Sentiment Analysis on the Shopee Application on Playstore Using the Random Forest Classification Method

Muhammad Rusdi Rahman[1], Ahmad Febri Diansyah[2], Hanafi[3*]

[1,2,3]*Informatics Department, Universitas Amikom Yogyakarta, Indonesia*

[1] muhrusdi17@students.amikom.ac.id, [2]ahmadfebri@students.amikom.ac.id

[3]hanafi@amikom.ac.id(*)

Received: 2022-11-30; Accepted: 2023-09-03; Published: 2023-11-29

Abstract— In analyzing customer or consumer satisfaction with company services, it is essential for companies to find service deficiencies and to know user expectations for the company. This study aims to build sentiment analysis on the Shoppe application on the Google Playstore. The method used includes TF-IDF as text vectorization, Random Forest as a classification model, and Evaluation Matrix as an evaluation model, providing accuracy, precision, recall, and F1-Score. Based on the results of this study, the model we used achieved an accuracy rate of 94%, a precision of 91%, a recall of 91%, and an F1-Score of 93%. The limitations of this study are identifying words in English and regional languages because the corpus module we use is literature, a special Indonesian corpus. In future research, we will try to build a new engine/algorithm and try to add datasets in the hope that the level of accuracy will be even better.

Keywords— Sentiment Analysis; Shopee; TF-IDF; Random Forest; Play Store.

I. Introduction

E-commerce, or online marketplace shopping, is accelerating [1]. The convenience and comfort of using the market or e-commerce and the many benefits of purchasing products without physically going to the store and doing it whenever and wherever you want are some of the factors driving its growth [2]. Another benefit of using the marketplace is the availability of numerous discounts, promotions, and even free shipping to the most well-liked retailers in the neighborhood [3]. According to data from the E-Commerce Map of Indonesia published by iPrice, Shopee was the most visited marketplace in Indonesia in Q1 2022, with 8 million monthly visitors [4]. If Shopee can provide services that meet consumer expectations, then Shopee will give a good impression in the eyes of consumers [5]. Google Play provides features that allow consumers to rate the products they buy. The results of consumer reviews, which we call sentiment analysis, can provide accurate results that make it easier for new users to make decisions [6].

Sentiment analysis evaluates people's attitudes, emotions, and opinions based on text representations combined with text mining and natural language processing (NLP) [7]. Sentiment analysis is used in politics (to predict election results in political forums), economics (to analyze online sentiment on social media for stock market predictions), and marketing (to predict sales of certain products) and is used in many fields [8]. Reviews of mobile applications may contain writing faults that make them difficult to understand. Numerous things contribute to this, including the proximity of the keyboard's letters, carelessness when typing, and failure to double-check [9]. This indicates that the term needs to be rethought to completely comprehend the function of user evaluations. After some time has passed, the word will be classified to ascertain the meaning of its meaning. A technique that is based on classification is required to investigate the reviews in this scenario.

The Random Forest method was utilized in this investigation because the research conducted utilizing this algorithm was not yet as effective as the research carried out [10][11]. At this stage, the data set will be divided into positive and negative classes, using existing patterns or documents taught to be recognized by the machine. Training data and test data will be generated using data that has successfully undergone text preparation. The categorization findings were 1668 for the positive and 1269 for the negative classes. The distribution of 80% training data and 20% test data results in accuracy, precision, memory, and f1 scores of 1.00%, 1.00%, 1.00%, and 1.00%, respectively.

This research on the Shopee application aims to classify user reviews into positive or negative sentiment categories to understand how users tend to feel satisfied or disappointed. If the user's expectations are not in accordance, they can be identified based on the user's negative comments so that a better evaluation of changes can be carried out. It helps companies maintain and improve service quality and user satisfaction. On the sentiment used, how optimal is the TF-IDF method for text in its use, the modeling used by Random Forest in classifying Shopee shop application reviews, the evaluation matrix results answer with 94% accuracy, 91% precision, and an F1 score of 93%. The problems were found in the limitations of the research, namely the identification of English words and regional languages because the corpus module we used was the Indonesian language corpus of literature.

II. Literatur Review

DOI : https://doi.org/10.25139/inform.v9i1.5465

Theories in previous research are used as a basic reference to limit the application of relevant theories. This research [5] aims to help Shopee manage the positive or negative opinions of application users and provide empirical evidence for related theories so that they can be used as a thought contribution to developing further theories. The results obtained using the Naive Bayes algorithm are 96.67% accuracy [5].

The second research is to identify opinions, ideas, or thoughts from online media. The results obtained using the SVM algorithm are 80.90% accuracy, and it can also be concluded from the use of the SVM algorithm in 2022 from January to March, a rating of 50% of users liked the app, and 5% of users didn't like the Shopee app [12].

The third study [3] processes user comments using the Support Vector Machine (SVM) algorithm and Decision Tree to see the algorithm's accuracy and the positive and negative reviews. This study uses the SVM algorithm to get an accuracy value of 82.19%. It can be concluded that if we compare the decision tree method and the support vector machine (SVM) method based on the accuracy of this study, the support vector machine (SVM) method has a higher prediction accuracy for sentiment analysis.

Using data from reviews on Google Playstore, the study [10] gauges how satisfied Shopee account holders are. This investigation obtained 89.0%, 89.4%, and 83.0% accuracy using the K-NN, Support Vector Machine, and Random Forest algorithms. The support vector machine (SVM) approach has a greater prediction accuracy for sentiment analysis if we compare the K-NN, Support Vector Machine, and Random Forest methods based on the accuracy of this study [10].

In the study [11], the goal was to learn about user feedback, which could be accomplished by examining the sentiments of each app review. In this investigation, the accuracy values for the Decision Tree and Random Forest algorithms were 54.38% and 60.08%, respectively. Based on the accuracy of this research, it can be said that when comparing the Decision Tree and Random Forest approaches, the Random Forest method has a higher prediction accuracy for sentiment analysis [11].

III. Research Methodology

This section will explain the research flow as in Figure 1. The first stage is entering the dataset that will be used. The dataset is processed first with the Cleaning, Case Folding, and Punctuation methods. After that, it enters into resistance vectorization using TF-IDF to proceed to the classification stage using Random Forest. Classification results will be tested using the Evaluation Matrix, and the results will be classification accuracy.

Sentiment analysis itself is textual context mining. The process is to identify and extract data from subjective textual information from the source material to express emotions, opinions, judgments, and attitudes and help you understand people's emotions. Emotion. This is done to understand the content of emotional information contained in an opinion statement about a person's dilemma or object and whether that person tends to have negative, positive, or neutral opinions [13].

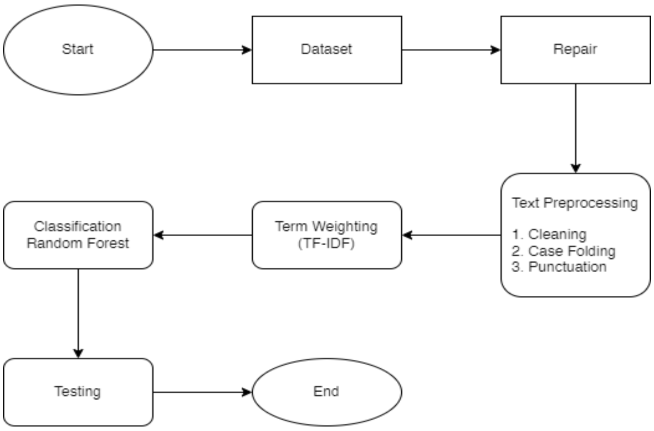

Figure 1. Research Flow

A. Dataset

The dataset comes from Kaggle under the title 'shopee reviews'. This dataset contains features including users, reviews, labels, dates, case folding, punctuation, and deleted emojis (clean). The dataset used is 2937 data in Table I, with 1668 data labeled positive and 1269 data labeled negative in Table II.

TABLE I
DATASET DESCRIPTION

Dataset	Number of Attributes	Amount of data
Shopee Review	7	2937

TABLE II
DATASET ATTRIBUTES

Attribute	Value
User	User
Review	Text
Sentiment	Positive and Negative
Date	Date
Case_folding	Changing the text to small or lowercase and the word separation process
Punctuation	As an aid to understanding and reading correctly, either silently or aloud, in handwriting and print.
Remove Emoji (Clean)	Cleaning data from unrelated attributes, such as hashtag numbers, website addresses, emails, usernames, and emoticons.
User	Application username

B. Random Forest

Random Forest is a classification algorithm. Random Forest builds several decision trees and combines them to get more sTable and accurate predictions [14]. A "forest" constructed by a Random Forest is a collection of decision trees typically trained using the bagging method [15]. The idea of a Random Forest is a classification algorithm. Random Forest builds several decision trees and combines them to get more sTable and accurate predictions. A "forest" constructed by a Random Forest is a collection of decision trees typically trained using the bagging method. The general idea of the bagging method is to combine learning models to improve overall results [16].

DOI : https://doi.org/10.25139/inform.v9i1.5465

The Random forest algorithm increases the randomness of the model as the tree grows. Instead of looking for the most important trait when splitting nodes, Random Forest looks for the best trait from a subset of random traits. As a result, these methods produce wide variations and generally lead to better models [17]. How the random forest algorithm works can be explained in the following steps:

- The algorithm selects a random sample from the provided data set.
- Create a decision tree for each selected sample. Then, get the predicted results from each decision tree obtained.
- The voting process is running for each prediction result. Classification problems use mode (the most frequently occurring value), and regression problems use the mean (average value). Algorithm
- Choose the prediction result with the most votes as the final prediction.

C. *Term-Weigthing (TF-IDF)*

Term Weighting is a system for assigning weights to each term in the text because the text classification method requires a word weighting system to transfer data forms and then change the text into a more effective model. The analysis process can be carried out [18]. How to calculate the term frequency-inverse document frequency (TF-IDF) for each term [19] using Equations (4) and (5). The *N* variable is a total of all documents, the *DF* variable is a total document containing the word, the *tf* variable is a term frequency, and the IDF variable is an inverse document frequency.

$$IDF = log/NDF \quad (4)$$

$$TF\text{-}IDF = tf*id \quad (5)$$

D. *Evaluation Matrix*

The confusion matrix determines whether the classification model's performance consists of the number of rows of data tested for true and false [20]. The goal is to determine the results of a classification model that can be read from the performance measurement parameters accuracy, recognition value, and precision based on Table III [21].

TABLE III
ACCURACY MEASUREMENT

Measurement	Definition	Formula
Accuracy (A)	Accuracy determines the accuracy of the algorithm in predicting instances	*A=(TP+TN) / (Total number of samples)*
Precision (P)	Classifier, correctness/accuracy is measured by precision	*P = TP / (TP+ FP)*
Recall (R)	To measure classifier completeness or sensitivity, using recall	*R =TP / (TP+FN)*
F-Measure (F)	F-Measure is the average of precision and recall.	*F=2*(P*R) / (P+R)*

IV. RESULT AND DISCUSSION

It was cleaning up preprocessing data in the dataset used at this point. The dataset collected has been expanded to include the cleaning procedure, case folding, and punctuation. Because all the words in this column have been cleaned, it will be used as the *X* variable. The Sentiment column will serve as the *Y* variable. The Sentiment column will undergo a labeling change with a positive label equal to 1 and a negative label equal to 0.

Data that has made it through the preprocessing stage and is ready for classification must be in numerical form. Transforming the data into a numerical format can be accomplished using the TF-IDF weighting approach. Combining the TF-IDF value of a word with its IDF value is the method by which the weight of a word is established. The TF-IDF method combines two ideas: the frequency at which a word appears in a text and the inverse frequency at which documents contain that word.

The random forest approach proposed by Breiman is a machine-learning algorithm with multiple decision trees. Random forest is a combination of bagging techniques and random subspaces. This method has proven valuable in regression and classification problems in recent years and is one of the best machine-learning algorithms used in various fields. The performance was evaluated from a random random forest using the confusion matrix in Figure 2.

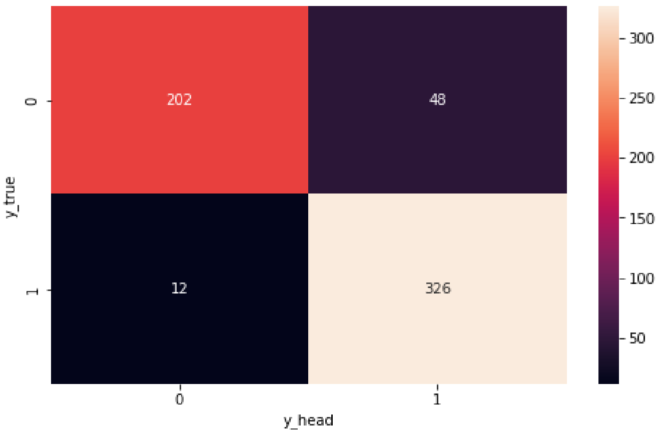

Figure 2. Confusion Matrix Random Forest

The random forest algorithm is used in this study. Accuracy, f value, recall, precision, and f value are used for classification in this study. Table II describes the dataset, and Table III provides the accuracy measurements.

TABLE IV
RANDOM FOREST ALGORITHM PERFORMANCE

Classification Algorithm	Precision	Recall	F-Measure	Accuracy
Random Forest	0.91	0.91	0.93	0.94

Table IV shows that the performance of the random forest classification algorithm has an accuracy rate of 94%. This represents a classification algorithm built to predict the level

DOI : https://doi.org/10.25139/inform.v9i1.5465

of usefulness of the Shopee application to the public. The closer to the value of one, the better the classification level that is built. A comparison of the two previous studies can be seen in Table V.

TABLE V
COMPARISON OF RESULTS WITH PREVIOUS STUDIES

Author	Algorithm	Accuracy
[10]	K Nearest Neighbor (K-NN) Random Forest Support Vector Machine (SVM)	89,0% 83,0% 89,4%
[11]	Random Forest	60.08%

Table V describes the accuracy of each research that has been done. In the [10] research, they used three algorithms to be compared. K-NN results are 89.0%, Random Forest 83.0%, and SVM 89.4%. This study has surpassed the results of the three algorithms. Likewise, research [11] using Random Forest only got 60.08% accuracy. The different strategies employed undoubtedly have an impact on the results. The general flow of previous studies is very similar to current research. At the preprocessing step, the discrepancies are noticeable. In research [10], a functionality that would eliminate all punctuation marks was not added. A language normalization tool was introduced in research [11] to convert slang, alay, or slang syllables back to regular language or normal forms, which we don't utilize the function for. Because of this, our data is more varied and has a higher potential for accuracy.

V. CONCLUSION

Taken from 2937 review data consisting of 1668 positive reviews and 1269 negative reviews from the testing results using the random forest algorithm using review data from the Shopee application on Playstore. The random forest algorithm produces 94% accuracy and is included in the very good classification. The random forest algorithm can solve the sentiment classification problem in the Shopee application. In future research, we will try to build a new engine/algorithm and try to add datasets in the hope that the level of accuracy will be even better.

ACKNOWLEDGMENT

We are grateful for the contribution of references from the journals Natural Language Processing, Machine Learning, and several other references that we used to complete this research. Thanks for the data support to Kaggle and Playstore. And especially thanks to Mr. M. Hanafi, S. Kom., M. Eng. Ph. D., who have provided funds for this research journal project so that it runs well and smoothly. We also do not forget to thank Amikom Yogyakarta University, which facilitated and provided meeting facilities, asking for suggestions and input from lecturers for us to complete this research. Hopefully, this research can be useful for us and all of us.

REFERENCES

[1] Hanafi, R. Widyawati, and A. S. Widowati, "Effect of service quality and online servicescape toward customer satisfaction and loyalty mediated by perceived value," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 704, no. 1, 2021, doi: 10.1088/1755-1315/704/1/012011.

[2] J. A. Josen Limbong, I. Sembiring, K. Dwi Hartomo, U. Kristen Satya Wacana, and P. Korespondensi, "Analisis Klasifikasi Sentimen Ulasan Pada E-Commerce Shopee Berbasis Word Cloud Dengan Metode Naive Bayes Dan K-Nearest Neighbor Analysis of Review Sentiment Classification on E-Commerce Shopee Word Cloud Based With Naïve Bayes and K-Nearest Neighbor Meth," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 2, pp. 347–356, 2019, doi: 10.25126/jtiik.202294960.

[3] Nurfaizah, T. Hariguna, and Y. I. Romadon, "The accuracy comparison of vector support machine and decision tree methods in sentiment analysis," *J. Phys. Conf. Ser.*, vol. 1367, no. 1, 2019, doi: 10.1088/1742-6596/1367/1/012025.

[4] Hanafi, N. Suryana, and A. S. Bashari, "Evaluation of e-Service Quality, Perceived Value on Customer Satisfaction and Customer Loyalty: a Study in Indonesia," *International Business Management*, vol. 11, no. 11. pp. 1892–1900, 2017, [Online]. Available: https://medwelljournals.com/abstract/?doi=ibm.2017.1892.1900.

[5] D. Pratmanto, R. Rousyati, F. F. Wati, A. E. Widodo, S. Suleman, and R. Wijianto, "App Review Sentiment Analysis Shopee Application in Google Play Store Using Naive Bayes Algorithm," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012043.

[6] J. Y. B. Yin, N. H. M. Saad, and Z. Yaacob, "Exploring Sentiment Analysis on E-Commerce Business: Lazada and Shopee," *TEM J.*, vol. 11, no. 4, pp. 1508–1519, 2022, doi: 10.18421/TEM114-11.

[7] D. Sánchez, L. Martínez-Sanahuja, and M. Batet, "Survey and evaluation of web search engine hit counts as research tools in computational linguistics," *Inf. Syst.*, vol. 73, pp. 50–60, 2018, doi: 10.1016/j.is.2017.12.007.

[8] S. Nanda, D. Mualfah, and D. A. Fitri, "Analisis Sentimen Kepuasan Pengguna Terhadap Layanan Streaming Mola Menggunakan Algoritma Random Forest," no. x, pp. 210–219, 2019.

[9] F. Gunawan, M. A. Fauzi, and P. P. Adikara, "Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan Naive Bayes dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi BCA Mobile)," *Syst. Inf. Syst. Informatics J.*, vol. 3, no. 2, pp. 1–6, 2017, doi: 10.29080/systemic.v3i2.234.

[10] S. Watmah, S. Suryanto, and M. Martias, "Komparasi Metode K-NN, Support Vector Machine Dan Random Forest Pada E-Commerce Shopee," *INSANtek*, vol. 2, no. 1, pp. 15–21, 2021, doi: 10.31294/instk.v2i1.419.

[11] S. W. Iriananda, R. P. Putra, and K. S. Nugroho, "Analisis Sentimen Dan Analisis Data Eksploratif Ulasan Aplikasi Marketplace Google Playstore," *4th Conf. Innov. Appl. Sci. Technol. (CIASTECH 2021)*, no. Ciastech, pp. 473–482, 2021.

[12] K. Hantoro, D. Handayani, and S. Setiawati, "A Implementation of Text Mining In Sentiment Analysis of Shopee Indonesia Using SVM," vol. 3, no. 2, pp. 115–120, 2022.

[13] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, "A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets," *Inf. Syst. Front.*, vol. 23, no. 6, pp. 1417–1429, 2021, doi: 10.1007/s10796-021-10135-7.

[14] N. Khasanah, R. Komarudin, N. Afni, Y. I. Maulana, and A. Salim, "Skin Cancer Classification Using Random Forest Algorithm," *Sisfotenika*, vol. 11, no. 2, p. 137, 2021, doi: 10.30700/jst.v11i2.1122.

[15] H. C. Morama, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogyakarta menggunakan Algoritma Random Forest Classifier," vol. 6, no. 4, pp. 1702–1708, 2022.

[16] W. Nugraha, "Prediksi Penyakit Jantung Cardiovascular Menggunakan Model Algoritma Klasifikasi," *J. Sigmata*, vol. 9, no. 2, pp. 78–84, 2021.

[17] U. Erdiansyah, A. Irmansyah Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 208, 2022, doi: 10.30865/mib.v6i1.3373.

[18] R. Kosasih and A. Alberto, "Sentiment analysis of game product on shopee using the TF-IDF method and naive bayes classifier," *Ilk. J. Ilm.*, vol. 13, no. 2, pp. 101–109, 2021, doi:

10.33096/ilkom.v13i2.721.101-109.

[19] S. Sudianto, A. D. Sripamuji, I. Ramadhanti, R. R. Amalia, J. Saputra, and B. Prihatnowo, “Penerapan Algoritma Support Vector Machine dan Multi-Layer Perceptron pada Klasisifikasi Topik Berita,” vol. 11, no. 2, pp. 84–91, 2022.

[20] D. Theckedath and R. R. Sedamkar, "Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks," *SN Comput. Sci.*, vol. 1, no. 2, pp. 1–7, 2020, doi: 10.1007/s42979-020-0114-9.

[21] N. Khoiruzzaman, R. D. Ramadhani, and A. Junaidi, "Hasil Klasifikasi Algoritma Backpropagation dan K-Nearest Neighbor pada Cardiovascular Disease," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 1, no. 1, pp. 17–27, 2021, doi: 10.20895/dinda.v1i1.141.