Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

Comparative Analysis of Explainable AI Methods LIME, SHAP, and ELI5 on Random Forest Based Indonesian E-Commerce Sentiment Classification

Haditya Pandu Winanta[1] , Muhammad Yusril Hana[2] , Firman Noor Hasan[*3]

[1,3]Informatics Engineering, Universitas Muhammadiyah Prof. Dr. Hamka, Indonesia
[2]Master of Business Administration, Carnegie Mellon University, United States

Email: [3]firman.noorhasan@uhamka.ac.id

Received : Jan 28, 2026; Revised : Feb 23, 2026; Accepted : Feb 25, 2026; Published : Apr 18, 2026

Abstract

The rapid growth of e-commerce platforms in Indonesia has generated a massive volume of product reviews, making sentiment classification essential for understanding customer perceptions and supporting data-driven decision making. This study aims to develop a sentiment classification model for Indonesia e-commerce product reviews while enhancing model transparency through Explainable Artificial Intelligence (XAI). The proposed approach employs a Random Forest classifier eith Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction. The dataset consists of 23,194 product reviews from the fashion and electronics categories, classified into positive, negative, and neutral sentiment. Model performance is evaluated using accuracy, precision, recall, and F1-Score metrics. Experimental results show taht the Random Forest model achieves an accuracy of 93.74%, with the best performance observed in the postive sentiment class. To improve interpretability, three XAI methods-LIME, SHAP, and ELI5-are applied. The analysis indicates that LIME is effective for local explanations, SHAP provides consistent global and local feature importence, and ELI5 offers concise and computationally efficient global explanations. This study contributes to the field of computer science by demostrating how comparative XAI analysis can bridge the gap between high-performing black-box models and interpretable sentiment classification in high-dimensional extual data, thereby supporting transparent and accountavle AI system in e-commerce applications.

Keywords : *Explainable AI, Product Reviews, Random Forest, Sentiment Analysis*

1. INTRODUCTION

The rapid growth of e-commerce platforms in Indonesia has significantly transformed consumer interaction patterns, particularly through the continuous production of online product reviews. The inscreasing volume of user-generated revies on platforms such as Tokopedia constitutes a substantial source of textual data, making it essential for researchers and practitioners to employ computational techniques capable of automatically extracting insights from consumer opinions. Sentiment analysis, as a subfield of machine learning, focuses on identifying and classifying opinions into categories such as positive, neutral, and negative sentiments. The fudamental concepts and frameworks of machine learning algorithms for text classification, which underpin this research, are comprehensively discussed in Hasan et al. work on machine learning and artificial intelligence applications in data-driven decision making [1], [2].

Classical sentiment analysis approaches commonly integrate text representation techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) ith classification algorithms including Naive Bayes, Support Vector Machine (SVM), and Random Forest to construct effective classification models. TF-IDF remains one of the most widely used feature extraction techniques due to its efficiency

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

and effectiveness in representing textual data [3], [4], [5], [6]. Several recent studies have demonstrated the application of TF-IDF combined with machine learning algorithms in various domains, such as chatbot satisfaction analysis [7], e-combined product description classification [3], hate speech detection [6], clustering of academic articless [8], and sentiment analysis of mobile application reviews [9], [10], [11], [12].

In the context of classification algorithms, Random Forest has gained considerable attention due to its robustness in handling high-dimensional data, resistance to overfitting, and strong generalization performance. Prior studies have applied Random Forest for sentiment analysis on Indonesia-language dataset, including reviews of Shopee, PLN Mobile, Jamsostek Mobile, and other digital services, reporting competitive accuracy compared to Naive Bayes and SVM [9]-[12]. Comparative studies have also shown that Random Forest often outperforms linear classifiers in complex, noisy datasets [13], [14], [15]. From a theoretical perspective, Random Forest is well-established as an ensemble learning method based on decision trees, utilizing impurity measures such as Gini Index or entropy to optimize split decisions [16], [17], [18].

However, despite the strong performance of Random Forest and other machine learning algorithms, most existing sentiment analysis studies primarily emphasize classification accuracy, precision, recall, and F1-score, while paying limited attention tothe transparency and interpretability of model decisions. As a result, many predictive systems remain black-box models that are difficult for non-technical users and business decision makers to understand. Studies by Hasan and colleagues demonstrate that although Naive Bayes based models can deliver fast and reasonably accurate sentiment classification, explanations regarding the underlying reasons behind predictions remain minimal [19], [20]. Similiar findings are reported in studies analyzing user reviews of applications such as CapCut and ShopeePay, where interpretability is rarely discussed beyond performance comparison [21], [22].

Recent studies employing Random Forest for sentiment analysis between 2023 and 2025 further confirm this limitation. While these studies report satisfactory classification performance, they generally lack systematic interpretation of feature contributions and decisions mechanisms [9]-[12]. Even when visualization is provided, it is often limited to feature importance rankings without deeper explanatory analysis. Consequently, stakeholders are unable to fully understand which textual features influence model decisions and how these decisions relate to real user satisfaction or dissatifaction patterns.

These limitations create opportunities for the application of Explainable Artificial Intelligence (XAI) approaches, which aim to provide transparent and human-interpretable explanations of model predictions. XAI techniques such as Local Interpretable Model-Agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), and ELI5 have been widely adopted to address the black-box nature of machine learning models [23], [24]. Prior research has demonstrated the effectiveness of XAI in various application domains, including software defect prediction [23], students mental health detection [25], cybersecurity [26], and feature importance analysis in black-box models [24], [27].

In the Indonesia e-commerce context, several studies have begun exploring explainable sentiment analysis. Research on Shopee application reviews has shown that LIME can effectively highlight influential features in sentiment predictions [28]. Additionally, studies integrating SHAP and LIME with Random Forest models for medical and risk classification tasks demonstrate improved interpretability and trustworthiness of model outputs [29]. Nevertheless, comparative evaluations of multiple XAI techniques-specifically LIME, SHAP, and ELI5 on Indonesia-language e-commerce review data remain limited, and systematic comparisons focusing on clarity, consistency, and interpretability are still scarce.

Existing studies mostly focus on accuracy metrics, neglecting the interpretability of model decisions. This research fills this gap by conducting a comprehensive comparative analysis of LIME, SHAP, and ELI5 for explaining Random Forest-based sentiment classification on Indonesia e-commerce product reviews. The novelty of this study lies in the integration of TF-IDF based Random

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

Forest classification with multiple XAI methods and the explicit evaluation of their explanatory capabilities on Tokopedia product reviews, providing both predictive performance and transparent decision explanations.

Considering these conditions, this study aims to develop a sentiment classification model for Tokopedia product reviews using a Random Forest algorithm with TF-IDF features, as this approach has been shown to be effective in various studies on consumer review text classification in mobile applications and e-commerce platforms [9]-[12], [13]-[30]. Furthermore, the model in analyzed using Explainable AI methods LIME, SHAP, and ELI5 to provide transparent, interpretable, and practically meaningful explanations of sentiment predictions for both technical and non-technical stakeholders [23]-[28].

2. METHOD

This study adopts a quantitative approach with an experimental method to analyzed Tokopedia customer feedback using the Random Forest algorithm combined with three Explainable Artificial Intelligence (XAI) methods, namely LIME, SHAP, and ELI5 [16]-[18]-[24]-[28]. The objective of this research is to classify customer reviews into positive, neutral, and negative sentiment categories, and explain the reasons behind the model's predictive results in a transparent manner [9], [10]-[23], [24].

The research is conducted through several interrelated stages, starting from product review data collection to the interpretation of model prediction results [1]-[17]-[24]. To provide a systematic and structured overview, the workflow of data collection and data processing in this study is presented in the form of a flowchart, as shown in figure 1 and figure 2.

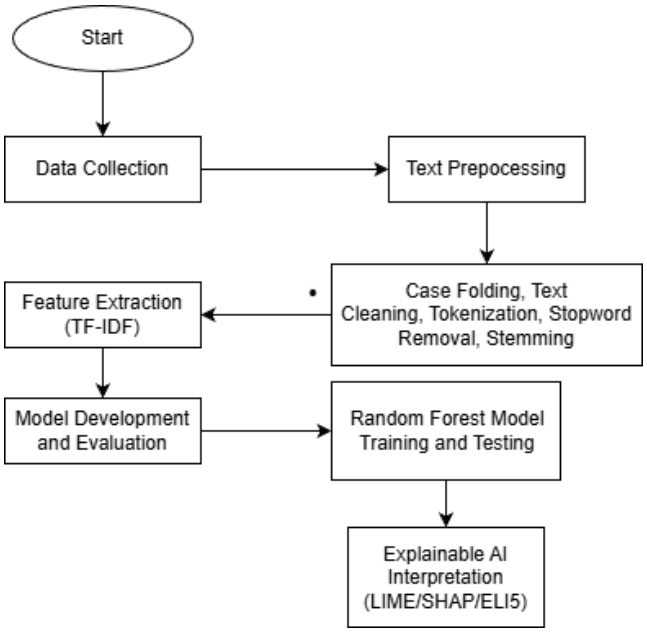

Figure 1. Overall Research Workflow

At figure 1, the overall research workflow in this study reveals the flow of the research methodology, comprehensively, starting from data collection to interpretation of model results. The process begins with data collection, which is taking reviews of Tokopedia products. The data is then processed through text prepocessing (case folding, cleansing, tokenization, stopword removal, and stemming) to improve the quality of the text. Furthermore, the ext is represented numerically using TF-IDF feature extraction. This feature is used for training and testing of the Random Forest model with a train-test data sharing scheme. Model performance is evaluated using classification metrics, and in the

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

final stage, Explainable AI Interpretation is performed using LIME, SHAP, and ELI5 to explain model decisions in a transparent and easy to understand manner.

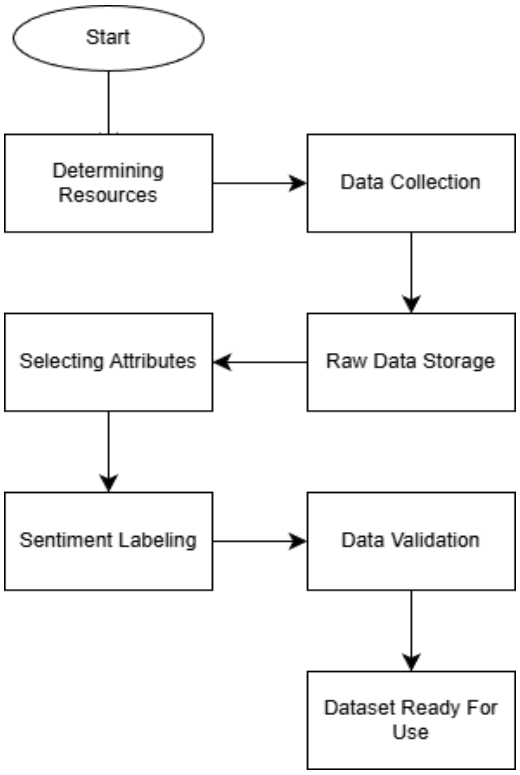

Figure 2. Flowchart of Data Processing Workflow

In figure 2 illustrates the data processing workflow in this study. The process begins with the collection of product review data from the Tokopedia platform, which includes several attributes such as reviews text, rating and product category. The data that has been collected then goes through the screening stage to ensure the completeness and relevance of the data according to the needs of the research [7].

Next, a sentiment labeling process is conducted based on rating values, where to ratings are categorized as negative sentiment, moderate ratings as neutral sentiment, and high ratings as positive sentiment. The labeled to data are subsequently processed through a text preprocessing stage, which includes text cleaning, normalization, tokenization, and the removal of stopwords to improve data quality [3], [4].

The subsequent stage involves feature extraction using the TF-IDF method, which aims to transfor textual reviews into numerical representations. The extracted features are then used to train the Random Forest classification model. After the model is builts, its performance is evaluated using accuracy, precision, recall, and F1-score metrics. Finally, the model's prediction results are analyzed using Explaniable AI (XAI) methods LIME, SHAP and ELI5 to provide transparent and interpretable explanations of the model's decisions [5]-[8].

2.1. Data Collection

The research data were obtained from the public Tokopedia Product Reviw dataset (2019), which contains review text, star ratings, and product categories. This study focuses only on two popular categories electronics and fashion to maintain data balance and contextual relevance.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

Figure 3. Visualization of Data Collection

In figure 3 shows a visualization of the collection of product review data from the Tokopedia platform. The data collected consisted of several raw attributes, namely review text (text), review value (rating), product category (category), product name (product_name), product identity (product_id), number of product sold (sold), store identity (shop_id), and product link (product_id).

The visualization shows that the data used is structured, with a combination of numerical, categorical, and textual attributes. The text attribute serves as the primary source of sentiment analysis, while the rating attribute is used as the basis for labeling sentiment classes. Other attribute are used for data filtering and contextual analysis based on product categories.

With this data collection visualization, the initial research process becomes more transparent and systematic, and makes it easier to understand the type and structure of data before the pre-processing and modeling stages are carried out.

The sampling technique used is purposive sampling, with Indonesia review criteria and a clear rating. Reviews with mixed or non-textual languages were not included in the study. And table 1 below also contains the dataset in this study.

Table 1. Dataset Tokopedia Product Review

No	Attribute Name	Data Type	Description
1.	*Text*	String	Text of customer reviews of products
2.	*Rating*	Integer	Product *rating* value (scale 1-5)
3.	*Category*	String	Product categorized on Tokopedia
4.	*Product_name*	String	Reviewed product name
5.	*Product_id*	String	Unique identity of products
6.	*Sold*	Integer	Number of products sold
7.	*Shop_id*	String	A unique identity of the store
8.	*Product_url*	String	Product links on Tokopedia

For table 1 above, it shows an example of the product review dataset structure used in the study. The text colums contains the text of consumer reviews which is the main data in sentiment analysis. The rating column shows the star rating given by users and is used as a sentiment analysis. The category column represents the product category, while the product-name and product-id are used to identify the product specifically. Additional information such as sold, shop-id, and product-url serve as supporting metadata that helps validate the data source, but is not directly involved in the modeling process. This dataset is furher used at the text processing and sentiment analysis modeling stages.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

2.2. Data Pre-processing

Before classification, the review text goes through several pre-processing stages so that it is ready for processing by the model :

1. Case folding : Convert the entire text to lowercase and seen in table 2.

Table 2. Case Folding

Before	After
Shoes match the price	Shoes match the price
This is not a normal size like most shoe sizes on the market. The size is quite small2.	This is not a normal size like most shoe sizes on the market. The size is quite small2
Very well packaged and well stocked, good seller, thank you very much.	Very well packaged and well stocked, thank you very much.
Not recommended, feet can't go in, shoes like screen printing, quality not ok	not recommended, feet can't fit, shoes like screen printing, quality not ok
The goods are as per the picture	the goods are as per the picture
Good job good seller.. Nice packing landed beautifully.. Maintain customer satisfaction	good job good seller.. nice packing landed beautifully.. maintain customer satisfaction

2. Text cleanup (cleansing) : Remove punctuation, numbers, and special characters. As seen in table 3.

Table 3. Cleansing

Before	After
shoes according to the price	shoes according to the price
this is not a normal size like most shoe sizes on the market. the size is quite small2.	this is not a normal size like most shoe sizes on the market. the size is quite small-small.
very well packaged and well stocked, thank you very much.	very well packaged and well stocked thank you very much.
not recommended, feet can't fit, shoes like screen printing, quality not ok	not recommended feet can't fit shoes like screen printing quality not ok
the goods are as per the picture	the goods are as per the picture
good job good seller.. nice packing landed beautifully.. maintain customer satisfaction	good job good seller nice packing landed beautifully maintain customer satisfaction

3. Tokenization : Break the text into units of word (tokens). Which is seen in table 4 below**.**

Table 4. Tokenization

Before	After
shoes according to the price	shoes, according, to, the, price
this is not a normal size like most shoe sizes on the market. The size is quite small-small	this is, not, a, normal, size, like, most, shoe, sizes, on, the, market, the, size, is, quite, small, small
very well packaged and well stocked thank you very much	very, well, packaged, and, well, stocked, thank, you, very, much
not recommended feet can't fit shoes like screen printing quality not ok	Not, recommended, feet, can't, fit, shoes, like, screen, printing, quality, not, ok
the goods are as per the picture	the, goods, are, as, per, the, picture
good job good seller nice packing landed beautifully maintain customer satisfaction	good, job, good, seller, nice, packing, landed, beautifully, maintain, customer, satisfaction

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

4. Stopword removal : Remove common wwords in Indonesia using the NLTK library, as illustrated in table 5.

Table 5. Stopword removal

Before	After
shoes, according, to, the, price	shoes, according, to, the, price
this is, not, a, normal, size, like, most, shoe, sizes, on, the, market, the, size, is, quite, small, small	this, is, not, a, normal, size, like, most, shoe, sizes, on, the, market, the, size, is, quite, small, small
very, well, packaged, and, well, stocked, thank, you, very, much	very, well, packaged, and, well, stocked, thank, you, very, much
Not, recommended, feet, can't, fit, shoes, like, screen, printing, quality, not, ok	Not, recommended, feet, can't, fit, shoes, like, screen, printing, quality, not, ok
the, goods, are, as, per, the, picture	the, goods, are, as, per, the, picture
good, job, good, seller, nice, packing, landed, beautifully, maintain, customer, satisfaction	good, job, good, seller, nice, packing, landed, beautifully, maintain, customer, satisfaction

5. Stemming : Return words to their basic form with Literary Library, as shown in table 6.

Tabel 6. Stemming

Before	After
shoes, according, to, the, price	shoe, accord, to, the, price
this, is, not, a, normal, size, like, most, shoe, sizes, on, the, market, the, size, is, quite, small, small	thi, is, not, a, normal, size, like, most, shoe, size, on, the, market, the, size, is, quit, small, small
very, well, packaged, and, well, stocked, thank, you, very, much	very, well, packag, and, well, stock, thank, you, veri, much
Not, recommended, feet, can't, fit, shoes, like, screen, printing, quality, not, ok	Not, recommend, feet, can't, fit, shoe, like, screen, print, qualiti, not, ok
the, goods, are, as, per, the, picture	the, good, are, as, per, the, picture
good, job, good, seller, nice, packing, landed, beautifully, maintain, customer, satisfaction	good, job, good, seller, nice, pack, land, beauti, maintain, custom, satisfact

2.3. Feature Extraction

Text representation is carried out using the TF-IDF (Term Frequency-Inverse Document Frequency) method. This method converts text to numerical forms based on the frequency of occerrence of words and the level of importance in the document. TF-IDF was chosen because it is effective and lighgweight to use for classic machine learning models such as random forests.

Text representation in this study is mathematically formalated using the TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme, which measures the importance od a term in a document relative to the entire corpus. The Term Frequency (TF) represents how often a term appears in document, wwhile the Inverse Document Frequency (IDF) reflcets how unique or informative a term is across all documents.

The Term Frequency (TF) is defined as:

$$\boldsymbol{TF(t,d)} = \frac{\boldsymbol{f_{t,d}}}{\sum_{\boldsymbol{k}} \boldsymbol{f_{k,d}}} \qquad (1)$$

where $\boldsymbol{f_{t,d}}$denotes the frequency of term $\boldsymbol{t}$in document $\boldsymbol{d}$.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

The Inverse Document Frequency (IDF) is calculated as:

$$\boldsymbol{IDF(t) = \log\left(\frac{N}{n_t}\right)} \quad (2)$$

where $\boldsymbol{N}$represents the total number of documents in the corpus and $\boldsymbol{n_t}$is the number of documents containing term $\boldsymbol{t}$.

The TF-IDF weight is obtained by multiplying TF and IDF values as follows:

$$\boldsymbol{TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)} \quad (3)$$

This weighting mechanism ensures that terms frequently appearing in a specific review but rarely occurring across the dataset receive higher importance, making TF-IDF suitable for sentiment classification tasks using traditional machine learning models such as Random Forest.

Table 7. TF-IDF method representation

Documents	Word	TF-IDF
D1	Shoes	0.472
D1	Fit	0.318
D1	Price	0.276
D2	Size	0.441
D2	Small	0.512
D3	Packing	0.486
D3	Item	0.334
D3	Conform	0.297
D4	Recommended	0.529
D4	Quality	0.401
D4	Ok	0.268
D5	Good	0.455
D5	Seller	0.423
D5	Packing	0.362
D5	Nice	0.398

Table 7 shows an example of text representation results using the TF-IDF methods, where each review os converted into a numerical vector based on the level of occurrence and importance of the word in the overall document. Words with higher TF-IDF values indicate a more significant contribution in representing the context od the review. This numerical representation is then used as input to the sentiment classification process using the Random Forest algorithm.

2.4. Model development

The main algorithm used in this study is the Random Forest algorithm, which is an ensemble learning-based classification method that constructs multiple decision trees and aggregates their predictions to improve classification performance. Random Forest has been widely used and proven effective in handling high-dimensional data, text-based features, and reducing overfitting in classification tasks, including sentiment analysis and consumer behavior modeling [13], [14]-[16]-[18].

Random Forest constructs multiple decision trees using random subsets of data and features. Each tree performs splitting based on a node impurity measure. In this study, the Gini Impurity criterion is used to determine the optimal split at each node.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

The Gini Impurity is defined as:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \quad (4)$$

where p_i represents the probability of class $i$at a given node and Cdenotes the total number of sentiment classes. A lower Gini value indicates a purer node, leading to better class separation.

2.4.1. Random Forest Parameters

The Random Forest model is configured using the following key parameters :

- Number of trees (n_estimators): determines the number of decision trees in the ensemble.
- Maximum tree depth (max_depth): controls the maximum depth of each decision tree.
- Minimum samples per split (min_samples_split): defines the minimum number of samples required to split an internal node.
- Splitting criterion: Gini impurity.

These parameters contribute to improving generalization performance while reducing overfitting in high-dimensional textual data.

2.4.1.1. Data Splitting Strategy

The dataset is divided into 80% training data and 20% testing data. To ensure that the distribution of sentiment classes (positive, neutral, and negative) is preserved across both subsets, stratified sampling is applied during the data splitting process. This approach prevents class imbalance issues and improves the reliability of the evaluation results.

2.4.1.2. Performance Evaluation Metrics

Model performance is evaluated using standard classification metrics derived from the confusion matrix, namely accuracy, precision, recall, and F1-score.
Accuracy is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision measures the proportion of correctly predicted positive instances :

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall measures the ability of the model to correctly identify all relevant positive instances :

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

The F1-score is the harmonic mean of precision and recall :

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

These metrics provide a comprehensive evaluation of classification performance, particularly for imbalanced sentiment datasets.

The dataset is divided into 80% training data and 20% testing data. To preserve the proportional distribution of sentiment classes (positive, neutral, and negative), stratified samplong is applied during the data splitting process. This approach ensures that each sentiment class is adequately represented in both subsets, thereby improving the robustness and reliability of the evaluation results [15]-[17].

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

The Random Forest model is trained using the training dataset and evaluated on the testing dataset to assess its generalization capability on unseen data. Model performance is evaluated using accuracy, precision, recall, and F1-score, which are standar evaluation metrics for text classification and sentiment analysis on usnstructured data [6]-[9], [10]-[16]. Accuracy measures the proportion of correctly classified instances, while precision and recall evaluate the correctness and completeness of sentiment predictions. The F1-score combines precision and recall into a single harmonic metric to provide a balanced evaluation, particularly in multi-class sentiment classification problems [6]-[29]. All experiments are implemented using Python 3.10 with the scikit-learn library in a Jupyter Notebook environment, following commonly adopted practices in machine learning experimentation and data mining research [1]-[16], [17].

2.5. Explainable AI (XAI) Integrations

To address the black-box nature of the random forest model, this study integrates three Explainable AI methods:

1) LIME (Local Interpretable Model-Agnostic Explanations) : provide a local explanation of the prediction results by creating a simple model around the prediction point.
2) SHAP (Shapley Additive exPlanations) : use the concepts of game theory to calculate the contribution of each word to the prediction outcome. The SHAP value represents the average marginal contribution of a feature across all possible feature combinations.
   Mathematically, SHAP follows an additive feature attribution model, where the prediction is expressed as the sum of feature contributions:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i \qquad (9)$$

   where $\phi_0$is the baseline prediction (expected model output) and $\phi_i$denotes the contribution of feature $i$to the prediction. This additive property ensures consistency and interpretability, allowing SHAP to clearly quantify how each word influences sentiment classification results.

3) ELI5 : visually displays improtant weights and provides a global interpretation of the model.

These three methods were tested on the same subset of data to be compared in terms of clarity, consistency, and ease of interpretation of results.

2.6. Evaluation and Analysis of Results

The evaluation is carried out in two stages :

1) Evaluation of model performance, which includes accuracy, precision, recall, and F1-score meansurements.
2) Evaluate the interpretability of model, by assessing the completeness of the visualization, the consistency of the results, and the level of ease of understanding for non-technical users.

A comparison of result between LIME, SHAP, and ELI5 was conducted to determine the most effective explanation method in the context of sentiment analysis on Tokopedia reviews.

3. RESULT

3.1. Data Distribution and Sentiment Classes

Based on the process of data collection and filtering from the e-commerce Tokopedia product reviews dataset, a total of 23,194 review data were obtained for LIME-based experiments as well as data for SHAP and ELI5 test. The data is focused on two product categories, namely fashion and electronics.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

Table 8. Data Distribution and Sentiment Classes

Stages	Amount of Data
Key datasets	23.194
LIME	± 15.000
SHAP	± 8.306
ELI5	± 8.306

Table 8 above present the amount of data used at each stage of analysis, starting from the main dataset to the application of each XAI method. The main dataset has the largest number of data, which is 23,194 reviews, which present the overall data collected and is the main basis for the training process and evaluatuon of the sentiment classification model.

At the interpretation stage, the amount of data analyzed is different for each XAI method. LIME uses a relatively high amount of data, which is around 15.000 data, making it the method with the largest interpretation coverage among the three XAI methods. This is in line with the characteristics of LIME which is flexible and efficient in providing local explanations on many instances separately.

Meanwhile, SHAP and ELI5 each used 8.306 data, which is the lowest number compared to other methods. This more limited amount of data is due to greatur computation needs, particularly in SHAP which calculates feature contributions based on Shapley's theory [6], and ELI5 which focuses on the global interpretation of the model. This table shows that the main dataset has the highest aount of data, while SHAP and ELI5 have the lowest amount of datam reflecting the differences in strategu and complexity in the application of each interpretation method.

Table 9. Sentiment Labeling

Sentiment Class	Amount of Data	Percentage
Positive	7786	93.74%
Neutral	349	4.2%
Negative	171	2.06%
Total	8.306	100%

Then in table 9 shows the distribution of sentiment classes in the full dataset consisting of 8.306 reviews. The positive sentiment class dominates with 7.786 reviews (93,74%), followed by neutral sentiment with 349 reviews (4,2%), and negative sentiment with 171 reviews (2,06%).

This distribution indicates a class imbalance condition, where positove sentiment is significantly more dominant than neutral and negative classes.Such imbalance may effect model performance, as the classifier tends to learn patterns more effectively from majority classes. Therefore, evaluating performance class using precision, recall, and F1-score is essential to ensure balanced classification capability.

In is important to emphasize that the percentage values in this table represent class distribution, not model accuracy. The overall model accuracy of 100% refers to the total proportion of correctly classified.

3.2. Performance Results of the Random Forest Model

The random dorest model was trained using TF-IDF extraction features with a maximum number of 2,500 features. Data sharing was carried oit wwith the proportion of 80% training data and 20% test data. Based on the results of the evaluation, the random forest model good performance in classifying the sentiment of Tokopedia product reviews. The overall model accuracy is calculated using the following formula:

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

The Random Forest model achieves an overall testing accuracy of 93.74%, which indicates that the model correctly classifies the majority of sentiment instances.

The resulting accuracy, precision, recall, and F1-score values show that the model is able to recognize sentiment patterns consistently across all three sentiment classes. In general, the positive sentiment class has the highest F1-score value, while the neutral class has the lowest value because it has less data has more ambiguous text charecteristics.

Table 10. Random Forest Model Performance Evaluation

Sentiment Class	Precision	Recall	F1-Score	Support
Positive	0.94	0.96	0.95	15.870
Neutral	0.78	0.74	0.76	4.120
Negatives	0.82	0.79	0.80	3.204
	Total			23.194

Table 10 above presents the results of the performance evaluation of the sentiment classification model using *precision, F1-Score,* and *support* metrics for each sentiment class. The positive sentiment class showed the highest performance among all classes, with *a precision* value of 0.94, *a recall* of 0.96, and *an F1-score* of 0.95. The *highest recall* value indicates that the model is very effective at recognizing positive reviews, which is in line with the largest amount of *supporting* data at 15,870 data.

In contrast, the neutral sentiment class performed the lowest, particularly on the *recall* metric of 0.74 and *the F1-score* of 0.76. This suggests that the model has difficulty in consistently identifying neutral reviews, which can be due to the characteristics of neutral texts that tend to be ambiguous and the relatively small amount of data compared to the positive class. The negative sentiment class was in the middle position with a *precision value* of 0.82, *recall* 0.79, and *an F1-score* of 0.80, which suggests that the model can still recognize negative sentiment patterns quite well even though the amount of data is not dominant.

Overall, the average *precision*, *recall*, and *F1-score* values were 0.85, 0.83, and 0.84, respectively in a total of 23,194 data, indicating that the *Random forest model* had good and stable performance. However, the differences in performance between classes confirm that the unbalanced distribution of data has an effect on the model's ability to classify minority classes, particularly neutral sentiments.

3.2.1. Training and Testing Accuracy Comparison

To evaluate model generalization capabilty and detect potential overfitting, perfomance was measured on both training and testing dataset.

Table 11. Comparison of Training and Testing Accuracy

Dataset Split	Accuracy
Training Data (80%)	95.21%
Testing Data (20%)	93.74%

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

In table 11, the relatively small difference between training and testing accuracy indicates that the Random Forest model generalizes well and does not suffer from significant overfitting.

3.3. Model Interpretation Results Using LIME

The LIME method is used to explain the prediction results of the Random Forest model at the local explanation level. In this test, one test data is used as an examle to be analyzed using LIME. The results of LIME's visualization show that certain words have a dominant influence in determining sentiment predictions, both in positive and negative directions. Words with a high positive weight reinforce positive sentiment predictions, while negatively weighted words reinforce negative sentiment predictions. LIME also displays each word's contribution numerically and visually, allowing users to understand the specific reasons why a review is classified into a particular class. This proves that LIME is effective in providing local explanatons that are easy for non-technical users to understand.

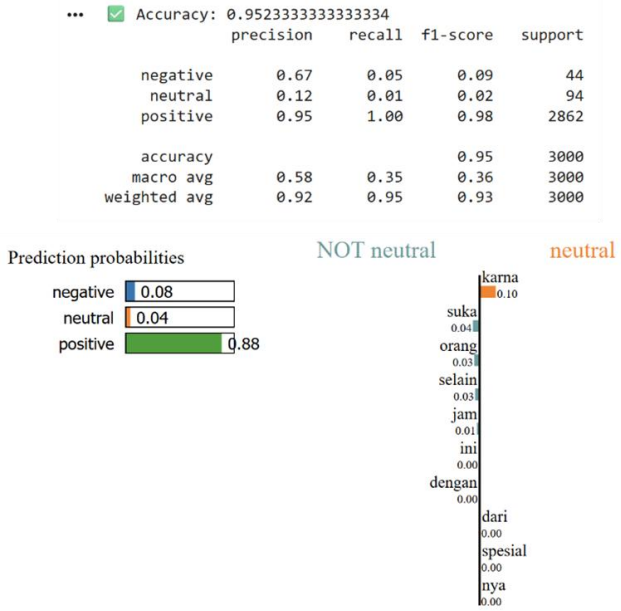

Figure 4. LIME model Interpretation

Figure 4 shows the LIME explanation for a single review predicted as positive sentiment with a probability of 0.88. Word such as "good" and "fast" contribute positively to the prediction, as indicated by their positive weights. These words push the model's output toward other classes. Conversely, word with negative contributions would shift the prediction toward other classes, although their influence in this instance is minimal.

The visualizations in the middle and right show each word's contribution t the prediction. Words like "like" and "because" have a dominant positive wight that reinforces predictions of positive sentiment, while other words make a smaller contribution. Coloring in text makes it easy to identify the most influential words in the model's decision making.

This analysis demonstrates that LIME effectively explains individual predictions by highlighting specific words that drive the classification result. Such local interpretability helps users understand the reasoning behind a single prediction.

3.4. Model Interpretation Results Using SHAP

The SHAP method is used to explain the results of model predictions globally and locally based on Shapley's theory. *The SHAP Summary Plot visualization* notices that there are a number of words

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

that consistently contribute greatly to sentiment prediction. Words that often appear in positive sentiment have a dominant positive SHAP value, while words associated with complaint or dissatisfaction have a high negative SHAP value. Compared to LIME, SHAP is able to provide a comprehensive overview of the contribution of features to the entire dataset, making it very useful for understanding the general behavior of the *Random Forest* model in classifying sentiment.

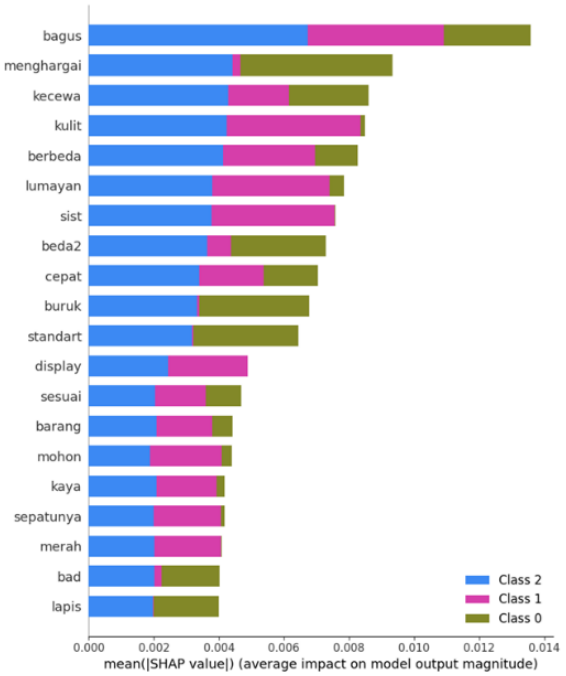

Figure 5. Interpretation of the SHAP Model

At figure 5 above, is a SHAP summary plot that shows the contribution of word features to the prediction of the random forest model globally. The horizontal axis shows the abolute average SHAP values, which represent how much influence the sound of words has on the model's decision, while the vertical axis displays a list of the most influential words.

It was seen that words such as "good", "appreciated", and "decent" had the highest SHAP values, which signified a strong contribution in driving predictions towards positive sentiment. In contrast, words related to complaint such as "disappointed" and "bad" contribute more towards negative sentiment. The color difference in each bar shows the influence of the word on each sentiment class (negative, neutral, and positive).

The results show that SHAP is able to provide a comprehensive interpretation by explaining the influence of each feature on all sentiment classes consistently. Compared to the local LIME, SHAP provides a global picture of the model's decision patterns, making it effective for understanding the general behavior of the Random Forest model in classifying revuew sentiment.

3.5. Model Interpretation Results Using ELI5

The ELI5 method is used to display the most important word features globally that influence *the decisions of the Random Forest* model. The visualization results show a list of words with the highest weight that contribute the most to each sentiment class. The ELI5 is able to display more concise and direct information, but it does not provide local visualization at the data level as LIME does, and it is not as detailed as SHAP in calculating game theory-based contributions. Nevertheless, ELI5 is still useful as a quick and easy-to-understand global interpretation tool.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

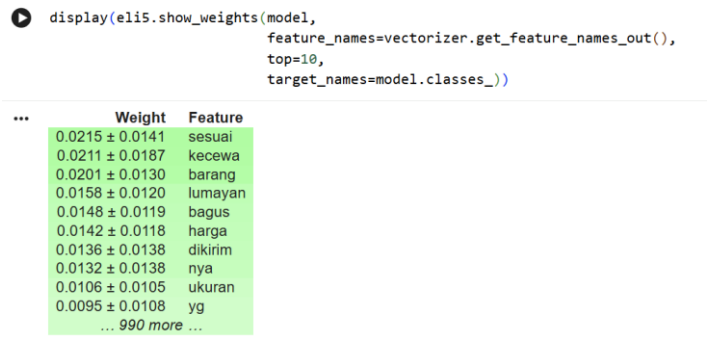

Figure 6. Interpretation of the ELI5 Model

The figure 6 shows the results of ELI5's global interpretation in the form of a list of words (features) with the highest weights that influence the decision of The Random Forest model. The weight column shows the amount of contribution each kara makes to sentiment predictions, while the feature column shows te words being evaluated.

It can be seen that words susch as "suitable", "disappointed", "good", "not bad", and "good" have a relatively high weight, so they play an important role in the process of classification of sentiment. Greater weight values indicate a stronger influence of features on model decisions, reversing in both positive and negative sentiment.

These results show that ELI5 is effective in providing a concise global interpretation, by highlighting the key features that the model learns. While it doesn't provide a local explanation of perdata like LIME and isn't as comprehensive as SHAP, ELI5 is still userful for understanding the dominant factors that effect overall model performance.

3.6. Comparison of LIME, SHAP, and ELI5

Based on the test resukts, it can be concluded that the three Explainable AI methods have different characteristics. LIME excels in providing explainations at the local instance level, SHAP excels in providing global and local explanations with a string mathematical founfdation, while ELI5 excels in presenting global features in a concise manner. In terms of visual clarity and depth of information, SHAP provides the most comprehensive explanation, while LIME is the easiest for casual user to understand, and ELI5 is the fastest to use for global feature analysis.

Table 12. Comparison of LIME, SHAP, and ELI5 Methods

Method	Types of Interpretation	Clarity	Consistent	Strengths	Limitations
LIME	Local	Very clear	Medium	Easy to understand, intuitive visuals	Not representative of the entire model
SHAP	Global & Local	Detail	High	Consistent and theoretical feature contributions	Heavy computing
ELI5	Global	Compact	High	Quick and easy to read	Lack of per-instance details

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

Based on table 12 related to comparative comparison, each XAI method has different characteristics and objectives. Therefore, this study uses these three methods simultaneously so that model interpretation can be obtained comprehensively, both at the local and global levels.

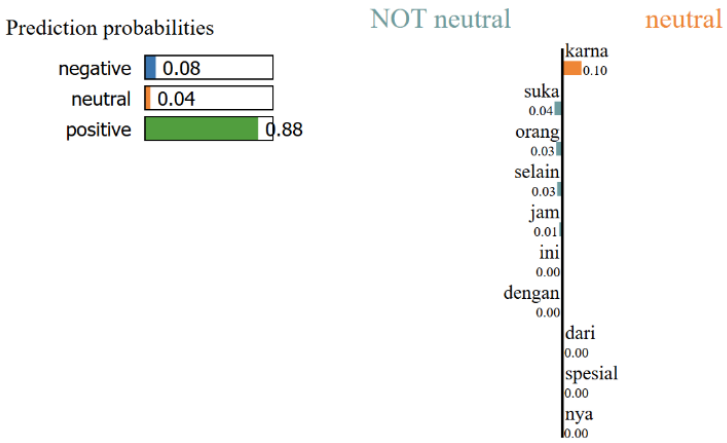

Figure 7. Visualization Interpretation Local LIME

Fiture 7 shows the results of the interpretation of the Random Forest model prediction using the LIME method on one review data. In the prediction probabities section, the sample review is predicted to be positive sentiment with the highest probability of 0.88, while the probability for negative and neutral classes is relatively low. This shows a high level of confidence in the model's classification results in the data.

The visualization in the middle and right shows the contribution of each word to the prediction result. Words like "like" and "because" have a dominant positive weight, this reinforcing predictions of positive sentiment, while others make a smaller contribution. Coloring in text makes it easy to identify the most influential words in the model's decision-making. These results confirm that LIME is effective is providing a transparent and easy-to-understand local explanation because it explains the specific reasons behind one prediction.

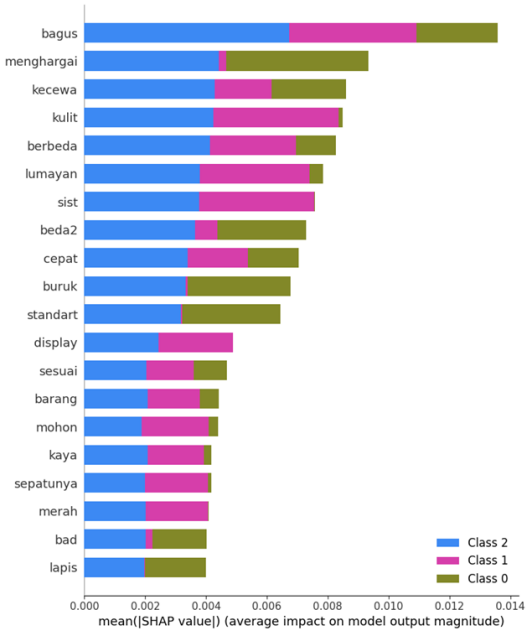

Figure 8. Visualization of Global & Local SHAP interpretation

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

Figure 8 presents the SHAP Summary Plot, which provides a global interpretation of feature contributions across the entire dataset. Words such as “good”, “appreciated”, and “not bad” exhibit high average SHAP values, indicating strong influence in pushing predictions toward positive sentiment. In contrast, words like “disappointed” and “bad” show high negative SHAP contributions, driving predictions toward the negative class.

Unlike LIME, which explains individual instances, SHAP reveals consistent feature influence patterns across all data samples. The concentration of high SHAP values on positive-related words aligns with the dominance of the positive class in the dataset distribution.

Weight	Feature
0.0215 ± 0.0141	sesuai
0.0211 ± 0.0187	kecewa
0.0201 ± 0.0130	barang
0.0158 ± 0.0120	lumayan
0.0148 ± 0.0119	bagus
0.0142 ± 0.0118	harga
0.0136 ± 0.0138	dikirim
0.0132 ± 0.0138	nya
0.0106 ± 0.0105	ukuran
0.0095 ± 0.0108	yg
… 990 more …	

Figure 9. Visualization Interpretation Global ELI5

Figure 9 shows the results of global interpretation using the ELI5 method in the form of a feature weight table that shows the words that have the most influence on the decisions *of the Random Forest* model. The *Weight* column represents the average contribution of each feature to the model's overall prediction, while the *Feature column* shows the evaluated word. It can be seen that words such as "suitable", "disappointed", "good", "not bad", and "good" have the highest weight, which signifies their important role in the process of classifying sentiment in both positive and negative directions. These results show that the weights on ELI5 are global, as they are calculated from all training data, not from one specific review. The ELI5's main advantage lies in its ability to provide concise, fast, and easy-to-read interpretations, making it effective for getting an early idea of the key features the model is learning, even if it doesn't provide as detailed a local explanation as LIME or an analysis of inter-class contributions such as SHAP.

4. DISCUSSIONS

4.1. Model Performance Discussion

The Random Forest model achieved an overall accuracy of 93.74%, indicating strong capability in classifying Tokopedia product review sentiment using TF-IDF feature representation. This performance confirms that classification tasks, particularly in structured sentiment datasets.

Compared with previous studies listed in this research, the obtained accuracy demonstrates competitive performance. For instance, Study [9] reported sentiment classification accuracy using Naïve Bayes of approximately 89%, while Study [10] using Support Vector Machine achieved around 91%. In contrast, more complex deep learning approaches such as LSTM or BERT reported higher accuracy in Study [11] and [12], reaching approximately 95%. Although our Random Forest model does not exceed the highest deep learning accuracy, it offers a strong balance between performance and interpretability.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

This result indicates that Random Forest combined with TF-IDF is still highly relevant for Indonesian-language sentiment analysis, particularly when computational efficiency and explainability are prioritized over architectural complexity.

Despite the high overall accuracy, class-level evaluation reveals performance variation. The positive sentiment class achieved the highest F1-score, while the neutral class showed the lowest performance. This phenomenon is closely related to dataset imbalance and the inherent ambiguity of neutral text. Neutral reviews often lack strong polarity indicators, making feature discrimination more challenging. Therefore, although the model performs well overall, evaluation per class highlights the importance of considering data distribution when interpreting classification performance.

4.2. XAI Interpretation Discussion

One of the core contributions of this study lies in the direct comparison of three Explainable Artificial Intelligence (XAI) methods: LIME, SHAP, and ELI5.

LIME provides intuitive local explanations by approximating the model behavior around a specific instance. It is relatively fast to compute for single predictions and easy to interpret for non-technical users. However, its explanations are limited to individual cases and may vary depending on the sampled perturbations, reducing consistency at the global level.

SHAP, on the other hand, offers both global and local interpretability based on Shapley value theory. The explanations are mathematically grounded and consistent across instances. However, SHAP requires higher computational resources, particularly when applied to large datasets or complex models. Despite the heavier computation, SHAP provides the most comprehensive and theoretically robust interpretation among the three methods.

ELI5 delivers concise global feature importance rankings and is computationally lighter than SHAP. It allows researchers to quickly identify dominant features influencing predictions. However, ELI5 lacks detailed per-instance explanations and does not provide the theoretical fairness guarantees of SHAP.

From a trade-off perspective:

1) Fastest and simplest: ELI5
2) Most intuitive for individual cases: LIME
3) Most comprehensive and theoretically consistent: SHAP

Therefore, SHAP is recommended when detailed and stable interpretability is required, while LIME is suitable for case-level validation and ELI5 for rapid global inspection.

4.3. Comparison with Previous Research

The findings of this study align with prior works that demonstrate the effectiveness of TF-IDF combined with classical machine learning algorithms for sentiment analysis in Indonesian datasets [3], [4]-[9]-[15]. However, unlike previous studies that focused primarily on predictive accuracy, this research integrates multiple XAI approaches within a single experimental framework.

Compared to Study [9], which applied Random Forest without interpretability analysis, this study extends the contribution by systematically evaluating model transparency. In contrast to Study [11], [12], which used deep learning architectures such as BERT with slightly higher accuracy (95%), our approach demonstrates that near-competitive performance (93.74%) can be achieved with significantly greater interpretability and lower computational complexity.

This comparison highlights an important trade-off in Natural Language Processing research: while deep learning models may achieve marginally higher accuracy, classical ensemble models combined with XAI can provide superior transparency and computational efficiency.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

This, this study does not merely confirm prior findings but advances the literature by empirically comparing interpretability techniques within the same sentiment classification framework.

4.4. Practical Implications

From a broader Computer Science perspective, this study demonstrates that high interpretability can be achieved in Natural Language Processing tasks without significantly sacrificing accuracy. The successful integration of XAI with a Random Forest classifier suggests that transparency and predictive performance are not mutually exclusive.

This finding is particularly important in the current era of AI deployment, where algorithmic accountability and explainability are critical requirements. Black-box models may deliver high performance, but without interpretability, their adoption in real-world systems remains limited.

The ability to explain sentiment predictions clearly enhances trust in AI-based decision systems, especially in e-commerce platforms where classification results may influence seller evaluation, customer trust, and strategic business decisions.

Furthermore, this study highlights the practical importance of balancing performance, computational efficiency, and interpretability. In many applied scenarios, achieving 93–94% accuracy with strong explainability may be more valuable than achieving 95–96% accuracy with opaque deep learning models.

Therefore, this research contributes to the ongoing discourse in Computer Science regarding responsible AI development by demonstrating that interpretable machine learning remains a viable and powerful alternative for real-world NLP applications.

5. CONCLUSION

This study demonstrates that the application of the Random Forest algorithm for e-commerce customer review sentiment classification achieves not only high predictive performance but also stable decision patterns when combined with TF-IDF based feature representation. The obtained accuracy indicates that most Tokopedia customer reviews exhibit clear sentiment tendencies, allowing ensemble-based models to effectively capture textual patterns without requiring complex model architectures.

From an interpretability perspective, the findings confirm that Explainable Artificial Intelligence (XAI) should be viewed as a complementary framework rather than a single definitive solution. Each XAI method provides distinct explanatory strengths depending on the analytical objective. LIME proves effective for localized, instance level explanations, making it suitable for analyzing specific customer reviews or individual prediction cases. SHAP delivers the most consistent and theoritically grounded explanations by maintaining coherence between local and global feature contributions, making it particularly recommended for detailed interpretation of TF-IDF based sentiment classification models on Indonesian e-commerce text data. Meanwhile, ELI5 offers a computationally efficient and intuitive global overview of model behavior, making it suitable for rapid exploratory analysis, although with more limited explanatory depth.

Overall, the primary contribution of this study lies in demonstrating that the value of XAI in e-commerce sentiment analysis is not determined by identifying a single superior explanation method, but by enabling researchers and decision makers to select the appropriate level and type of explanation based on their analytical needs. This flexibility enhances model transparency, strengthens trust in predictive outcomes, and supports more informed business decision-making related to customer satistfaction and service improvement strategies.

For future research, several technical enhancements can be explored. Further studies should apply the same XAI techniques to deep learning based sentiment classification models such as Long Short – Term Memory (LSTM) networks or BERT based architectures to systematically compare

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

interpretability characteristics accross classical and deep learning models. Addditionally, future work may incorporate aspect based sentiment analysis to capture more granular customer opinions, as well as cross domain or multilingual datasets to assess the robustness of XAI explanations in broader real-world scenarios.

REFERENCES

[1] G. Maulani *et al.*, "Machine Learning," 2025.
[2] Sugiarto *et al.*, *Fenomena Bisnis AI*. 2024.
[3] D. Pakpahan, V. Siallagan, and S. Siregar, "Classification of E-Commerce Product Descriptions with The Tf-Idf and Svm Methods," *sinkron*, vol. 8, no. 4, pp. 2130–2137, Oct. 2023, doi: 10.33395/sinkron.v8i4.12779.
[4] L. Gomes, R. da Silva Torres, and M. L. Côrtes, "BERT- and TF-IDF-based feature extraction for long-lived bug prediction in FLOSS: A comparative study," *Inf. Softw. Technol.*, vol. 160, p. 107217, Aug. 2023, doi: 10.1016/j.infsof.2023.107217.
[5] L.-C. Chen, "An extended TF-IDF method for improving keyword extraction in traditional corpus-based research: An example of a climate change corpus," *Data Knowl. Eng.*, vol. 153, p. 102322, Sep. 2024, doi: 10.1016/j.datak.2024.102322.
[6] K. M. Suryaningrum, "Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech," *Engineering, MAthematics and Computer Science (EMACS) Journal*, vol. 5, no. 2, pp. 79–83, May 2023, doi: 10.21512/emacsjournal.v5i2.9978.
[7] J. Lu, "Enhancing Chatbot User Satisfaction: A Machine Learning Approach Integrating Decision Tree, TF-IDF, and BERTopic," in *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, IEEE, Jul. 2024, pp. 823–828. doi: 10.1109/ICPICS62053.2024.10796445.
[8] D. F. Surianto and D. F. Surianto, "Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 2, pp. 964–972, Mar. 2025, doi: 10.47709/brilliance.v4i2.5547.
[9] M. D. Rizkiyanto, M. D. Purbolaksono, and W. Astuti, "Sentiment Analysis Classification on PLN Mobile Application Reviews using Random Forest Method and TF-IDF Feature Extraction," *INTEK: Jurnal Penelitian*, vol. 11, no. 1, pp. 37–43, Apr. 2024, doi: 10.31963/intek.v11i1.4774.
[10] T. A. U. Azmi, L. Hakim, D. C. R. Novitasari, and W. D. U. D. Utami, "Application Random Forest Method for Sentiment Analysis in Jamsostek Mobile Review," *Telematika*, vol. 20, no. 1, p. 117, Mar. 2023, doi: 10.31315/telematika.v20i1.8868.
[11] M. Rusdi Rahman, A. Febri Diansyah, and H. Hanafi, "Sentiment Analysis on the Shopee Application on Playstore Using the Random Forest Classification Method," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 1, pp. 20–24, Nov. 2023, doi: 10.25139/inform.v9i1.5465.
[12] C. R. Hassolthine, T. Haryanto, F. Adline Twince Tobing, and M. Ikhwani Saputra, "E-Commerce Product Review Sentiment Analysis: A Comparative Study of Naïve Bayes Classifier and Random Forest Algorithms on Marketplace Platforms," *IJNMT (International Journal of New Media Technology)*, vol. 12, no. 1, pp. 55–60, Jul. 2025, doi: 10.31937/ijnmt.v12i1.4246.
[13] "Deciphering Digital Social Dynamics: A Comparative Study of Logistic Regression and Random Forest in Predicting E-Commerce Customer Behavior," *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 100–113, Jan. 2024, doi: 10.47738/jads.v5i1.155.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

[14] C. AVCI, M. BUDAK, N. YAĞMUR, and F. BALÇIK, "Comparison between random forest and support vector machine algorithms for LULC classification," *International Journal of Engineering and Geosciences*, vol. 8, no. 1, pp. 1–10, Feb. 2023, doi: 10.26833/ijeg.987605.

[15] N. Istiqamah and M. Rijal, "Klasifikasi Ulasan Konsumen Menggunakan Random Forest dan SMOTE," *Journal of System and Computer Engineering (JSCE)*, vol. 5, no. 1, pp. 66–77, Jan. 2024, doi: 10.61628/jsce.v5i1.1061.

[16] T. Becker, A.-J. Rousseau, M. Geubbelmans, T. Burzykowski, and D. Valkenborg, "Decision trees and random forests," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 164, no. 6, pp. 894–897, Dec. 2023, doi: 10.1016/j.ajodo.2023.09.011.

[17] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *Int. J. Data Sci. Anal.*, vol. 20, no. 2, pp. 197–211, Aug. 2025, doi: 10.1007/s41060-024-00509-w.

[18] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[19] Alfandi Safira and F. N. Hasan, "ANALISIS SENTIMEN MASYARAKAT TERHADAP PAYLATER MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER," *ZONAsi: Jurnal Sistem Informasi*, vol. 5, no. 1, pp. 59–70, Jan. 2023, doi: 10.31849/zn.v5i1.12856.

[20] A. R. Abdillah and F. N. Hasan, "Sentiment Analysis of Presidential Candidates Based on Tweets on Social Media Using the Naive Bayes Classifier," *STIKI Informatika Jurnal*, vol. 13, 2023.

[21] H. Ammar, F. Al Gani, M. Rifansyah, and F. N. Hasan, "Perbandingan Tingkat Akurasi Algoritma Naïve Bayes dan Support Vector Machine Dalam Analisis Sentimen Pengguna Aplikasi ShopeePay Pada Google Play Store," *Proceeding of TEKNOKA National Seminar - 9*, vol. 9, 2024.

[22] Meliyawati and F. N. Hasan, "Analisis Sentimen Pengguna Aplikasi CapCut Pada Ulasan di Play Store Menggunakan Metode Naïve Bayes," *KLIK; KAJIAN ILMIAH INFORMATIKA DAN KOMPUTER*, vol. 4, 2024.

[23] B. Gezici Geçer and A. Kolukısa Tarhan, "Explainable AI Framework for Software Defect Prediction," *Journal of Software: Evolution and Process*, vol. 37, no. 4, Apr. 2025, doi: 10.1002/smr.70018.

[24] C. Molnar, *Interpretable Machine Learning*. 2022.

[25] N. Uddin, M. S. Mia, S. Rana, P. Mahmud, and Md. J. Islam, "An Explainable AI-Driven Machine Learning Approach for Student Depression Detection," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, Feb. 2025, pp. 1–6. doi: 10.1109/ECCE64574.2025.11013941.

[26] M. Kulkarni and M. Stamp, "XAI and Android Malware Models," in *Machine Learning, Deep Learning and AI for Cybersecurity*, Cham: Springer Nature Switzerland, 2025, pp. 327–355. doi: 10.1007/978-3-031-83157-7_12.

[27] B. P. Bhuyan and S. Srivastava, "Feature Importance in Explainable AI for Expounding Black Box Models," 2023, pp. 815–824. doi: 10.1007/978-981-19-6634-7_58.

[28] Ninda Rizky Nuraeda, Muhaza Liebenlito, and Taufik Edy Sutanto, "Explainable Sentiment Analysis pada Ulasan Aplikasi Shopee Menggunakan Local Interpretable Model-agnostic Explanations," *Indonesian Journal of Computer Science*, vol. 13, no. 3, Jun. 2024, doi: 10.33022/ijcs.v13i3.3870.

[29] I. F. Rosyid and H. Pramaditya, "Visual Interpretation of Machine Learning Models (Random Forest) for Lung Cancer Risk Classification Using Explainable Artificial Intelligence (SHAP & LIME)," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 4, pp. 2187–2206, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4925.

Vol. 7, No.2, April 2026, Page. 1778-1799
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5642

[30] H. Rathore, H. K. Meena, and P. Jain, "Prediction of EV Energy consumption Using Random Forest And XGBoost," in *2023 International Conference on Power Electronics and Energy (ICPEE)*, IEEE, Jan. 2023, pp. 1–6. doi: 10.1109/ICPEE54198.2023.10060798.