# E-Commerce Product Review Sentiment Analysis: A Comparative Study of Naïve Bayes Classifier and Random Forest Algorithms on Marketplace Platforms

Cian Ramadhona Hassolthine[1], Toto Haryanto[2],
Fenina Adline Twince Tobing[3], Muhammad Ikhwani Saputra[4]
[123]School of Data Science, Mathematics and Informatics, IPB University
hassolthinecian@apps.ipb.ac.id, totoharyanto@apps.ipb.ac.id,
feninatobing@apps.ipb.ac.id, ikhwanisaputra@apps.ipb.ac.id

Accepted 16 June 2025
Approved 14 July 2025

***Abstract*— Achieving customer satisfaction and trust is a major challenge for success in the business world. Entrepreneurs must identify problems that arise from reviews given by customers. However, reading and sorting each review is time-consuming and considered inefficient. In order to overcome this, a study was conducted that aims to analyze sentiment on products sold in the Shopee marketplace using the Naïve Bayes Classifier and Random Forest algorithms. The focus of this study is on product reviews from XYZ Store. The main objective of this study is to determine a more accurate and efficient algorithm in classifying review sentiment, which can help companies in marketing strategies and product development. The results of this study can provide insight for companies about consumer responses to marketed products, so that they can be used as a basis for making strategic decisions to improve the quality of services and products. The results of the Random Forest method classification produce superior predictions compared to the Naïve Bayes Classifier method with an accuracy value of 92.5%, precision of 93%, Recall of 92.5% and F1-Score of 90%.**

***Index Terms*— Marketplace, Naïve Bayes Classifier Algorithm; Product Review; Random Forest Algorithm; Sentiment Analysis**

## I. Introduction

Marketplaces have become the primary platform for consumers to search for and purchase products, making customer reviews an important source of information [1]. Sentiment analysis of these reviews helps understand customer perceptions, which is essential for improving service quality and making better business decisions. Various machine learning algorithms have been applied in sentiment analysis, including Naïve Bayes and Random Forest [2]. However, the effectiveness of these algorithms can vary depending on the characteristics of the data and different e-commerce platforms. Comparative research comparing the performance of these two algorithms on a particular marketplace platform is still needed to identify the best approach to understanding customer sentiment [3]. Sentiment analysis is a discipline of machine learning and Natural Language Processing (NLP) that functions to identify and extract opinions, sentiments, reviews, attitudes, and emotions contained in text, thus providing deeper insight into responses and views in a particular context [4]. The Naïve Bayes Classifier algorithm is a probability-based classification method that predicts an event based on historical data using Bayes' theorem [5]. Its main advantages are ease of implementation, computational speed, and effectiveness on high-dimensional datasets. The Random Forest algorithm was developed as an evolution of the CART (Classification and Regression Trees) method by utilizing the bagging technique (combining random samples) and random feature selection [6].

Research on sentiment analysis of e-commerce product reviews using the Naïve Bayes and Random Forest algorithms has been widely conducted, but the focus and context vary. In research [7] conducted sentiment analysis on Shopee e-commerce using the KNN and Support Vector Machine (SVM) algorithms. Another study by [8] analyzed sentiment reviews on Shopee using Naïve Bayes and SVM. The differences in context, dataset, and algorithms used indicate that there is still room for further research that focuses on comparing Naïve Bayes and Random Forest specifically on different marketplace platforms or with unique data characteristics. This study aims to conduct a comparative study between the Naïve Bayes Classifier and Random Forest algorithms in sentiment analysis of product reviews on e-commerce marketplace platforms. The results of this study are expected to provide guidance for e-commerce practitioners in choosing the most appropriate algorithm for sentiment analysis. In addition, this study also contributes to the literature by providing empirical evidence regarding the performance of both algorithms in the context of sentiment analysis of e-commerce product reviews.

## II. Methodology

The methodology that will be conducted in this research is as follows.

### A. Literature Study

Literature studies are carried out by taking and studying information from various literary sources such as journals, books or other scientific sources to support research according to existing theories.

### B. Research Flow Diagram

The sentiment analysis system is designed to process and analyze buyer reviews from the Shopee marketplace on XYZ Store. This study begins by collecting a dataset of product reviews to be analyzed.

Consumer review data was taken from a product named "*Meja Belajar Polos A*" sold by the XYZ store on the Shopee Indonesia marketplace. A total of 844 reviews available on the review page were taken from the selected product page on the Shopee marketplace. Figure 1 presents a flow diagram in this research.

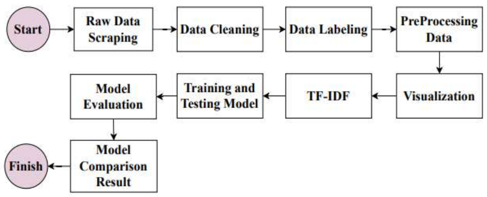
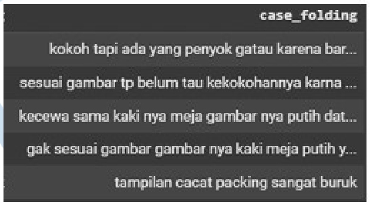


Fig 1. Research Flow Diagram

- Raw Data Scraping : The review data successfully collected from the product review page on the Shopee website amounted to 844 review data. This review data is called raw data which will then go through a data cleaning process.
- *Data Cleaning* : The results of raw data scraping are then continued to the data cleaning process. this stage involves data from csv format to xlsx format, Sort data based on "Transaction Time" by taking data from a time range, Cleaning noise in the data, namely "comment ID", "item ID", "Shop ID", "username", "User Name", "Anonymous", "Region", "Item Name", and "Transaction Time".
- *Data Labeling* : Data labeling on the dataset used is based on 2 categories, positive and negative sentiments. Positive sentiment if the review gets a rating of 4-5, while negative sentiment if the review gets a rating of 1-3. Labeling process is conducted manually and separately using tools of Microsoft Excel application. From the labeling results, a total of 351 positive sentiment records were obtained, while 42 negative sentiment records, a total of 393 records for the dataset used. So that there is an imbalance in the data, therefore the SMOTE (Synthetic Minority Over-sampling Technique) technique is used to understand the distribution of words in the data and identify dominant words. This happens because the dataset used does have a more dominant number of positive sentiments compared to the number of negative sentiments. The SMOTE oversampling method overcomes the problem of data imbalance by generating synthetic data for negative sentiments. Thus, the data distribution becomes more balanced, so that the machine learning model that will be applied in the next stage can learn more effectively and avoid bias that leads to positive sentiments.
- *Pre-processing Data* : The successful pre-processing process is carried out in several stages which can be explained as follows: Text Cleaning, Lowercase Folding, Tokenizing, Slang Word Conversion, Stopword Removing and Stemming [9].

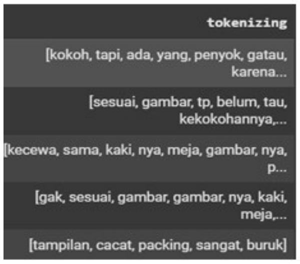
1. Text Cleaning : The process of cleaning noise and removing symbols in the review. This process involves a series of steps to clean and process raw text, remove noise, and format text to make it more structured and relevant. Text cleaning aims to improve data quality so that analysis or modeling performed on text data can produce more accurate and meaningful results [10].
2. Lowercase Folding : Standardize all letters written in the review into lower case letters to have a standard form [11]. Figure 2 present Lowercase Folding.



| case_folding |
|---|
| kokoh tapi ada yang penyok gatau karena bar... |
| sesuai gambar tp belum tau kekokohannya karna ... |
| kecewa sama kaki nya meja gambar nya putih dat... |
| gak sesuai gambar gambar nya kaki meja putih y... |
| tampilan cacat packing sangat buruk |

Fig 2. Lowercase Folding

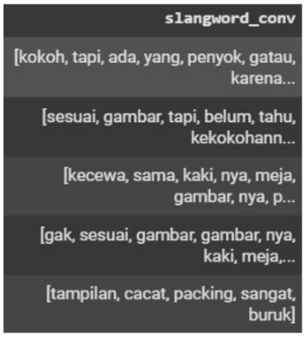
3. Tokenizing : Changing sentences into words that are in accordance with the rules of the Indonesian dictionary so that sentences are more meaningful and are converted into token arrays [12]. The tokenizing stage is carried out using the help of the "nltk" library in the Google Collab application. Figure 3 present tokenizing result.



| tokenizing |
|---|
| [kokoh, tapi, ada, yang, penyok, gatau, karena... |
| [sesuai, gambar, tp, belum, tau, kekokohannya,... |
| [kecewa, sama, kaki, nya, meja, gambar, nya, p... |
| [gak, sesuai, gambar, gambar, nya, kaki, meja,... |
| [tampilan, cacat, packing, sangat, buruk] |

Fig 3. Tokenizing

4. Slang Word Conversion : perform conversion so that reviews containing slang words can be normalized into a more formal language. This is done to ensure that the text processing model can understand and process text more accurately, especially in the context of sentiment analysis, text classification, or other natural language processing applications [13]. Figure 4 slang word conversion



| slangword_conv |
|---|
| [kokoh, tapi, ada, yang, penyok, gatau, karena... |
| [sesuai, gambar, tapi, belum, tahu, kekokohann... |
| [kecewa, sama, kaki, nya, meja, gambar, nya, p... |
| [gak, sesuai, gambar, gambar, nya, kaki, meja,... |
| [tampilan, cacat, packing, sangat, buruk] |

Fig 4. Slang Word Conversion

- *Visualization* : After the data has successfully gone through the stemming process, the next step is to present the visuals in the form of a word cloud which helps to make it easier to show the main words or topics that often appear in reviews that are grouped into positive and negative sentiments [14]. Frequently occurring words or topics will be large, while less common ones will be small. Figure 5 present Word Cloud for positive review.

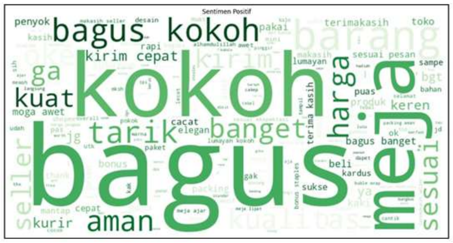


Fig 5 Word Cloud or positive review

Figure 5 some words that often appear in positive sentiment reviews according to text size include "kokoh", "meja", "bagus", "barang", etc. In addition to positive sentiment reviews, here are the results of word cloud visualization on negative sentiment. Figure 6 present Word Cloud for negative reviews.

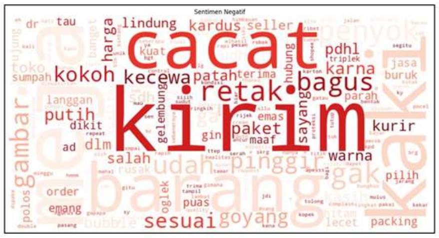


Fig 6. Negative Reviews Word Cloud

Figure 6 shows several words that often appear in negative sentiment reviews according to text size, including "kirim", "cacat", "retak", "kecewa", and "patah", etc. This word cloud visualization can make it easier to understand the main topic according to the sentiment in the data.

- *TF-IDF* : The TF-IDF stage begins by taking stemming data. One of the methods used to measure how important a word is in a document in a collection or corpus of documents. This method is often used in text analysis, information retrieval, and document modeling, such as in recommendation systems and text classification. [15]. This stage divides the data into 20% test data: 80% training data, then calculations are performed to find and display the frequency of the 10 most common words based on TF-IDF values. The goal is to understand the distribution of words in the data and identify dominant words.
- *Training and Testing Model* : The modeling process is carried out using the Naïve Bayes Classifier and Random Forest algorithms [16]. In this process, the data used is 393 data. The modeling process begins by dividing the training and test data. The ratio for training and test data is divided into 3 ratios, namely 90:10, 80:20, and 70:30. The purpose of this division is to find the optimal data division proportion for the model to be used, besides that it is also used to prevent overfitting.
- *Model Evaluation* : After the modeling process is complete, a model evaluation can be carried out from the confusion matrix produced by the model [17].

1. Accuracy : A measure of how well an SVM model classifies the given data [18]. Accuracy is calculated by comparing the number of correct predictions to the total number of data points tested

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

2. Precision : a metric used to measure how many positive predictions are actually correct (relevant) out of all the positive predictions made by the model [19].

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

3. Recall : a metric used to measure how many positives the model actually detected out of all the actual positive data [20].

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

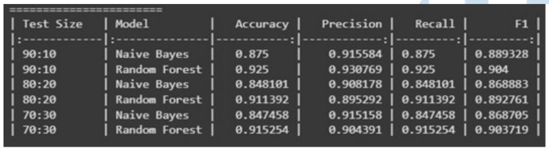
4. F1-Score : A metric that combines two important metrics in evaluating classification models, namely precision and recall, into one number that provides an overview of the balance between the two[20].

$$F1score = 2 \times \frac{Precision \ \times Recall}{Precision + Recall} \quad (4)$$

- *Model Comparison Result* : The results of an analysis that compares the performance of several machine learning models or algorithms in a particular task, with the aim of determining which model is the most effective and efficient based on relevant evaluation metrics.

## III. Result and Discussion

After getting the confusion matrix results according to the data division during modeling, the next step is to calculate the accuracy, precision, recall, and F1 Score values of each algorithm model. The calculation results can be seen in Figure 7.



```
========================
| Test Size    | Model         |  Accuracy |   Precision |   Recall |       F1 |
|:-------------|:--------------|----------:|------------:|---------:|---------:|
| 90:10        | Naive Bayes   |  0.875    |    0.915584 | 0.875    | 0.889328 |
| 90:10        | Random Forest |  0.925    |    0.930769 | 0.925    | 0.904    |
| 80:20        | Naive Bayes   |  0.848101 |    0.908178 | 0.848101 | 0.868883 |
| 80:20        | Random Forest |  0.911392 |    0.895292 | 0.911392 | 0.892761 |
| 70:30        | Naive Bayes   |  0.847458 |    0.915158 | 0.847458 | 0.868705 |
| 70:30        | Random Forest |  0.915254 |    0.904391 | 0.915254 | 0.903719 |
```

Fig 7. Model Evaluation Report

The figure 7 shows the evaluation results of the Naive Bayes and Random Forest models using various test data sizes for classification. The models used were evaluated based on four main metrics: Accuracy, Precision, Recall, and F1-Score. From the data shown, it can be seen that the Random Forest model consistently gives better results compared to Naive Bayes, especially at larger test sizes. For example, at a test size of 90:10, Random Forest has an accuracy of 0.925, while Naive Bayes only reaches 0.875. In general, Random Forest shows more stable performance across all metrics, with higher precision, recall, and F1 values than Naive Bayes. This shows that Random Forest is more effective in processing data with various training and test set sizes, while Naive Bayes tends to be slightly less optimal in this case. Below is a graphical image of the trend based on the data distribution ratio.

The figure 8 shows the performance trends of two classification models, Naive Bayes and Random Forest, based on four different metrics: Accuracy, Precision, Recall, and F1 Score. The graph compares the two models at different training and testing data split ratios. In general, Random Forest tends to give better results in Precision and Recall, especially at larger data split ratios, but there is a decrease in performance in F1 Score at a 70:30 split ratio. On the other hand, Naive Bayes shows more stable consistency in Accuracy, but with lower values in Precision and F1 Score compared to Random Forest. This trend indicates that although Random Forest is superior in terms of recall and precision, Naive Bayes can be a good choice if consistency in Accuracy is more important. A comparison of the evaluation metrics generated by each algorithm can be seen in the table 1.

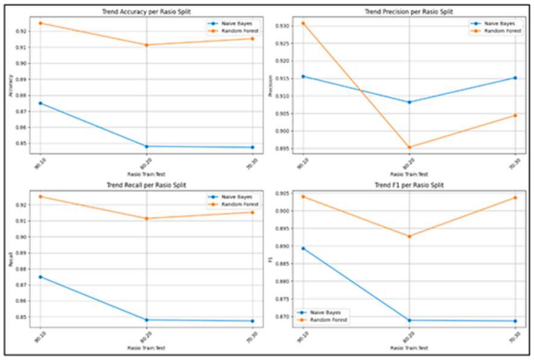


Fig 8. Data Share Ratio Trend Chart

TABLE I. COMPARISON OF OVALUATION METRICS RESULT

| Evaluation Metrics | Naïve Bayes Classifier | Random Forest |
|---|---|---|
| Accuracy | 87.5 % | 92.5 % |
| Precision | 91 % | 93 % |
| Recall | 87.5 % | 92.5 % |
| F-1 Score | 89 % | 90 % |

According to the table I, it shows that the performance of the Random Forest algorithm is better in predicting the majority class (positive). Based on the results of the model classification performance, it can be seen that the Random Forest algorithm has the highest accuracy value. Accuracy explains the extent to which the testing model can classify data correctly. When viewed from other evaluation metrics, the results also explain that the Random Forest algorithm is still superior. The confusion matrix for the Naive Bayes Classifier algorithm can be seen in Figure 9.

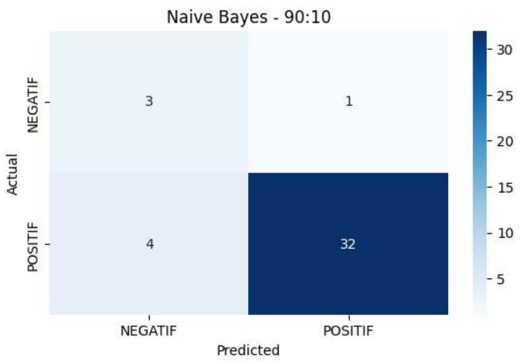


Fig 9. Confusion Matrix Naïve Bayes Classifier

The figure 9 shows the confusion matrix of the classification results using the Naive Bayes algorithm with a 90:10 data sharing scheme. From the matrix, it can be seen that the model successfully classified 3 negative data correctly (true negative) and 32 positive data correctly (true positive). However, there was 1 negative data that was incorrectly classified as positive (false positive) and 4 positive data that were incorrectly classified as negative (false negative). The model shows quite good performance in recognizing positive data, as seen from the high number of true positives. However, the model still has weaknesses in detecting negative data, because the number of false negatives and false positives still exists. This could be an indication that positive data is more dominant or the model is more sensitive to the positive class, so further evaluation is needed, for example by analyzing precision, recall, and F1-score to get a more comprehensive picture of model performance.The confusion matrix for the Random Forest algorithm can be seen in Figure 10.

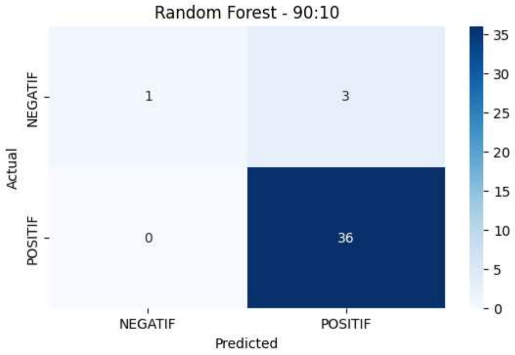


Fig 10. Confusion Matrix Random Forest

The confusion matrix in Figure 10 shows the classification results using the Random Forest algorithm with a 90:10 data sharing scheme. From the matrix, the model successfully classified 36 positive data correctly (true positive) and 1 negative data correctly (true negative). However, there were 3 negative data that were incorrectly classified as positive (false positive), while no positive data were incorrectly classified as negative (false negative). The Random Forest model is very good at recognizing positive data, as seen from the absence of false negatives and the high value of true positives. However, the model is still less than optimal in detecting negative data, as evidenced by the higher number of false positives compared to true negatives. This can be a consideration for adjusting the threshold or balancing the data so that the model's performance in the negative class can be improved.

Precision measures how many of the predicted positive cases are actually correct. In this case, Random Forest has a precision of 93%, while Naïve Bayes has 91%. This suggests that Random Forest has a slightly better ability to avoid false positives compared to Naïve Bayes. Recall measures how many actual positive cases were correctly identified by the model. Both models have the same recall value of 87.5% for Naïve Bayes and 92.5% for Random Forest, showing that Random Forest is better at identifying actual positive cases, leading to fewer false negatives. he F-1 Score balances precision and recall. Random Forest has a slightly better F-1 Score (90%) than Naïve Bayes (89%), reinforcing that it is more balanced in terms of both precision and recall.

## IV. Conclusion

This research compares the Naïve Bayes Classifier and Random Forest algorithms in analyzing sentiment of e-commerce product reviews in the marketplace. The stages carried out include pre-processing, classification, and testing with data division at ratios of 90:10, 80:20, and 70:30. The best results were obtained at a ratio of 90:10. At the model evaluation stage, a confusion matrix was used to measure model performance, followed by the calculation of evaluation metrics. Random Forest produced an accuracy of 92.5%, higher than the Naïve Bayes Classifier which only reached 88%. In addition, Random Forest also excels in precision metrics (93%), recall (92.5%), and F1-Score (90%) in the positive class, indicating that this algorithm is more effective in predicting positive reviews. The results show that Random Forest is better than Naïve Bayes Classifier in classifying sentiment of e-commerce product reviews, especially in predicting positive reviews consistently. This research can be used by e-commerce players to improve marketing strategies, identify potential product problems, and optimize product development based on customer feedback. In addition, the use of sentiment analysis can help e-commerce automate the process of assessing and monitoring product reviews, thereby increasing operational efficiency and customer satisfaction.

## References

[1] A. Sharma, S. Kumar Mishra, dan V. Kant Srivastav, "The Evolution And Impact Of E-Commerce," 2023.

[2] S. Rashiq Nazar dan T. Bhattasali, "SENTIMENT ANALYSIS OF CUSTOMER REVIEWS," *Azerbaijan Journal of High Performance Computing*, vol. 4, no. 1, hlm. 113–125, Jun 2021, doi: 10.32010/26166127.2021.4.1.113.125.

[3] L. R. Krosuri dan R. S. Aravapalli, "Feature level fine grained sentiment analysis using boosted long short-term memory with improvised local search whale optimization," *PeerJ Comput Sci*, vol. 9, 2023, doi: 10.7717/PEERJ-CS.1336.

[4] K. Alemerien, A. Al-Ghareeb, dan M. Z. Alksasbeh, "Sentiment Analysis of Online Reviews: A Machine Learning Based Approach with TF-IDF Vectorization," *Journal of Mobile Multimedia*, vol. 20, no. 5, hlm. 1089–1116, 2024, doi: 10.13052/jmm1550-4646.2055.

[5] S. Ruan, H. Li, C. Li, dan K. Song, "Class-specific deep feature weighting for naïve bayes text classifiers," *IEEE Access*, vol. 8, hlm. 20151–20159, 2020, doi: 10.1109/ACCESS.2020.2968984.

[6] W. Deng, Y. Guo, J. Liu, Y. Li, D. Liu, dan L. Zhu, "A missing power data filling method based on improved

random forest algorithm," *Chinese Journal of Electrical Engineering*, vol. 5, no. 4, hlm. 33–39, Des 2019, doi: 10.23919/CJEE.2019.000025.

[7] A. Nurian, M. S. Ma'arif, I. N. Amalia, dan C. Rozikin, "ANALISIS SENTIMEN PENGGUNA APLIKASI SHOPEE PADA SITUS GOOGLE PLAY MENGGUNAKAN NAIVE BAYES CLASSIFIER," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 1, Jan 2024, doi: 10.23960/jitet.v12i1.3631.

[8] Tania Puspa Rahayu Sanjaya, Ahmad Fauzi, dan Anis Fitri Nur Masruriyah, "Analisis sentimen ulasan pada e-commerce shopee menggunakan algoritma naive bayes dan support vector machine," *INFOTECH : Jurnal Informatika & Teknologi*, vol. 4, no. 1, hlm. 16–26, Jun 2023, doi: 10.37373/infotech.v4i1.422.

[9] H. He, G. Zhou, dan S. Zhao, "Exploring E-Commerce Product Experience Based on Fusion Sentiment Analysis Method," *IEEE Access*, vol. 10, hlm. 110248–110260, 2022, doi: 10.1109/ACCESS.2022.3214752.

[10] H. Huang, A. A. Zavareh, dan M. B. Mustafa, "Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions," 2023, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2023.3307308.

[11] S. Zhang dan H. Zhong, "Mining Users Trust From E-Commerce Reviews Based on Sentiment Similarity Analysis," *IEEE Access*, vol. 7, hlm. 13523–13535, 2019, doi: 10.1109/ACCESS.2019.2893601.

[12] S. A. A. Shah, M. Ali Masood, dan A. Yasin, "Dark Web: E-Commerce Information Extraction Based on Name Entity Recognition Using Bidirectional-LSTM," *IEEE Access*, vol. 10, hlm. 99633–99645, 2022, doi: 10.1109/ACCESS.2022.3206539.

[13] B. M. Shoja dan N. Tabrizi, "Customer Reviews Analysis with Deep Neural Networks for E-Commerce Recommender Systems," *IEEE Access*, vol. 7, hlm. 119121–119130, 2019, doi: 10.1109/ACCESS.2019.2937518.

[14] D. Chehal, P. Gupta, dan P. Gulati, "Evaluating Annotated Dataset of Customer Reviews for Aspect Based Sentiment Analysis," 2022, *River Publishers*. doi: 10.13052/jwe1540-9589.2122.

[15] L. Huang, Z. Dou, Y. Hu, dan R. Huang, "Textual analysis for online reviews: A polymerization topic sentiment model," 2019, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2019.2920091.

[16] H. Tufail, M. U. Ashraf, K. Alsubhi, dan H. M. Aljahdali, "The Effect of Fake Reviews on e-Commerce during and after Covid-19 Pandemic: SKL-Based Fake Reviews Detection," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3152806.

[17] A. Qazi, N. Hasan, R. Mao, M. E. M. Abo, S. K. Dey, dan G. Hardaker, "Machine Learning-Based Opinion Spam Detection: A Systematic Literature Review," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3399264.

[18] A. Mardjo dan C. Choksuchat, "HyVADRF: Hybrid VADER-Random Forest and GWO for Bitcoin Tweet Sentiment Analysis," *IEEE Access*, vol. 10, hlm. 101889–101897, 2022, doi: 10.1109/ACCESS.2022.3209662.

[19] S. Gupta, B. Kishan, dan P. Gulia, "Comparative Analysis of Predictive Algorithms for Performance Measurement," *IEEE Access*, vol. 12, hlm. 33949–33958, 2024, doi: 10.1109/ACCESS.2024.3372082.

[20] Y. Huang, R. Wang, B. Huang, B. Wei, S. L. Zheng, dan M. Chen, "Sentiment Classification of Crowdsourcing Participants' Reviews Text Based on LDA Topic Model," *IEEE Access*, vol. 9, hlm. 108131–108143, 2021, doi: 10.1109/ACCESS.2021.3101565.