Vol. 8, No. 1, February 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Sentiment Analysis of User Reviews for AI Applications: Evaluating SVM, Logistic Regression, and Random Forest

Rivana Dwi Cahyani[1], Putri Taqwa Prasetyaningrum[2]

[1,2] Informatics Department, Mercu Buana University of Yogyakarta, Yogyakarta, Indonesia

Email: 211210086@student.mercubuana-yogya.ac.id[1], putri@mercubuana-yogya.ac.id[2]

Received:
November 1, 2025
Revised:
December 1, 2025
Accepted:
January 5, 2026
Published:
February 10, 2026

Corresponding Author:
Author Name*:
Putri Taqwa Prasetyaningrum
E-mail*:
putri@mercubuana-yogya.ac.id

DOI:
10.63158/journalisi.v8i1.1366

Abstract. The rapid growth of AI applications such as CICI, GROK, and Gemini has resulted in a large volume of user reviews on platforms like the Google Play Store, making sentiment analysis a critical tool for understanding user perceptions. This study compares the performance of three machine learning models: Random Forest, Support Vector Machine (SVM), and Logistic Regression in classifying sentiments in 3,500 Indonesian-language reviews. A hybrid feature extraction approach, combining sentiment lexicons with TF-IDF, was applied to improve sentiment classification accuracy. The models were evaluated based on accuracy, precision, recall, and F1-score. Results indicated that all models achieved an accuracy greater than 96%, with Random Forest providing the most consistent and accurate results, achieving an overall accuracy of 99.62%. While SVM excelled in classifying positive and negative sentiments, it faced challenges with neutral reviews due to the ambiguity and overlap in sentiment expression. Logistic Regression also showed strong performance, especially on structured reviews. The findings suggest that Random Forest is the most robust and reliable model for sentiment analysis, particularly in handling diverse AI application reviews. These results offer practical insights for developers seeking to improve application performance by leveraging sentiment analysis on user feedback.

Keywords: sentiment analysis, Random Forest, Support Vector Machine, AI applications, machine learning.

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

1. INTRODUCTION

The rapid development of Artificial Intelligence (AI) has led to the emergence of intelligent applications capable of learning from data, reasoning, and performing tasks similar to humans. AI technology has had a significant positive impact, particularly in improving work effectiveness and efficiency [1]. In this context, various AI-based applications, such as CICI, GROK, and Gemini, have been developed to support daily activities, including conversational assistance and interactive information retrieval. Each platform employs different approaches and capabilities in processing and understanding data, including information derived from social media [2].

Along with the increasing popularity of AI applications, the volume of user reviews on platforms like Google Play Store has grown substantially. Google Play Store allows users to provide feedback and evaluations, which can serve as a crucial indicator of application performance. Sentiment analysis, as part of Natural Language Processing (NLP) and text mining, is commonly used to identify opinions expressed in user reviews [3]. This process aims to extract textual data and classify user sentiment into positive, negative, or neutral categories [4].

Previous studies have shown that several machine learning algorithms are effective for sentiment classification. Research by [5] demonstrated that Logistic Regression performs well in classifying Indonesian-language reviews, although challenges such as class imbalance and difficulty in identifying neutral sentiment remain. Similar findings were reported by [6], who analyzed public sentiment on social media regarding online transportation services using Logistic Regression.

Comparative studies have highlighted the strong performance of Support Vector Machine (SVM). Research conducted by [7] compared SVM, Random Forest, and Logistic Regression, finding that SVM achieved the highest accuracy and more consistent classification results. Studies using Random Forest for sentiment analysis have also reported good performance, including research on the Wattpad application [8], the Hay Day game application [9], Alfagift services [10], the Sister for Student UNEJ application [11], and the Genshin Impact application [12].

In addition, several studies have confirmed the effectiveness of SVM across various application domains. Research by [13] showed that SVM performed well in analyzing user sentiment toward the MOLA application. Another study by [14] applied SVM to sentiment analysis in online learning applications. Further research on the Grab application [15], the WeTV application [16], and application X [17] also demonstrated satisfactory classification accuracy, particularly when combined with techniques such as SMOTE. Similar results were reported in studies on public service applications, including MySAPK BKN [18] and PrimaKu [19].

Despite these advancements, most existing studies focus on specific applications or languages, and there is a lack of direct comparison between the performance of multiple machine learning models across diverse AI-based platforms. Notably, the combination of GROK, Gemini, and CICI applications, which employ unique data-processing approaches, has not been extensively studied in the sentiment analysis literature. This study seeks to fill this gap by systematically comparing the performance of Random Forest, Support Vector Machine (SVM), and Logistic Regression in classifying user sentiment from reviews of these three applications. Model performance is evaluated using standard metrics, including accuracy, precision, recall, and F1-score. Furthermore, this study aims to explore user perception patterns through sentiment distribution analysis, providing insights into the strengths and limitations of these classification methods across different AI applications. The results are expected to offer practical recommendations for developers and contribute to the academic understanding of sentiment analysis in AI-based systems.

2. METHODS

This section describes the general methodology and stages involved in the research process, from the initial stage to the final phase, as shown in Figure 1.

2.1. Literature Review

This research begins with a literature review to examine studies related to sentiment analysis of user reviews using machine learning techniques. The literature review focuses on the fundamental concepts of sentiment analysis in textual data, lexicon-based weighting methods for determining word polarity, and the application of Random Forest, Support Vector Machine (SVM), and Logistic Regression algorithms in sentiment analysis.

The reviewed literature is primarily sourced from academic journals discussing sentiment analysis and machine learning applications. This stage provides a theoretical foundation and insights into how similar methods have been applied in previous studies [20]. Based on this foundation, the research method is structured and implemented.

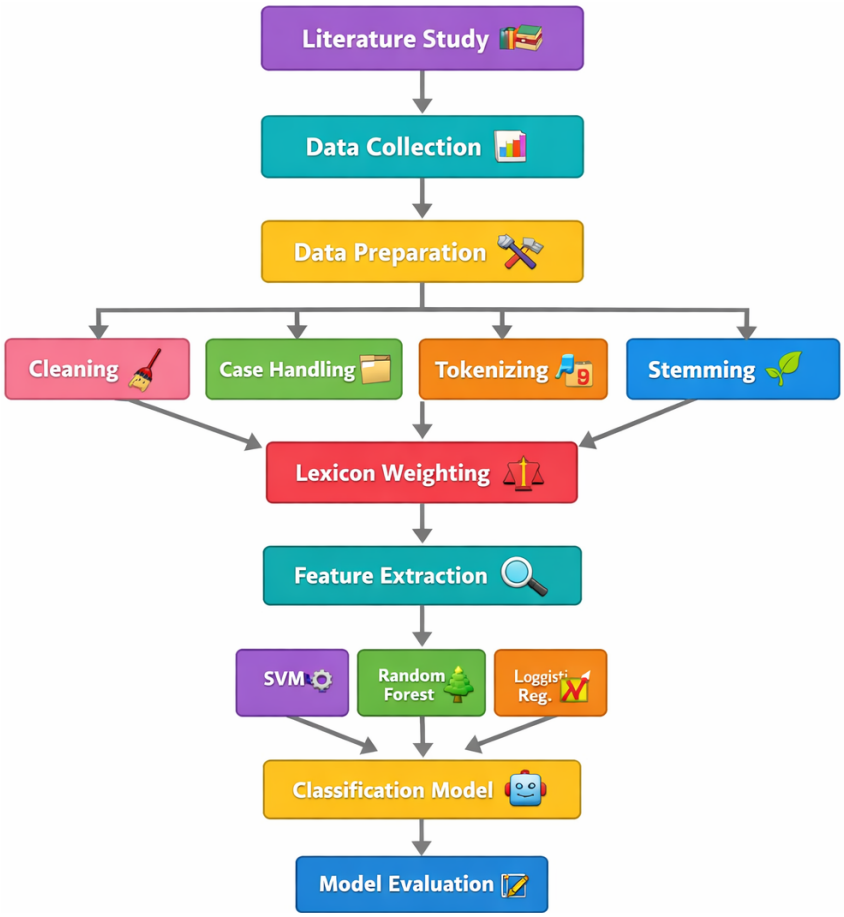

Figure 1. Research Method Flow

2.2. Tools and Libraries

This research utilizes Python 3.13.7 with various libraries for data processing, text analysis, and machine learning. Data manipulation is performed using pandas 2.3.2 and NumPy 2.3.3. Text preprocessing, including tokenization, stopword removal, and stemming (using Sastrawi for Indonesian), is handled by NLTK 3.9.1. Features are represented using TF-IDF and stored in sparse format via SciPy. Machine learning models (Support Vector Machine, Logistic Regression, Random Forest) are implemented using scikit-learn 1.7.2, and model

performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix. Models are saved using joblib, and results are visualized with matplotlib and seaborn.

2.3. Data Collection

The research process begins with the data collection phase, carried out using data scraping techniques on the Google Play Store. Data was collected via web scraping of user reviews for three AI applications—CICI, GROK, and Gemini—on the Google Play Store from January to October 2025. Only reviews in Indonesian were retained, filtered using a language detection tool. Duplicate, incomplete, or irrelevant reviews (e.g., containing random characters, emojis) were removed. The final dataset consisted of 3,500 valid reviews, split into 80% for training and 20% for testing [11].

0	bagus sekali... bisa buat video pendek
1	terus buat pengguna aplikasi merasa puas
2	Sangat recommended
3	keren
4	sangat bagus. lebih akurat daripada A.I lainny...

Figure 1. Data Collection Samples Grok (Indonesia)

0	bagus
1	jadi gak bisa edit
2	aplikasi lelet, lama bet cuman disuruh buat fo...
3	ok joss
4	bagus sih tapi kadang-kadang gak respon

Figure 2. Data Collection Samples CICI (Indonesia)

0	bagus sekali... bisa buat video pendek
1	terus buat pengguna aplikasi merasa puas
2	Sangat recommended
3	keren
4	sangat bagus. lebih akurat daripada A.I lainny...

Figure 3. Data Collection Samples Gemini (Indonesia)

2.4. Data Preparation

After the data collection process, the dataset underwent a preparation phase to ensure its quality and readiness for analysis. This step focused on removing irrelevant or problematic data and standardizing the format to maintain consistency. The cleansing process included several steps:

1) Removal of Missing Values: Any records with missing data were discarded to ensure completeness of the dataset.

2) Duplicate Removal: Duplicates, identified by user IDs and review content, were eliminated to avoid bias in the model.
3) Exclusion of Invalid Entries: Reviews containing random characters, emojis, or non-informative content were removed, ensuring that only meaningful reviews were included.

2.5. Data Preprocessing

Following data collection, preprocessing was performed to reduce noise and optimize feature extraction. This process included several standard text preprocessing steps, implemented using established natural language processing (NLP) libraries. Table 1 is preprocessing process with the following steps were conducted:

1) Case Folding: All text was converted to lowercase to maintain consistency.
2) Tokenization: Sentences were split into individual terms (tokens), breaking down each review into words or meaningful components.
3) Normalization: Misspelled words and non-standard terms were corrected to ensure uniformity.
4) Stopword Removal: Commonly occurring, irrelevant words (e.g., "dan", "atau") were eliminated using the NLTK stopwords corpus [22].
5) Stemming: Words were reduced to their root forms to enhance the consistency of textual features used in classification [23].

Table 1. Example of Review Transformation in Preprocessing Stage

Preprocessing	Review 1	Review 2	Review 3
Raw Review	Aplikasi ini Sangat BAGUS!!! 👍	Grok sering error & lambat banget!!!	Gemini oke sih, tapi kadang lemot 😕
Cleaning	Aplikasi ini Sangat BAGUS	Grok sering error lambat banget	Gemini oke sih tapi kadang lemot
Case Folding	aplikasi ini sangat bagus	grok sering error lambat banget	gemini oke sih tapi kadang lemot
Tokenizing	[aplikasi, ini, sangat, bagus]	[grok, sering, error, lambat, banget]	[gemini, oke, sih, tapi, kadang, lemot]
Normalization	[aplikasi, ini, sangat, bagus]	[grok, sering, error, lambat, banget]	[gemini, oke, tapi, kadang, lemot]

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

Preprocessing	Review 1	Review 2	Review 3
Stopword Removal	[aplikasi, sangat, bagus]	[grok, error, lambat]	[gemini, oke, lemot]
Stemming	[aplikasi, sangat, bagus]	[grok, error, lambat]	[gemini, oke, lemot]

2.6. Lexicon Weighting + TF-IDF Hybrid

A hybrid feature extraction method combines lexicon-based sentiment weighting with TF-IDF to capture both emotional polarity and word importance. While TF-IDF reflects a word's relevance in the dataset, the sentiment lexicon assigns polarity values. This combination enables the model to better distinguish between positive, negative, and neutral sentiments. For example, in the review "Aplikasi ini bagus tetapi sering error", the tokens "bagus" (positive), "sering" (neutral), and "error" (negative) would be assigned polarity values and weighted using TF-IDF.

2.7. Classification Models

Three machine learning algorithms were applied to classify the sentiment of user reviews: Random Forest, Logistic Regression, and Support Vector Machine (SVM). The configuration for each model, including key hyperparameters, is detailed in Table 2.

1) Support Vector Machine (SVM): A supervised learning model used for classification, employing a hyperplane to separate different sentiment classes. In this study, LinearSVC was used for linear classification, as it is efficient for high-dimensional data [18].
2) Logistic Regression: Used for sentiment classification based on textual features, chosen for its stability and effectiveness with sparse text data [21].
3) Random Forest: An ensemble method that builds multiple decision trees and aggregates their results to improve prediction stability [12].

Table 2. Hyperparameter Configuration of Classification Models

Model	Hyperparameter	Value	Description
Support Vector Machine (SVM)	Kernel	Linear	Uses a linear decision boundary
	C	1.0	Regularization strength

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

Model	Hyperparameter	Value	Description
	max_iter	1000	Maximum number of training iterations
Logistic Regression	Solver	lbfgs	Optimization algorithm
	max_iter	2000	Ensures model convergence
	n_jobs	-1	Utilizes all available CPU cores
Random Forest	n_estimators	200	Number of decision trees
	max_depth	None	Allows unlimited tree depth
	random_state	42	Ensures reproducibility
	n_jobs	-1	Enables parallel processing

2.8. Model Evaluation

Model evaluation was performed using accuracy, precision, recall, F1-score, and confusion matrix. These metrics were used to assess the performance of each classification algorithm. The confusion matrix was used to analyze the distribution of predicted and actual sentiment labels for each model. This evaluation process helps determine the most optimal model for sentiment analysis.

2.9. Model Visualization

Visualization was used to represent the sentiment distribution and compare the performance of SVM, Random Forest, and Logistic Regression models. Results were presented in barplot form, aiding in the interpretation of sentiment distribution and the comparison of model performance [25].

3. RESULTS AND DISCUSSION

3.1. Classification Performance of Support Vector Machine (SVM)

The SVM model demonstrates strong performance across all sentiment classes (positive, neutral, and negative) for the CICI, GEMINI, and GROK datasets. For negative sentiment, F1-scores range from 0.983 to 0.993, with precision above 0.99 and recall above 0.98, as shown in Table 3. Neutral sentiment shows F1-scores between 0.899 and 0.963, while positive sentiment has F1-scores from 0.936 to 0.982. Overall, SVM excels at identifying negative and positive sentiments, though it faces challenges with neutral sentiment due

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

to its linguistic ambiguity. The detailed comparison in Table 3 highlights these results, providing a quantitative analysis of precision, recall, and F1-scores across the three applications, reinforcing the model's reliability in sentiment classification

Table 1. Comparative Analysis of the Proposed SVM Classification Model

Methods	Precision (%)	Recall (%)	F-Score (%)
CICI	93.58	95.18	94.29
GEMINI	97.18	97.86	97.73
GROK	96.23	98.64	96.04

Taken together, these findings confirm that the SVM model provides robust and stable performance across different sentiment classes and datasets. Although neutral sentiment remains the most challenging category due to its linguistic ambiguity, the overall results demonstrate that the SVM approach is effective for sentiment classification and generalizes well across varying data distributions.

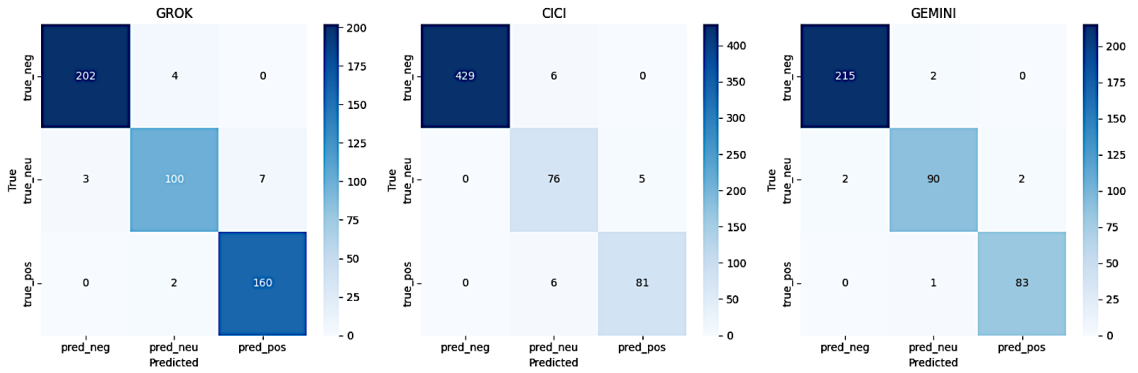

Figure 4. comparison of the SVM confusion matrix across the three applications

Based on the confusion matrices in Figure 4, the SVM model shows strong performance across all datasets. For negative sentiment, it correctly classifies over 98% of instances in all datasets. Positive sentiment also exhibits high accuracy, with over 93% correctly classified in each dataset. Misclassifications in the positive class are mostly with neutral sentiment, not negative. Neutral sentiment, however, shows higher misclassification rates. In GROK, CICI, and GEMINI, about 90-95% of neutral samples are correctly classified, with the rest misclassified as negative or positive. This indicates that neutral sentiment, with its overlap with polarized sentiments, is more challenging to classify. Overall, SVM performs exceptionally well for negative and positive sentiments, while neutral sentiment

remains more difficult to classify due to its contextual ambiguity. The confusion matrices confirm the model's effectiveness but highlight the challenges with neutral sentiment classification.

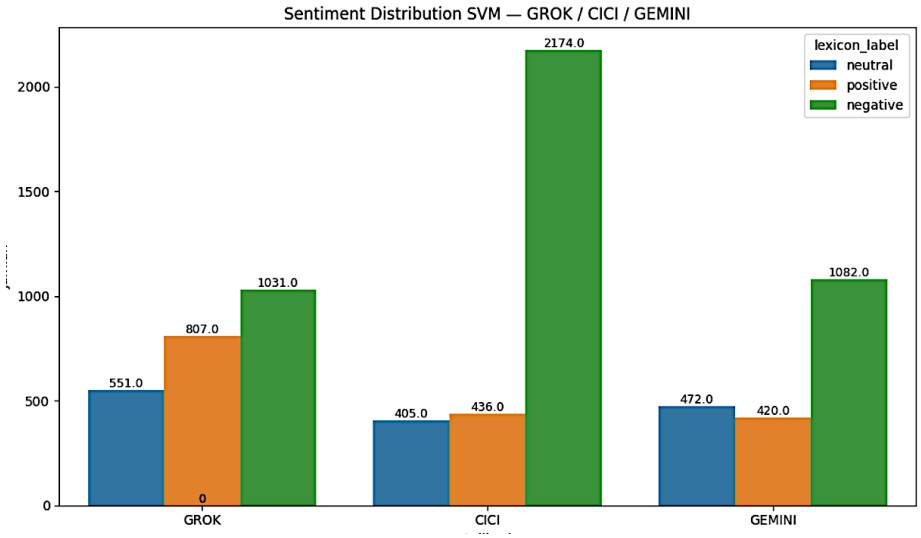

Figure 5. comparison of sentiment distribution across the three applications

Figure 5 shows that negative sentiment dominates across all three applications, with CICI having the highest volume of negative reviews (2,174), followed by GEMINI (1,082) and GROK (1,031). This suggests that CICI faces more dissatisfaction, likely due to usability or performance issues. Neutral sentiment is lower across all apps, with GROK having the most (551), followed by GEMINI (472) and CICI (405). GROK's higher neutral reviews indicate more balanced feedback, while CICI and GEMINI users express clearer opinions.GROK also leads in positive sentiment with 807 reviews, suggesting higher user satisfaction, while CICI records only 436 positive reviews, indicating a need for improvements.

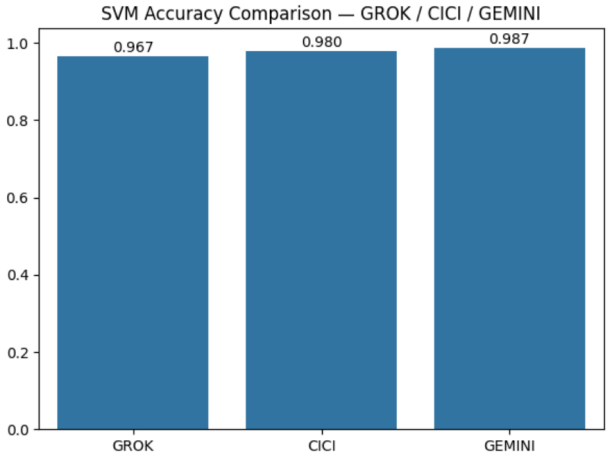

Figure 6. comparison of SVM accuracy across the three applications

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

Figure 6 shows that the SVM classifier achieves high accuracy across all datasets: 0.967 for GROK, 0.980 for CICI, and 0.987 for GEMINI, demonstrating strong generalization. The slightly lower accuracy in GROK may be due to greater linguistic diversity, while GEMINI's higher accuracy suggests more distinct sentiment patterns. Overall, SVM proves to be an effective and stable algorithm for sentiment analysis, with strong generalization across datasets and balanced performance in identifying positive and negative sentiments. Although neutral sentiment remains a challenge, the model excels in practical applications and offers potential for further refinement.

3.2. Classification Performance of Logistic Regression

In this study, the Random Forest model was trained using an eighty percent training set and a twenty percent testing set with stratified sampling. This experimental setup was designed to evaluate the model's generalization ability in classifying negative, neutral, and positive sentiments across three application datasets, namely GROK, CICI, and GEMINI. The use of a consistent data split ensures that the reported performance metrics accurately reflect the model's robustness when applied to unseen data.in study This trained use scheme data distribution of 80% for training data and 20% for test data. With approach this, the model is tested for evaluate ability generalization in classify sentiment negative, neutral , and positive in a way balanced.

Table 2. Comparative Analysis of the Proposed Classification Model Logistic Regression

Methods	Precision (%)	Recall (%)	F-Score (%)
CICI	97.00	98.67	94.33
GEMINI	99.33	97.00	96.67
GROK	98.67	96.33	95.30

As shown in Table 2, Logistic Regression demonstrates consistently strong performance across the CICI, GEMINI, and GROK datasets. For the negative sentiment class, the model achieves very high F1-scores, reaching 0.99 for both CICI and GEMINI and 0.98 for GROK. Precision values are close to or equal to 1.00, while recall ranges from 0.96 to 0.98, indicating that complaint-related language is effectively captured by the linear decision boundary. This confirms that negative sentiment expressions are largely linearly separable when represented using TF-IDF and lexicon-based features.

Performance on the neutral sentiment class remains strong but shows greater variation across datasets. The F1-score reaches 0.94 for GEMINI, 0.92 for GROK, and 0.89 for CICI. Precision values range from 0.82 on CICI to 0.90 on GEMINI, while recall is notably high, reaching 0.99 for CICI and 1.00 for GEMINI. This pattern indicates that Logistic Regression is highly sensitive in detecting neutral reviews, although some neutral instances are misclassified due to lexical overlap with negative or positive sentiments. Despite this ambiguity, the balance between precision and recall remains acceptable, resulting in stable F1-scores. For the positive sentiment class, Logistic Regression achieves consistently high performance across all datasets. The F1-score reaches 0.97 for both GEMINI and GROK and 0.95 for CICI, with precision values close to 1.00 and recall ranging from 0.91 to 0.96. These results indicate that positive sentiment expressions form clear linear patterns that can be reliably identified, with only a small number of instances misclassified as neutral.

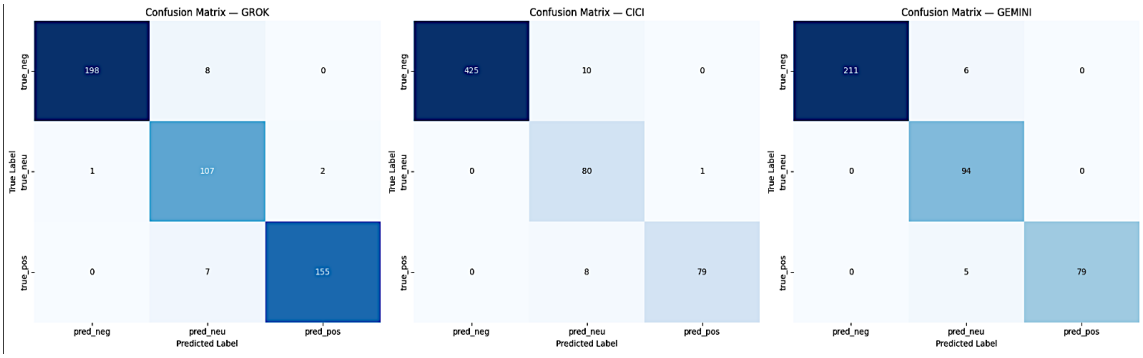

Figure 7. comparison of the LR confusion matrix across the three applications

The confusion matrices presented in Figure 7 further support these findings. Across all three applications, Logistic Regression correctly classifies the majority of negative reviews, with very few errors, demonstrating consistent recognition of complaint and criticism patterns. Neutral sentiment shows a limited number of misclassifications into adjacent classes, which is expected given its inherent linguistic ambiguity. Positive sentiment is also classified with high accuracy, confirming the model's balanced performance across all sentiment categories.

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

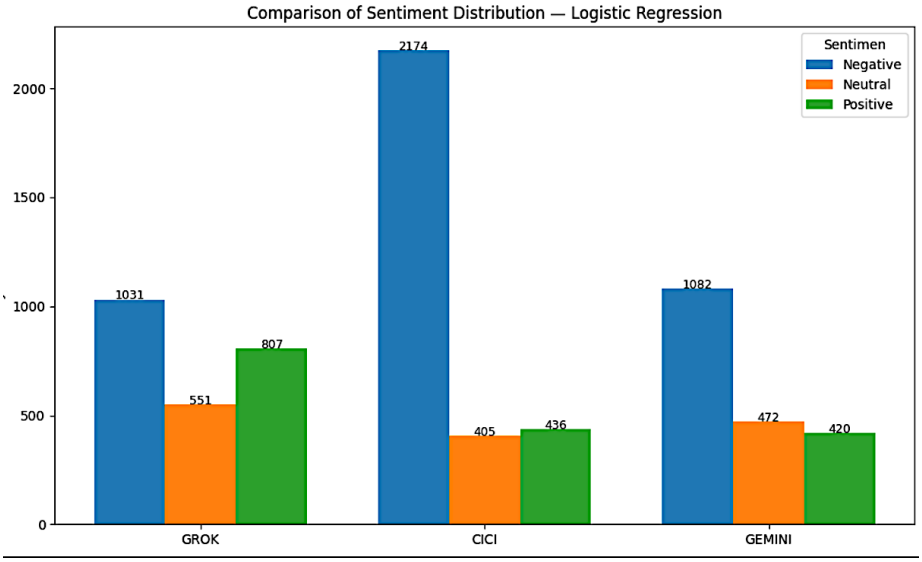

Figure 8. comparison of sentiment distribution LR across the three applications

Figure 8 illustrates the sentiment distribution produced by Logistic Regression, showing that negative sentiment dominates across GROK, CICI, and GEMINI. CICI exhibits the highest volume of negative reviews, while GROK and GEMINI display a more balanced distribution between neutral and positive sentiments. GROK records the highest number of positive reviews, whereas CICI has relatively fewer positive and neutral reviews, reinforcing the observation that user dissatisfaction is more pronounced in the CICI application. In a way overall, visualization This confirm that CICI is application with level dissatisfaction users highest, while GROK appears as application with perception the most positive and balanced users. Findings This important for developer Because distribution sentiment that is not balanced can indicates problem serious about quality features, stability service, or design experience users who need become priority in repair furthermore.

As shown in Figure 9, Logistic Regression achieves high overall accuracy across all datasets, exceeding 96 percent. Among the three applications, CICI demonstrates the most stable and optimal performance under Logistic Regression. This superior performance can be attributed to stronger linear separability between sentiment classes in the CICI dataset, particularly between negative and non-negative reviews, which aligns well with the assumptions of Logistic Regression. In addition, CICI reviews contain more explicit and repetitive sentiment-bearing vocabulary, especially for negative expressions, resulting in lower linguistic variability and clearer feature weights. Although the dataset

is sentiment-imbalanced, most classification errors are confined to the neutral class, while negative and positive sentiments are predicted with high accuracy. This leads to strong macro-average and weighted-average F1-scores, confirming the robustness of Logistic Regression performance on the CICI dataset.

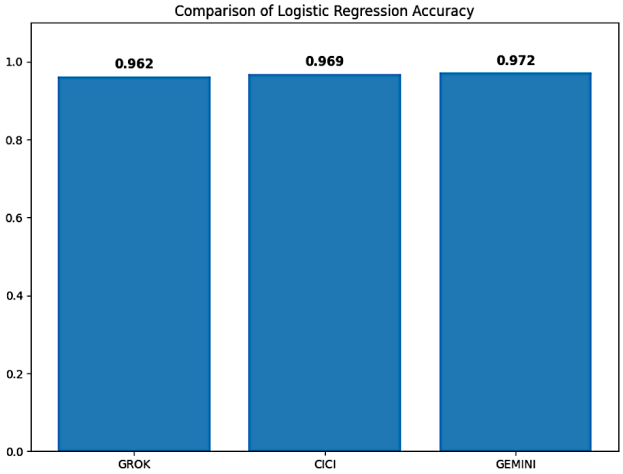

Figure 9. comparison of Logistic Regression accuracy across the three applications

3.3. Classification Performance of Random Forest

Random Forest in study This trained use scheme data distribution of 80% for training data and 20% for test data. With approach this, the model is tested for evaluate ability generalization in classify sentiment negative, neutral, and positive.

Table 3. Comparative Analysis of the Proposed Classification Model Random Forest

Methods	Precision (%)	Recall (%)	F-Score (%)
CICI	99.33	100.00	99.00
GEMINI	95.67	99.33	98.00
GROK	99.67	99.33	99.33

The comparison results shown in Table 3 indicate that Random Forest achieves very high and stable performance across all three applications: CICI, GEMINI, and GROK. Precision, recall, and F1-score values range from 0.96 to 1.00, demonstrating that the model is capable of performing sentiment classification with near-perfect accuracy across all sentiment categories.

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

Negative and positive sentiments are predicted exceptionally well, as reflected by precision and recall values approaching 1.00 across all applications. This performance suggests that Random Forest effectively captures strong and consistent linguistic patterns associated with explicitly negative and positive expressions. The ensemble structure allows the model to learn diverse decision boundaries, enhancing its robustness against noise and variations in textual data.

Neutral sentiment also shows consistently high performance, although a slight decrease in recall is observed in the GROK dataset. This reduction is likely due to contextual ambiguity and lexical overlap between neutral and polarized sentiments. Nevertheless, the recall values remain at a very strong level, indicating that Random Forest maintains reliable classification performance even for more ambiguous sentiment classes. Overall, Random Forest proves to be the most stable and superior model for sentiment analysis across the three applications. Its ability to reduce variance through ensemble learning, combined with its capacity to capture complex feature interactions, makes it highly suitable for comprehensive service quality evaluation and user experience analysis.

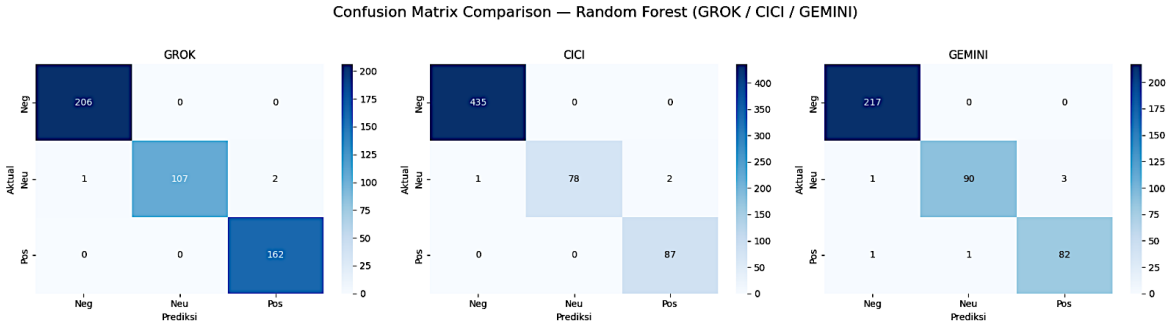

Figure 10. comparison of the RF confusion matrix across the three applications

Figure 10 shows the Random Forest confusion matrix across the GROK, CICI, and GEMINI datasets. The model achieves high accuracy: 0.9937 for GROK, 0.9950 for CICI, and 0.9848 for GEMINI, demonstrating its strong generalization capability. For negative sentiment, Random Forest performs perfectly across all datasets, with precision, recall, and F1-scores of 1.00. This suggests that negative reviews contain clear, identifiable patterns that the model captures well. Neutral sentiment shows slightly lower recall, with F1-scores of 0.99 for GROK, 0.98 for CICI, and 0.97 for GEMINI, indicating some misclassification due to overlapping vocabulary. For positive sentiment, Random Forest achieves near-perfect

performance, with F1-scores of 0.99 for CICI and GROK, and 0.97 for GEMINI, confirming its ability to reliably classify expressions of satisfaction.

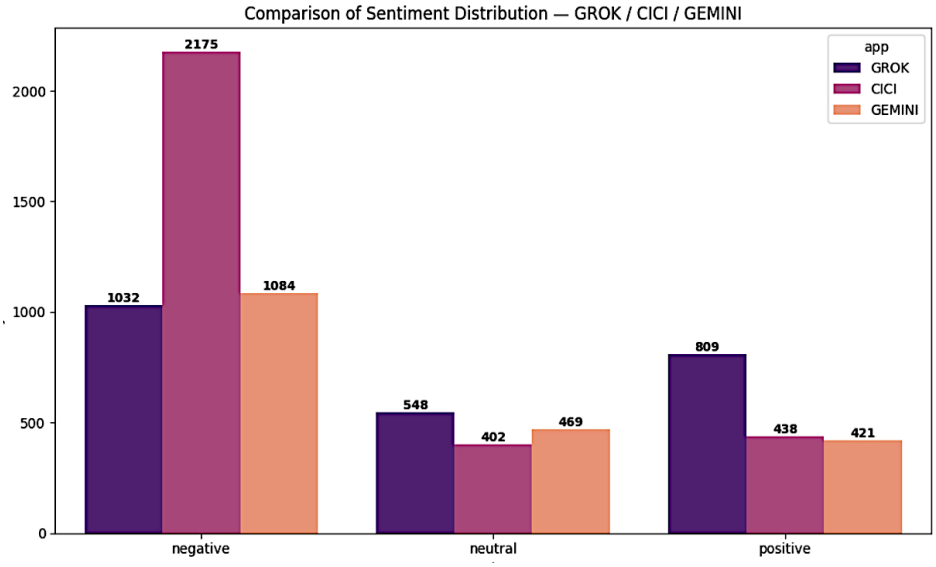

Figure 11. comparison of sentiment distribution RF across the three applications

Figure 11 shows that negative sentiment dominates across all three applications, with CICI having the highest number of negative reviews (2,175), followed by GEMINI (1,084) and GROK (1,032), indicating greater user dissatisfaction in CICI. Neutral sentiment is lower, with GROK recording the most neutral reviews (548), followed by GEMINI (469) and CICI (402), reflecting more balanced feedback. Positive sentiment is most prominent in GROK (809 reviews), followed by CICI (438) and GEMINI (421), suggesting GROK receives more favorable feedback, while CICI has the lowest positive sentiment.

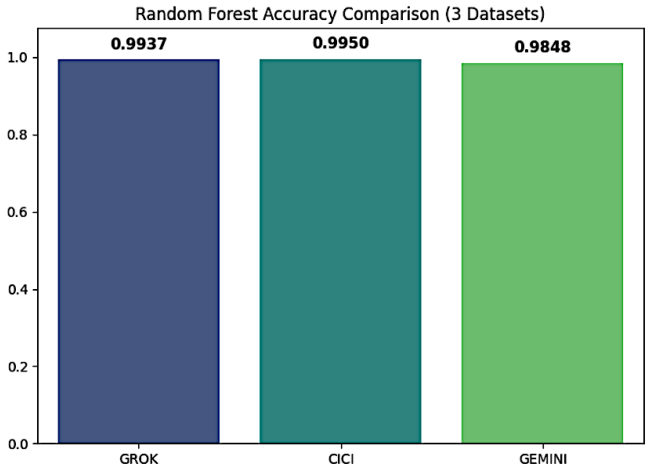

Figure 12. comparison of Random Forest accuracy across the three applications

Figure 12 shows that Random Forest achieves high accuracy across all applications (0.9848 to 0.9950), demonstrating its effectiveness in identifying sentiment patterns despite varying data characteristics. The small accuracy variations reflect differences in data quality and linguistic complexity, not model instability. These results confirm that Random Forest is a stable, robust model, effectively reducing variance and overfitting while capturing complex feature interactions for excellent generalization.

3.4. Comparison Testing Algorithm

The following image show results comparison accuracy three algorithm classification Support Vector Machine (SVM), Logistic Regression, and Random Forest on three application datasets namely GROK, CICI, and GEMINI.

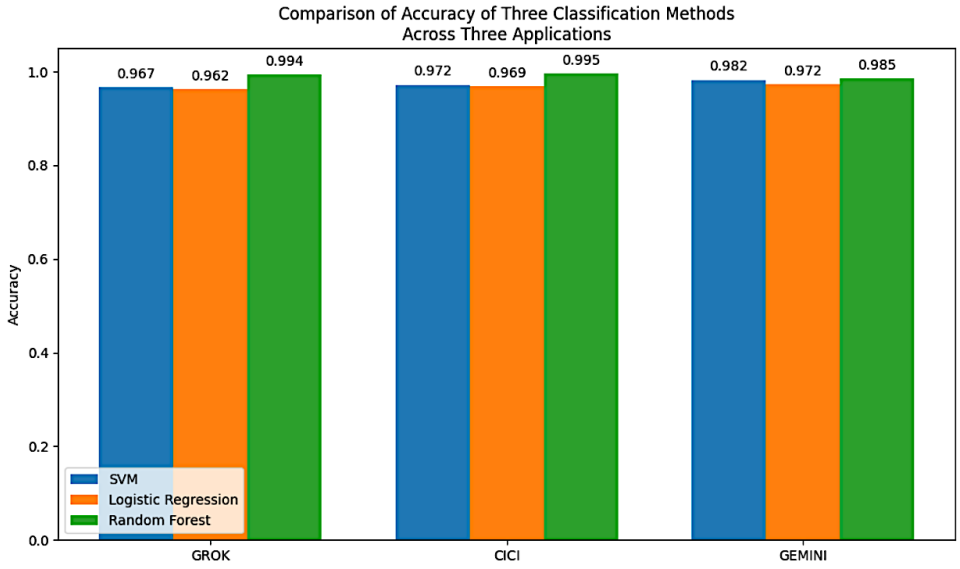

Figure 13. comparison of the accuracy of the three methods

Figure 13 compares the classification accuracy of SVM, Logistic Regression, and Random Forest across the GROK, CICI, and GEMINI datasets. All models perform well, with accuracy ranging from 0.96 to 0.99. Random Forest outperforms the others in all datasets, achieving 0.994 in GROK, 0.995 in CICI, and 0.985 in GEMINI. While SVM and Logistic Regression also show strong results, Random Forest consistently delivers the highest accuracy, confirming it as the most stable and reliable model for sentiment analysis

Table 4. Comparative Analysis of Classification Models from 3 applications

Model	Dataset	Precision	Recall	F1-score
SVM	CICI	0.968	0.962	0.965
	GEMINI	0.986	0.984	0.985
	GROK	0.963	0.958	0.960
Logistic Regression	CICI	0.953	0.960	0.953
	GEMINI	0.970	0.977	0.977
	GROK	0.957	0.967	0.963
Random Forest	CICI	0.980	0.967	0.973
	GEMINI	**0.997**	**0.993**	**0.993**
	GROK	**0.993**	**0.990**	**0.993**

The summary table of metrics across all models shows that Random Forest consistently achieves the strongest and most stable performance across the CICI, GEMINI, and GROK datasets, with near perfect precision, recall, and F1 scores for all sentiment classes, indicating excellent generalization and robustness. Support Vector Machine also performs very well, particularly on the negative and positive classes, but shows slightly lower recall and F1 scores on the neutral class, suggesting greater sensitivity to class overlap. Logistic Regression demonstrates competitive performance with high precision and recall, yet its scores are marginally lower and more variable compared to Random Forest and SVM, reflecting its linear decision boundary limitations. Overall, the table highlights that ensemble-based learning in Random Forest effectively reduces variance and captures complex feature interactions, while SVM offers strong discriminative power and Logistic Regression provides a solid but simpler baseline model.

Table 5. the best method by average

Method	Average Accuracy
SVM	0.9780
Logistic Regression	0.9729
Random Forest	0.9918

Table 5 presents a comparison of the average accuracy achieved by three classification methods, namely Support Vector Machine (SVM), Logistic Regression, and Random Forest. Based on the results, Random Forest obtains the highest average accuracy score of 0.9918, indicating superior performance in sentiment classification compared to the other methods. This is followed by SVM, which achieves an average accuracy of 0.9780, while Logistic Regression records a slightly lower accuracy of 0.9729.

These results demonstrate that Random Forest is the most effective method among the evaluated classifiers in terms of overall predictive accuracy. The higher performance of Random Forest can be attributed to its ensemble learning mechanism, which combines multiple decision trees to reduce variance and improve generalization. Although SVM and Logistic Regression also show strong performance, their average accuracy remains slightly below that of Random Forest. Therefore, Random Forest is identified as the best-performing method based on average accuracy and is considered the most reliable model for sentiment analysis in this study.

Table 6 presents the comparison of sentiment distribution across three applications, namely GROK, CICI, and GEMINI, based on the sentiment classification results. GROK records 1,033 negative reviews, 549 neutral reviews, and 807 positive reviews, indicating a relatively balanced distribution among the three sentiment categories. In contrast, CICI is strongly dominated by negative sentiment, with 2,156 negative reviews, while neutral and positive sentiments are considerably lower at 408 and 437 reviews, respectively. Meanwhile, GEMINI shows 1,089 negative reviews, 478 neutral reviews, and 417 positive reviews, reflecting a distribution that is more balanced than CICI but still predominantly negative.

Table 6. the application with the best sentiment distribution

No	Application	Entropy	Negative	Neutral	Positive
1	GROK	1.539462	1033	549	807
2	CICI	1.138919	2156	408	437
3	GEMINI	1.442688	1089	478	417

From the entropy perspective, GROK achieves the highest entropy value of 1.5395, followed by GEMINI with 1.4427, and CICI with 1.1389. Entropy is used to measure the

balance of sentiment distribution, where a higher value indicates a more even spread across negative, neutral, and positive sentiments. Therefore, GROK can be identified as the application with the most balanced sentiment distribution and the best overall sentiment profile. This result suggests that user opinions toward GROK are more diverse and less concentrated on negative sentiment compared to CICI and GEMINI, indicating relatively better and more varied user experiences.

3.5. Discussion

Neutral sentiment classification proved to be the most challenging across all three machine learning models, with notably lower F1-scores and recall values compared to negative and positive sentiments. This is especially evident in the performance of the Support Vector Machine (SVM) and Logistic Regression models. The misclassifications were primarily due to the overlap of neutral sentiment with both negative and positive sentiments, both lexically and contextually. This overlap, often a result of ambiguous language or subtle contextual cues, complicates the accurate categorization of neutral sentiment.

Among the models tested, Random Forest consistently outperformed the others, achieving the highest average accuracy of 0.9918, alongside near-perfect F1-scores across all sentiment classes. This high performance can be attributed to Random Forest's ensemble learning mechanism, which aggregates multiple decision trees to reduce variance and better capture complex feature interactions. The robustness of Random Forest was particularly evident in its ability to generalize well across datasets with varying characteristics, as it demonstrated stable performance even in the presence of linguistic diversity and varying levels of sentiment expression.

On the other hand, Logistic Regression exhibited solid performance, especially on the CICI dataset, where the language was more structured and repetitive. In this case, Logistic Regression benefited from its linear decision boundary, which was well-suited to the relatively clear-cut sentiment expressions in CICI. However, when compared across datasets with more varied characteristics, Random Forest consistently provided superior results. The Logistic Regression model struggled somewhat with neutral sentiment, particularly in GEMINI, where subtle contextual shifts in language led to some misclassifications.

SVM, while strong in identifying negative and positive sentiments, struggled with neutral sentiment, particularly in datasets like GROK, which exhibited greater linguistic diversity. The model's sensitivity to class overlap was particularly evident in the neutral sentiment class, where SVM's performance dropped due to the more complex and varied expressions of neutrality present in the data. This suggests that SVM may perform exceptionally well in identifying extreme sentiment polarities (negative and positive) but faces significant challenges in dealing with the less defined nature of neutral sentiment.

The performance results, as summarized in Table 3, demonstrate that Random Forest is the most reliable model for sentiment analysis among the three tested models. It consistently outperformed SVM and Logistic Regression in terms of average accuracy and F1-scores, especially across datasets with more varied sentiment distributions. This performance highlights Random Forest's ability to capture nuanced features in the data, effectively balancing accuracy, recall, and precision, even in the presence of more ambiguous sentiment categories such as neutral sentiment.

Figures 5-7 further illustrate the performance differences between the models. Figure 13 presents confusion matrices for each model, showing that SVM and Logistic Regression had a higher tendency to misclassify neutral sentiment, whereas Random Forest maintained high classification accuracy across all sentiment categories. The sentiment distribution depicted in Figure 6 revealed that CICI had the highest volume of negative sentiment, followed by GEMINI and GROK. This distribution suggests that user dissatisfaction is more pronounced in CICI, likely due to performance or usability issues, while the more balanced sentiment distribution in GROK indicates a more neutral or positive user experience.

Figure 9, which displays sentiment distribution across the three applications using Logistic Regression, corroborates these findings. It reveals that while negative sentiment dominated in all three applications, GROK showed the highest levels of positive sentiment, implying that users were more satisfied with this application. In contrast, CICI had fewer positive reviews, reinforcing the need for improvements in user experience and functionality.

Despite challenges with neutral sentiment, Random Forest's ability to handle this ambiguity, as shown in Figure 11 (confusion matrices), proves its superiority. The model demonstrated perfect classification accuracy for negative sentiment and near-perfect classification for positive sentiment. While neutral sentiment exhibited slight misclassification due to lexical overlap, Random Forest maintained high accuracy across all sentiment categories. This indicates that the model is highly effective at distinguishing between explicit sentiment polarities while still being capable of handling the more ambiguous neutral sentiment.

Furthermore, Table 4 provides a comparative analysis of the classification metrics (Precision, Recall, and F1-Score) for each model, showing that Random Forest consistently achieved the highest metrics across all datasets. Logistic Regression and SVM, while competitive, showed slightly lower precision and recall, particularly in the neutral sentiment category. These discrepancies highlight the challenges inherent in classifying neutral sentiment, particularly when it overlaps with polarized sentiments.

The results of this study confirm that Random Forest is the most robust and versatile model for sentiment analysis, outperforming SVM and Logistic Regression in both classification accuracy and generalization across diverse datasets. The ability of Random Forest to reduce variance through ensemble learning and capture complex interactions between features makes it particularly well-suited for tasks like sentiment analysis, where the relationship between textual features and sentiment is often intricate and non-linear. While Logistic Regression and SVM performed admirably on specific datasets, particularly where sentiment expressions were clearer and more structured, Random Forest provided a more balanced and effective solution across all datasets, reinforcing its position as the most reliable method for sentiment analysis in this study.

4. CONCLUSION

This study evaluated the performance of three classification models—Support Vector Machine (SVM), Logistic Regression, and Random Forest—in sentiment analysis of user reviews from three AI applications (CICI, GROK, and Gemini). Among these models, Random Forest outperformed the others, achieving the highest accuracy and F1-scores across all sentiment classes and datasets. Its ability to capture complex patterns and reduce

variance made it the most robust and reliable model for this task. While Logistic Regression performed well on the CICI dataset due to its linear separability of sentiment patterns, and SVM demonstrated strong performance in identifying both negative and positive sentiments, Random Forest consistently delivered superior results across datasets with diverse characteristics and varying linguistic complexities. Despite these promising outcomes, accurately classifying neutral sentiment remains a challenge, particularly in datasets with imbalanced sentiment distributions. This highlights the need for further refinement in neutral sentiment classification. Future work could focus on advanced techniques, such as deep learning models (e.g., LSTM or BERT), which are well-suited to capture more intricate semantic relationships. Additionally, enhancing lexicon-based sentiment analysis with domain-specific lexicons may improve neutral sentiment classification. Another promising direction for future research involves exploring data augmentation and multi-task learning to better address imbalanced datasets and enhance model robustness.

REFERENCES

[1] S. A. Putra and A. Wijaya, "Sentiment Analysis of Artificial Intelligence (AI) on Twitter Social Media Using Lexicon-Based Method (Analisis Sentimen Artificial Intelligence (Ai) Pada Media Sosial Twitter Menggunakan Metode Lexicon Based)," *JuSiTik: J. Sist. Teknol. Inform. Komunik.*, vol. 7, no. 1, pp. 21–28, 2023.

[2] B. H. Nugroho, "Comparison of AI Capabilities of Grok, ChatGPT, and Gemini in Social Media Content Analysis (Perbandingan Kemampuan AI Grok, ChatGPT, dan Gemini dalam Analisis Konten Media Sosial)," *LogicLink*, vol. 2, no. 1, pp. 56–69, 2025.

[3] A. G. Budianto, A. Trisno, E. Suryo, and G. Rudi, "Comparison of the Performance of Support Vector Machine (SVM) and Logistic Regression Algorithms for Sentiment Analysis of Retail Application Users on Android (Perbandingan Performa Algoritma Support Vector Machine (SVM) dan Logistic Regression untuk Analisis Sentimen Pengguna Aplikasi Retail di Android)," *J. Sains Dan Informatika*, vol. 10, no. November, pp. 1–10, 2024.

[4] I. T. Julianto and L. Lindawati, "Sentiment Analysis of the Academic Information System at the Garut Institute of Technology (Analisis Sentimen terhadap Sistem Informasi Akademik Institut Teknologi Garut)," *J. Algoritma*, vol. 19, no. 1, pp. 458–468, 2022.

[5] S. F. Kadir and A. Fairuzabadi, "Sentiment Analysis of Shopee Reviews on Google Play Using TF-IDF and Logistic Regression (Analisis Sentimen Ulasan Shopee di Google Play dengan TF-IDF dan Logistic Regression)," *RIGGS: J. Artif. Intell. Digital Bus.*, vol. 4, no. 2, pp. 7940–57945, 2025.

[6] B. Kholifah, I. Thoib, N. Sururi, and N. D. Kurnia, "Sentiment Analysis of Public Opinion on Online Transportation Service Issues Using Lexicon-Based InSet with Logistic Regression (Analisis Sentimen Warganet terhadap Isu Layanan Transportasi Online Berbasis InSet Lexicon menggunakan Logistic Regression)," *KLIK-KUMPULAN JURNAL ILMU KOMPUTER*, vol. 11, no. 1, pp. 14–25, 2024.

[7] S. Butsianto and A. M. Rifa'i, "Sentiment Analysis of Jamsostek Application Reviews Using SVM, Random Forest, and Logistic Regression (Analisis Sentimen Ulasan Aplikasi Jamsostek dengan SVM, Random Forest, dan Logistic Regression)," *J. Informatika Ekonomi Bisnis*, pp. 700–706, 2025, doi: 10.37034/infeb.v7i3.1266.

[8] S. N. Adhan, G. N. A. Wibawa, D. C. Arisona, I. Yahya, and R. Ruslan, "Sentiment Analysis of Wattpad Application Reviews on Google Play Store Using Random Forest (Analisis Sentimen Ulasan Aplikasi Wattpad Di Google Play Store Dengan Metode Random Forest)," *AnoaTIK: J. Teknol. Inform. Komp.*, vol. 2, no. 1, pp. 6–15, 2024.

[9] P. A. Effendi and T. Ernawati, "Sentiment Analysis of Hay Day Game Application Reviews Using Random Forest Algorithm (Analisis Sentimen Ulasan Aplikasi Game Hay Day Menggunakan Algoritma Random Forest)," *J. Informatika Dan Teknik Elektro Terapan*, vol. 13, no. 3S1, 2025.

[10] M. D. Hendriyanto, A. A. Ridha, and U. Enri, "Sentiment Analysis of Mola Application Reviews on Google Play Store Using Support Vector Machine Algorithm (Analisis Sentimen Ulasan Aplikasi Mola Pada Google Play Store Menggunakan Algoritma Support Vector Machine)," *J. Inf. Technol. Comput. Sci.*, vol. 5, no. 1, pp. 1–7, 2022.

[11] A. A. Munandar, F. Farikhin, and C. E. Widodo, "Sentiment Analysis of Online Learning Applications Using SVM Classification (Sentimen Analisis Aplikasi Belajar Online Menggunakan Klasifikasi SVM)," *JOINTECS (J. Inf. Technol. Comput. Sci.)*, vol. 8, no. 2, p. 77, 2023, doi: 10.31328/jointecs.v8i2.4747.

[12] M. B. Prayogi and G. Masitoh, "Sentiment Analysis of Alfagift Application User Reviews Using Random Forest (Analisis Sentimen Ulasan Pengguna Aplikasi Alfagift Menggunakan Random Forest)," *JISKA (J. Informatika Sunan Kalijaga)*, vol. 10, no. 2, pp. 158–170, 2025.

[13] A. F. Anjani, D. Anggraeni, and I. M. Tirta, "Implementation of Random Forest Using SMOTE for Sentiment Analysis of Sister for Students UNEJ Application Reviews (Implementasi Random Forest Menggunakan SMOTE untuk Analisis Sentimen Ulasan Aplikasi Sister for Students UNEJ)," *Jurnal Nasional Teknologi Dan Sistem Informasi*, vol. 9, no. 2, pp. 163–172, 2023.

[14] N. O. Adiwijaya, M. F. Al Abror, T. Dharmawan, and M. A. Hidayat, "Optimizing the Thesis Topic Recommendation Model Based on Student Academic Performance Using SMOTE (Optimasi Model Rekomendasi Topik Skripsi berdasarkan Performa Akademik Mahasiswa menggunakan SMOTE)," *Proc. Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika (SNESTIK)*, vol. 5, no. 1, pp. 83–90, Jul. 2025.

[15] R. Wahyudi et al., "Sentiment Analysis on Grab Application Reviews on Google Play Store Using Support Vector Machine (Analisis sentimen pada review aplikasi grab di google play store menggunakan support vector machine)," *Jurnal Informatika*, vol. 8, no. 2, pp. 200–207, 2021.

[16] U. Kulsum, M. Jajuli, and N. Sulistiyowati, "Sentiment Analysis of WeTV Application on Google Play Store Using Support Vector Machine Algorithm (Analisis Sentimen Aplikasi WETV di Google Play Store Menggunakan Algoritma Support Vector Machine)," *J. Appl. Informatics Comput.*, vol. 6, no. 2, pp. 205–212, 2022.

[17] E. Eskiyaturrofikoh and R. R. Suryono, "Sentiment Analysis of Application X on Google Play Store Using Naive Bayes and Support Vector Machine (SVM) Algorithms (Analisis sentimen aplikasi x pada google play store menggunakan algoritma naïve bayes dan support vector machine (svm))," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 3, pp. 1408–1419, 2024.

[18] R. I. Alhaqq, I. Made, K. Putra, Y. Ruldeviyani, I. M. K. Putra, and Y. Ruldeviyani, "Sentiment Analysis of MySAPK BKN Application Usage on Google Play Store (Analisis Sentimen terhadap Penggunaan Aplikasi MySAPK BKN di Google Play Store)," *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, vol. 11, no. 2, 2022.

[19] T. Tinaliah and T. Elizabeth, "Sentiment Analysis of PrimaKu Application Reviews Using Support Vector Machine Method (Analisis Sentimen Ulasan Aplikasi PrimaKu Menggunakan Metode Support Vector Machine)," *JATISI (J. Teknik Informatika Dan Sistem Informasi)*, vol. 9, no. 4, pp. 3436–3442, 2022.

[20] M. Fauzi et al., "Implementation of Machine Learning for Weather Prediction Using Support Vector Machine (Implementasi Machine Learning Untuk Memprediksi Cuaca Menggunakan Support Vector Machine)," *J. Ilm. Komputasi*, vol. 23, no. 1, pp. 45–50, 2024, doi: 10.32409/jikstik.23.1.3499.

[21] K. A. Rokhman, B. Berlilana, and P. Arsi, "Comparison of Support Vector Machine and Decision Tree Methods for Sentiment Analysis of Reviews on Online Transportation Applications (Perbandingan metode support vector machine dan decision tree untuk analisis sentimen review komentar pada aplikasi transportasi online)," *J. Inf. Syst. Manag. (JOISM)*, vol. 2, no. 2, pp. 1–7, 2021.

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

[22] F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, "Sentiment Analysis of Dana Application Reviews Using Random Forest Method (Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest)," *J. Pengembang. Teknol. Inform. Ilmu Komput.*, vol. 6, no. 9, pp. 4305–4313, 2022.

[23] O. I. Gifari, M. Adha, I. R. Hendrawan, and F. F. S. Durrand, "Sentiment Analysis of Film Reviews Using TF-IDF and Support Vector Machine (Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine)," *J. Inf. Technol.*, vol. 2, no. 1, pp. 36–40, 2022.

[24] P. A. Nugroho, N. Sucahyo, and I. Kurniati, "Sentiment Analysis on Twitter Social Media to Assess Public Response to the Prakerja Card Selection (Sentimen Analisis pada Sosial Media Twitter untuk Menilai Respon Masyarakat terhadap Seleksi Kartu Prakerja)," *J. Teknol. Inform. dan Komput. MH. Thamrin*, vol. 9, no. 1, pp. 72–83, 2023.

[25] S. A. Putra and A. Wijaya, "Sentiment Analysis of Artificial Intelligence (AI) on Twitter Social Media Using Lexicon-Based Method (Analisis Sentimen Artificial Intelligence (Ai) Pada Media Sosial Twitter Menggunakan Metode Lexicon Based)," *JuSiTik: J. Sist. Teknol. Inform. Komunik.*, vol. 7, no. 1, pp. 21–28, 2023.