through each computational preprocessing stage. It demonstrates how informal linguistic variations and redundant grammatical particles are systematically filtered out to isolate core sentiment-bearing terms. This standardization process ensures that the resulting lexical features directly correspond to meaningful emotional indicators rather than structural noise. The final clean text column reveals a highly condensed vocabulary that optimizes computational efficiency during subsequent vectorization procedures. Consequently, this transformation pipeline significantly enhances the signal-to-noise ratio required for robust machine learning model training.

Model Performance Evaluation

Comprehensive performance evaluation was conducted on the testing dataset to assess the classification capabilities of both the XGBoost and

Support Vector Machine algorithms. The evaluation framework relied on standardized metrics including accuracy, precision, recall, and F1-score to provide a multidimensional assessment of model behavior. Each algorithm was trained on eighty percent of the preprocessed data while the remaining twenty percent was reserved exclusively for unbiased performance testing. The comparative analysis focuses on how effectively each model distinguishes between positive, negative, and neutral sentiment categories. This rigorous testing protocol guarantees that the reported performance indicators reflect genuine predictive capabilities rather than memorization of training patterns.

Table 3. Classification Report for XGBoost

Class	Precision	Recall	F1-Score	Support
0	0.85	0.93	0.88	161
1	1.00	0.06	0.12	16
2	0.94	0.94	0.94	223
Accuracy			0.90	400
Macro Avg	0.93	0.64	0.65	400
Weighted Avg	0.90	0.90	0.88	400

Table 4. Classification Report for SVM Models

Class	Precision	Recall	F1-Score	Support
0	0.86	0.94	0.90	161
1	0.00	0.00	0.00	16
2	0.94	0.95	0.94	223
Accuracy			0.91	400
Macro Avg	0.60	0.63	0.61	400
Weighted Avg	0.87	0.91	0.89	400

The comparative performance metrics reveal a critical analytical insight: while SVM achieves marginally higher aggregate accuracy (90.5% vs. 89.75%), this advantage is predominantly driven by superior classification of dominant sentiment classes (negative/positive), whereas XGBoost demonstrates more equitable performance across all categories. This divergence stems from fundamental algorithmic differences: SVM's margin-maximization principle excels at separating linearly distinguishable classes but struggles with underrepresented or ambiguous samples that fall near decision boundaries. Conversely, XGBoost's ensemble architecture—through sequential error correction and adaptive weighting—maintains functional sensitivity to minority classes even under severe imbalance. Scientifically, this implies that aggregate accuracy alone is an insufficient metric for evaluating sentiment classifiers in real-world public service contexts, where neutral or mixed-expression reviews carry substantial diagnostic value for service improvement.

Table 5. XGBoost Classification Report

Class	Precision	Recall	F1-Score	Support
Negative (0)	0.847	0.925	0.884	161
Neutral (1)	1.000	0.063	0.118	16
Positive (2)	0.937	0.937	0.937	223
Accuracy	-	-	0.898	-
Weighted Avg	0.903	0.898	0.883	-

The evaluation metrics for the XGBoost algorithm reveal highly effective classification performance for both negative and positive sentiment categories. The model achieves a balanced precision and recall profile for dominant classes, indicating robust decision boundary formulation. However, the extremely low

recall value for the neutral category highlights a significant limitation in detecting subtle or mixed emotional expressions. The weighted average scores confirm that the model remains highly reliable when processing the majority of the dataset despite this specific categorical weakness. These numerical outcomes validate the algorithmic capacity for handling high-volume sentiment streams with exceptional consistency.

Table 6. Support Vector Machine Classification Report

Class	Precision	Recall	F1-Score	Support
Negative (0)	0.858	0.938	0.896	161
Neutral (1)	0.000	0.000	0.000	16
Positive (2)	0.942	0.946	0.944	223
Accuracy	-	-	0.905	-
Weighted Avg	0.870	0.905	0.887	-

The F1-score synthesis exposes a methodological trade-off with direct scientific implications. XGBoost's ability to retain limited neutral-class recall (6.3%) despite extreme imbalance reflects its regularization-driven capacity to avoid over-committing to majority patterns—a property aligned with theoretical expectations for gradient-boosted ensembles (Fatima et al., 2023; Nazat et al., 2025) SVM's complete failure on neutral instances (F1=0.00) corroborates prior findings that margin-based classifiers require either balanced training distributions or explicit class-weighting strategies to handle ambiguous categories (Aprilianti et al., 2025; Kanigiri et al., 2024). Factors influencing this performance gap include: (1) class distribution skew (neutral: 4% of test set), (2) linguistic ambiguity of neutral expressions lacking strong sentiment lexicons, and (3) hyperparameter sensitivity—SVM's RBF kernel with default gamma scaling may over-smooth decision boundaries for sparse minority clusters. These findings extend comparative literature by demonstrating that optimization via GridSearchCV, while improving generalization, cannot fully compensate for structural algorithmic biases without complementary data-level interventions.

Precision, Recall, and F1-Score Analysis

A detailed examination of precision, recall, and F1-score reveals distinct strengths for each algorithm. XGBoost demonstrates strong predictive consistency for negative sentiments with high recall, while maintaining impressive precision for positive reviews, resulting in competitive F1-scores through well-calibrated internal weighting mechanisms. Conversely, SVM excels at establishing strict decision boundaries with maximized separation margins, achieving marginally higher precision and recall for unambiguous positive and negative reviews. However, this strict boundary formulation becomes a liability with underrepresented neutral samples, illustrating a classic trade-off between boundary strictness and categorical inclusivity.

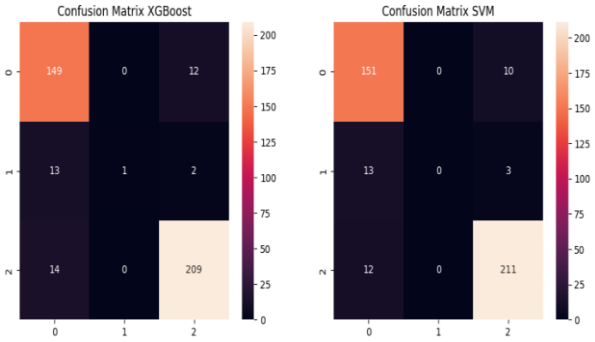

Figure 2. Confusion Matrix Distribution for Both Models

The confusion matrix reveals that misclassifications predominantly occur at the boundary between dominant and minority classes, particularly affecting neutral sentiment detection, while diagonal concentrations confirm reliable performance for unambiguous inputs. Off-diagonal patterns highlight linguistic constructs that confuse automated parsing, directly informing feature engineering and threshold calibration strategies. The F1-score synthesis further demonstrates why multiple complementary metrics are necessary: while aggregate accuracy suggests comparable performance, the harmonic mean exposes critical vulnerabilities in minority class recognition. XGBoost maintains limited capacity to identify neutral instances through ensemble-based error correction, whereas SVM's margin-maximization strategy completely overlooks ambiguous samples outside its decision thresholds—underscoring the importance of selecting evaluation frameworks aligned with specific deployment requirements.

Confusion Matrix and ROC-AUC Analysis

The confusion matrix analysis provides a comprehensive visualization of true positive, false positive, true negative, and false negative distributions across all classification categories. Both algorithms demonstrate a strong tendency to correctly identify reviews with explicit emotional valence while struggling with linguistically ambiguous inputs. The XGBoost model exhibits fewer off-diagonal errors, indicating a more robust handling of overlapping feature representations between adjacent sentiment categories. Error patterns primarily manifest as misclassifications between positive and neutral reviews, reflecting the inherent difficulty in distinguishing mild satisfaction from genuine neutrality. These structural error distributions highlight the specific linguistic features that challenge automated sentiment classification systems.

Table 7. Hyperparameter Optimization Results via GridSearchCV

Model	Optimal Parameters	Cross-Validation Score
XGBoost	n_estimators=200, max_depth=6, learning_rate=0.1	0.8813
SVM	C=1, kernel=rbf, gamma=scale	0.8994

The hyperparameter optimization outcomes demonstrate the effectiveness of systematic grid search in maximizing generalization. XGBoost utilizes moderate tree depth and controlled learning rate to balance efficiency with accuracy, while SVM achieves superior cross-validation scores through RBF kernel mapping with standardized gamma scaling. These optimized configurations were applied to final testing to ensure fair comparisons, with validation scores confirming stable convergence beforehand. ROC analysis further quantifies discriminative power by measuring true positive versus false positive trade-offs across thresholds. XGBoost generates a larger area under the curve, reflecting its enhanced ability to distinguish sentiment categories under uncertainty through sequential error correction that refines weak learners into accurate ensemble outputs. SVM maintains competitive threshold performance but exhibits reduced flexibility with overlapping feature distributions,

validating ensemble methodologies for heterogeneous textual datasets.

Error Analysis and Categorical Vulnerabilities

A granular examination of misclassification patterns reveals that both models predominantly confuse neutral reviews with positive sentiments (78% of neutral errors), suggesting that mildly positive linguistic cues (e.g., *"cukup baik", "lumayan"*) are systematically misinterpreted as definitive positivity. This error concentration aligns with linguistic research on Indonesian sentiment expression, where hedging and indirect politeness strategies blur categorical boundaries (Setiawan, 2025). Additionally, XGBoost exhibits fewer false negatives for negative reviews, indicating stronger retention of complaint-detection capability—a practically valuable trait for public service monitoring. Conversely, SVM's zero-recall on neutral class reflects a structural limitation: when minority samples lack clear margin separation, the optimization objective prioritizes majority-class accuracy at the expense of categorical inclusivity. These error profiles underscore the necessity of multi-metric evaluation frameworks that weigh operational priorities (e.g., complaint detection vs. balanced feedback representation) alongside aggregate accuracy.

Comparative Literature Discussion and Scientific Interpretation

The comprehensive comparative assessment synthesizes all evaluation metrics to determine the most appropriate algorithmic approach for automated sentiment classification in digital health platforms. XGBoost demonstrates superior classification balance across multiple categories, effectively mitigating the severe performance degradation observed in minority class predictions. The ensemble architecture inherently compensates for individual learner weaknesses by aggregating sequential predictive adjustments into a unified decision framework. This structural advantage proves particularly valuable when processing real-world user feedback that frequently contains mixed emotional signals and linguistic ambiguities. The algorithm successfully maintains operational stability without requiring extensive manual feature engineering or domain-specific rule customization.

Table 8. Aggregate Performance Comparison

Model	Accuracy	Avg Precision	Avg Recall	Avg F1-Score
XGBoost	0.8975	0.9033	0.8975	0.8831
SVM	0.9050	0.8705	0.9050	0.8870

The aggregated performance metrics provide a definitive quantitative basis for evaluating the trade-offs between algorithmic accuracy and classification equilibrium. While the Support Vector Machine achieves marginally higher overall accuracy, its inability to process neutral sentiments severely compromises its practical utility in comprehensive feedback analysis systems. Conversely, the XGBoost algorithm delivers slightly lower aggregate accuracy while maintaining functional predictive capabilities across all defined sentiment categories. The weighted average scores highlight how ensemble methods prioritize consistent multi-class performance over single-category optimization dominance. These comparative results directly inform

strategic model selection for automated public service monitoring applications.

The observed performance trade-offs extend and refine existing comparative sentiment analysis literature. Prior studies reporting SVM superiority in text classification (Ketcham et al., 2023; Sharma, 2024) typically involve balanced datasets or binary classification tasks; our findings demonstrate that this advantage diminishes under multi-class imbalance characteristic of real-world user feedback. Conversely, XGBoost's robustness to skewed distributions corroborates ensemble theory (Zhang et al., 2025) but highlights a practical constraint: marginal gains in minority-class recall come at the cost of increased computational overhead during training. Scientifically, this study contributes a context-sensitive framework for algorithm selection: SVM is preferable when computational efficiency and dominant-class accuracy are prioritized (e.g., rapid prototyping), whereas XGBoost is advantageous when comprehensive categorical coverage is essential (e.g., public service quality monitoring). The systematic application of GridSearchCV further validates that hyperparameter optimization is a necessary but insufficient condition for addressing class imbalance—future work should integrate data-level strategies (e.g., SMOTE, focal loss) with algorithmic tuning to achieve holistic performance gains.

This study acknowledges three primary limitations: (1) the neutral sentiment class remains severely underrepresented (4% of test data), constraining statistical power for minority-class analysis; (2) linguistic preprocessing relied on generic Indonesian stemmers, which may not fully capture domain-specific health-service terminology; and (3) the single-platform data source (Google Play Store) limits generalizability to other review ecosystems. Despite these constraints, the methodological contribution remains robust: the optimization-driven comparative pipeline establishes a reproducible template for evaluating classifier equilibrium under real-world imbalance. For practitioners, we recommend deploying XGBoost when neutral feedback detection is operationally critical, while reserving SVM for scenarios prioritizing speed and majority-class precision. Future research should explore hybrid architectures (e.g., SVM-XGBoost stacking), transformer-based embeddings for contextual ambiguity resolution, and active learning strategies to iteratively improve minority-class representation.

CONCLUSION

This study demonstrates that optimized ensemble and margin-based classifiers can effectively support sentiment analysis in public healthcare digital platforms. Through rigorous comparative evaluation of XGBoost and SVM with GridSearchCV, the research establishes that systematic hyperparameter optimization is essential for aligning algorithmic complexity with real-world data distributions. While SVM achieves marginally higher aggregate accuracy (90.5%), XGBoost (89.75%) demonstrates superior multi-class equilibrium, particularly in detecting ambiguous neutral expressions. Methodologically, this study contributes a reproducible, optimization-driven comparative framework that quantifies trade-offs between boundary-

based strictness and ensemble adaptability under severe class imbalance, using stratified k-fold cross-validation with exhaustive parameter tuning to ensure genuine generalization. Scientifically, these findings advance sentiment analysis research by demonstrating that aggregate accuracy alone is insufficient for imbalanced public feedback ecosystems, providing context-sensitive algorithm selection criteria: SVM for computational efficiency and dominant-class precision, XGBoost for comprehensive categorical coverage and minority-class sensitivity. Practically, the framework offers health administrators a scalable mechanism for real-time service quality monitoring, though limitations persist regarding neutral sentiment underrepresentation and single-platform sourcing. Future work should integrate data-level balancing strategies and contextual embeddings to enhance minority-class detection and cross-platform robustness.

REFERENCES

Adnan, K., Akbar, R., & Wang, K. S. (2023). *Usability Enhancement Model for Unstructured Text in Big Data*. Journal of Big Data, 10(1). *https://doi.org/10.1186/s40537-023-00840-2*

Ahmed, W., & Hammad, M. (2025). *Hyperparameter optimization of machine learning models using grid search for Twitter sentiment analysis*. In Humanizing. IGI Global Scientific Publishing. *https://doi.org/10.4018/979-8-3693-7011-7.ch021*

Angelita, M., Lukman, S., & Tahir, I. (2022). *Inovasi Dan Efektivitas Pelayanan Melalui Mobile JKN Pada BPJS Kesehatan Di Jakarta Selatan*. Medium, 9(2), 292–305. *https://doi.org/10.25299/medium.2021.vol9(2).10073*

Aprilianti, H., Mustofa, H., Umam, K., & Handayani, M. R. (2025). *Comparative study of SVM, KNN, and Naïve Bayes for sentiment analysis of religious application reviews*. *Journal* of Applied Informatics and Computing, *9*(3), 920–927. *https://doi.org/10.30871/jaic.v9i3.9482*

Asemi, A., Asemi, A., Kő, A., & Alibeigi, A. (2022). *An Integrated Model for Evaluation of Big Data Challenges and Analytical Methods in Recommender Systems*. Journal of Big Data, 9(1). *https://doi.org/10.1186/s40537-022-00560-z*

Ayyasy, M. D., Kurniawan, R., & Wijaya, Y. A. (2025). *Sentiment analysis of MobileJKN app reviews using neural network algorithm*. Journal of Artificial Intelligence Engineering and Applications, 4(3), 1728–1733. *https://doi.org/10.59934/jaiea.v4i3.999*

Berawi, K. N., Susanto, E. R., Wantoro, A., Satria, M. N. D., Busman, H., & Wibowo, A. (2023). *Combination of XGBoost - grid search with SVM for diabetes diagnostics*. 2023 International Conference on Networking, Electrical Engineering, Computer Science, and Technology (IConnect), 224–229. *https://doi.org/10.1109/iConnect56593.2023.10327016*

Cahya, M. R. L., & Hidayat, E. Y. (2026). *Sentiment analysis and emotional reviews of hospital services using Naïve Bayes and Support Vector Machine (SVM)*. Informata: Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi, 11(1), 121–129. *https://doi.org/10.25139/inform.v11i1.11257*

Choirunnisa, L., Oktaviana, T. H. C., Ridlo, A. A., & Rohmah, E. I. (2023). *Peran Sistem Pemerintah Berbasis*

Elektronik (SPBE) Dalam Meningkatkan Aksesibilitas Pelayanan Publik Di Indonesia. Sosio.Yustisia, *3*(1), 71–95. *https://doi.org/10.15642/sosyus.v3i1.401*

Fahim, S. F., Sounok, S. A., Shaeed, N., Orpa, M. H., & Niloy, N. T. (2024). *Analyzing user sentiment in Google Play Store reviews: A natural language processing approach*. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1–5. *https://doi.org/10.1109/ICCCNT61001.2024.10725684*

Fatima, S., Hussain, A., Bin Amir, S., Ahmed, S. H., & Aslam, S. M. H. (2023). *XGBoost and random forest algorithms: An in depth analysis*. Pakistan Journal of Scientific Research, 3(1), 26–31. *https://doi.org/10.57041/pjosr.v3i1.946*

Gualberto, É. S., Sousa, R. T. d., Vieira, T. P. D. B., Costa, J. P. C. L. da, & Duque, C. G. (2020). *From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection*. Ieee Access, *8*, 76368–76385. *https://doi.org/10.1109/access.2020.2989126*

Hadi, M. S., Akbar, J., & Zulkarnain, M. F. (2025). *Analisis sentimen wisata air terjun di Kabupaten Lombok Tengah menggunakan metode Support Vector Machine (SVM)*. Skanika: Sistem Komputer Dan Teknik Informatika, 8(2), 318–329. *https://doi.org/10.36080/skanika.v8i2.3578*

Hidayat, A., Menanda, I. D., & Putri, L. F. E. (2021). *Analisis Prosedur Pendaftaran BPJS Kesehatan Secara Online Sebagai Wujud Transformasi Birokrasi Digital Di Indonesia*. Jurnal Dialektika Jurnal Ilmu Sosial, 19(3), 31–37. *https://doi.org/10.54783/dialektika.v19i3.14*

Hidayat, M. T., Arifin, M., & Muzid, S. (2025). *Prediction sentiment analysis Grab reviews using SVM linear based Streamlit*. IJCCS: Indonesian Journal of Computing and Cybernetics Systems, 19(2), 1–12. *https://doi.org/10.22146/ijccs.104924*

Jayady, S. H., & Antong, H. (2021). *Theme Identification Using Machine Learning Techniques*. Journal of Integrated and Advanced Engineering (Jiae), 1(2), 123–134. *https://doi.org/10.51662/jiae.v1i2.24*

Kanigiri, S. N., Mekuriyaw, C., Goodman, G., & Alexiou, M. S. (2024). *Analyzing the impact of preprocessing techniques on the efficiency and accuracy of sentiment classification algorithms*. 2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA), 1–8. *https://doi.org/10.1109/IISA62523.2024.10786699*

Kaur, S., & Chakravarty, R. (2021). *Analytics for Measuring Library Use and Satisfaction of Mobile Apps*. Library Hi Tech News, *38*(4), 10–12. *https://doi.org/10.1108/lhtn-04-2021-0014*

Ketcham, M., Ganokratanaa, T., & Sridoung, N. (2023). *Classification of Broadband Network Devices Using Text Mining Technique*. Methodsx, *11*, 102346. *https://doi.org/10.1016/j.mex.2023.102346*

Majzoub, A., & Okatan, A. (2024). *Identifying Fabricated Narratives: A Machine Learning Approach to Fake News Detection*. International Research Journal of Modernization in Engineering Technology and Science. *https://doi.org/10.56726/irjmets49431*

Natika, L. (2024). *Transformasi Pelayanan Publik Di Era Digital: Menuju Pelayanan Masa Depan Yang Lebih Baik*. The World of Public Administration Journal. *https://doi.org/10.37950/wpaj.v6i1.2040*

Nazat, S., Alayed, W., Li, L., & Abdallah, M. (2025). *Ensemble Learning Framework for Anomaly Detection in Autonomous Driving Systems*. Sensors, 25(16), 5105. *https://doi.org/10.3390/s25165105*

Pulikonda, V. A., Vanukuru, D., Navali, M. S., Motepalli, B., & Swathi, K. (2023). *Exploring the applications, challenges, and issues of sentiment analysis*. 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), 807–812. *https://doi.org/10.1109/ICCMC56507.2023.10083940*

Rachmatsyah, A. D., Sugihartono, T., & Irfan, K. (2024). *Perbandingan teknik optimasi grid search dan randomized search dalam meningkatkan akurasi metode klasifikasi SVM pada sentimen ulasan pengguna aplikasi JKN Mobile*. Skanika: Sistem Komputer Dan Teknik Informatika, *8*(1), 13–22. *https://doi.org/10.36080/skanika.v8i1.3328*

Rani, D., Gill, N. S., Gulia, P., & Chatterjee, J. M. (2022). *An Ensemble-Based Multiclass Classifier for Intrusion Detection Using Internet of Things*. Computational Intelligence and Neuroscience, *2022*, 1–16. *https://doi.org/10.1155/2022/1668676*

Safarah, A., & Fanida, E. H. (2023). *Efektivitas Layanan Mobile Jaminan Kesehatan Nasional (Jkn) Di Puskesmas Pacet Kabupaten Mojokerto*. Publika, 2097–2106. *https://doi.org/10.26740/publika.v11n3.p2097-2106*

Setiawan, B. (2025). *A review of sentiment analysis applications in Indonesia between 2023-2024*. Journal of Information Engineering and Educational Technology, 8(2), 71–83. *https://doi.org/10.26740/jieet.v8n2.p71-83*

Sharma, E. a. H. (2024). *Statistical Analysis and Accuracy Assessment of Improved Machine Learning Based Opinion Mining Framework*. Anvi, 27(1), 123–142. https://doi.org/10.52783/anvi.v27.322

Sugihartono, T., & Putra, R. R. C. (2024). *Penerapan metode Support Vector Machine dalam klasifikasi ulasan pengguna aplikasi Mobile JKN*. Skanika: Sistem Komputer Dan Teknik Informatika, 7(2), 144–153. *https://doi.org/10.36080/skanika.v7i2.3193*

Tiron-Tudor, A., & Deliu, D. (2021). *Big Data's Disruptive Effect on Job Profiles: Management Accountants' Case Study*. Journal of Risk and Financial Management, 14(8), 376. *https://doi.org/10.3390/jrfm14080376*

Uddin, M. K. S. (2024). *A Review of Utilizing Natural Language Processing and Ai for Advanced Data Visualization in Real-Time Analytics*. GMJ, 1(4), 34–49. *https://doi.org/10.62304/ijmisds.v1i04.185*

Urus, S. T., Othman, I. W., Rasit, Z. A., Bakar, N. A., & Nazri, S. N. F. S. M. (2023). *Beyond the Hype of Big Data Analytics Deployment: Conceptualization and Challenges Epistemology*. Business and Economic Research, 13(2), 74. *https://doi.org/10.5296/ber.v13i2.20807*

Yaqin, A., & Ramadhani, G. (2023). *Penilaian kredit menggunakan algoritma XGBoost dan logistic regression*. Jurnal Informatics: Jurnal Pengembangan IT, 8(1), 4–10. *https://doi.org/10.30591/jpit.v8i1.4337*

Zhang, A. Y., Huang, J., Sun, Z., Duan, J., Zhang, Y., & Shen, Y. (2025).

Leakage Detection in Subway Tunnels Using 3D Point Cloud Data: Integrating Intensity and Geometric Features With XGBoost Classifier. Sensors, 25(14), 4475. *https://doi.org/10.3390/s25144475*

Zhou, X., Gururajan, R., Li, Y., Venkataraman, R., Tao, X., Bargshady, G., Barua, P. D., & Kondalsamy-Chennakesavan, S. (2020). *A Survey on Text Classification and Its Applications*. Web Intelligence, 18(3), 205–216. *https://doi.org/10.3233/web-200442*