Paradigma, Vol. 27, No. 1, March 2025, P-ISSN 1410-5063, E-ISSN: 2579-3500, Page. 11-20
Published: LPPM Universitas Bina Sarana Informatika
PARADIGMA
website: http://jurnal.bsi.ac.id/index.php/paradigma/

Prediction Customer Loyalty Using Random Forest Algorithm on Shopee Reviews

Ferdi Saputra[1], Fersellia[2]

[1,2] Informatics Engineering, Universitas Ma'arif Nahdlatul Ulama Kebumen, Kebumen, Indonesia

ARTICLE INFORMATION

Artikel History:
Received: January 3, 2025
Revised: February 5, 2025
Accepted: February 25, 2025
Available Online: March 6, 2025

Keyword:

Customer Loyalty
E-commerce
Prediction
Random Forest Algorithm
Shopee

ABSTRACT

This research develops a Shopee customer loyalty prediction model using Random Forest algorithm, utilizing customer reviews from Google Play Store. One of the key issues in e-commerce is maintaining customer loyalty amidst intense competition, so it is important to identify loyal customers and understand the factors that influence their commitment. This study involves data collection through web scraping, data cleaning, loyalty labeling, and Random Forest-based prediction model building and evaluation. The evaluation process was conducted using a confusion matrix to measure accuracy, precision, recall, and F1-score. The model classified customers into loyal, neutral, and disloyal categories, with an overall accuracy of 97%. The model showed precision, recall, and F1-score of 0.98 for loyal customers, and 0.99, 1.00, and 0.99 for disloyal customers. However, identification of neutral customers is still a challenge, with precision, recall, and F1-score of 0.92, 0.85, and 0.88, respectively. The results of this study provide strategic insights for Shopee in improving customer retention strategies and demonstrate the effectiveness of the Random Forest algorithm in analyzing review data.

Corresponding Author:

Ferdi Saputra,
Informatics Engineering,
Universitas Ma'arif Nahdlatul Ulama Kebumen,
Jl. Kutoarjo No.Km.05, Wonoboyo, Jatisari, Kec. Kebumen, Kabupaten Kebumen, Jawa Tengah, 54317,
Email: ferdisaputra755@gmail.com

INTRODUCTION

In the growing digital era, the utilization of information technology (IT) is the main key in driving business progress, especially in the field of e-commerce (Tambunan et al. 2023). One of the largest and most recognized e-commerce platforms in Indonesia is Shopee. One of the largest and most recognized e-commerce platforms in Indonesia is Shopee. Based on SimilarWeb data, Shopee became the e-commerce marketplace category with the highest number of site visits in Indonesia throughout 2023, recording around 2.3 billion visits from January to December (Adi Ahdiat, 2024). Shopee's success in attracting and retaining customers cannot be separated from effective marketing strategies, responsive customer service, and a variety of products at competitive prices (Viona et al., 2021).

However, along with the increase in the number of users and transactions, Shopee also faces challenges that are no less severe in maintaining customer loyalty. Customer loyalty is a commitment based on a positive attitude towards a brand, store, or supplier, which is reflected in repeat purchases (Yolanda et al. 2021). This loyalty reflects the customer's commitment to continue buying a particular product or service in the future, despite changes in the situation or marketing influences (Sulistyawati & Munawir, 2024). While customer reviews are often used to measure loyalty, positive reviews do not necessarily reflect continued loyalty, and negative reviews do not necessarily indicate disloyalty. Therefore, a more comprehensive approach is needed, including analysis of customer behavior patterns to complement review data.

The most valuable data source in understanding and measuring customer loyalty is customer reviews. These reviews provide direct insight into the shopping experience on e-commerce

DOI: https://doi.org/10.31294/p.v27i1.7940

platforms, covering aspects of product quality, delivery speed, customer service, and shopping convenience. (Oktavia et al. 2022). By analyzing customer reviews, patterns and trends in customer behavior and sentiment can be revealed, which ultimately helps in predicting customer loyalty.

In this context, the Random Forest algorithm becomes relevant to use. Random Forest is a powerful and flexible machine learning algorithm, capable of handling large and complex data involving many variables. It works by dividing the data into subsets and using decision trees on each subset, which are then combined to produce the final prediction. One of the main advantages of Random Forest is its ability to overcome overfitting (Nurohanisah et al. 2024). Overfitting is a common problem in machine learning that results in the model overfitting itself to the training data, resulting in degraded performance on test data (Alkhairi et al. 2024). With these characteristics, Random Forest becomes very effective in identifying important variables that affect customer loyalty and in predicting loyalty levels based on customer reviews.

Despite the advantages of the Random Forest algorithm, previous studies have shown significant research gaps. As in the research of Ferdyandi et al. (2022) in their study showed that although Random Forest provides adequate accuracy, namely 0.78 on test data, data limitations and subjectivity of reviews are the main obstacles (Ferdyandi et al. 2022). In addition, research by Masripah & Wulandari (2024) used the Naive Bayes algorithm based on Particle Swarm Optimization (PSO) and achieved 79.09% accuracy, but required additional optimization to further improve model performance (Masripah & Wulandari, 2024). Meanwhile, Mustafa et al. (2024) showed that Random Forest excels in predicting new user retention with 73.36% accuracy, but the complexity of the model is still a challenge (Mustafa et al. 2024). Another study by Muktafin et al. (2020) used a Natural Language Processing (NLP) approach for sentiment analysis, which increased prediction accuracy to 76.92%, but required more computational resources (Muktafin et al. 2020). And finally, Nafisyah & Sulistiyowati (2024) highlighted that although Naive Bayes is simple and effective, the analysis results rely heavily on review data which is often biased (Nafisyah & Sulistiyowati, 2024).

From these various studies, it can be concluded that although algorithms such as Naive Bayes, Random Forest, and NLP approaches have proven effective in some cases, each method has advantages and disadvantages that should be considered based on the complexity of the data and the purpose of the analysis. Random Forest was chosen in this study due to its superior ability to handle complex and diverse data, such as customer review text, as well as its ability to overcome overfitting by combining results from multiple decision trees (Fitri & Damayanti, 2024). It is also known to have more stable accuracy and can handle outliers more effectively than algorithms such as Naive Bayes. By using Random Forest, this research aims to build a more accurate and reliable customer loyalty prediction model, so as to provide deeper insight into the main factors that influence customer loyalty in Shopee. The results of this model will provide strategic recommendations that can be implemented by Shopee and other e-commerce platforms in improving customer satisfaction and loyalty.

This research aims to develop a more accurate customer loyalty prediction model by utilizing Shopee customer review data and an optimized Random Forest algorithm. The process includes data collection and pre-processing, model building and evaluation. In addition, this research will identify key factors that influence customer loyalty, such as product quality, price, delivery speed, and customer service. It is expected that the results of this research can significantly contribute to e-commerce, helping Shopee and other players retain customers amid fierce competition.

RESEARCH METHOD

Figure 1 shows a flow chart of the research process that includes several stages: business understanding, data understanding, data correction, data preparation, modeling, enhancement, evaluation, and deployment. There is iteration between the data correction and data preparation stages, and between the enhancement and modeling stages. This diagram is particularly relevant as it provides a clear picture of the steps to be taken in the research process, from understanding the business problem to applying the research results.

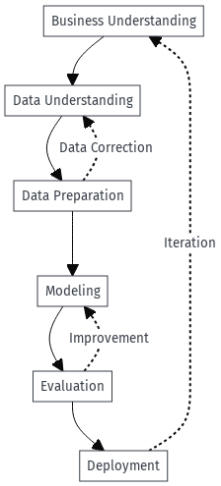

Figure 1. Research Process Flow

Figure 2 below shows the research flow which is divided into several stages, starting from problem identification and research objectives. Next, data is collected by web scraping Shopee app reviews on the Google Play Store. After the data is collected, a data cleaning process is carried out to eliminate irrelevant or incorrect information. The next stage is data pre-processing, which includes several steps such as case folding, stopword removal (filtering), tokenization, and

stemming. This pre-processed data is then labeled through the Loyalty Labeling process, which involves sentiment analysis using Lexicon Vader and classification based on review scores. In the next stage, a customer loyalty prediction model is built using the Random Forest algorithm. Evaluation of the model's performance is done using a confusion matrix, which includes metrics such as accuracy, precision, recall, and F1 score. Finally, the researcher will interpret the research results and draw conclusions in accordance with the research objectives that have been formulated at the initial stage.

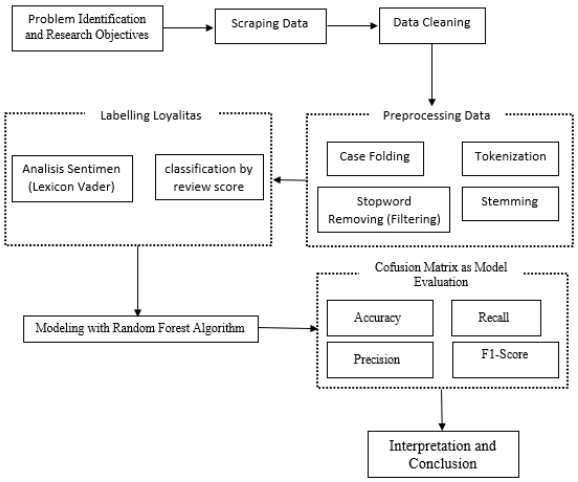

Figure 2. Research Flow for Predicting Shopee Customer Loyalty Using Random Forest Algorithm

1. Problem Identification and Research Objectives

In the initial stage, the researcher identified the problem that became the background of this research. In the context of e-commerce such as Shopee, maintaining customer loyalty is one of the biggest challenges faced by companies. Satisfied customers tend to return to make repeat purchases, while customers who are dissatisfied or have negative experiences tend to move to other platforms, leading to churn or loss of customers. However, in order to accurately determine the level of customer loyalty based on their behavior, a more comprehensive method is needed than simply analyzing the rating or review score given by the customer.

This problem arises from the fact that reviews given by users on platforms like Google Play Store do not always reflect loyalty directly. A positive review does not necessarily mean that the customer will use the same service again, and a negative review does not necessarily mean that the customer is not loyal. For example, a 5-star review does not necessarily reflect high loyalty, nor does a 3-star review necessarily mean a neutral customer. Therefore, this research combines review score analysis with text sentiment analysis to provide more comprehensive predictions.

The purpose of this research is to develop a customer loyalty prediction model based on reviews of the Shopee application on the Google Play Store using the Random Forest algorithm. This research also aims to identify factors that influence customer loyalty through sentiment analysis and loyalty labeling on the reviews obtained. In addition, this research will evaluate the performance of the prediction model using confusion matrix to determine the level of accuracy, precision, recall, and F1-score in predicting customer loyalty.

By understanding the problem and purpose of this study, the researcher focuses on how reviews collected from Shopee can be appropriately processed, labeled, and analyzed to build a predictive model capable of classifying loyal and disloyal customers. The results of this study are expected to help companies in improving customer retention strategies and reducing churn by utilizing publicly available review data.

2. Scraping Data

The collection of customer review data from the Google Play Store is done using web scraping techniques by utilizing the google-play-scraper tool (Larasati et al. 2022). A total of 10,000 reviews were collected in the period August to October 2024. This scraping process includes several main steps. First, the target data is determined, where the data collected includes Shopee customer reviews on the Google Play Store. Relevant information such as the username, the content of the review, the score or rating given on a scale of 1 to 5, as well as the date the review was left became the main focus. In addition, additional data such as the number of 'thumbs up' on the review, the version of the app used when the review was given, as well as the sentiment of the review generated after being analyzed by the Lexicon Vader method were also collected (Asri et al. 2022).

The next stage is the implementation of scraping, which is done using google-play-scraper, a Python library designed to extract app reviews from the Google Play Store. This process involves automatically extracting data by utilizing the HTML tags present on the review page. Once the scraping process is complete, the data that has been retrieved through the tool will go through a validation stage to ensure that there are no errors or formatting mistakes in the data collected. Once validated, the data is saved in CSV format or stored in a database for further processing in the preprocessing stage.

3. Data Cleaning

After the scraping process is complete, the next step is data cleaning. Data that has been collected through scraping often contains duplication or irrelevant elements, so it needs to be cleaned for more accurate analysis. In this stage, duplication removal is first performed, which removes duplicated reviews based on the same content to ensure each review is unique (Apriliansyah et al. 2025). Next, text cleaning is performed by removing special characters, excessive punctuation, and extra spaces, so that the text is cleaner and ready for further processing. In addition, a missing data check is also performed to ensure there is no missing data in the required columns. If blank data is found, a decision will be made to either delete it or fill

it with appropriate values, depending on the context and needs of the research. This data cleaning process is very important to ensure the quality of the data used in the subsequent analysis. After data cleaning is complete, the dataset will also be translated into English, to facilitate data preprocessing.

4. Preprocessing

Figure 3 below shows the flow of data preprocessing which includes the stages of case folding, tokenization, stopword removal, and stemming. This process aims to clean and simplify the customer review text so that it is more ready to be used in customer loyalty prediction analysis and modeling.

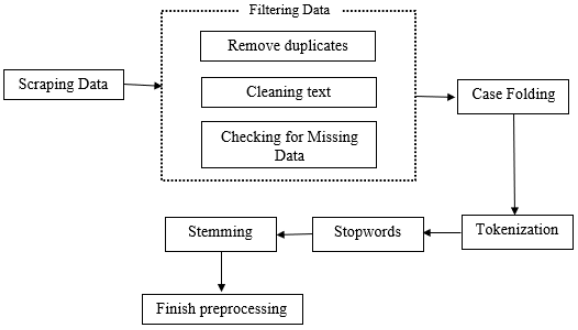

Figure 3. Pre-processing Flowchart

Figure 3 illustrates the steps taken during data preprocessing to ensure data quality before further analysis. These steps include case folding, tokenization, stopword removal, and stemming. Each stage plays a crucial role in preparing the data for use in the prediction model. The preprocessing begins with case folding, which converts all review text to lowercase letters to maintain consistency (Rahmadani et al. 2024). This is important so that words like “Shopee” and “shopee” are treated the same. For example, the sentence “Pelayanan Shopee Sangat Bagus” will be converted to “pelayanan shopee sangat bagus”. The next step is tokenization, which breaks the review text into word units (tokens) that will be analyzed individually (Permana & Bunyamin, 2024). For example, “pelayanan shopee sangat bagus” will be broken down into [“pelayanan”, “shopee”, “sangat”, “bagus”].

Following tokenization, stopword removal is performed. This stage removes common words, such as “dan,” “yang,” and “untuk,” that do not contribute significantly to the analysis (Rahmadani et al., 2024). For example, the tokens [“pelayanan”, “shopee”, “sangat”, “bagus”] are simplified to [“pelayanan”, “shopee”, “bagus”]. Lastly, stemming reduces words to their root forms to simplify analysis (Suryawan et al. 2024). This step minimizes variations of the same word, such as converting “memesan” to “pesan”. As a result, the tokens [“pelayanan”, “shopee”, “bagus”] are further simplified to [“layan”, “shopee”, “bagus”]. To summarize, the table below illustrates the changes in text throughout the preprocessing steps:

Table 1. Example of Text Change in the Data Preprocessing Stage

Step	Original Text	Processed Text
Original	"Pelayanan Shopee Sangat Bagus"	"Pelayanan Shopee Sangat Bagus"
Case Folding	"Pelayanan Shopee Sangat Bagus"	"pelayanan shopee sangat bagus"
Tokenization	"pelayanan shopee sangat bagus"	[“pelayanan”, “shopee”, “sangat”, “bagus”]
Stopword Removal	[“pelayanan”, “shopee”, “sangat”, “bagus”]	[“pelayanan”, “shopee”, “bagus”]
Stemming	[“pelayanan”, “shopee”, “bagus”]	[“layan”, “shopee”, “bagus”]

5. Labelling Loyalitas

After preprocessing, the next step is sentiment analysis and classification based on review scores. Sentiment distribution analysis provides an overview of the number of reviews in the positive, negative and neutral categories. Using Lexicon Vader, the sentiment score is calculated and weighted based on the words contained in the review text. Based on the calculated weight, if the total weight is more than 0.05, the review falls into the positive sentiment category. Conversely, if the total weight is less than -0.05, the review is categorized as a negative sentiment. Reviews with a weight between -0.05 to 0.05 will be categorized as neutral sentiment (Fathoni et al. 2024). However, sentiment alone is insufficient to fully capture customer loyalty. Therefore, the reviews are further analyzed based on their scores: reviews with scores of 4 or 5 are considered indicative of loyal customers, especially when supported by behavioral data such as high purchase frequency. Reviews with a score of 3 are classified as neutral, representing moderate satisfaction but lacking strong loyalty indicators, while reviews with scores of 1 or 2 are categorized as disloyal, often indicating dissatisfaction.

The model faces difficulties in identifying the neutral category, which is reflected in the low F1-Score in this class. This could be due to the ambiguity of the neutral category, where reviews with a score of 3 or 4 may reflect a positive attitude but not strong enough to be considered loyal. In addition, the uneven distribution of data between the loyal, neutral and disloyal categories reduces the representation of patterns in the neutral category. Some 5-star reviews may also not reflect high loyalty, for example because there were criticisms despite the 5-star score. Separation based on sentiment, especially with Lexicon Vader analysis, sometimes does not accurately capture the emotional nuances in the neutral category.

To improve the performance of the model, it is necessary to adjust the definition of neutral categories, improve sentiment analysis with more sophisticated models, and add review data for neutral categories for better representation.

6. Modeling with Random Forest Algorithm

Modeling with the Random Forest algorithm is carried out after the preprocessing and labeling process of customer review data is complete. This study uses Random Forest due to its robustness and flexibility in handling complex datasets with multiple features, such as customer reviews that include text, scores, and sentiment. Random Forest is known for its ability to mitigate overfitting by combining predictions from multiple decision trees, thereby producing more accurate and stable results (Fitri & Damayanti, 2024). Additionally, its effectiveness in identifying important variables, such as review sentiment and score, makes it well-suited for predicting customer loyalty (Nurohanisah et al. 2024).

In this study, the data was divided into several scenarios of training and testing data proportions, namely 70%-30%, 80%-20%, and 90%-10%, to evaluate the effect of data distribution on model accuracy (Azmi et al. 2023). Each scenario maintains a balance between loyal, neutral, and disloyal classes to ensure representative and accurate training results. The model building process involves generating multiple decision trees based on random subsets of the training data and randomly selected features, such as review scores, sentiment, and review length. Each tree independently predicts customer loyalty, and their results are aggregated to produce the final prediction.

This research emphasizes the importance of effectively distinguishing between loyal, neutral, and disloyal customers, with particular attention to improving the identification of neutral customers—a known challenge in customer loyalty prediction. By leveraging Random Forest's strength in handling imbalanced data and complex relationships, this study aims to provide reliable predictions that support data-driven decision-making in customer retention strategies.

7. Cofusion Matrix as Model Evaluation

Model evaluation is performed using several metrics such as accuracy, precision, recall, and F1-Score, as well as confusion matrix to describe the performance of customer loyalty prediction.

a. Accuracy

Accuracy measures the proportion of correct predictions among all predictions made by the model. Calculated by the formula:

$$Accurracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where TP (True Positive) is the number of correct predictions for loyal customers, TN (True Negative) is the number of correct predictions for non-loyal customers, FP (False Positive) is the number of incorrect predictions for loyal customers, and FN (False Negative) is the number of incorrect predictions for non-loyal customers (Wardani et al., 2022).

b. Precision

Precision measures how many of the loyal predictions are actually loyal. This is important to understand how many of the positive predictions are relevant. Calculated with the formula:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

High precision indicates that the model can be trusted in identifying loyal customers.

c. Recall

Recall measures how many of the truly loyal customers are identified by the model. This is important to reduce the risk of losing loyal customers. Calculated with the formula:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

A high recall indicates that the model is able to capture most of the loyal customers.

d. F1-Score

F1-Score is a combination of precision and recall, giving an overview of the balance between the two. Calculated with the formula:

$$F1\ Score = 2x\ \frac{Recall\ X\ Precission}{Recall+Precission} \quad (4)$$

F1-Score is particularly useful when there is an imbalance between positive and negative classes in the dataset.

Results The results of the model evaluation will be presented in the form of a confusion matrix, which is an important tool for model performance analysis. Confusion matrix depicts the number of correct and incorrect predictions in a clear format (Anggarda et al. 2023). The components of the confusion matrix include:

1) True Positive (TP): The number of correct predictions for loyal customers.
2) True Negative (TN): The number of correct predictions for non-loyal customers.
3) False Positive (FP): The number of false predictions for loyal customers (i.e., the model identifies non-loyal customers as loyal).
4) False Negative (FN): The number of false predictions for non-loyal customers (i.e., the model identifies loyal customers as non-loyal).

8. Interpretation and Conclusion

After model evaluation, the accuracy shows that the model works well overall, but it is important to pay attention to the balance between the prediction of loyal and non-loyal customers. High precision indicates that the model is effective in recognizing loyal customers, while high recall indicates the model's ability to capture most loyal customers. F1-Score illustrates the balance between precision and recall. From the confusion matrix, the model can be identified more clearly, for example if there are many non-loyal customers categorized as loyal (False Positive) or vice versa (False Negative), which could have an impact on Shopee's business strategy.

RESULTS AND DISCUSSION

1. Data Collection

The data used in this study was collected

through a web scraping process of Shopee app reviews on the Google Play Store. A total of 10,000 reviews were collected in the period August to October 2024.

Table 2. Initial results of scraping shopee customer review data that has been done data cleaning and

userName	score	at	content
Putri Mela	5	2024-08-20 09:49:47	I really like shopping for anything at Shopee,...
night fury	1	2024-08-28 16:54:23	This application advertisement is very annoyin..
Ray ronald Mashudi	1	2024-08-08 12:03:31	The expedition is being repaired, but it has n...
Soekamto Antok	5	2024-09-12 06:00:04	Okay, both are profitable, cool, thank you Sho...
Wana Waruwu	5	2024-09-14 21:12:38	want to return now it's easier refund is also ...

Table 2 shows the initial results of the Shopee customer review data scraping process on the Google Play Store. And has gone through the data cleaning and translating process.

2. Preprocessing

After data collection, a preprocessing stage is performed which consists of several important steps. First, case folding is performed to convert the entire text into lowercase letters.

Table 3. First Stage Preprocessing Results: Case Folding

userName	score	at	content
Putri Mela	5	2024-08-20 09:49:47	i really like shopping for anything at shopee,...
night fury	1	2024-08-28 16:54:23	this application advertisement is very annoyin.....
Ray ronald Mashudi	1	2024-08-08 12:03:31	the expedition is being repaired, but it has n...
Soekamto Antok	5	2024-09-12 06:00:04	okay, both are profitable, cool, thank you sho...
Wana Waruwu	5	2024-09-14 21:12:38	want to return now it's easier refund is also ...

Table 3 shows the results of the first preprocessing stage, case folding. The case folding process converts all text in the content column into lowercase letters. This is done to ensure consistency in data analysis, so that variations in the use of upper or lower case letters do not affect the analysis results.

Next, tokenization is performed to break the review text into individual words.

Table 4. Tokenization Preprocessing Result

content	content_tokens
i really like shopping for anything at shopee,...	[i, really, like, shopping, for, anything, at,...
this application advertisement is very annoyin...	[this, application, advertisement, is, very, a...
the expedition is being repaired, but it has n...	[the, expedition, is, being, repaired, ,, but,...
okay, both are profitable, cool, thank you sho...	[okay, ,, both, are, profitable, ,, cool, ,, t...
want to return now it's easier refund is also ...	[want, to, return, now, it, 's, easier, refund...

Table 4 shows the results of the next preprocessing process, namely tokenization. Tokenization is the process of breaking down the review text in the content column into word units or tokens. This process aims to enable individual word analysis, which is very useful in studying text patterns.

With tokenization, each word in the review can be analyzed separately to detect certain key words or patterns relevant to customer loyalty. This stage forms an important basis for subsequent preprocessing steps, such as stopword removal.

Table 5. Stopword Removal Preprocessing Results (Filtering)

content_tokens	content_tokens_filtered
[i, really, like, shopping, for, anything, at,...	[really, like, shopping, anything, shopee, ,, ...

content_tokens	content_tokens_filtered
[this, application, advertisement, is, very, a...	[application, advertisement, annoying, ,, ever...
[the, expedition, is, being, repaired, ,, but,	[expedition, repaired, ,, delivered, yet, ,, p...
[okay, ,, both, are, profitable, ,, cool, ,, t...	[okay, ,, profitable, ,, cool, ,, thank, shope...
[want, to, return, now, it, 's, easier, refund...	[want, return, 's, easier, refund, also, faste...

Table 5 shows the results of the next stage of preprocessing, which is stopword removal or removal of common words that are not significant for analysis. This process ensures that only important words are used for further analysis, such as in customer loyalty determination or text pattern identification. This stage also helps prepare the data for the next step, which is stemming.

Table 6. Stemming Preprocessing Result

content_tokens_filtered	content_tokens_stemmed
[really, like, shopping, anything, shopee, ,, ...	[really, like, shopping, anything, shopee, , n...
[application, advertisement, annoying, ,, ever...	[application, advertisement, annoying, , every...
[expedition, repaired, ,, delivered, yet, ,, p...	[expedition, repaired, , delivered, yet, , pro...
[okay, ,, profitable, ,, cool, ,, thank, shope...	[okay, , profitable, , cool, , thank, shopee, ...
[want, return, 's, easier, refund, also, faste...	[want, return, s, easier, refund, also, faster...

Table 6 shows the results of the last stage of preprocessing, stemming. Stemming is the process of converting words into their basic or root form. This process is done to simplify text analysis by eliminating variations of words that have the same meaning. This stemming process is important to ensure that word variations, such as “shopping” and “shop”, are considered as the same word in the analysis. This stage results in simpler and more relevant text data, allowing the model to understand consistent patterns in customer reviews.

3. Labelling Loyalitas

After data preprocessing, the next step is to perform sentiment analysis and loyalty classification based on customer review scores in the Google Play Store. Sentiment analysis was conducted using the Lexicon Vader method to identify three sentiment categories: positive, negative and neutral. Positive sentiments indicate a high level of satisfaction, while negative sentiments indicate dissatisfaction with the product or service. The following are the results of sentiment analysis and loyalty classification based on customer review scores on the Google Play Store can be seen in Figure 4.

Furthermore, the loyalty classification is done based on the review score, where a score of 4 or 5 is considered to indicate customer loyalty, while scores of 1 and 2 are considered disloyal customers and a score of 3 is considered neutral. A high score reflects a customer's desire to shop again on Shopee, whereas a low score may indicate dissatisfaction that may lead to a move to another platform. Figure 5 below shows the results of the data analysis using score.

Figure 4. Results of Data Analysis Using Lexicon Vader

Figure 5. Results of Data Analysis Using Score

4. Modeling and Model evaluation

The customer loyalty prediction model was evaluated using three data sharing scenarios (70:30, 80:20, and 90:10). Table 6 below shows a comparison of the model evaluation metrics, including precision, recall, and F1-score for each category (Loyal, Neutral, Not Loyal), as well as overall accuracy.

Table 7. Comparison of Model Performance in Three Testing Scenarios

70%-30% data scenario				
Category	precision	recall	f1-score	accuracy
Loyal	0.98	0.98	0.98	97%
Neutral	0.92	0.85	0.88	
Not Loyal	0.99	1.00	0.99	
80%-20% data scenario				
Category	precision	recall	f1-score	accuracy
Loyal	0.96	0.98	0.97	97%
Neutral	0.92	0.77	0.84	
Not Loyal	0.99	1.00	0.99	
90%-10% data scenario				
Category	precision	recall	f1-score	accuracy
Loyal	0.96	0.98	0.97	97%
Neutral	0.92	0.80	0.86	
Not Loyal	0.99	1.00	0.99	

The comparison of the prediction model evaluation results in Table 7 above shows variations in accuracy and other evaluation metrics in each scenario. In the first scenario with 30% test data and 70% training data, the model achieved 97% accuracy. In the “Loyal” category, the model recorded a precision of 0.98, recall of 0.98, and F1-score of 0.98. However, the model's performance in identifying the “Neutral” category is still lower than the other categories, with a

precision of 0.92, recall of 0.85, and F1-score of 0.88. In the "Not Loyal" category, the model showed optimal performance with precision 0.99, recall 1.00, and F1-score 0.99.

In the second scenario with 20% test data and 80% training data, accuracy remained stable at 97%. The model performance for the "Loyal" category decreased slightly, with a precision of 0.96, recall of 0.98, and F1-score of 0.97. Meanwhile, the "Neutral" category saw a decrease in recall to 0.77, and F1-score to 0.84, indicating challenges in recognizing neutral customers in this scenario. The "Disloyal" category still showed the best performance with precision 0.99, recall 1.00, and F1-score 0.99.

In the third scenario, with 10% test data and 90% training data, the accuracy remained at 97%. The model performance for the "Loyal" category was consistent with a precision of 0.96, recall of 0.98, and F1-score of 0.97. Meanwhile, the "Neutral" category showed improvement, with recall of 0.80, and F1-score increasing to 0.86. In the "Disloyal" category, performance remained optimal with a precision of 0.99, recall of 1.00, and F1-score of 0.99, confirming that the model has an excellent ability to recognize disloyal customers.

Overall, the prediction model showed stable performance in each scenario with 97% accuracy. However, the best performance in recognizing the "Neutral" category occurred in the first scenario with a precision of 0.92, recall of 0.85, and F1-score of 0.88. These results indicate that the 70%-30% scenario provides a good balance between the size of the training and test data, resulting in more accurate and stable predictions.

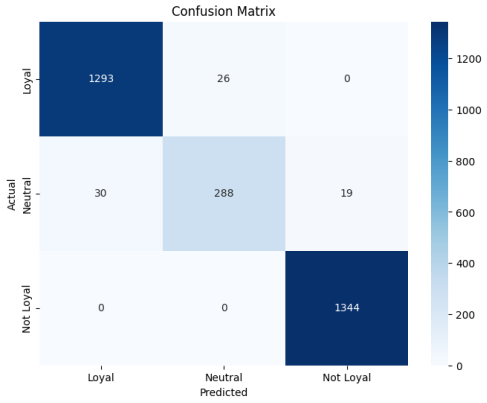

Figure 6. Confusion Matrix on Test Data for the 70%-30% Scenario

Figure 6 depicts the confusion matrix on the test data for the 70%-30% scenario which illustrates the performance of the classification model in predicting three categories: "Loyal," "Neutral," and "Not Loyal." The vertical axis represents the actual categories, while the horizontal axis shows the predicted categories. For the Loyal category, the model recorded 1,293 true positives, correctly identifying Loyal data, with 30 false negatives (Loyal data misclassified as Neutral) and 26 false positives (data incorrectly predicted as Loyal). In the Neutral category, the model achieved 288 true positives but struggled with 30 false positives (Loyal data misclassified as Neutral) and 19 false negatives (Neutral data misclassified as Not Loyal). For the Not Loyal category, the model performed exceptionally well with 1,344 true positives and no misclassifications. Overall, the model demonstrated strong accuracy in predicting the Loyal and Not Loyal categories but showed some difficulty in distinguishing Neutral data, indicating the need for further refinement to enhance performance in this category.

5. Achievement of Research Objectives

This research successfully developed a customer loyalty prediction model for the Shopee e-commerce platform using an optimized Random Forest algorithm. The model analyzes review data from the Google Play Store and classifies customers into three categories: loyal, neutral, and disloyal. The results show that the developed model consistently achieves 97% accuracy across different data splitting scenarios. These findings demonstrate the effectiveness of the Random Forest algorithm in processing complex and unstructured review data, providing accurate predictions regarding customer loyalty trends.

In addition, this study successfully identified several key factors influencing customer loyalty on e-commerce platforms: product quality, product price, delivery speed, and customer service. Product quality plays a crucial role, as positive reviews about high-quality products significantly boost loyalty. Similarly, competitive pricing attracts customers who seek the best deals. Fast delivery enhances customer satisfaction, whereas delays often lead to dissatisfaction and customer churn. Lastly, responsive customer service strengthens loyalty, as customers who receive prompt and effective solutions to their complaints are more likely to stay loyal.

Customer reviews further support these findings. For example, Pus Iyah mentioned, "The goods are complete, cheap, and come with free shipping." Similarly, Aska Mutiara stated, "The delivery is fast, and the goods match the order." Meanwhile, Wana Waruwu highlighted the importance of customer service, saying, "Returning items is now easier, and refunds are also faster. Overall, it's good for now." These testimonials confirm that quality, price, delivery, and service are essential factors in driving customer loyalty.

By achieving these objectives, this research is expected to make a meaningful contribution to data-driven customer retention strategies in e-commerce. The findings can help Shopee maintain customer loyalty amid fierce market competition by focusing on these critical factors.

6. Loyalty Prediction Model Success Factors

Several important factors support the success of the loyalty prediction model developed in this study. One of the main factors is the selection of the right algorithm. The Random Forest algorithm was chosen due to its ability to handle complex and diverse text data. The advantage of Random Forest lies in its ability

to reduce the risk of overfitting through combining predictions from various decision trees, which results in more stable and accurate predictions. The reliability of this algorithm in handling unstructured review data is one of the reasons why this model can achieve quite good results.

In addition, effective data preprocessing also plays an important role in the success of this model. Preprocessing steps, such as case folding, tokenization, stopword removal, and stemming, successfully clean and simplify complex review texts, so that the data used by the model is of better quality. This process ensures that the model can process the data more efficiently, which in turn improves the resulting prediction results. The evaluation and validation of the model also showed satisfactory results. The use of confusion matrix in the evaluation shows the good performance of the model, especially in recognizing loyal and disloyal customers. For loyal customers, the model showed a precision of 0.98, recall of 0.98, and F1-score 0.98, while for the disloyal category, the model achieved a precision of 0.99, recall of 1.00 and F1-score 0.99. This shows that the model is effective in predicting customer loyalty trends, with fairly stable and accurate results.

7. Strategic Recommendations Based on Prediction Results

This research recommends strategies for Shopee to increase customer loyalty based on the Random Forest model which achieves 97% accuracy, higher than the research of Ferdyandi et al. (2022) (78%) and Mustafa et al. (2024) (73.36%). These results confirm the effectiveness of Random Forest in predicting customer behavior.

Fast customer service is proven to have a major effect on loyalty, as found in the studies of Masripah & Wulandari (2024) and Nafisyah & Sulistiyowati (2024). In addition, delivery speed and product quality are also major factors, according to the findings of Muktafin et al. (2020) and Tambunan et al. (2023). Shopee can utilize customer review analysis to improve on these aspects to increase customer satisfaction and retention.

Real-time monitoring of customer reviews is also crucial. Larasati et al. (2022) and Fathoni et al. (2024) proved that machine learning-based sentiment analysis can increase customer satisfaction with faster responses to complaints. Therefore, Shopee can implement automated monitoring to respond to negative reviews and strengthen customer loyalty. Compared to previous research, this model is superior in accuracy and effectiveness, so it can be the basis for Shopee's customer experience improvement strategy.

CONCLUSION

This research successfully developed a Shopee customer loyalty prediction model using Random Forest with 97% accuracy, showing high performance in identifying loyal and disloyal customers, but still faces difficulties in classifying neutral customers, which is reflected in the lower Recall and F1-score of the neutral category. The main limitations of this study are the imbalance of data as well as the limitations of Vader's Lexicon-based sentiment analysis, which may not adequately capture the nuances of emotions in customer reviews. In addition, this study only uses review data from the Google Play Store, without considering direct customer behavior data such as frequency of purchase or interaction with customer service. Therefore, future research is recommended to use more sophisticated approaches such as transformer-based models (BERT) or deep learning, as well as incorporating transaction data and customer interactions to improve classification accuracy, especially in the neutral category. With these findings, this study contributes to the development of data-driven strategies in e-commerce, helping Shopee increase customer loyalty, reduce churn, and improve user experience.

REFERENCES

Adi Ahdiat. (2024, January 15). Retrieved January 8, 2025, from Katadata.co.id website: https://databoks.katadata.co.id/infografik/2024/01/15/5-e-commerce-dengan-pengunjung-terbanyak-sepanjang-2023

Alkhairi, P., Windarto, A. P., & Efendi, M. M. (2024). Optimasi LSTM Mengurangi Overfitting untuk Klasifikasi Teks Menggunakan Kumpulan Data Ulasan Film Kaggle IMDB. *Technology and Science (BITS)*, *6*(2). https://doi.org/10.47065/bits.v6i2.5850

Anggarda, M. F., Kustiawan, I., Nurjanah, D. R., & Hakim, N. F. A. (2023). Pengembangan Sistem Prediksi Waktu Penyiraman Optimal pada Perkebunan: Pendekatan Machine Learning untuk Peningkatan Produktivitas Pertanian. *Jurnal Budidaya Pertanian*, *19*(2), 124–136. https://doi.org/10.30598/jbdp.2023.19.2.124

Apriliansyah, R. D. R., Astuti, R., Prihartono, W., & Hamonangan, R. (2025). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Pengunjung Di Pantai Kejawanan. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(1). https://doi.org/10.23960/jitet.v13i1.5774

Asri, Y., Suliyanti, W. N., Kuswardani, D., & Fajri, M. (2022). Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile. *PETIR*, *15*(2), 264–275. https://doi.org/10.33322/petir.v15i2.1733

Azmi, B. N., Hermawan, A., & Avianto, D. (2023). Analisis Pengaruh komposisi data training dan data testing Pada penggunaan PCA Dan Algoritma decision tree untuk KLASIFIKASI Penderita Penyakit liver. *JTIM: Jurnal Teknologi Informasi dan Multimedia*, *4*(4), 281-290. https://doi.org/10.35746/jtim.v4i4.298

Fathoni, M. F. N., Puspaningrum, E. Y., & Sihananto, A. N. (2024). Perbandingan Performa Labeling

Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM. *Jurnal Informatika Dan Sains Teknologi*, *1*(3), 62–76. https://doi.org/10.62951/modem.v1i3.112

Ferdyandi, M., Setiawan, N. Y., & Bachtiar, F. A. (2022). Prediksi Potensi Penjualan Makanan Beku berdasarkan Ulasan Pengguna Shopee menggunakan Metode Decision Tree Algoritma C4.5 dan Random Forest (Studi Kasus Dapur Lilis) *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *6*(2), 588-596. Retrieved from https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10560

Fitri, D. A., & Damayanti. (2024). Komparasi Algoritma Random Forest Classifier Dan Support Vector Machine Untuk Sentimen Masyarakat Terhadap Pinjaman Online Di Media Sosial. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *9*(4), 2018–2029. https://doi.org/10.29100/jipi.v9i4.5608

Larasati, F. A., Ratnawati, D. E., & Hanggara, B. T. (2022). Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *6(9), 4305-4313.* Retrieved from https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/11562

Masripah, S., & Wulandari, D. A. N. (2024). Analisa Online Customer Review (OCR) Menggunakan Algoritma Naive Bayes berbasis Particial Swarm Optimization (PSO). In *Jurnal* (Vol. 6). Retrieved from http://ejournal.bsi.ac.id/ejurnal/index.php/infortech59

Muktafin, E. H., Kusrini, & Luthfi, E. T. (2020). Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing. *Jurnal Eksplora Informatika*, *10*(1), 32–42. https://doi.org/10.30864/eksplora.v10i1.390

Mustafa, W. F., Hidayat, S., & Fudholi, D. H. (2024). Prediksi Retensi Pengguna Baru Shopee Menggunakan Machine Learning. *JURNAL Media Informatika Budidarma*, *8(1), 612-623.* https://doi.org/10.30865/mib.v8i1.7074

Nafisyah, S., & Sulistiyowati, R. (2024). Analisis Sentimen Ulasan Produk Toko Online Esrocte untuk Peningkatan Pelayanan Menggunakan Algoritma Naïve Bayer *Blantika: Multidisciplinary Journal*, *2*(8). https://doi.org/10.57096/blantika.v2i8.189

Nurohanisah, S., Astuti, R., & Basysyar, F. M. (2024). Deteksi Berita Palsu Menggunakan Algoritma Random Forest. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 422-428. https://doi.org/10.36040/jati.v8i1.8418

Oktavia, V. D., Sarsono, & Marwati, F. S. (2022). Loyalitas pelanggan ditinjau dari pelayanan, kepuasan dan kepercayaan pada CV cipta kimia sukoharjo. *Jurnal ilmiah edunomika*, *6(1), 540.* https://doi.org/10.29040/jie.v6i1.4656

Permana, N. A., & Bunyamin, H. (2024). Perbandingan Logistic Regression dengan Random Forest dalam Memprediksi Sentimen Pada IMDb Moview Review. *Jurnal STRATEGI-Jurnal Maranatha, 6(2), 391-399.* Retrieved from https://www.strategi.it.maranatha.edu/index.php/strategi/article/view/538

Rahmadani, R., Rahim, A., & Rudiman. (2024). Analisis Sentimen Ulasan "Ojol The Game" Di Google Play Store Menggunakan Algoritma Naive Bayes dan Model Ekstraksi Fitur Tf-Idf Untuk Meningkatkan Kualitas Game. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3). https://doi.org/10.23960/jitet.v12i3.4988

Sulistyawati, U. S., & Munawir. (2024). Membangun Keunggulan Kompetitif melalui Platform E-Commerce: Studi Kasus Tokopedia. *Jurnal Manajemen Dan Teknologi (JMT)*, *1*(1). https://doi.org/10.35870/jmt.vxix.776

Suryawan, M. A., Israwan, L. M. F., & Arland, F. (2024). Penerapan Algoritma Stemming Nazief-Adriani dengan Metode Cosine Similarity Dalam Aplikasi Ujian Esai. *Prosiding SISFOTEK*, *8*(1), 237-243. Retrieved from https://seminar.iaii.or.id/index.php/SISFOTEK/article/view/495

Tambunan, S. F. A., Charos, W. A., & Nurbaiti. (2023). Analisis Perbandingan Sebelum Dan Sedudah Menggunakan Teknologi Informasi Dalam Bidang E-Commerce. *Jurnal Akuntansi Keuangan Dan Bisnis*, *1*(3), 2023. https://doi.org/10.47233/jakbs.v1i3

Viona, V., Yohanes, K., Mega, L. S., Kurniawati, W., Farady Marta, R., & Isnaini, D. M. (2021). Narasi Shopee Dalam Mengembangkan Ekonomi Kreatif Berbasis Teknologi E-Commerce Di Era Moderen. *AGUNA: Jurnal Ilmu Komunikasi*, *2*(1), 46-65. Retrieved from http://ejournal.amikompurwokerto.ac.id/index.php/AGUNA

Wardani, N. W., Arnidya, D. J., Putra, I. N. A. S., Desmayani, N. M. M. R., Nugraha, P. G. S. C., Hartono, E., & Mahendra, G. S. (2022). Prediksi Pelanggan Loyal Menggunakan Metode Naïve Bayes Berdasarkan Segmentasi Pelanggan dengan Pemodelan RFM. *Jurnal Manajemen dan Teknologi Informasi*, *12*(2), 113-124. Retrieved from Online website: https://ojs.mahadewa.ac.id/index.php/jmti

Yolanda, R., Hardilawati, W. L., & Hinggo, H. T. (2021). Pengaruh Perceived Quality, Customer Relationship Marketing Dan Store Atmosphere Terhadap Loyalitas Konsumen. *ECOUNTBIS: Economics, Accounting and Business Journal*, *1*(1), 146-156. Retrieved from https://jom.umri.ac.id/index.php/ecountbis/article/view/224