Journal of Information Systems and Informatics
Vol. 7, No. 2, June 2025 e-ISSN: **2656-4882** p-ISSN: **2656-5935**
DOI: 10.51519/journalisi.v7i2.1099 Published By DRPM-UBD

Sentiment Analysis and Classification of User Reviews of the 'Access by KAI' Application Using Machine Learning Methods to Improve Service Quality

Hildegardis Kristina Saka[1], Putri Taqwa Prasetyaningrum[2]

[1,2] Information Systems, Mercu Buana University Yogyakarta, Yogyakarta, Indonesia
Email: [1]211210041@student.mercubuana-yogya.ac.id, [2]putri@mercubuana-yogya.ac.id

Abstract

This research applies sentiment analysis to understand user perceptions of the Access by KAI application, especially specific aspects such as speed, payment process, and user interface (UI/UX). User reviews are collected and processed through preprocessing stages, balancing using the SMOTE method, and classified using three machine learning algorithms, namely Support Vector Machine (SVM), Decision Tree, and Logistic Regression. The SVM model achieved the highest accuracy of 89.33%, followed by Logistic Regression at 88%, and Decision Tree at 86.67%. Precision, recall, and F1-scores for each model were also evaluated, showing strong performance in detecting negative sentiments but lower performance for neutral and positive sentiments. In addition, keyword-based analysis revealed that negative sentiment was most commonly found in the aspects of the payment process and speed. WordCloud visualization also strengthens the results by showing the dominance of negative words in user reviews. The results of this study provide important suggestions and input for application developers to improve aspects of the service that are considered less satisfactory by users. Thus, this study can be used as a practical guide in making strategic decisions to improve the quality of service and user satisfaction of the Access by KAI application.

Keywords: Analysis Sentiment, Machine Learning, Access by KAI, Service Quality, Speed, Payment, UI/UX

1. INTRODUCTION

The rapid development of information and communication technology (ICT) has significantly transformed how people interact with public services, especially in the transportation sector. The "Access by KAI" application, developed by PT Kereta Api Indonesia, is designed to facilitate efficient train ticket booking, cancellations, and e-boarding pass issuance, aiming to provide convenience for passengers in planning their trips. Despite these features, challenges in user satisfaction remain, highlighting the need for in-depth analysis to identify and address user concerns [1].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Sentiment analysis has emerged as an effective method to understand user perceptions by analyzing app reviews. Through this analysis, developers can identify specific aspects requiring improvement, such as application speed, payment processing, and user interface design. Although sentiment analysis has been widely applied to improve service quality across many domains, there is a noticeable gap in research specifically focusing on the Access by KAI application. Existing studies tend to emphasize general sentiment classification without deeply investigating the critical service aspects that influence user satisfaction in this particular context [2]. Sentiment analysis has proven to be an effective method to understand user perceptions across various domains, including transportation applications. However, specific research on the Access by KAI application remains limited

Previous studies have shown that sentiment analysis can provide valuable insights for companies to improve service quality and meet user expectations. For example, research by [3] shows that sentiment analysis can provide deep insights into user experience, helping developers design features that are more responsive to user needs. In a similar context, research by [4] regarding the prediction of Indihome services using the K-Nearest Neighbor method also shows how data mining-based analysis can help service providers understand customer needs and preferences. By utilizing historical data classification, the study successfully predicted the most popular types of services with a high accuracy of 99.99% " Research by[5] Emphasizing that an accurate understanding of user sentiment can increase customer satisfaction and improve the quality of application services. In the context of the 'Access by KAI' Application, the application of machine learning methods in classifying user reviews is very important. This method not only allows grouping reviews into positive, 1 negative, or neutral categories, but can also identify specific problems faced by users. In addition, the data mining approach can also be used to improve service efficiency through mining transaction patterns, as shown by[6] which utilizes the FP-Growth algorithm to form the best sales package in a coffee shop based on customer shopping habits. By analyzing historical data, the study is able to provide product recommendations based on association rules that help make more precise decisions. Another relevant approach can be seen in the implementation of the Support Vector Machine (SVM) algorithm for sentiment analysis of comments on digital signature applications. The study by [7] shows that the use of RBF and linear kernels, combined with feature selection and n-grams, produces high accuracy in classifying user sentiment. These findings support the importance of machine learning-based classification in understanding user perceptions of digital services automatically and efficiently.14

In the ever-evolving digital era, analyzing app user reviews is a crucial approach in understanding user perceptions and experiences. Many previous studies have

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

shown that reviews on platforms such as Google Play can provide in-depth insights into the advantages and disadvantages of digital services. One relevant study is by [8], which conducted a sentiment analysis of Natasha Clinic reviews using the SVM algorithm. This study proves the importance of automatic classification in helping service managers understand public perception and improve service quality. Not only in the context of healthcare, a similar approach is also used to understand consumer behavior in the digital banking environment. [9] utilizes clustering and classification methods to analyze the behavior of users of gamification-based mobile banking applications. The results of this study indicate that SVM is the most consistent and accurate algorithm in grouping user behavior based on interactions with gamification features. Furthermore, in the context of education, [10] using Multilayer Perceptron (MLP) to predict student graduation based on historical academic data. This study highlights how machine learning algorithms can provide high-accuracy predictions and assist educational institutions in strategic decision making. From a customer service strategy perspective, implementing website-based e-CRM has also been shown to increase customer satisfaction. [11] in his research in the Nusantara Tea business, showed that the use of a digital CRM system is able to provide more personal services and increase customer loyalty in the face of tight business competition. Other research that is also important to note is by [12] which compares the performance of RBF and Linear kernels on the SVM algorithm in the analysis of mobile banking service user sentiment. This study found that the selection of the right kernel and parameters has a significant effect on the accuracy of user review classification, thus supporting continuous service improvement.

Furthermore, several studies specifically examine the Access by KAI application. In research conducted by [13] used is SEMMA (Sample, Explore, Modify, Model, Assess). The results showed that the Logistic Regression model achieved the highest accuracy of 84%, followed by Random Forest with an accuracy of (78%), and Naive Bayes of (73%). This study highlights the dominance of negative sentiment in reviews, indicating the need for improvements to the application. In research conducted by [14] focuses on analysis sentiment review Access by KAI application uses word embedding with Word2Vec model and algorithm Naive Bayes classification and Logistic Regression. The results showed that the Logistic Regression method showed better performance in terms of accuracy and precision compared to Naive Bayes, with an accuracy of 68.83% and a precision of 75.49%. Research by [15] do analysis sentiment for understand perception user to KRL Access application on Google Play Store. This study explores the influence of various classification algorithms, such as Naïve Bayes, Random Forest, Logistic Regression, SVM, and K-Nearest Neighbors (KNN), as well as data preprocessing steps on sentiment analysis results. The results show that the Naïve Bayes method has high precision, while Random Forest provides a good balance between precision and recall. Research by [16] conducted sentiment analysis on user reviews

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

of the Access by KAI application using the K Nearest Neighbor (K-NN) algorithm. The results showed that the KNN model can achieve the highest accuracy of 87% with optimal K parameters, indicating the effectiveness of classifying user sentiment.

However, the novelty of this research lies in the integration of three key approaches that are rarely used together in the context of transportation apps in Indonesia: aspect-based sentiment analysis, WordCloud visualization, and data balancing with the SMOTE method. This combination allows for a more in-depth analysis of user perceptions, while improving classification accuracy by systematically addressing data imbalance.

This study aims to bridge this gap by applying machine learning techniques to classify user sentiments in Access by KAI reviews and identify key areas for service improvement. By focusing on specific aspects like UI/UX, speed, and payment process, the study intends to provide actionable insights that can guide developers in enhancing the overall quality of the application and improving customer satisfaction.

Based on the explanation above, this study aims to apply the Machine Learning method in analyzing sentiment and classifying user reviews of the 'Access by KAI' application, with a focus on identifying specific aspects that need to be improved, such as user interface, application speed, and payment process. The results of this study are expected to provide recommendations that can be followed up to improve customer satisfaction and the quality of 'Access by KAI' services.

2. METHODS

Machine Learning model development method for sentiment analysis. Machine learning methods have proven effective in classifying customer reviews and providing actionable insights that guide improvements in service quality [17]. The methodology flowchart can be seen in Figure 1. Research Methodology Flowchart.

2.1. Literature Study

At this stage, the researcher will conduct a literature review to understand the basic concepts of sentiment analysis, classification methods, and machine learning techniques that are relevant to this research, including the use of SVM, Decision Tree, and Logistic Regression for text classification. Similar methodological steps have also been applied in previous studies involving sentiment analysis on application reviews, such as the study by [18], which analyzed sentiment distribution on KAI Access reviews using Naive Bayes and KNN, though without aspect-level classification.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

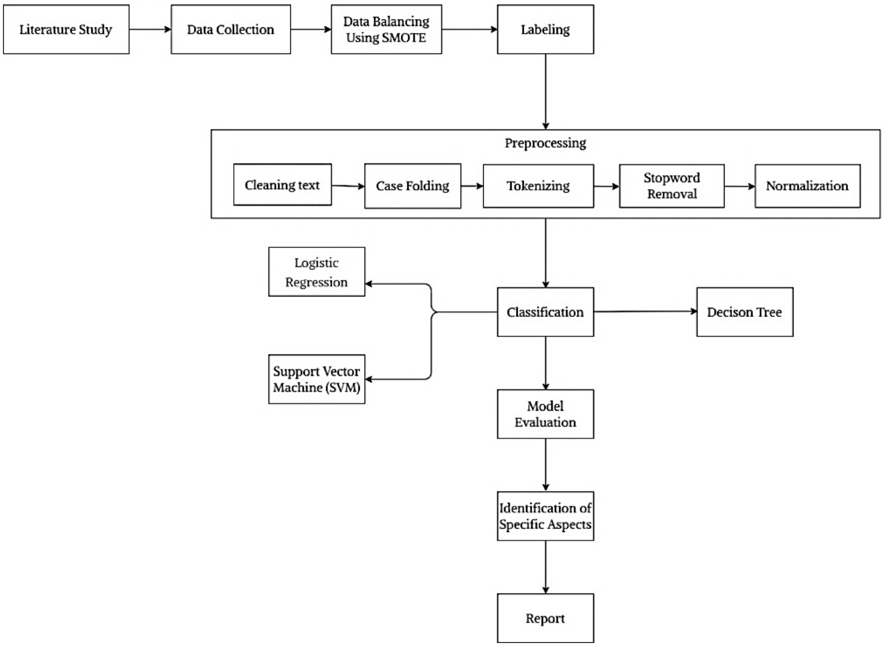
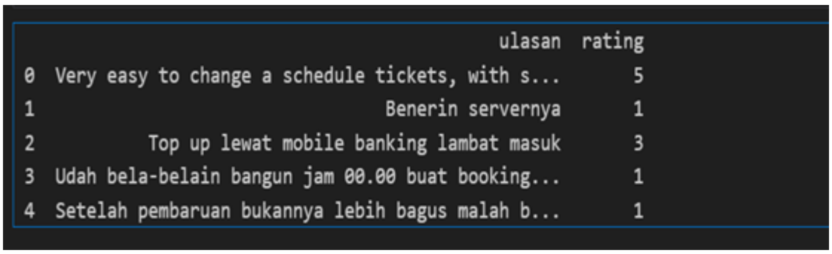


Figure 1. Research Flow

2.2. Data Collection

User review data of the application "Access by KAI" was obtained from the App Store using the Python programming language and the App Store scraper for information collection. The dataset taken amounted to 1500 datasets. This number was chosen because it was considered sufficient for user perception as well as the processing and training capabilities of the model. To avoid data bias, the data collection process was carried out randomly by maintaining the diversity of sentiments (positive, neutral, negative), as well as the aspects discussed (speed, UI / UX, payment). The resulting dataset is illustrated in Figure 2, which shows a sample of the collected user reviews in the file access by kai appstore.csv.



	ulasan	rating
0	Very easy to change a schedule tickets, with s...	5
1	Benerin servernya	1
2	Top up lewat mobile banking lambat masuk	3
3	Udah bela-belain bangun jam 00.00 buat booking...	1
4	Setelah pembaruan bukannya lebih bagus malah b...	1

Figure 2. Data Collection

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2.3. Data Balancing Using SMOTE

To address the imbalance in sentiment classes within the collected user reviews, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE artificially generates new samples for minority classes by interpolating between existing minority class instances. This approach helps prevent model bias towards majority classes and improves classification performance across all sentiment categories. This method has also been successfully applied in previous sentiment analysis research, showing significant improvements in minority class recall[19].

In this study, the oversampling target was set to 1500 reviews. This number was chosen based on preliminary experiments and computational constraints, aiming to sufficiently balance the dataset without excessively inflating minority class samples, which could lead to overfitting. By increasing the minority class instances to approximately equal levels with the majority class, the dataset became more balanced, enabling the classifiers—Support Vector Machine, Decision Tree, and Logistic Regression—to better learn distinctive patterns from all classes.

The SMOTE parameters, including the number of nearest neighbors (k) used for sample interpolation, were set to the default value of 5, which is widely accepted for text data augmentation. This configuration was considered appropriate given the nature of the dataset and the sentiment classification task.

2.4. Labeling

User review sentences are labeled positive, negative and neutral through a labeling process. In this study, the labeling process is carried out manually which comes from the user rating value. The labeling process is determined as follows: rating >3 is included in the positive sentiment category (1), rating =3 as neutral sentiment (0), and rating <3 as negative sentiment (-1). However, this manual labeling based on rating scores can introduce bias, especially in ambiguous cases such as ratings right at 3 that are categorized as neutral. To address this, reviews with ratings around the neutral threshold are carefully reviewed or additional validation is performed to ensure labeling consistency and minimize misclassification.

2.5. Preprocessing

The following are the preprocessing stages used in sentiment analysis to prepare text data before entering the classification stage. Details of each preprocessing process can be seen in Table 1.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Table 1. Data Preprocessing Steps

Preprocessing	Description	Objective
Cleaning text	Removes punctuation, numbers, and emoticons from text.	Removes unnecessary characters from data.
Case Folding	Change each line's sentences into lower case sentences.	Standardizes text formatting so that it does not differentiate between uppernand lower case letters.
Tokenization	Breaks text into words (tokens).	Facilitates word-by-word analysis.
Stopword Removal	Remove text that is meaningless but frequently appears in sentences.	Reduce noise and focus on meaningful words.
Normalization	Change the words in a sentence into basic words.	Standardize words for consistency of analysis.

2.6. Classification

In this study, three classification algorithms were used, namely Support Vector Machine (SVM), Decision Tree, and Logistic Regression. [20] SVM was chosen because of its superior ability to handle high-dimensional text data and provide high accuracy even though the amount of data is limited, but it is sensitive to the selection of kernels and parameters [21]. Support Vector Machines (SVM) are known for their ability to handle high-dimensional data, making them an ideal choice for sentiment analysis in applications with large and complex datasets. Decision Tree was used because of its easy-to-understand interpretation and clear visualization of results, although it tends to experience overfitting if pruning is not done. Meanwhile, Logistic Regression was chosen because of its simplicity of the model, efficiency, and ability to predict binary output effectively, although limited to linear relationships [22].

2.7. Model Evaluation

Before evaluating the calcification model, the dataset is divided into training and testing sets with a ratio of 70:30 to ensure unbiased assessment of model performance. In addition, k-fold cross validation with k=5 is applied during the training phase to optimize model parameters and reduce overfitting. Accuracy, precision, recall, and f1-score are some of the methods used to evaluate classification performance.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2.8. Identify Specific Aspects

Identify specific aspects that need improvement such as user interface, application speed, and payment process based on the results of aspect-based sentiment analysis. This step helps to highlight which features receive the most negative sentiment from users and can be prioritized for service improvement.

2.9. Report

Presenting research results in the form of word clouds, graphs and data visualizations to provide a clearer interpretation of sentiment distribution and aspect-based analysis outcomes. These visualizations help illustrate the frequency of keywords, sentiment polarity trends, and performance comparisons between classification models.

3. RESULTS AND DISCUSSION

In this study, the primary objective is to evaluate key aspects influencing user experience—specifically UI/UX design, system speed, and the payment process. After completing the preprocessing phase to clean and balance the dataset, the next step focuses on classifying user review sentiments using three prominent machine learning algorithms: Support Vector Machine (SVM), Decision Tree, and Logistic Regression.

3.1. Support Vector Machine (SVM) Classification Performance

The Support Vector Machine was trained and tested on a balanced dataset of 450 user reviews, covering negative, neutral, and positive sentiments. The SVM model achieved an overall accuracy of 89.33%, suggesting that it was generally effective in predicting the correct sentiment classes across the dataset. However, accuracy alone does not tell the whole story; a closer examination of class-specific metrics is essential to understand where the model excels and where it struggles.

Table 2. Classification of Support Vector Machine Test Data

Sentiment Class	Precision	Recall	F1-Score
Negative (-1)	90%	99%	94%
Neutral (0)	67%	8%	14%
Positive (1)	67%	8%	14%

As shown in Table 2, the precision, recall, and F1-score for negative sentiments (-1) were remarkably high, with a precision of 90% and a near-perfect recall of 99%. This indicates that the model almost always identifies negative reviews accurately, missing very few of them. The high F1-score of 94% confirms the model's robustness in classifying negative sentiments, making it particularly reliable for detecting dissatisfaction or problems users may encounter with the UI, system speed, or payment process. However, the model's performance dramatically declines when it comes to neutral (0) and positive (1) sentiments. For both of these classes, precision stands at 67%, which might seem acceptable at first glance, but the recall drops sharply to just 8%. This means that while two-thirds of the neutral and positive reviews the model predicts are correct, it fails to identify more than 90% of actual neutral and positive reviews in the dataset. The F1-score for both classes, consequently, is very low at only 14%, highlighting a critical weakness.

These findings indicate a significant imbalance in the model's performance across different sentiment classes. The SVM is highly sensitive to negative reviews, but it practically ignores or misclassifies the vast majority of neutral and positive feedback. Such skewed predictions can lead to misleading conclusions if the model is deployed without additional adjustments, as it may exaggerate the prevalence of negative user experiences while underreporting neutral or positive ones. This performance disparity could stem from several factors, including the linguistic characteristics of the reviews themselves. Negative reviews often use strong, distinctive words or phrases that make them easier for the model to detect. In contrast, neutral and positive reviews might share overlapping vocabulary or more subtle expressions of sentiment, making them harder for the SVM to distinguish.

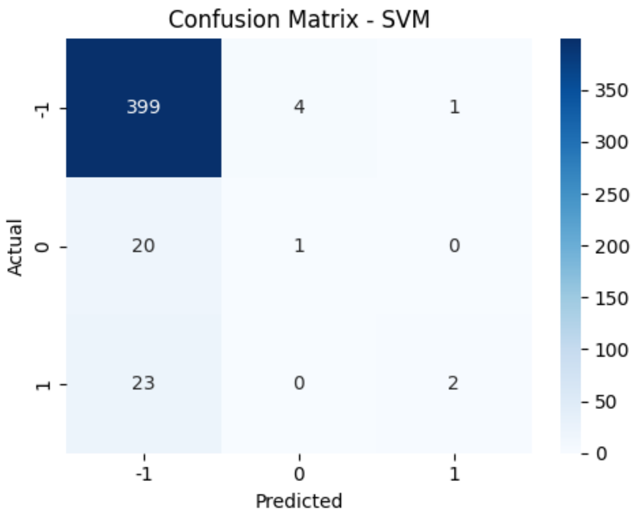


Figure 3. Confusion matrix of SVM test data

The confusion matrix presented in Figure 3 underscores this point visually, showing that 399 out of 404 negative reviews were correctly identified, while the vast majority of neutral and positive reviews were misclassified as either negative

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

or neutral. This clear pattern of misclassification demonstrates the SVM's heavy reliance on features unique to negative sentiment, at the expense of properly learning the nuances of the other two classes.

From an application perspective, this has important implications. If this model were to be integrated into a real-time user feedback analysis system—such as monitoring reviews for a mobile app or an e-commerce website—it would likely generate biased insights by overemphasizing negative experiences. This could lead developers or business stakeholders to focus resources disproportionately on fixing perceived problems, even when user sentiment might be more balanced or even positive overall.

Addressing this imbalance is crucial for improving the model's practical utility. Several strategies could help, such as collecting additional labeled data with a stronger representation of neutral and positive reviews, using data augmentation techniques, or applying advanced algorithms like class-weighted SVMs to penalize misclassification of underrepresented classes more heavily. Additionally, exploring ensemble methods or deep learning-based classifiers might improve performance by capturing more subtle patterns in neutral and positive sentiments.

Beyond algorithmic adjustments, feature engineering could also play a key role in enhancing model performance. Incorporating contextual features—such as the presence of emojis, user ratings, or syntactic patterns—might provide more distinguishing signals for neutral and positive reviews. Fine-tuning preprocessing steps, like lemmatization and stop-word removal, could further improve sentiment differentiation by reducing noise in the text data.

In summary, while the Support Vector Machine demonstrates excellent capability in detecting negative sentiments—which can be valuable for flagging urgent issues—the significant shortfall in recognizing neutral and positive feedback points to a critical limitation. Improving the model's balance across sentiment classes is essential to achieve a more accurate and holistic understanding of user perceptions related to UI/UX design, performance speed, and payment processes.

3.2. Decision Tree Classification Performance

The Decision Tree classifier was tested on the same balanced dataset of 450 user reviews, achieving an overall accuracy of **86.67%**, which, while respectable, falls slightly below the SVM's performance. However, like the SVM, the Decision Tree also demonstrates a clear bias towards negative sentiment detection, with markedly weaker performance in recognizing neutral and positive reviews. Table

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3 summarizes the Decision Tree's precision, recall, and F1-score across each sentiment class.

Table 3. Classification of Decision Tree Test Data

Sentiment Class	Precision	Recall	F1-Score
Negative (-1)	91%	95%	93%
Neutral (0)	11%	11%	11%
Positive (1)	26%	20%	22%

The model demonstrates excellent performance in detecting negative sentiments, with a precision of 91% and a recall of 95%. These figures mean that not only does the model correctly identify most negative reviews, but it also rarely mislabels non-negative reviews as negative. The high F1-score of 93% further confirms the model's consistency in classifying negative feedback, making it reliable for highlighting areas where users express dissatisfaction with UI/UX, loading times, or payment issues.

In stark contrast, the model's ability to detect neutral (0) and positive (1) sentiments is significantly compromised. For neutral reviews, the Decision Tree achieves only 11% precision and 11% recall—indicating that it correctly identifies just a small fraction of neutral reviews, while most are misclassified as negative. The positive class fares only slightly better, with a precision of 26% and a recall of 20%, revealing that the majority of positive sentiments are also incorrectly categorized, most often as negative.

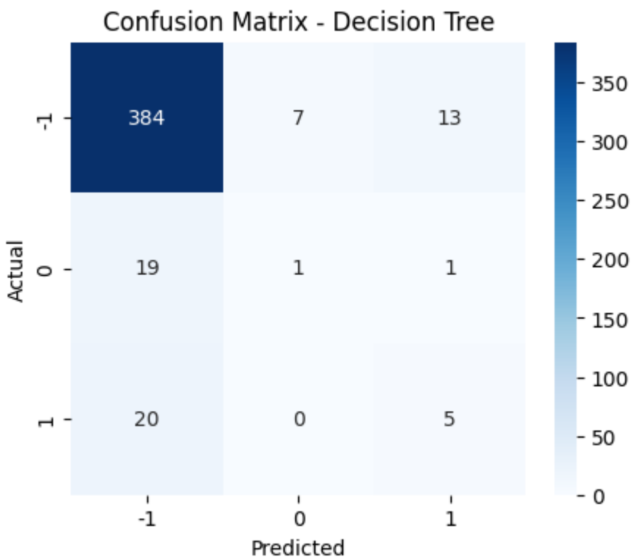


Figure 4 .Confusion matrix of Decision Tree test data

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

The confusion matrix in Figure 4 offers a detailed breakdown of these misclassifications. For the 404 negative sentiment reviews, 384 are correctly classified, while 7 are incorrectly labeled as neutral and 13 as positive. This shows that the model is highly adept at spotting negative feedback, which can be valuable in quickly surfacing critical issues in user reviews. However, the performance sharply declines in the other classes: among 21 neutral reviews, only 1 is correctly classified, with 19 misclassified as negative and 1 as positive; for the 20 positive reviews, just 5 are correctly identified, while 15 are incorrectly labeled as negative. These patterns illustrate a significant class imbalance issue. The Decision Tree's tendency to over-predict the negative class can lead to misleading analytics in real-world scenarios. For example, if integrated into a customer experience dashboard, the system might trigger unnecessary alarms about widespread dissatisfaction even if many users are actually neutral or happy, potentially misguiding business decisions.

Several factors may contribute to the model's shortcomings. Decision Trees can be highly sensitive to feature distributions and imbalanced training data, often leading to overfitting on dominant patterns—in this case, linguistic cues tied to negative sentiment. Furthermore, the rigid, rule-based splits in Decision Trees can struggle to capture the subtleties and contextual nuances present in neutral and positive sentiments, especially when reviews use overlapping vocabulary or less emphatic language.

Improving this model's performance will likely require a combination of strategies. First, enhancing the dataset with more diverse examples of neutral and positive reviews can help the model learn better distinctions. Second, employing ensemble methods like Random Forest or Gradient Boosting could improve generalization by combining multiple trees, each capturing different aspects of the data. Third, refining feature engineering by adding semantic features, sentiment lexicons, or even embeddings from transformer models may help the classifier better recognize subtle expressions of neutral or positive sentiment. Despite its current limitations, the Decision Tree's excellent performance on negative sentiments means it could still serve a role in systems where identifying dissatisfaction is the top priority—such as alerting developers about UI bugs or slow payment processing. However, without significant improvements, relying solely on this model risks missing out on valuable insights hidden in neutral or positive user feedback.

3.3. Logistic Regression Classification

The Logistic Regression model was evaluated on the same balanced dataset of 450 user reviews, achieving an overall accuracy of 88%. This places it between the SVM (89.33%) and the Decision Tree (86.67%), indicating it performs competitively overall. However, like the other models, its performance varies significantly across

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

sentiment classes, showing a clear strength in negative sentiment detection while struggling with neutral and positive classes. Table 4 provides a detailed breakdown of precision, recall, and F1-score for each sentiment label.

Table 4 . Classification of Logistic Regression Test Data

Sentiment Class	Precision	Recall	F1-Score
Negative (-1)	91%	96%	94%
Neutral (0)	9%	5%	6%
Positive (1)	50%	24%	32%

The model's performance on negative reviews is outstanding, with a precision of 91% and a recall of 96%, resulting in an impressive F1-score of 94%. This demonstrates that Logistic Regression is highly effective at identifying negative feedback accurately, minimizing both false positives and false negatives. In practical applications, this means it would reliably flag complaints or issues in user reviews about slow performance, UI bugs, or payment failures, helping teams quickly identify and address critical user concerns. However, the model's effectiveness sharply declines for neutral and positive reviews. For the neutral class, Logistic Regression achieves an alarmingly low precision of 9% and a recall of only 5%. This suggests that not only does the model rarely predict the neutral class correctly, but it also misclassifies the vast majority of neutral reviews as either negative or positive, failing to capture the subtleties of neutral sentiment. The positive class fares slightly better, with a precision of 50% and a recall of 24%, but these metrics are still insufficient for reliable classification, given that more than three-quarters of positive reviews remain misclassified.

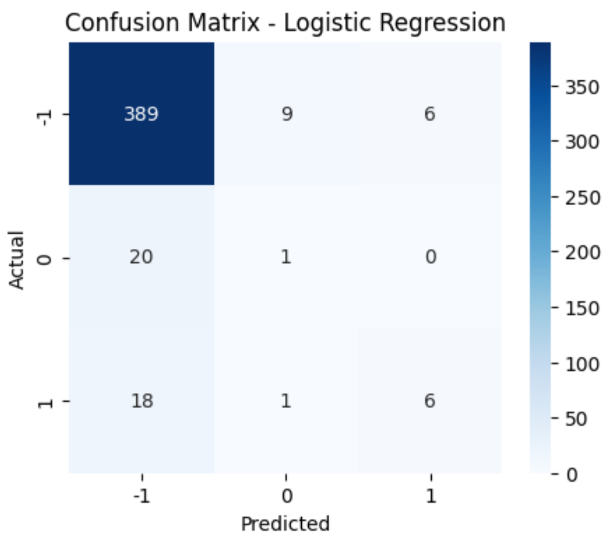


Figure 5. Confusion matrix of Logistic regression test data

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

The confusion matrix presented in Figure 5 illustrates these patterns visually. It highlights the model's excellent performance in identifying negative reviews—almost all negative samples are classified correctly. However, most neutral and positive reviews are misclassified, particularly as negative. Specifically, in the neutral sentiment group, only 1 out of 21 reviews is correctly classified, while the majority are incorrectly labeled as negative. For positive reviews, just 5 out of 20 are predicted correctly, with the rest misclassified, again mostly as negative.

This strong skew toward detecting negative sentiment aligns with findings from previous studies, such as the research by [23], which noted that Logistic Regression achieved high accuracy when classifying negative sentiments in reviews of KAI Access, but consistently underperformed in recognizing neutral sentiments. This consistency suggests that Logistic Regression, by its nature as a linear classifier, tends to focus on the most dominant patterns in the data—which, in this case, are the distinctive cues signaling negative sentiment—at the expense of more subtle or overlapping patterns present in neutral or positive feedback.

The practical implication of this imbalance is clear: deploying this model in production without addressing its shortcomings could result in dashboards or analytics systems that overemphasize user dissatisfaction. This could mislead developers or decision-makers into believing that issues are more widespread than they truly are, potentially diverting attention from opportunities for enhancing positive experiences or understanding nuanced feedback.

Several strategies could help improve the Logistic Regression model's performance on underrepresented classes. Techniques like resampling to balance class distributions, applying class weighting during model training, or exploring more complex nonlinear models could reduce misclassification. Additionally, enriching the feature set with sentiment-specific linguistic features, contextual embeddings, or metadata (e.g., star ratings or review lengths) might help the model better distinguish between neutral, positive, and negative sentiments.

Moreover, combining Logistic Regression with other classifiers in an ensemble approach—such as stacking or voting classifiers—could leverage its strong performance on negative classes while compensating for its weaknesses on neutral and positive sentiments. Fine-tuning preprocessing steps, like expanding contractions or accounting for negations, could also add valuable nuance that Logistic Regression might otherwise miss. Ultimately, while the Logistic Regression model proves to be a reliable tool for identifying negative user experiences, it requires significant improvement to provide a balanced and comprehensive view of user sentiment across all classes. Only then can it support more accurate and actionable insights for teams focused on enhancing user experience in areas such as UI/UX, system speed, and payment processing.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

3.4. Comparison of Algorithm Testing

After individually evaluating the performance of Support Vector Machine (SVM), Decision Tree, and Logistic Regression on the balanced review dataset, a comprehensive comparison was conducted to better understand each model's strengths and weaknesses. This comparison, based on cross-validation results, provides a holistic view of how these algorithms perform relative to each other in real-world sentiment classification scenarios.

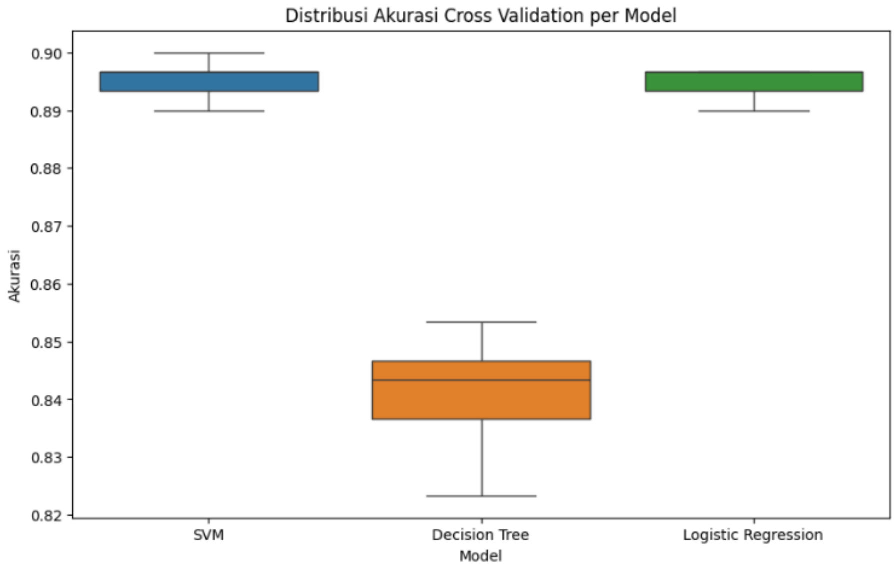


Figure 6. Comparison of Algorithms

The cross-validation results, illustrated in Figure 6, reveal that the SVM model achieved the highest average accuracy of 0.8953, closely followed by Logistic Regression at 0.8947, with the Decision Tree lagging behind at 0.8407. This suggests that both SVM and Logistic Regression offer consistent performance across different data folds, providing greater reliability for practical applications where variations in incoming user reviews are expected. In contrast, the Decision Tree exhibits greater variance and generally lower accuracy, indicating less stability and robustness when faced with diverse review content.

Figure 7, the comparison diagram, visualizes this trend clearly: SVM consistently outperforms the other two models, maintaining the highest accuracy across multiple folds. Logistic Regression tracks closely behind but still shows occasional dips in performance. The Decision Tree, meanwhile, displays the most fluctuation and the lowest average accuracy overall. These visual insights reinforce that while all three algorithms can identify negative sentiments effectively, only SVM and

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Logistic Regression demonstrate the consistency required for dependable deployment in sentiment analysis tasks.

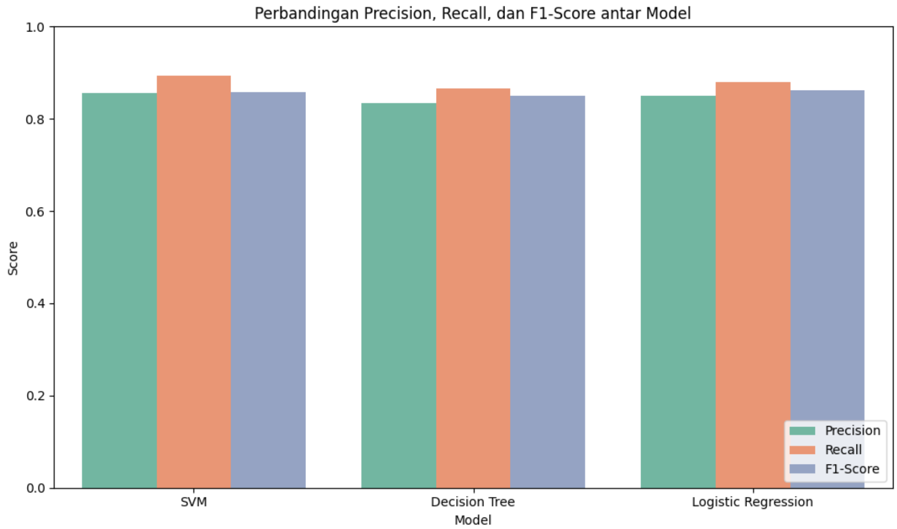


Figure 7. Comparison diagram of Algorithms

Supporting these findings, research by [24] has shown that SVM achieves superior precision and recall in real-world text classification applications, such as categorizing user feedback in digital library systems. This external validation strengthens confidence in SVM's reliability for sentiment analysis in areas like UI/UX evaluation, performance feedback, and payment process reviews, making it a strong candidate for production use. The evaluation results summarized in Table 5 provide a detailed breakdown of each algorithm's performance metrics, particularly in recognizing negative, neutral, and positive sentiments.

Table 5. Performance Metrics of Sentiment Classification Models

Algorithm	Accuracy (%)	Precision Negative	Recall Negative	Precision Neutral	Recall Neutral	Precision Positive	Recall Positive
SVM	89.33	90%	99%	67%	8%	67%	8%
Decision Tree	86.67	91%	95%	11%	11%	26%	20%
Logistic Regression	88.00	91%	96%	9%	5%	50%	24%

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

These results highlight several key insights:

1) SVM stands out as the best-performing model overall, offering the highest accuracy and exceptional performance in identifying negative sentiments, but it struggles with recognizing neutral and positive sentiments, reflected in low recall for those classes.
2) Logistic Regression provides competitive accuracy and excellent negative sentiment detection, but like SVM, it performs poorly with neutral reviews and only moderately with positive sentiments.
3) Decision Tree achieves the weakest overall performance, especially for neutral and positive classes, though it still performs reasonably well on negative sentiments.

The consistent shortcomings across all three models in recognizing neutral and positive sentiments point to a fundamental challenge in the dataset or feature representation: negative reviews are often more explicit and easier for algorithms to classify, while neutral and positive sentiments can be more nuanced and linguistically similar, leading to misclassifications. Addressing these limitations will require targeted improvements. Strategies like enhancing dataset balance with more neutral and positive examples, adopting advanced feature engineering with context-aware representations (e.g., word embeddings or transformer-based models), and exploring ensemble methods or hybrid deep learning architectures could significantly improve recognition of underrepresented sentiment classes. In conclusion, while SVM emerges as the most reliable model for identifying negative user experiences, both it and Logistic Regression show significant weaknesses in detecting neutral and positive feedback—gaps that must be bridged to build a comprehensive sentiment analysis system capable of supporting balanced, actionable insights for improving user experience in UI/UX design, performance optimization, and payment process enhancement.

3.5. Identification of Aspects

This study explores specific aspects that users often mention in Access by KAI reviews, including Speed, Payment Process and UI/UX. Identification of these aspects is done through a keyword-based matching approach from the review text. The following keywords are used, UI/UX_keywords : ["appearance", "design", "menu", "interface", "userfriendly", "easy", "design", "navigation", "bug", "logout", "data", "filter", "complicated", "face", "features", "recognition"] Speed_keywords : [" lemot ", " slow ", " fast ", " speed ", " loading ", " responsive ", "lag ", "delay ", "server ", "timeout ", " ngebug ", "error ", "buffering ", "crash ", "overload"] Payment_keywords : [" pay ", "payment", " transaction ", " failed ", " balance ", " method ", " payment ", " topup ", "refund ", " otp ", " qris ", " ovo ",

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

" shopeepay ", " dana ", " kaipay ", "virtual ", "account"]. And the results of sentiment distribution based on aspects in the following Figure 8.

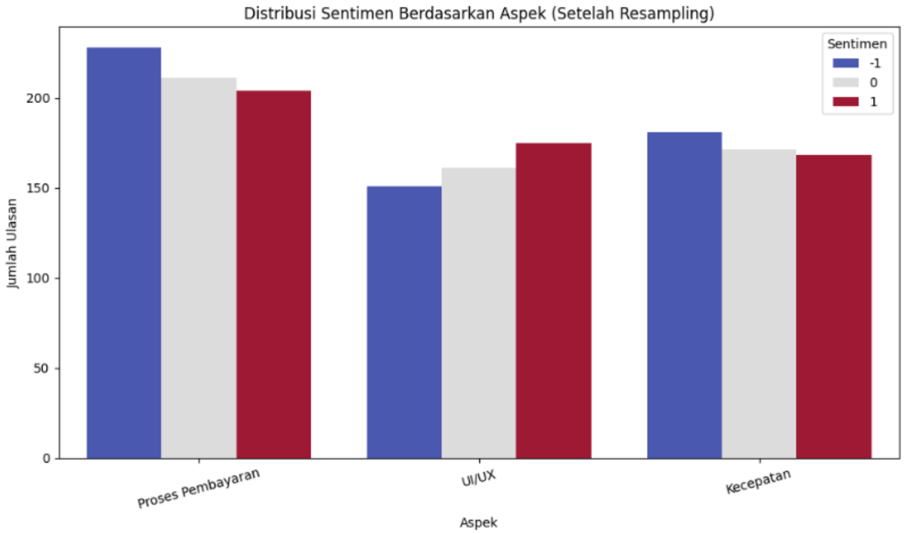


Figure 8. Sentiment graph based on aspect

Analysis results distribution sentiment show payment process aspects has the highest negative sentiment, followed by the aspect speed. The UI/UX aspect shows a diversity of sentiments, with the highest number of positive sentiments compared to the other two aspects. Based on these results, the payment process and speed features are important parts that need to be considered in improving service quality.

3.6. WordCloud Reviews

3.6.1. Negative Sentiment

Figure 9 shows a negative WordCloud, the words most often used by users in their comments. Words such as "but", "already", "even", "no", and "again" are most often used to indicate complaints related to technical issues, access issues, or disappointment with the service.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 9. WordCloud Sentimen Negatif

3.6.2. Positive Sentiment

In figure 10, words such as “application”, “train”, “me”, “kai”, and “again” have the largest size. This shows that these words appear frequently. Words such as “many”, “very”, “good”, and “very” also show that clear user satisfaction with the performance of the application or service, such as ease of access, satisfying user experience and the benefits of the service felt by the user.

Figure 10. WordCloud Sentimen Positif

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.6.3. Neutral Sentiment

In text analysis, the WordCloud visualization shown in figure 11 the most frequently occurring words in user comments or reviews that are categorized as neutral sentiment. Words like "update", "please", "often", "ticket", and "login" have larger font sizes indicating that these are words that users mention more frequently.

Figure 11. WordCloud Sentimen Netral

The aspect-based analysis highlights that the payment process and application speed are the primary sources of negative sentiment, while the UI/UX aspect shows a more balanced sentiment distribution with relatively more positive feedback. The WordCloud visualization further supports these findings by displaying dominant negative words in problematic reviews such as 'slow', 'failed', and 'refund', alongside positive words like 'easy', 'good', and 'comfortable' that appear frequently in the appreciated UI/UX aspect.

3.7. Discussion

The results of this study clearly demonstrate that all three classifiers—Support Vector Machine, Decision Tree, and Logistic Regression—perform strongly when it comes to identifying negative sentiments, achieving consistently high precision and recall scores in this category. Among them, the SVM model emerges as the top performer, slightly surpassing both Logistic Regression and Decision Tree in overall accuracy and precision for negative sentiment detection. This indicates that SVM's ability to capture clear, polarized expressions of dissatisfaction makes it highly effective for flagging critical issues reported by users. However, a striking trend emerges across all models: consistent difficulty in classifying neutral and

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

positive sentiments, evidenced by their uniformly low recall scores. This means the models frequently fail to identify when a review expresses a neutral or positive sentiment, often misclassifying these as negative. Such misclassifications skew the overall sentiment analysis, leading to an inaccurate representation of user satisfaction.

Several factors contribute to this shortcoming. First, there's a noticeable class imbalance in the dataset: negative reviews significantly outnumber neutral and positive ones, which limits the models' exposure to diverse examples of non-negative feedback during training. Second, the subtlety and ambiguity inherent in neutral sentiments make them difficult for traditional machine learning models to capture reliably. Unlike strongly negative or positive language, neutral statements often lack distinct markers, increasing the likelihood of misclassification. Third, feature overlap between sentiment classes—where words used in neutral or mildly positive contexts resemble those in negative reviews—further complicates accurate separation. This challenge in recognizing neutral feedback is not unique to this study. As highlighted by [25], Naive Bayes models, despite performing well on highly polarized datasets, also struggle to accurately detect neutral sentiments. This consistency across studies underscores a broader limitation in traditional machine learning approaches when applied to nuanced sentiment analysis tasks.

To address these issues, it's clear that improving detection of neutral and positive sentiments requires both data and methodological enhancements. Expanding the dataset to include more balanced examples across all sentiment categories would provide the models with richer training signals, helping them learn the nuances of each class. Additionally, employing advanced modeling techniques, such as deep learning or transformer-based architectures like BERT, could significantly improve performance. These models excel at capturing context and subtle semantic cues, enabling more precise differentiation between sentiments that traditional algorithms often conflate. Another important consideration is how the data is preprocessed and represented. Re-evaluating preprocessing strategies to preserve sentiment-relevant details—like negations, intensifiers, or contextual phrases—can prevent loss of critical information during tokenization or feature extraction. Techniques like contextual embeddings can also enable models to understand the meaning of words within specific review contexts, reducing misclassifications stemming from ambiguous or overlapping vocabulary.

This study also took care to guide readers through each figure with narrative explanations. For instance, prior to Figure 3, the confusion matrix of SVM results was introduced to clearly highlight where the model excelled and where it fell short. This approach ensures that visualizations are not only informative but also accessible, helping readers grasp the practical implications of the performance metrics.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Ultimately, these findings highlight a critical takeaway: while high accuracy in detecting negative sentiments is valuable—particularly for identifying pain points and urgent issues—overlooking neutral and positive feedback risks missing important opportunities to understand what users appreciate or find satisfactory. This incomplete view can skew business strategies, leading to missed chances to reinforce strengths or address less obvious weaknesses. For organizations seeking balanced, actionable insights, future work should focus on refining sentiment models to deliver consistent performance across all sentiment categories. Techniques such as data augmentation for minority classes, contextual embedding for semantic understanding, and fine-tuning large pre-trained language models hold promise for achieving more nuanced and accurate sentiment classification. By bridging these gaps, businesses and developers can gain a holistic understanding of user experiences, empowering them to make informed, user-centered decisions that truly elevate service quality and customer satisfaction.

5. CONCLUSION

Based on the research findings, the Support Vector Machine (SVM) algorithm demonstrated the best overall performance in classifying sentiments from user reviews of the Access by KAI application, achieving an accuracy of 89.33% and showing excellent capability in detecting negative reviews. However, all three tested algorithms—SVM, Decision Tree, and Logistic Regression—exhibited significant limitations in accurately identifying neutral and positive sentiments, as reflected by their low recall scores in these categories. Aspect-based analysis using a keyword-matching approach revealed that the payment process received the highest proportion of negative sentiment, followed by the application's speed. In contrast, the UI/UX aspect showed a more balanced sentiment distribution, with a relatively higher proportion of positive feedback. These results were further corroborated by WordCloud visualizations, which highlighted the prevalence of complaint-related words in negative reviews and appreciation-related words in positive reviews. Enhancing user satisfaction by improving critical aspects such as the payment process and user interface has been recognized in previous studies as essential for delivering a better customer experience.

This study has several limitations, including the relatively small dataset size and the use of a simple keyword-based model for aspect identification, which may affect the depth and generalizability of the findings. For future research, it is recommended to employ more advanced techniques, such as deep learning or transformer-based models like BERT, to improve sentiment classification and aspect extraction accuracy. Additionally, expanding the dataset, potentially including multilingual reviews, would increase the robustness and representativeness of the analysis. As an actionable strategic recommendation, this sentiment analysis system can be further developed by building a monthly feedback

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

dashboard to monitor user sentiment trends and activate an early warning system to detect spikes in negative sentiment towards certain features. In addition, this system also has the potential to be extended to other transportation applications and can integrate review data from various platforms and social media to increase the scope and depth of analysis.

REFERENCES

[1] G. Radiena and A. Nugroho, "Analisis Sentimen Berbasis Aspek Pada Ulasan Aplikasi KAI Access Menggunakan Metode Support Vector Machine," 2023.
[2] H. Indrawan, B. Irawan, and T. Suprapti, "Klasifikasi Ulasan Pengguna Aplikasi Access By KAI Berbasis Aspek Dengan Algoritma Naïve Bayes Dan SVM," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3541–3548, 2023.
[3] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," 2021.
[4] A. U. Haspriyanti and P. W. Prasetyaningrum, "Penerapan Data Mining Untuk Prediksi Layanan Produk Indihome Menggunakan Metode K-Nearst Neighbor," 2021.
[5] Irbah salsabila and Yuliant Sibaroni, "Multi Aspect Sentiment of Beauty Product Reviews using SVM and Semantic Similarity," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 520–526, Jun. 2021, doi: 10.29207/resti.v5i3.3078.
[6] W. Wahyuningsih and P. T. Prasetyaningrum, "Enhancing Sales Determination for Coffee Shop Packages through Associated Data Mining: Leveraging the FP-Growth Algorithm," *Journal of Information Systems and Informatics*, vol. 5, no. 2, pp. 758–770, May 2023, doi: 10.51519/journalisi.v5i2.500.
[7] M. Oktafani and P. T. Prasetyaningrum, "Implementasi Support Vector Machine Untuk Analisis Sentimen Komentar Aplikasi Tanda Tangan Digital," *Jurnal Sistem Informasi dan Bisnis Cerdas*, vol. 15, no. 1, Mar. 2022, doi: 10.33005/sibc.v15i1.2697.
[8] M. Rizqi, A. Rustiawan, and P. T. Prasetyaningrum, "Analisis Sentimen Terhadap Klinik Natasha Skincare di Yogyakarta Dengan Metode Google Review," *Journal of Information Technology Ampera*, vol. 5, no. 1, pp. 2774–2121, 2024, doi: 10.51519/journalita.v5i1.556.
[9] P. T. Prasetyaningrum, P. Purwanto, and A. F. Rochim, "Consumer Behavior Analysis in Gamified Mobile Banking: Clustering and Classifier Evaluation," *Online) Journal of System and Management Sciences*, vol. 15, no. 2, pp. 290–308, 2025, doi: 10.33168/JSMS.2025.0218.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

[10] M. Windarti and P. T. Prasetyaninrum, "Prediction Analysis Student Graduate Using Multilayer Perceptron," 2020.

[11] P. T. Prasetyaningrum, A. R. Wicaksono, and H. Nurrofiq, "Transformasi Pelayanan Pelanggan: Implementasi E-CRM Pada Bisnis Teh Nusantara Berbasis Website," *Technologia : Jurnal Ilmiah*, vol. 14, no. 4, p. 368, Oct. 2023, doi: 10.31602/tji.v14i4.12157.

[12] P. T. Prasetyaningrum, I. Pratama, N. T. Kadir, and A. Y. Chandra, "Comparison Of Support Vector Machine Radial Base And Linear Kernel Functions For Mobile Banking Customer Satisfaction Analysis."

[13] M. A. S. Nugroho, D. Susilo, and D. Retnoningsih, "Analisis Sentimen Ulasan Aplikasi "Access By KAI" Menggunakan Algoritma Machine Learning," *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 7, no. 2, p. 820, Dec. 2024, doi: 10.37600/tekinkom.v7i2.1854.

[14] R. Damanhuri and V. A. Husein, "Analisis Sentimen pada Ulasan Aplikasi Access by KAI Berbahasa Indonesia Menggunakan Word-Embedding dan Classical Machine Learning," *Jurnal Masyarakat Informatika*, vol. 15, no. 2, pp. 97–106, Nov. 2024, doi: 10.14710/jmasif.15.2.62383.

[15] N. D. Septiyanti, M. I. Luthfi, and N. T. Romadloni, "Komparasi Metode Klasifikasi Dalam Analisis Sentimen Ulasan Pengguna Aplikasi KRL Access Di Google Play Store," *Journal Computer Science and Information Systems : J-Cosys*, vol. 4, no. 1, pp. 64–75, Mar. 2024, doi: 10.53514/jco.v4i1.495.

[16] A. Rhamadanti, A. Rifa'i, F. Dikananda, and K. Anam, "Analisis Sentimen Pada Ulasan Access By Kereta Api Indonesia Dengan K-Nearest Neighbor," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 1, pp. 2830–7062, doi: 10.23960/jitet.v12i1.3691.

[17] A. A. Muhammad, E. Ermatita, and D. S. Prasvita, "Analisis Sentimen Pengguna Aplikasi Dana Berdasarkan Ulasan pada Google Play Menggunakan Metode Support Vector Machine," in *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya*, 2022, pp. 194–204.

[18] N. B. Sidauruk and N. Riza, "Sentimen Analisis Data Pengguna Terhadap KAI Access Systematic Literature Review," 2023.

[19] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 16, no. 2, p. 113, Apr. 2022, doi: 10.22146/ijccs.68903.

[20] Y. Sibaroni, "Multi aspect sentiment of beauty product reviews using SVM and semantic similarity," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 520–526, 2021.

[21] O. S. D. Fadhillah, J. H. Jaman, and C. Carudin, "Perbandingan Naive Bayes, Support Vector Machine, Logistic Regression Dan Random Forest Dalam Menganalisis Sentimen Mengenai Tiktokshop," *Jurnal Informatika*

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

dan Teknik Elektro Terapan, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5746.

[22] N. C. Ramadani, “Analisis Sentimen Untuk Mengukur Ulasan Pengguna Aplikasi Mobile Legend Menggunakan Algoritma Naive Bayes, SVM, Random Fores, Decision Tree, dan Logistic Regression,” *JSI : Jurnal Sistem Informasi (E-Journal*, vol. 16, no. 1, 2024.

[23] M. A. S. Nugroho, D. Susilo, and D. Retnoningsih, “Analisis Sentimen Ulasan Aplikasi ”Access By KAI” Menggunakan Algoritma Machine Learning,” *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 7, no. 2, p. 820, Dec. 2024, doi: 10.37600/tekinkom.v7i2.1854.

[24] S. D. Alyusi and I. Yuadi, “Implementation Of Machine Learning In Improving Website User Experience And Satisfaction,” *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 781–788, Mar. 2025, doi: 10.29100/jipi.v10i1.7439.

[25] R. A. Hidayah, R. Insani, and B. R. Lidiawaty, “Sentiment Classification of User Reviews for KAI Access Application Using Naive Bayes Method,” *Journal of Dinda Data Science, Information Technology, and Data Analytics*, vol. 4, no. 2, pp. 91–97, 2024.