Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2869~2878
e-ISSN: 2548-6861

Sentiment Analysis of E-Commerce Product Reviews on Tokopedia Using Support Vector Machine

Azna Alaiya [1]*, Nurdin [2], Cut Agusniar [3]***
* Departement of Informatics, Universitas Malikussaleh, Lhokseumawe, Indonesia
** Departement of Information Technology, Universitas Malikussaleh, Lhokseumawe, Indonesia
azna.210170050@mhs.unimal.ac.id [1], nurdin@unimal.ac.id [2], cut.agusniar@unimal.ac.id [3]

Article Info

Article history:

Received 2025-08-30
Revised 2025-09-10
Accepted 2025-09-19

Keyword:

E-commerce,
Sentiment Analysis,
Support Vector Machine,
Text Mining,
Tokopedia.

ABSTRACT

This research aims to analyze the performance of Support Vector Machine (SVM) algorithm in classifying sentiment of e-commerce product reviews on the Tokopedia platform using web scraping data of 571 reviews from the 2024 period. The data includes review text variables, publication dates, and usernames processed through text preprocessing (text cleaning, stopword removal, stemming with Sastrawi), auto-labeling using a lexicon-based approach, and TF-IDF feature extraction with optimal parameters (max_features=5000, ngram_range=(1,2)) resulting in 1,187 features. Data splitting was performed using stratified method with proportions of training (80%) and testing (20%) on 461 reviews from binary classification filtering (positive vs negative). The research results demonstrate that Support Vector Machine with linear kernel achieved excellent performance with accuracy 95.70%, precision 95.89%, recall 95.70%, and F1-score 94.89% on the testing set. Despite the imbalanced dataset characteristics (92.4% positive vs 7.6% negative), SVM effectively handled the classification task by identifying negative sentiment with 100% precision and 42.86% recall, demonstrating its robustness in handling skewed data distribution. TF-IDF feature analysis identified the highest discriminative words such as "suitable", "goods", and "good" that are relevant for classifying consumer sentiment towards e-commerce products. The results indicate that SVM algorithm is highly effective for sentiment classification of e-commerce product reviews, making it suitable for practical implementation in automated sentiment analysis systems for online marketplaces.

.

I. Introduction

The online marketplace in Indonesia is rapidly growing and has now become one of the main focuses for internet users and online entertainment platforms. Various types of market services are available and accessible to most Indonesian people. This demonstrates a shift in consumer behavior that increasingly relies on online shopping services as the primary option to fulfill their needs and desires. One of the online shopping platforms that is popular among Indonesian society is Tokopedia [1] [2].

Online sales are currently developing very rapidly because transactions have become increasingly easy with just the palm of our hands, we can already obtain the desired products. However, this development also presents various challenges, such as problems in service, product variation, and product quality. As a result, many customers feel disappointed with conditions like this [3].

Consumer reviews in e-commerce systems are generally considered as important resources that reflect users' experiences, feelings, and desires in purchasing a product. This information can include consumer perspectives that reveal their interests, sentiments, and opinions. Various studies show that people tend to trust fellow users who have similar views on something more. Data mining can be utilized to analyze social media user data that visits e-commerce platforms. This research uses data mining techniques that aim to compare classification results in sentiment analysis from e-commerce customer perspectives that have been written on Tokopedia. Opinion Mining, also known as sentiment analysis, is a process that analyzes text data generated in the form of suggestions, feedback, tweets, and comments. E-commerce portals generate large amounts of data every day in the form of customer reviews. Analyzing data from e-commerce platforms can help online resellers understand customer expectations, provide better shopping experiences, and increase sales [4].

The Support Vector Machine (SVM) algorithm is known to have high performance in text classification, including sentiment analysis, especially when working with data that has a large number of features, such as customer review texts. In sentiment analysis, there are often reviews that have mixed sentiments, compared to several other algorithms, SVM provides stable results in binary classification tasks, such as in determining whether reviews are positive or negative. This is very important to maintain sentiment quality so that it corresponds with the reality in the field [5].

The main problem driving this research is the abundance of unstructured reviews on Tokopedia, including informal language, mixed expressions, and emoticons, which are difficult to analyze manually to understand positive, negative, or neutral sentiment. Additionally, there are challenges in selecting the most accurate algorithm, such as SVM, to process large data with variations of informal language. The purpose of this research is to conduct sentiment analysis on product reviews on the Tokopedia e-commerce platform using the Support Vector Machine (SVM) algorithm. This research aims to identify and classify sentiment from customer reviews, whether they are positive or negative. By using SVM, this research can determine an effective algorithm for processing review data and providing accurate sentiment classification results. The results of this analysis are expected to provide insights for Tokopedia regarding customer perceptions of products, help improve service quality, and provide suggestions for developing better marketing strategies. The use of the Support Vector Machine (SVM) algorithm in this research has several advantages in certain aspects such as accuracy and classification precision, processing speed, ability to handle high dimensions, and suitability for large-scale data.

II. METHODS

This research was conducted on the Tokopedia platform on January 1, 2025, within a processing time of approximately 3 hours. Data collection was performed online using the web scraping method, with the assistance of Python software.

A. Algotrithm Schema SVM

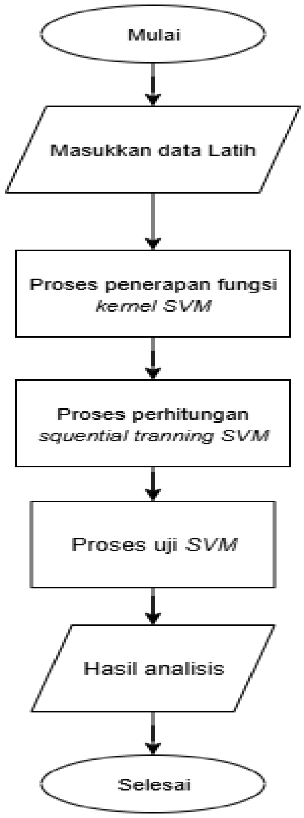

Figure 1. Schema Algorithm SVM

The Support Vector Machine (SVM) algorithm process begins with data preparation for model training, followed by inputting the training data which consists of labeled data used to train the SVM model to recognize patterns within the data. Next, the SVM kernel function is applied to map the input training data into a higher-dimensional space, transforming the data so it can be linearly separated more easily. The process continues with sequential training calculations where the SVM model is trained using sequential learning algorithms to find the optimal hyperplane that separates the data most effectively. After the SVM model is trained, it undergoes testing using previously unseen test data to evaluate the model's performance. The results are then analyzed by examining metrics such as accuracy, precision, and recall to assess the model's effectiveness. Finally, the process concludes, marking the end of the SVM classification procedure.

B. Data Collection Techniques

The data collection process was conducted using the web scraping method to obtain information related to products sold on Tokopedia [6] [7]. The collected data includes dates, usernames, and customer reviews. The scraping script was written in Python using the BeautifulSoup and Selenium libraries. The data successfully collected consisted of 570 review data that were stored in CSV format for further analysis.

C. Data Mining Method

Data is a collection of facts or information that can be measured, calculated, or processed. Data can be in the form of numbers, text, images, or other formats. In the field of computing, data is a digital representation of information that can be stored, processed, and transmitted by computer systems [8]. The data exploration process involves in-depth analysis of data archives to reveal relationships, certain patterns, or regularities in large datasets. The insights generated from this analysis play an important role in improving decision-making processes [9].

As part of Knowledge Discovery in Databases, data mining aims to transform raw data into knowledge through identification of significant patterns in historical data archives [10]. This process enables organizations to make more accurate data-driven decisions. Methodologically, data mining implementation integrates various disciplines including statistical analysis, mathematical models, intelligent systems, and machine learning techniques to filter crucial information from the ocean of data [11].

The data mining process can be divided into several interrelated and interactive stages [9]:

1. Data Cleaning: The dataset cleaning process aims to remove invalid, inaccurate, or unnecessary data. Meanwhile, data consolidation (data integration) involves collecting data from various sources into one integrated database.
2. Data Selection: To make analysis more efficient and precise, only relevant and necessary data will be selected from the database, considering that not all stored information will be used.
3. Data Transformation: To be analyzed effectively in data mining, data needs to be converted or unified into a compatible and standardized format.
4. Mining Process: This process becomes the heart of the activity where various algorithms are implemented to extract potential knowledge contained in the data. Several techniques are available for use, adapted to the type of analysis in data exploration.
5. Pattern Evaluation: Its main function is to detect meaningful relationships in data for further integration into the formed knowledge repository.
6. Knowledge Presentation: This process involves graphical representation and insight delivery about the approaches used, so that analysis results can be presented to users in an intuitive and easily digestible format.

D. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a computational-based analysis process to identify opinions, sentiments, or feelings. This analysis is useful for evaluating the tendency of a sentiment or opinion, whether it tends to be positive or negative [12]. Sentiment analysis is one of the research branches in text mining, which is related to broader fields such as data processing for various specific activities [13]. Sentiment analysis is a field in machine learning that focuses on processing opinions conveyed through text [14]. The purpose of this analysis is to identify sentiment by classifying text based on its polarity, so it can be categorized as positive or negative sentiment [15] [16].

E. Support Vector Machine

The Support Vector Machine (SVM) method is one of the techniques frequently used in the classification process [17]. Support Vector Machine (SVM) is a supervised learning method that analyzes data and recognizes patterns, and is used for classification and regression [18]. SVM is a classification algorithm developed by Boser, Guyon, and Vapnik to separate data into two different classes using an optimal separating line (hyperplane) [19]. Support Vector Machine (SVM) is a supervised learning method used for classification and analysis, which analyzes data and recognizes patterns. Support Vector Machine (SVM) is a classification algorithm that is quite widely used. Linear classifier is the basis of SVM which only classifies two classes, but cases in the real world are generally more than two classes, so it was developed to work on non-linear problems with kernel tricks and mapping data into high-dimensional space [20]. In the SVM algorithm, there is a separating line called hyperplane that is used as a separator for positive and negative sentiment. Support Vector Machine (SVM) is a machine learning method that is based on the principle of Structural Risk Minimization (SRM), with the aim of finding the best hyperplane that separates two classes in the input space [21]. The formula for determining the hyperplane:

$$w \cdot x + b = 0 \rightarrow w_0 + w_1 x_1 + w_2 x_2 \qquad (1)$$

F. Model Evaluation Methodology

This research employs stratified train-test split methodology for model evaluation to ensure representative class distribution in both training and testing sets. The stratified splitting approach was chosen over cross-validation due to the limited dataset size (461 reviews after binary classification) and the need to maintain consistent class proportions given the highly imbalanced nature of the data (92.4% positive vs 7.6% negative sentiment).

The data splitting process uses an 80:20 ratio, resulting in 370 reviews for training (79.9%) and 93 reviews for testing (20.1%). Stratified sampling ensures that both subsets maintain the original class distribution, preventing bias in model evaluation. The random_state parameter was set to ensure reproducibility of results.

Model performance is evaluated using multiple metrics including accuracy, precision, recall, F1-score, and confusion matrix analysis. This single holdout validation approach is appropriate for this exploratory study, though future research with larger datasets could benefit from k-fold cross-validation to provide more robust performance estimates.

III. Result and Discussion

Analyzing sentiment in e-commerce product reviews on the Tokopedia platform presents a complex challenge due to various influencing factors, such as product quality, price, seller service, delivery speed, packaging, customer service experience, consumer expectations, and market trends. This process requires a deep understanding to identify patterns and emotional nuances in review texts, providing accurate insights into consumer perceptions and preventing discrepancies between product quality and consumer expectations. However, conventional approaches relying on manual analysis or simple methods are often inaccurate and unable to handle the complexity of textual data, making timely business decision-making difficult. Therefore, applying the Support Vector Machine (SVM) method offers an appropriate solution, as it can systematically process textual data and deliver objective and reliable sentiment classification based on complex review data patterns. This study focuses on building a sentiment classification model using SVM algorithm on a Tokopedia case study to evaluate its effectiveness in analyzing consumer sentiment from e-commerce product reviews.

A. Data Description Analyst

The research data used in this analysis consist of e-commerce product reviews on the Tokopedia platform collected through a web scraping process in 2024, with a total of 571 observations comprising three main variables. These variables include the product review text (Review), the date the review was written (Date), and the username of the reviewer (Username). This initial dataset provides a comprehensive overview of consumer feedback dynamics on Skintific products sold on the Tokopedia platform during the data collection period, covering various review characteristics and diverse user engagement patterns.

The data underwent systematic filtering through multiple stages: first, text preprocessing reduced the dataset from 571 to 568 reviews due to the removal of three reviews that became empty strings after cleaning (containing only numbers, punctuation, emojis, or stopwords without meaningful content). Subsequently, sentiment classification filtering removed 107 neutral sentiment reviews, resulting in a final binary classification dataset of 461 reviews (427 positive and 34 negative) used for SVM model training and evaluation.

```
📈 STATISTIK DESKRIPTIF
==================================================
Rata-rata panjang ulasan: 63.44 karakter
Ulasan terpendek: 2 karakter
Ulasan terpanjang: 174 karakter
```

	Ulasan	Tanggal	Username
0	Langsung beli 2 karena cocok	2/2/2024	J***l
1	Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% O...	2/3/2024	F***i
2	Nyampenya cepet. Baru nyobain sih jd belum tau...	2/4/2024	Lucy
3	Ringan. Packaging bagus dan rasanya manis bang...	2/5/2024	G***y
4	Pertama kali pake produk Skintific langsung co...	2/6/2024	D***n
...	...	...	...
566	Repeat order. Bagus banget kualitas produk nya...	8/21/2024	Fritzca
567	Bagus bgt coveragenya!! Ga perlu pake foundati...	8/22/2024	Putri
568	paling cocok sama produk ini, harus selalu sto...	8/23/2024	Yuningsih
569	cocok banget ama pelembab yang satu ini, dapat...	8/24/2024	Yuningsih
570	oke banget	8/25/2024	M***i

571 rows × 3 columns

Figure 2. Research Dataset

Based on the descriptive data analysis, it was observed that the length of product reviews shows significant variation, with an average of 63.44 characters per review, ranging from the shortest review with 2 characters to the longest review with 174 characters. This indicates the diversity in the way consumers express their opinions, from short reviews such as "oke banget" to more detailed and comprehensive reviews explaining various aspects of the product. The review data were collected during the period from February to August 2024, reflecting the dynamics of consumer feedback on e-commerce products on the Tokopedia platform during the research period. This variation in review length also reflects different levels of consumer engagement and satisfaction, where longer reviews tend to provide more detailed information about the user's experience with the purchased products.

B. Data Scrapping Analyst

The web scraping process utilized Chrome WebDriver with a maximized browser configuration to ensure all web elements could be accessed optimally. Automatic scrolling was implemented using JavaScript to load dynamic content (infinite scroll) commonly used on modern e-commerce platforms, where the system scrolls the page to the bottom and waits 2 seconds to load new content. An automatic pagination strategy was also applied by detecting and programmatically clicking the "Next Page" button, combined with a random delay of 2–4 seconds to avoid detection as a bot by Tokopedia's anti-scraping system.

Data extraction was performed by identifying specific HTML elements using the appropriate CSS selectors and data attributes. Review data was extracted from <span> elements with the attribute data-testid='lblItemUlasan', usernames from elements with the class name or css-1sxk7zv, and dates from <time> elements or <p> elements with the data-unify attribute. The system was equipped with robust exception handling to manage variations in HTML structure and missing elements, ensuring the scraping process continued even when encountering inconsistencies in the HTML markup.

Date format conversion was a crucial aspect of the scraping process because the Tokopedia platform uses relative date formats such as "Today," "Yesterday," "X days ago," "X weeks ago," "X months ago," and "X years ago." The system implemented the convert_date() function to transform these relative date formats into the standard DD-MM-YYYY format using regular expressions to extract numerical values and calculate backward from the scraping date. This approach ensured temporal consistency in the data, which is essential for time series analysis and consumer sentiment trend tracking.

Data validation and deduplication were performed in real time during the scraping process by comparing the combination of username, date, and review text to prevent duplicate entries. An integrated monitoring system displayed the data collection progress in real time, including the number of reviews found per page and the cumulative total collected. A failsafe mechanism was implemented to stop the process if no new data was found over several iterations, indicating that all available data had been collected or that technical obstacles had been encountered.

The final web scraping process produced a dataset stored in CSV format with three columns: Username, Date, and Review. The scraping process successfully collected 571 reviews covering a wide time period with varying dates, providing a comprehensive representation of consumer sentiment toward Skintific products on the Tokopedia platform. This success confirmed the effectiveness of the applied web scraping methodology and provided a high-quality dataset for the subsequent preprocessing and sentiment analysis stages using the Support Vector Machine (SVM.

C. Dataset Processing Analysis

This sub-section comprehensively explains all stages of data analysis and processing carried out in this research. All computational processes and data analyses were conducted using machine learning software that enables flexible text processing and calculation execution. The main objective of this stage is to prepare the collected e-commerce product review data for sentiment classification, identify patterns and sentiment characteristics within consumer reviews, and apply the Support Vector Machine (SVM) method to build a sentiment classification model for product reviews on the Tokopedia platform.

The data analysis and processing involve several key stages, starting from data import, examination and handling of potential duplicate or noisy text data, to text preprocessing steps such as tokenization, stopword removal, stemming, and transformation of textual data into numerical representations using TF-IDF feature extraction. Subsequently, sentiment labeling and data splitting into training and testing sets are performed, followed by the application of the SVM classification method and evaluation of its performance using metrics such as accuracy, precision, recall, and F1-score. Each of these steps is designed to ensure that the resulting model is accurate, relevant, and capable of delivering reliable sentiment classification to support the analysis of consumer perceptions toward e-commerce products on the Tokopedia platform.

1) Import Library

The first stage in dataset processing is importing all the libraries required for sentiment analysis of e-commerce product reviews. This process includes importing libraries for data manipulation such as Pandas and NumPy, libraries for text preprocessing such as re and string, as well as data visualization libraries using Plotly, which enables the creation of interactive charts. In addition, machine learning libraries from Scikit-learn are also imported, including the Support Vector Machine (SVM) algorithm, evaluation metrics such as accuracy, precision, recall, and F1-score, as well as the TF-IDF Vectorizer for transforming text into numerical representations. For preprocessing Indonesian-language text, the Sastrawi library is used, providing stemming and stopword removal functions tailored to the characteristics of the Indonesian language, ensuring the text cleaning process can be carried out optimally and accurately before sentiment classification is performed using the SVM algorithm.

2) Dataset Labelling

```
CONTOH HASIL LABELING
==================================================
10 Baris Pertama dengan Label:
```

	Ulasan	sentiment_score	sentimen
0	Langsung beli 2 karena cocok	2	Positif
1	Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% O...	1	Positif
2	Nyampenya cepet. Baru nyobain sih jd belum tau...	2	Positif
3	Ringan. Packaging bagus dan rasanya manis bang...	-1	Negatif
4	Pertama kali pake produk Skintific langsung co...	3	Positif
...	...	...	...
566	Repeat order. Bagus banget kualitas produk nya...	3	Positif
567	Bagus bgt coveragenya!! Ga perlu pake foundati...	1	Positif
568	paling cocok sama produk ini, harus selalu sto...	4	Positif
569	cocok banget ama pelembab yang satu ini, dapat...	3	Positif
570	oke banget	2	Positif

571 rows × 3 columns

Figure 3. Labeling Dataset

Based on the results of the auto-labeling process, it can be observed that the lexicon-based system successfully assigned sentiment labels to all 571 product reviews by calculating sentiment scores for each review text. Sample labeling results show variations in sentiment scores, such as the review *"Langsung beli 2 karena cocok"* (Directly bought 2 because it suits me) which received a score of +2 and was labeled as *'Positive'*, the review *"Ringan. Packaging bagus dan rasanya manis banget"* (Lightweight. The packaging is good, and it tastes very sweet) which received a score of -1 and was labeled as *'Negative'*, and the review *"Pertama kali pake produk Skintific langsung cocok"* (First time using Skintific products, and it suits me right away) which received a score of +3 and was labeled as *'Positive'*. This automatic labeling process produced a labeled dataset that is ready to be used in the next stage, namely text preprocessing and the development of classification model using the Support Vector Machine (SVM) algorithm, where the generated sentiment labels will serve as the target variable in the supervised learning process.

3) Processing Text

```
CONTOH HASIL PREPROCESSING
================================================================================

Contoh 1:
Original    : Langsung beli 2 karena cocok...
Cleaned     : langsung beli karena cocok...
Preprocessed: langsung beli cocok...
--------------------------------------------------------------------------------

Contoh 2:
Original    : Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% O...
Cleaned     : ori ori ori ori ori ori ori ori ori ori ori ori ori ori ori ori ori ori...
Preprocessed: ori ori ori ori ori ori ori ori ori ori ori ori ori ori ori ori ori ori...
--------------------------------------------------------------------------------

Contoh 3:
Original    : Nyampenya cepet. Baru nyobain sih jd belum tau efeknya. Smoga cocok...
Cleaned     : nyampenya cepet baru nyobain sih jd belum tau efeknya smoga cocok...
Preprocessed: nyampenya cepet baru nyobain sih jd tau efek smoga cocok...
--------------------------------------------------------------------------------
```

Figure 4. Processing Text

Based on the examples of preprocessing results presented, the effectiveness of each text cleaning stage in preparing the data for sentiment classification can be clearly observed. The first example shows the transformation from *"Langsung beli 2 karena cocok..."* to *"langsung beli cocok..."* after removing numbers and punctuation marks, while the second example demonstrates how repetitive text like *"Ori 100% Ori 100%..."* becomes simpler as *"ori ori ori..."* after the cleaning process. The stopword removal and stemming processes are also evident in the third example, where *"Nyampenya cepet. Baru nyobain sih jd belum tau efeknya. Smoga cocok..."* is transformed into *"nyampenya cepet baru nyobain sih jd tau efek smoga cocok..."* by eliminating irrelevant words such as *"belum"* and unnecessary punctuation. Overall, the preprocessing results show that the text becomes more consistent and structured, with significant noise reduction while preserving key terms essential for identifying sentiment, ensuring the data is ready for the next stage, namely feature extraction using TF-IDF and the implementation of the Support Vector Machine (SVM) algorithm.

```
STATISTIK PREPROCESSING
==================================================
Rata-rata panjang teks original: 63.44 karakter
Rata-rata panjang teks clean: 60.85 karakter
Rata-rata panjang teks preprocessed: 50.36 karakter
Jumlah data setelah menghapus teks kosong: 568 dari 571 ulasan
```

	Ulasan	ulasan_clean	ulasan_preprocessed	sentimen
0	Langsung beli 2 karena cocok	langsung beli karena cocok	langsung beli cocok	Positif
1	Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% O...	ori ori ori ori ori ori ori ori ori ori ori or...	ori ori ori ori ori ori ori ori ori ori ori or...	Positif
2	Nyampenya cepet. Baru nyobain sih jd belum tau...	nyampenya cepet baru nyobain sih jd belum tau ...	nyampenya cepet baru nyobain sih jd tau efek s...	Positif
3	Ringan. Packaging bagus dan rasanya manis bang...	ringan packaging bagus dan rasanya manis bange...	ringan packaging bagus rasa manis banget kalo ...	Negatif
4	Pertama kali pake produk Skintific langsung co...	pertama kali pake produk skintific langsung co...	pertama kali pake produk skintific langsung co...	Positif
...	...	...	...	...
566	Repeat order. Bagus banget kualitas produk nya...	repeat order bagus banget kualitas produk nya ...	repeat order bagus banget kualitas produk nya ...	Positif
567	Bagus bgt coveragenya!! Ga perlu pake foundati...	bagus bgt coveragenya ga perlu pake foundation...	bagus bgt coveragenya ga perlu pake foundation...	Positif
568	paling cocok sama produk ini, harus selalu sto...	paling cocok sama produk ini harus selalu stoc...	paling cocok sama produk harus selalu stock ru...	Positif
569	cocok banget ama pelembab yang satu ini, dapat...	cocok banget ama pelembab yang satu ini dapat ...	cocok banget ama pelembab satu dapat harga murah	Positif
570	oke banget	oke banget	oke banget	Positif

568 rows × 4 columns

Figure 5. Text Processing Statistics

Preprocessing statistics demonstrate the effectiveness of each text-cleaning stage, where the average text length progressively decreased from 63.44 characters in the original text to 60.85 characters after cleaning, and finally to 50.36 characters after complete preprocessing. The reduction in the number of data points from 571 reviews to 568 reviews was due to three reviews becoming empty strings after the comprehensive preprocessing process. These reviews likely contained only elements such as numbers, punctuation marks, emojis, or stopwords that were removed, without any meaningful words left for further processing. The removal of empty data is crucial to ensure the quality of the dataset to be used in the feature extraction and machine learning stages, as empty texts provide no useful information for sentiment classification and may negatively affect the performance of the Support Vector Machine (SVM) algorithm in identifying sentiment patterns from e-commerce product reviews.

4) Filtering Data

	Ulasan	ulasan_preprocessed	sentimen
0	Langsung beli 2 karena cocok	langsung beli cocok	Positif
1	Ori 100% Ori 100% Ori 100% Ori 100% Ori 100% O...	ori ori ori ori ori ori ori ori ori ori ori or...	Positif
2	Nyampenya cepet. Baru nyobain sih jd belum tau...	nyampenya cepet baru nyobain sih jd tau efek s...	Positif
3	Ringan. Packaging bagus dan rasanya manis bang...	ringan packaging bagus rasa manis banget kalo ...	Negatif
4	Pertama kali pake produk Skintific langsung co...	pertama kali pake produk skintific langsung co...	Positif
...	...	...	...
566	Repeat order. Bagus banget kualitas produk nya...	repeat order bagus banget kualitas produk nya ...	Positif
567	Bagus bgt coveragenya!! Ga perlu pake foundati...	bagus bgt coveragenya ga perlu pake foundation...	Positif
568	paling cocok sama produk ini, harus selalu sto...	paling cocok sama produk harus selalu stock ru...	Positif
569	cocok banget ama pelembab yang satu ini, dapat...	cocok banget ama pelembab satu dapat harga murah	Positif
570	oke banget	oke banget	Positif

463 rows × 3 columns

Figure 6. Data Filtering Result

The data filtering process for binary sentiment classification involved systematic removal of neutral sentiment reviews to create a focused positive-negative classification task. From the initial 568 preprocessed reviews, the lexicon-based auto-labeling system identified three sentiment categories: 427 positive reviews (75.2%), 107 neutral reviews (18.8%), and 34 negative reviews (6.0%). For binary classification purposes, the 107 neutral sentiment reviews were excluded from the analysis to create a clear distinction between positive and negative sentiments, resulting in a final dataset of 461 reviews. This filtering approach is standard practice in binary sentiment analysis research, where neutral sentiment often represents ambiguous cases that can introduce classification noise and reduce model performance clarity. The final binary dataset exhibits significant class imbalance with 427 positive reviews (92.6%) and 34 negative reviews (7.4%), reflecting the typical distribution pattern in e-commerce product reviews where satisfied customers tend to provide more feedback than dissatisfied ones. This imbalance characteristic presents both challenges and opportunities for evaluating the SVM algorithm's robustness in handling skewed data distributions commonly encountered in real-world e-commerce sentiment analysis applications.

5) Splitting Data

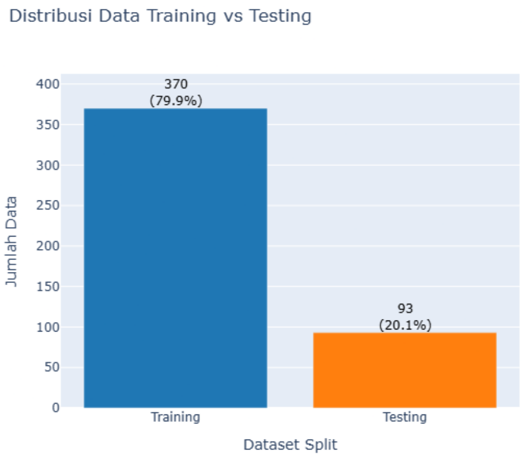

Figure 7. Splitting Data

The stratified train-test split was implemented to ensure optimal evaluation methodology while maintaining class distribution integrity. The splitting process divided the 461 filtered reviews into training (370 reviews, 79.9%) and testing (93 reviews, 20.1%) sets using stratified sampling to preserve the original class proportions in both subsets. from sklearn.model_selection import train_test_split # Stratified split to maintain class distribution X_train, X_test, y_train, y_test = train_test_split(X_tfidf, y_binary, test_size=0.2, random_state=42, stratify=y_binary) The stratified approach ensures that the severe class imbalance (92.4% positive vs 7.6% negative) is consistently represented across both training and testing datasets, preventing evaluation bias that could occur with random sampling. This methodology provides reliable performance assessment while acknowledging the constraints of the available dataset size.

6) TF-IDF Vectorizer Inisialization

```
📊 HASIL FEATURE EXTRACTION
==================================================
Jumlah fitur yang diextract: 1187
Shape training data: (370, 1187)
Shape testing data: (93, 1187)
Sparsity training data: 98.94%
```

Figure 8. TF-IDF Vectorizer Inisialization

Based on the displayed feature extraction results, the TF-IDF process successfully extracted 1,187 unique features from the product review dataset vocabulary, which is far below the maximum limit of 5,000 features that had been set, indicating that the review data has limited but still informative word diversity. The training data dimension is (370, 1,187) and the testing data dimension is (93, 1,187), confirming that the feature transformation is consistent across both data subsets with the same number of features. The training data sparsity of 98.83% indicates that most elements in the TF-IDF matrix have zero values, which is a normal characteristic in text mining where each document contains only a small portion of the total available vocabulary. This high sparsity level demonstrates data representation efficiency and confirms that the TF-IDF parameters have been properly configured to eliminate noise while retaining relevant information for sentiment classification using the Support Vector Machine (SVM) algorithm.

```
📈 TOP 20 FITUR DENGAN TF-IDF SCORE TERTINGGI
============================================================
 1. cocok                  : 0.056809
 2. barang                 : 0.041493
 3. bagus                  : 0.039447
 4. sesuai                 : 0.034306
 5. cepat                  : 0.033855
 6. kulit                  : 0.033359
 7. moga                   : 0.031194
 8. kirim                  : 0.029954
 9. moga cocok             : 0.023089
10. nya                    : 0.020769
11. baik                   : 0.020714
12. banget                 : 0.020650
13. kirim cepat            : 0.020375
14. aman                   : 0.020267
15. terima                 : 0.019092
16. packing                : 0.018193
17. produk                 : 0.017451
18. barang sesuai          : 0.015452
19. mantap                 : 0.015260
20. cocok kulit            : 0.014774
```

Figure 9. Top 20 Fiture with Highest TF IDF Score

Analysis of the top 20 features with the highest TF-IDF scores provides deep insights into the most discriminative words in classifying the sentiment of e-commerce product reviews. Top features such as "cocok" (0.056809), "barang" (0.041493), and "bagus" (0.039447) indicate words with high discriminative value for distinguishing positive and negative sentiments in the context of product reviews. The presence of words such as "sesuai," "cepat," "kulit," and bigrams such as "moga cocok," "kirim cepat," and "barang sesuai" confirms that TF-IDF successfully captures not only single words but also word combinations that provide more specific sentiment context. The dominance of words related to product quality ("bagus," "sesuai"), purchase experience ("cepat," "kirim," "packing"), and customer satisfaction ("cocok," "mantap") in the top feature list shows that the vectorizer has successfully identified the key aspects influencing consumer sentiment, making these feature representations informative inputs for the Support Vector Machine (SVM) algorithm in classifying review sentiments.

7) Training Result SVM Model

```
📈 CONFUSION MATRIX SVM (Testing Set)
========================================
                 Prediksi
               Pos    Neg
Aktual   Pos     3      4
         Neg     0     86

📋 CLASSIFICATION REPORT SVM
============================================================
              precision    recall  f1-score   support

     Negatif       1.00      0.43      0.60         7
     Positif       0.96      1.00      0.98        86

    accuracy                           0.96        93
   macro avg       0.98      0.71      0.79        93
weighted avg       0.96      0.96      0.95        93
```

Figure 10. Confusion Matrix & Evaluation Matrix

The model evaluation follows a rigorous single holdout validation approach using the stratified train-test split described previously. This methodology ensures unbiased performance assessment by maintaining consistent class distribution across training and testing phases, which is crucial given the dataset's imbalanced characteristics.

The evaluation results of the Support Vector Machine (SVM) model show excellent performance in classifying the sentiment of e-commerce product reviews with an overall accuracy of 96% on the testing set. The confusion matrix indicates that out of 7 actual negative reviews, 4 reviews were incorrectly predicted as positive and 3 reviews were correctly predicted as negative, while out of 86 actual positive reviews, all were correctly classified as positive with no false negatives. The classification report shows that the SVM model achieved perfect precision (1.00) for the negative class and very high precision (0.96) for the positive class, with recall of 0.43 for the negative class and perfect recall (1.00) for the positive class, resulting in an F1-score of 0.60 for the negative class and 0.98 for the positive class. The imbalanced performance between the two classes reflects the characteristics of the imbalanced dataset, where the model tends to be more accurate in identifying positive sentiment than negative sentiment. However, overall, the SVM model demonstrates solid capability with a weighted average F1-score of 0.95, which will serve as the baseline for comparison with the Naive Bayes algorithm.

D. System Implementation

The implementation includes the development of a real-time analysis interface that allows input of new product reviews and produces outputs in the form of sentiment classification (positive and negative,), probability distribution for each sentiment class, and a confidence score to support business decision-making. The system is equipped with a text preprocessing mechanism that includes tokenization, stopword removal, stemming, and feature extraction using TF-IDF or Bag of Words, as well as robust error handling to accommodate various review input formats. The following sections will elaborate on each component of the sentiment analysis system, explaining its functionality in the context of e-commerce applications, and how the visualization of classification results and model performance evaluation from both the SVM and Naive Bayes algorithms is presented to provide in-depth insights into each algorithm’s capability in analyzing consumer sentiment based on the available Tokopedia product review data.

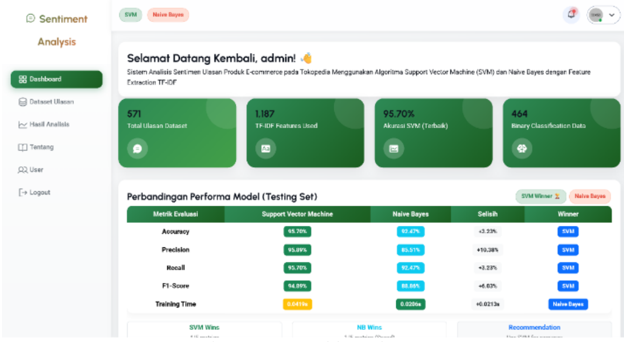

Figure 11. Dashboard Page

The system dashboard page serves as the central control and monitoring hub, providing a comprehensive overview of the sentiment analysis system's performance through an intuitive and informative interface, where the top section displays a welcome message to the administrator and a complete system description, followed by four key performance indicator (KPI) cards showing essential metrics including the total dataset of 571 reviews, the number of TF-IDF features used (1,187), the best SVM accuracy (95.70%), and the total number of binary classification data points (464), while the main component of the dashboard is the "Model Performance Comparison (Testing Set)" table presenting a comparative evaluation between the Support Vector Machine and Naive Bayes algorithms across multiple evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and Training Time, with a Difference column highlighting the performance gap and a Winner column identifying the best-performing algorithm for each metric, and the bottom section summarizing the total wins for each algorithm along with system recommendations, whereas the left sidebar provides a navigation menu for accessing various system features including Dashboard, Review Dataset, Analysis Results, About, Users, and Logout, with visual indicators showing which algorithm (SVM or Naive Bayes) is currently active via a toggle button at the top.

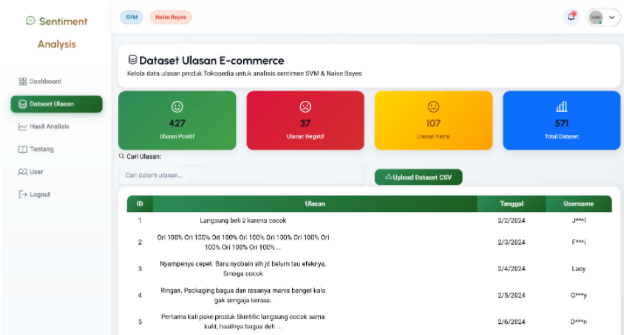

Figure 12. Dataset Page

The "E-commerce Review Dataset" page functions as a data management interface providing a comprehensive overview and access to the Tokopedia product review dataset used for training and testing the sentiment analysis model, where the top section displays four color-coded statistical cards showing the sentiment distribution in the dataset—427 positive reviews (green), 37 negative reviews (red), 107 neutral reviews (yellow), and a total of 571 datasets (blue)—offering a clear visual representation of data composition and balance, while the interface includes a "Search Review" feature enabling administrators to filter or search for specific review content and an "Upload Dataset CSV" button to facilitate the addition of new data into the system in Comma Separated Values format, with the main component being the data table displaying the dataset structure through ID, Review, Date, and Username columns, where review content appears in full text reflecting the original language and writing styles of Tokopedia users, usernames are anonymized to maintain user privacy, and the date column provides timestamps for each review, enabling temporal analysis of consumer sentiment trends.

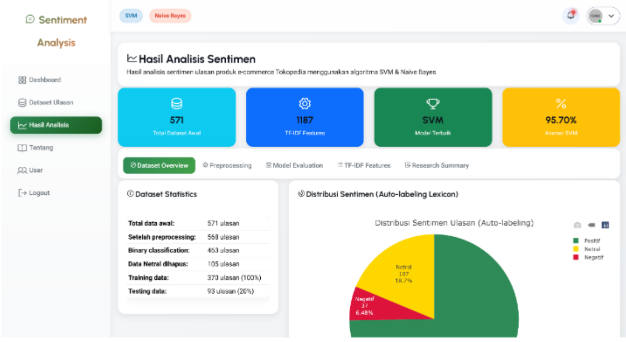

Figure 13. Analytics Page

The "Sentiment Analysis Results" page is a comprehensive analytics dashboard that presents the processing and evaluation results of the sentiment analysis model for Tokopedia e-commerce product reviews with structured and informative data visualizations, where the top section displays four key metrics cards covering the total initial dataset (571), the number of TF-IDF features used (1187), the best-performing model (SVM), and the highest accuracy achieved (95.70%) to provide a quick overview of the system's overall performance, while the interface is equipped with tab navigation that allows users to explore various aspects of the analysis such as "Dataset Overview," "Preprocessing," "Model Evaluation," "TF-IDF Features," and "Research Summary," each presenting specific information on the stages and results of the analysis, with the left section showing "Dataset Statistics" in a table format detailing the number of data points at each preprocessing stage starting from the total initial data, after preprocessing, binary classification, removed neutral data, to the training and testing data split with clear percentages, while the right section presents "Sentiment Distribution (Auto-labeling Lexicon)" in a colored pie chart visualizing the proportions of positive (green) and negative (red) sentiments along with specific percentages to provide an easy-to-understand visual understanding of the sentiment composition in the dataset.

IV. CONCLUSION

This study concludes that the Support Vector Machine (SVM) algorithm with a linear kernel demonstrates excellent performance in classifying e-commerce product review sentiments on Tokopedia, achieving 95.70% accuracy, 95.89% precision, 95.70% recall, and 94.89% F1-score through a rigorous stratified train-test split evaluation methodology that maintained class distribution integrity across training and testing phases. Despite the highly imbalanced dataset characteristics (92.4% positive vs 7.6% negative), SVM effectively handled the classification challenge by identifying negative sentiment with 100% precision and 42.86% recall, demonstrating its robustness in managing skewed data distribution, though future research should incorporate additional evaluation metrics such as AUC-ROC scores and macro-averaged metrics along with techniques like SMOTE or ensemble methods for more comprehensive performance assessment. The comprehensive preprocessing pipeline using Sastrawi and TF-IDF successfully extracted 1,187 discriminative features from the systematically filtered dataset (reduced from 571 initial reviews to 461 binary classified reviews through preprocessing and neutral sentiment removal), while the lexicon-based auto-labeling system facilitated efficient data labeling with appropriate sentiment categories. This research provides valuable insights for Tokopedia and other Indonesian e-commerce platforms in understanding customer satisfaction, developing targeted marketing strategies, and improving service quality based on consumer feedback patterns, contributing empirical evidence of SVM's effectiveness in handling imbalanced text classification tasks and demonstrating the practical potential for automated sentiment analysis systems in real-world e-commerce applications requiring accurate sentiment classification of customer reviews.

REFERENCES

[1] M. Aulia and A. Hermawan, "Analisis Perbandingan Algoritma SVM, Naïve Bayes, dan Perceptron untuk Analisis Sentimen Ulasan Produk Tokopedia," *Jurnal Media Informatika Budidarma*, vol. 7, no. 4, p. 1850, 2023, doi: 10.30865/mib.v7i4.6839.

[2] S. S. Muna, Nurdin, and Taufiq, "Tokopedia and Shopee Marketplace Performance Analysis Using Metrix Google Lighthouse," *Int. J. Eng. Sci. Inf. Technol.*, vol. 2, no. 3, pp. 106–110, 2022, doi: 10.52088/ijesty.v1i4.312.

[3] A. Ernawati, A. O. Sari, S. N. Sofyan, M. Iqbal, and R. F. W. Wijaya, "Implementasi Algoritma Naïve Bayes dalam Menganalisis Sentimen Review Pengguna Tokopedia pada Produk Kesehatan," *Bulletin of Information Technology (BIT)*, vol. 4, no. 4, pp. 533–543, 2023, doi: 10.47065/bit.v4i4.1090.

[4] R. A. E. V. T. Sapanji, D. Hamdani, and P. Harahap, "Sentiment Analysis of the Top 5 E-commerce Platforms in Indonesia using Text Mining and Natural Language Processing (NLP)," *Journal of Applied Informatics and Computing*, vol. 7, no. 2, pp. 202–211, 2023, doi: 10.30871/jaic.v7i2.6517.

[5] N. K. Putri, A. V. Vitianingsih, S. Kacung, A. L. Maukar, and V. Yasin, "Sentiment Analysis of Brand Ambassador Influence on Product Buyer Interest Using KNN and SVM," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 7, no. 2, p. 327, 2024, doi: 10.24014/ijaidm.v7i2.29469.

[6] Nurdin, Bustami, M. Hutomi, M. Elveny, and R. Syah, "Implementation of the bfs algorithm and web scraping techniques for online shop detection in Indonesia," *J. Theor. Appl. Inf. Technol.*, vol. 99, no. 12, pp. 2878–2889, 2021.

[7] Nurdin, M. Hutomi, M. Qamal, and B. Bustami, "Sistem Pengecekan Toko Online Asli atau Dropship pada Shopee Menggunakan Algoritma Breadth First Search," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 6, pp. 1117–1123, 2020, doi: 10.29207/resti.v4i6.2514.

[8] D. Astika and N. Nurdin, "Penerapan Data Mining Untuk Menganalisis Penjualan Barang Dengan Menggunakan Metode Apriori Pada Supermarket Sejahtera Lhokseumawe," 2019.

[9] B. W. K. Nurdin, "Implementasi Data Mining Untuk Mengklasifikasi Data Nasabah Pt. Adira Finance Aceh Tengah Menggunakan Algoritma C4.5," *Jurnal Sistem Informasi Kaputama (JSIK)*, vol. 1, no. 1, 2017.

[10] Fauziah, Dedy Hartama, and Irfan Sudahri Damanik, "Analisa Kepuasan Pelanggan Menggunakan Klasifikasi Data Mining," *Jurnal Penerapan Kecerdasan Buatan*, 2020.

[11] N. Nurdin, M. Suhendri, Yesy Afrilia, and R. Rizal, "SISTEMASI: Jurnal Sistem Informasi Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (Nbc)," 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[12] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 2, no. 5, pp. 178–185, 2022, doi: 10.30865/klik.v2i5.362.

[13] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional," *Jurnal Tekno Kompak*, vol. 15, no. 1, p. 131, 2021, doi: 10.33365/jtk.v15i1.744.

[14] V. A. Zahrah, Nurdin, and Risawandi, "Sentiment Analysis of Google Maps User Reviews on the Play Store Using Support Vector Machine and Latent Dirichlet Allocation Topic Modeling," *Int. J. Eng. Sci. Inf. Technol.*, vol. 4, no. 4, pp. 87–100, 2024, doi: 10.52088/ijesty.v4i4.580.

[15] B. Samudera, N. Nurdin, and H. A. Aidilof, "Sentiment Analysis of User Reviews on BSI Mobile and Action Mobile Applications on the Google Play Store Using Multinomial Naive Bayes Algorithm," vol. 4, no. 4, pp. 101–112, 2024.

[16] A. Anggara, Nurdin, and R. Meiyanti, "Sentiment Analysis of the MK Decision Trial of the Result of the 2024 President and Vice President General Election on Social Media X Using the Support Vector Machine Method," *Int. J. Eng. Sci. Inf. Technol.*, vol. 4, no. 4, pp. 125–134, 2024, doi: 10.52088/ijesty.v4i4.591.

[17] Styawati, Andi Nurkholis, Zaenal Abidin, and Heni Sulistiani, "Optimasi Parameter Support Vector Machine Berbasis Algoritma Firefly Pada Data Opini Film," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 5, pp. 904–910, 2021, doi: 10.29207/resti.v5i5.3380.

[18] Oryza Habibie Rahman, Gunawan Abdillah, and Agus Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 17–23, 2021, doi: 10.29207/resti.v5i1.2700.

[19] A. P. Ruise, A. S. Mashuri, M. Sulaiman, and F. Rahman, "Studi Komparasi Metode Svm, Logistic Regresion Dan Random Forest Clasifier Untuk Mengklasifikasi Fake News di Twitter," *J I M P - Jurnal Informatika Merdeka Pasuruan*, vol. 7, no. 2, p. 64, Sep. 2023, doi: 10.51213/jimp.v7i2.472.

[20] A. Ginting, Nurdin, and C. Agusniar, "Performance Analysis of SVM and Linear Regression for Predicting Tourist Visits in North Sumatera," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 101–108, 2025, doi: 10.52088/ijesty.v5i1.667.

[21] I. M. D. P. Asana and N. P. D. T. Yanti, "Sistem Klasifikasi Pengajuan Kredit Dengan Metode Support Vector Machine (SVM) I Made Dwi Putra Asana," *Jurnal Sistem Cerdas*, vol. 6, no. 2, pp. 123–133, 2023