JTH: Journal of Technology and Health *ISSN 3025-1168*
Vol. 4, No. 1, 2026, pp. 610 ~ 632
Open Acces: *https://doi.org/10.61677/jth.v4i1.891*

BENCHMARKING MACHINE LEARNING MODELS FOR INDONESIAN E-COMMERCE SENTIMENT CLASSIFICATION

Muhammad Fairuzabadi[1], Indo Intan[2*], Sitti Suhada[3]

[1]Study Program of Informatics, Faculty of Science and Technology, Universitas PGRI Yogyakarta, Indonesia
[2]*Study Program of Informatics Engineering, Universitas Dipa Makassar, Indonesia
[3]Department of Informatics Engineering, Faculty of Engineering, Universitas Negeri Gorontalo, Indonesia
fairuz@upy.ac.id[1], indo.intan@undipa.ac.id[2], sittisuhada@ung.ac.id[3]*
**Corresponding author*
Received June 30, 2026; Revised July 9, 2026; Accepted July 25, 2026; Published July 31, 2026

ABSTRACT

Indonesia's expanding e-commerce sector generates a growing volume of customer-written product reviews that can reveal both satisfaction and dissatisfaction. Automatically determining sentiment in these reviews is nevertheless difficult because marketplace language commonly includes informal wording, inconsistent spelling, brief statements, and domain-specific terms. This research benchmarks conventional machine learning methods for classifying the sentiment of Indonesian e-commerce reviews in the PRDECT-ID dataset. The data were obtained from Tokopedia and contain sentiment and emotion annotations. Following preprocessing, the experiment used 5,305 reviews, comprising 2,752 negative and 2,553 positive instances. The processing pipeline included case folding, text cleaning, normalization, tokenization, selective removal of stopwords, and Term Frequency-Inverse Document Frequency (TF-IDF) feature construction. Multinomial Naive Bayes, Support Vector Machine, and Random Forest were then evaluated under the same experimental configuration. The TF-IDF and Support Vector Machine combination produced the strongest results, reaching 0.9595 accuracy, 0.9594 macro-F1, and 0.9595 weighted-F1. These findings establish a reproducible reference point for sentiment classification in Indonesian e-commerce reviews.

Keywords: *Sentiment analysis, e-commerce reviews, PRDECT-ID, TF-IDF, machine learning.*

INTRODUCTION

The continued growth of electronic commerce has made product reviews a major source of information created directly by customers. Through these reviews, users communicate satisfaction or dissatisfaction, describe product and service experiences, and indicate purchase-related intentions. Computational studies generally examine such content through sentiment analysis, a natural language processing task concerned with recognizing subjective opinions and assigning polarity labels, typically positive or negative (Mao et al., 2024; Medhat et al., 2014; Pang & Lee, 2008; Wankhade et al., 2022). Because the number of reviews is increasing rapidly, manual examination is no longer efficient or scalable. Automated sentiment classification is therefore important for summarizing customer opinions, tracking satisfaction, and informing data-driven decisions in online marketplaces (Q. Li et al., 2022; Mao et al., 2024; Wankhade et al., 2022).

Approaches to sentiment analysis have progressed from lexicon-oriented techniques toward machine learning, deep learning, and transformer architectures. Initial work relied heavily on sentiment dictionaries, statistical term weighting, and supervised classifiers, whereas later studies sought to model semantic and contextual relationships through neural and transformer-based representations (Minaee et al., 2021; Taboada et al., 2011; Zhang et al., 2018). Even with the growing use of deep learning, conventional machine learning remains valuable for text classification because it requires relatively modest computation, is straightforward to reproduce, and can perform effectively on moderately sized datasets when paired with an appropriate feature representation (Kowsari et al., 2019; J. Li et al., 2022; Sebastiani, 2002).

Indonesian-language sentiment classification introduces additional linguistic difficulties. Reviews posted in Indonesian marketplaces often contain colloquial wording, spelling inconsistencies, slang, abbreviations, mixed languages, reduplicated forms, and vocabulary specific to online shopping. Unless preprocessing and feature construction account for these characteristics, predictive performance may decline. Prior findings indicate that preprocessing choices can materially influence classification accuracy, particularly for high-dimensional text representations (Kowsari et al., 2019; Q. Li et al., 2022; Uysal & Gunal, 2014). To address this issue, the present study compares several conventional classifiers within one controlled experimental environment. Every model uses the same dataset, preprocessing sequence, TF-IDF representation, splitting procedure, and evaluation criteria, allowing the most dependable baseline for Indonesian e-commerce sentiment classification to be identified more fairly.

The empirical analysis is based on PRDECT-ID, an openly accessible dataset of Indonesian product reviews collected from Tokopedia. It covers several product categories and supplies both sentiment and emotion annotations, enabling research on either classification task (Harmandini & L, 2024). Unlike datasets dominated by general social media posts, PRDECT-ID consists of marketplace-specific review text. It consequently offers a more appropriate domain setting for assessing sentiment classifiers intended for Indonesian e-commerce applications.

Multinomial Naive Bayes, Support Vector Machine, and Random Forest are well-established algorithms for text categorization. Naive Bayes commonly serves as an efficient probabilistic reference model for frequency-based features. Support Vector Machine is particularly effective when observations are represented in sparse, high-dimensional spaces such as those generated by TF-IDF (Cortes & Vapnik, 1995). Random Forest follows an ensemble strategy by aggregating multiple decision trees to improve predictive robustness (Breiman, 2001). TF-IDF remains widely adopted in information retrieval and text classification because it quantifies each term according to its relevance within a document and its distribution across the corpus (Forman, 2003; Salton & Buckley, 1988).

Although neural models have achieved strong sentiment-analysis results, conventional classifiers should still be assessed as interpretable and reproducible baselines. A credible baseline is necessary for determining whether the additional

complexity of an advanced model yields a meaningful gain. Evaluation must also extend beyond accuracy, particularly when class frequencies differ. Precision, recall, F1-score, macro-F1, weighted-F1, and the confusion matrix collectively describe performance more comprehensively than a single accuracy value (He & Garcia, 2009; Saito & Rehmsmeier, 2015; Sokolova & Lapalme, 2009).

Whereas much prior work has emphasized broad sentiment datasets, social media language, or deep learning architectures, this study constructs a reproducible benchmark specifically for Indonesian e-commerce reviews in PRDECT-ID. Its contribution is a controlled comparison of probabilistic, margin-based, and tree-ensemble classifiers under an identical TF-IDF representation and validation protocol. The analysis additionally examines n-gram sensitivity, confusion-matrix patterns, influential features, and misclassified examples. Accordingly, the study contributes both comparative performance evidence and a transparent baseline for subsequent research on Indonesian sentiment analysis.

The objective of this study is to benchmark Multinomial Naive Bayes, Support Vector Machine, and Random Forest for sentiment classification of Indonesian e-commerce reviews in the PRDECT-ID dataset.

The study is guided by the following research questions:

1. How effectively do Multinomial Naive Bayes, Support Vector Machine, and Random Forest classify sentiment in Indonesian e-commerce product reviews?
2. Which classifier performs best according to accuracy, precision, recall, F1-score, macro-F1, and weighted-F1?
3. Which misclassification patterns are observable in the resulting confusion matrices?

RESEARCH METHOD

A quantitative experimental design was adopted to compare conventional machine learning algorithms for sentiment classification of Indonesian e-commerce reviews. This design enables model outcomes to be measured objectively through numerical performance indicators. It is appropriate for benchmarking because competing algorithms can be evaluated under controlled conditions and assessed with identical metrics. Consequently, differences among the models can be interpreted from measurable classification evidence rather than subjective judgment.

The research was organized as a controlled benchmark in which every classifier was trained and assessed with an identical dataset, preprocessing pipeline, feature representation, data split, validation procedure, and collection of evaluation measures. Standardizing these elements reduces the possibility that performance differences arise from inconsistent experimental settings. It also improves reproducibility because other researchers can repeat the workflow with the same data and parameter configuration.

The workflow comprised six principal stages: acquiring the dataset, understanding its contents, preprocessing the text, constructing features, training the classifiers, and evaluating their outputs. PRDECT-ID, a public Indonesian product-review dataset, served as the data source (Harmandini & L, 2024). Review text was represented through Term

Frequency-Inverse Document Frequency (TF-IDF), and classification was performed using Multinomial Naive Bayes, Support Vector Machine, and Random Forest. These conventional algorithms were selected to establish an interpretable and repeatable baseline before future work considers more computationally complex neural or transformer models. Figure 1 summarizes the complete research framework.

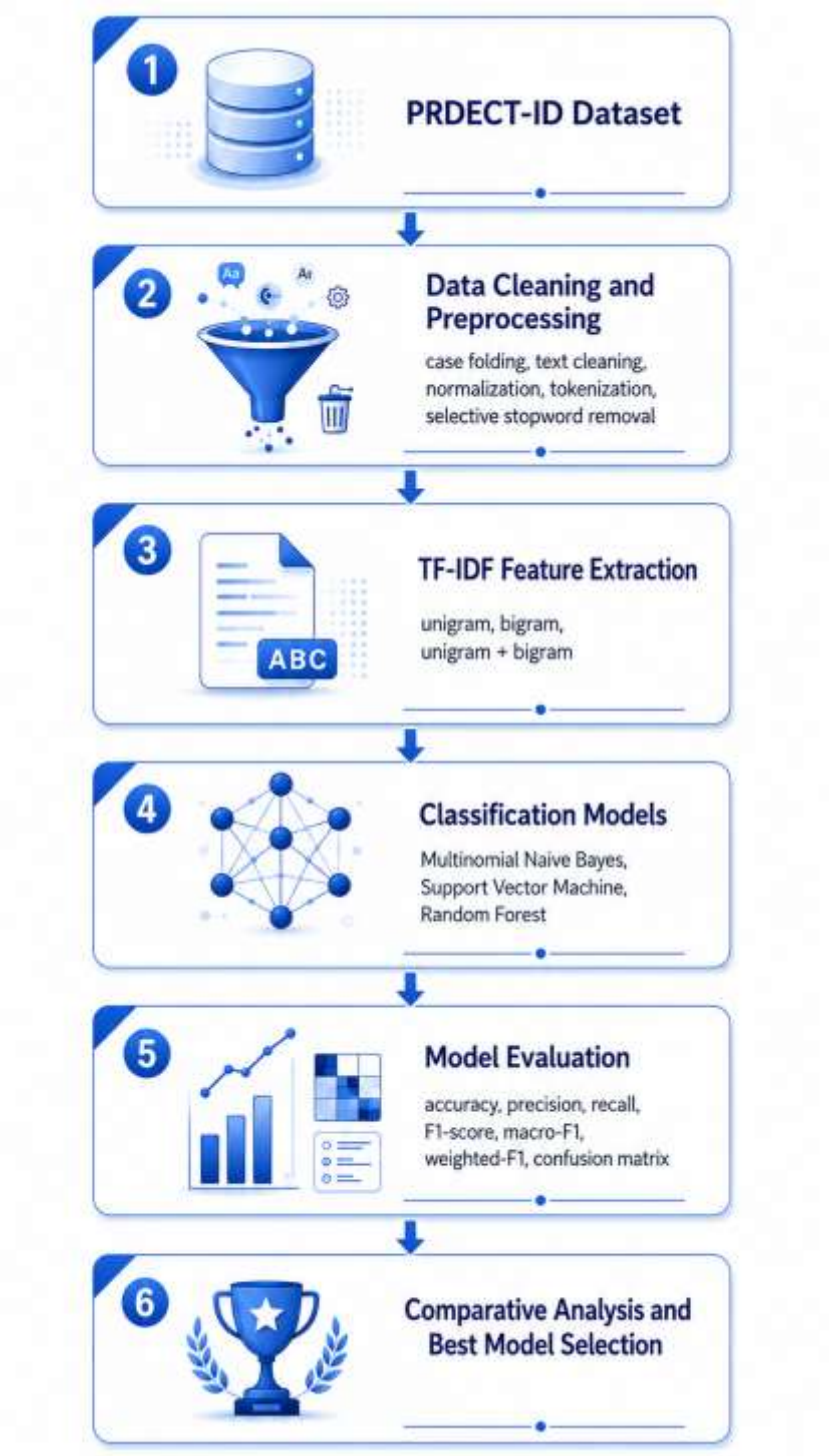

Figure 1. Experimental Workflow for Sentiment Classification

Figure 1 presents the experiment as a sequence of connected stages. The procedure starts with the PRDECT-ID dataset and continues with cleaning and preprocessing to remove noise and harmonize the review text. The resulting corpus is converted into TF-IDF vectors under unigram, bigram, and combined unigram-bigram configurations. Multinomial Naive Bayes, Support Vector Machine, and Random Forest are subsequently trained as the candidate classifiers. In the final stage, their predictions are compared using accuracy, precision, recall, F1-score, macro-F1, weighted-F1, and confusion-matrix analysis to identify the strongest model.

Dataset

PRDECT-ID was used as the empirical dataset. It contains Indonesian product reviews gathered from Tokopedia across multiple product categories and includes annotations for sentiment and emotion (Harmandini & L, 2024). For the present binary classification task, customer-review text functioned as the predictor input and the sentiment annotation as the target. Emotion annotations were retained in the dataset but were not treated as the primary outcome because the analysis focused exclusively on sentiment polarity. As summarized in Table 1, the dataset comprises five principal attributes, each of which was assigned a specific analytical role in this study.

Table 1. Dataset Fields and Their Analytical Roles

No.	Attribute	Description	Usage
1	Product Category	Category of the reviewed product	Descriptive analysis
2	Product Name	Name of the reviewed product	Not used as the main feature
3	Customer Review	Review text written by customers	Main input feature
4	Sentiment	Sentiment label of the review	Target variable
5	Emotion	Emotion label of the review	Optional descriptive analysis

Because PRDECT-ID is publicly available, independent researchers can replicate the experiment using the same underlying records. Such reproducibility is essential in text-classification research, where results may change in response to preprocessing choices, feature-extraction parameters, or evaluation protocols (Kowsari et al., 2019; Q. Li et al., 2022).

Research Variables

Three types of variables were specified for the experiment.

1. The predictor variable consisted of Indonesian product-review text. Every review was handled as a separate document containing customer statements about product quality, delivery, packaging, price, and overall satisfaction.
2. The outcome variable was the sentiment label associated with each review. This annotation expresses the polarity of the customer's opinion and serves as the ground-truth class for model training and evaluation.

3. Each algorithm generated a predicted sentiment class as its output. These predictions were compared with the corresponding ground-truth labels to compute classification performance.

Text Preprocessing

Preprocessing was performed before feature construction to standardize the reviews and limit textual noise. Indonesian marketplace comments frequently include casual expressions, inconsistent spelling and punctuation, repeated characters, emoticons, and shopping-specific vocabulary. Earlier research demonstrates that decisions made at this stage can substantially affect classification results, especially when lexical machine learning features are used (Kowsari et al., 2019; Uysal & Gunal, 2014).

The pipeline covered case folding, text cleaning, whitespace normalization, tokenization, selective stopword filtering, and optional stemming. Case folding converted all characters to lowercase so that capitalization variants did not create duplicate vocabulary entries. Cleaning removed URLs, HTML markup, excessive punctuation, special characters, and unrelated symbols. Repeated spaces and unnecessary line breaks were reduced to one space, after which tokenization separated each review into individual word units for numerical representation.

Stopword filtering was applied conservatively. Common Indonesian function words were discarded, whereas negators such as 'tidak', 'bukan', 'kurang', and 'belum' were preserved because they can reverse sentiment polarity. Stemming was treated as optional and used cautiously, since overly aggressive reduction to root forms may eliminate useful affective cues from informal e-commerce language.

Exactly the same preprocessing sequence was used for every classifier. This produced a standardized corpus that was subsequently converted into TF-IDF features, thereby maintaining comparability across models.

Feature Extraction Using TF-IDF

Term Frequency-Inverse Document Frequency was used to transform text into predictive features. TF-IDF is a term-weighting method frequently applied in information retrieval and classification because a term receives a larger value when it is prominent in a particular document but comparatively uncommon in the full corpus (Salton & Buckley, 1988; Forman, 2003). The method therefore represents each review as a numerical vector suitable for machine learning.

The weight assigned to term t in document d is defined as follows:

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

In this expression, TF(t,d) denotes the frequency of term t in document d, whereas IDF(t) is the inverse document frequency of term t in the corpus.

Inverse document frequency is obtained from the following equation:

$$\text{IDF}(t) = \log\left(\frac{N}{\text{DF}(t)}\right)$$

Here, N is the total number of documents and DF(t) is the number of documents in which term t occurs.

Both unigram and bigram features were included because sentiment in Indonesian reviews can be conveyed by single tokens or by combinations of words. Terms such as 'bagus', 'murah', and 'cepat' may independently carry polarity, whereas phrases such as 'tidak sesuai', 'barang rusak', 'sangat bagus', and 'cepat sampai' often provide stronger contextual evidence. The specific TF-IDF feature-extraction configuration used in this study is summarized in Table 2.

Table 2. Configuration of TF-IDF Feature Extraction

Parameter	Configuration
Analyzer	Word
N-gram range	Unigram and bigram
Minimum document frequency	2
Maximum document frequency	0.90
Maximum features	10,000
Term weighting	TF-IDF
Sublinear TF scaling	True

Classification Algorithms

The benchmark included Multinomial Naive Bayes, Support Vector Machine, and Random Forest. Together, these methods represent three distinct learning principles: probabilistic estimation, maximum-margin separation, and ensemble decision trees.

Multinomial Naive Bayes served as the probabilistic reference classifier. It is commonly selected for text applications because it is computationally economical and works effectively with frequency-oriented representations. The method assumes conditional independence among features when assigning a class. Although natural-language features rarely satisfy this assumption completely, Multinomial Naive Bayes often remains competitive in text categorization (Kowsari et al., 2019).

Support Vector Machine represented the margin-based approach. Its objective is to identify a separating hyperplane that maximizes the distance between observations from different classes (Cortes & Vapnik, 1995). This property is advantageous for TF-IDF text data, which typically form sparse, high-dimensional feature spaces. A linear configuration was therefore adopted because linear boundaries are often effective for such representations.

Random Forest represented ensemble learning. The algorithm builds multiple decision trees and aggregates their outputs, improving robustness and reducing the tendency of a single tree to overfit (Breiman, 2001). It was included to determine whether a tree-based ensemble could match probabilistic and margin-based methods when all three operate on the same TF-IDF vectors.

Experimental Scenario

Three complementary experimental scenarios were implemented. The first compared the classifiers under one shared TF-IDF representation. The second

investigated sensitivity to the feature unit by testing unigram, bigram, and combined unigram-bigram settings. The third assessed robustness through stratified splitting and cross-validation. The classifier configurations and their corresponding evaluation purposes are summarized in Table 3.

Table 3. Experimental Scenarios and Evaluation Purposes

Scenario	Description	Purpose
Scenario 1	TF-IDF + Multinomial Naive Bayes	Probabilistic baseline evaluation
Scenario 2	TF-IDF + Support Vector Machine	Margin-based classifier evaluation
Scenario 3	TF-IDF + Random Forest	Ensemble classifier evaluation

All classifiers received identical preprocessed text, TF-IDF parameters, training and testing partitions, and evaluation measures so that their results could be compared on equal terms.

Data Splitting and Validation Strategy

A stratified procedure divided the records into training and testing subsets. Stratification maintained approximately the same positive-to-negative ratio in both partitions. Preserving class proportions is important because an uneven split can distort performance estimates, particularly when accuracy is considered in isolation (He & Garcia, 2009).

The principal hold-out configuration allocated 80% of the reviews to model training and 20% to final testing. Classifiers were fitted on the training portion, whereas the test portion remained unseen until final evaluation. Because the split was stratified, the proportions of positive and negative reviews in each subset closely followed those of the complete dataset.

Model stability was also examined through stratified five-fold cross-validation. The corpus was partitioned into five folds while retaining the original class distribution. Each fold served once as the validation subset, and the other four were used for training, producing five evaluation cycles for every classifier. Mean performance and standard deviation were then calculated across the folds. This procedure limits dependence on one favorable train-test division and provides a more dependable comparison among algorithms. The data-partitioning procedure and validation settings applied in this study are summarized in Table 4.

Table 4. Data Partitioning and Validation Configuration

Procedure	Configuration
Train-test split	80:20
Split type	Stratified
Cross-validation	Stratified 5-fold cross-validation
Random seed	42
Evaluation focus	Accuracy, precision, recall, F1-score, macro-F1, weighted-F1

Evaluation Metrics

Performance was assessed through accuracy, precision, recall, F1-score, macro-F1, weighted-F1, and the confusion matrix. Relying on several measures is preferable because accuracy by itself can conceal weaknesses when class frequencies are unequal (Saito & Rehmsmeier, 2015; Sokolova & Lapalme, 2009).

Accuracy summarizes the proportion of all predictions that are correct. Precision expresses how many items predicted as positive are truly positive, whereas recall indicates how many actual positive items are successfully retrieved. F1-score combines precision and recall through their harmonic mean.

The mathematical definitions of the evaluation measures are given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In these equations, TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Macro-F1 was calculated by averaging the class-specific F1 values without weighting them by class size, thereby giving equal importance to both sentiments. Weighted-F1 instead incorporates the number of observations in each class. Confusion matrices were additionally examined to determine the direction and frequency of classification errors.

RESULTS AND DISCUSSION

This section reports and interprets the outcomes of the three-classifier experiment. The discussion covers the class distribution, preprocessing results, comparative model performance, cross-validation stability, sensitivity to n-gram configuration, confusion-matrix behavior, influential SVM features, and observed error cases.

Dataset Overview

After records with missing content, empty text, or duplicate entries had been removed, 5,305 reviews remained for model development. Each record was assigned to one of two sentiment categories: negative or positive.

Table 5. Distribution of Sentiment Classes

Sentiment Label	Number of Reviews	Percentage
Negative	2,752	51.88%
Positive	2,553	48.12%
Total	5,305	100.00%

Table 5 shows 2,752 negative records and 2,553 positive records, corresponding to 51.88% and 48.12% of the corpus, respectively. The modest difference between the two groups indicates an approximately balanced dataset. This distribution reduces the likelihood that a classifier will be driven primarily by one dominant class.

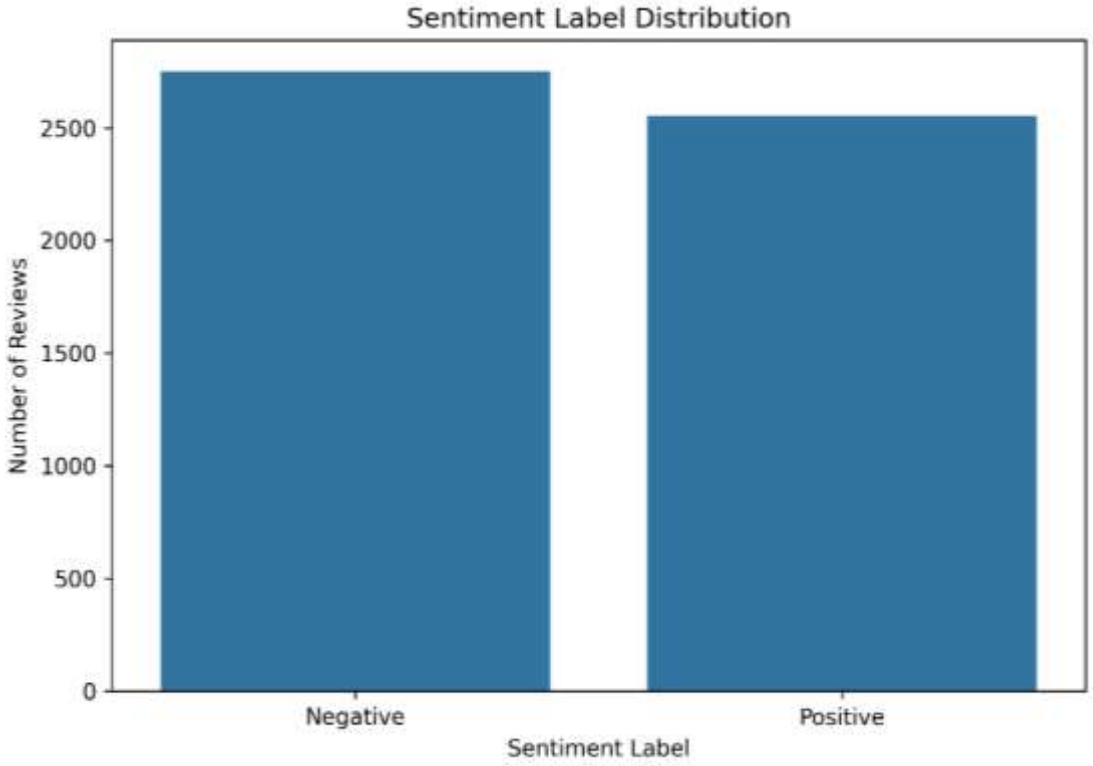

Figure 2. Distribution of Sentiment Labels

Figure 2 visually confirms that the negative and positive classes occur at similar frequencies, although negative reviews are slightly more numerous. Negative observations constitute 51.88% of all reviews, compared with 48.12% for positive observations. Because this imbalance is minor, accuracy remains useful as a summary measure. Nevertheless, precision, recall, F1-score, macro-F1, and weighted-F1 are also reported to capture performance from several complementary perspectives.

The near-balanced class structure supports an equitable comparison of the three algorithms. Macro-F1 is still particularly relevant because it gives equal weight to the two sentiment categories regardless of their sample counts. Model selection therefore considers not only overall accuracy but also whether performance remains comparable for negative and positive reviews.

Text Preprocessing Results

Following the removal of invalid and duplicated records, the remaining reviews passed through the complete preprocessing pipeline. The resulting dataset contained 5,305 entries and retained three principal fields: the original review, its cleaned form, and

the sentiment label. The cleaning strategy sought to remove noise without discarding expressions that convey polarity in Indonesian marketplace language.

Processing included conversion to lowercase, punctuation filtering, normalization of repeated characters and informal words, selective stopword removal, and whitespace standardization. Several English marketplace terms were also mapped to Indonesian forms; for instance, 'packaging' became 'kemasan' and 'seller' became 'penjual'. This step addressed the frequent mixing of Indonesian and English terminology in e-commerce reviews.

Table 6. Illustrative Results of Text Preprocessing

No.	Original Review	Clean Review	Sentiment
1	Alhamdulillah berfungsi dengan baik. Packaging...	alhamdulillah berfungsi baik kemasan aman...	Positive
2	barang bagus dan respon cepat, harga bersaing...	barang bagus respon cepat harga bersaing...	Positive
3	barang bagus, berfungsi dengan baik, seler ramah...	barang bagus berfungsi baik seler ramah...	Positive
4	bagus sesuai harapan penjual nya juga ramah...	bagus sesuai harapan penjual nya ramah...	Positive
5	Barang Bagus, pengemasan Aman, dapat Berfungsi...	barang bagus pengemasan aman berfungsi baik	Positive
6	barang bagus, seller ramah..	barang bagus penjual ramah	Positive
7	mantap paten joss	mantap paten joss	Positive
8	Works fine. Respon seller cepat, barang berfungsi...	works fine respon penjual cepat barang berfungsi...	Positive
9	barang bagus.. segel.. utuh.. original..	barang bagus segel utuh original berfungsi...	Positive
10	Packing aman, pengiriman cepat dan barang berfungsi...	kemasan aman pengiriman cepat barang berfungsi...	Positive

The examples in Table 6 indicate that unnecessary textual elements were reduced while polarity-bearing words, including 'bagus', 'aman', 'cepat', 'ramah', 'mantap', and 'berfungsi', remained available to the classifier. These terms capture favorable judgments concerning product condition, seller responsiveness, packaging, and delivery.

The samples also demonstrate code-mixing through expressions such as 'works fine', 'seller', 'packing', and 'original'. Some were translated or normalized, whereas others were retained when they could still contribute useful semantic evidence. Effective preprocessing for Indonesian marketplace reviews must therefore account for informal constructions and vocabulary that is specific to the e-commerce domain.

Stopwords were filtered selectively so that words influencing polarity were not removed. The negators 'tidak', 'bukan', 'belum', and 'kurang' were kept because they can invert the interpretation of surrounding terms. For example, deleting 'tidak' from 'tidak bagus' would incorrectly convert a negative expression into a positive one. This treatment aligns with prior evidence that preprocessing decisions strongly affect lexical text classifiers such as TF-IDF-based models (Uysal & Gunal, 2014; Kowsari et al., 2019; Li et al., 2022).

TF-IDF Feature Representation

The cleaned reviews were encoded as TF-IDF vectors. Under this scheme, a term's weight reflects both its relevance to a specific review and its relative rarity throughout the corpus (Salton & Buckley, 1988; Forman, 2003). Incorporating unigrams and bigrams allows the representation to preserve polarity conveyed by individual terms as well as by short phrases.

Bigram information is especially useful for Indonesian reviews because many evaluations are expressed through two-word constructions. Phrases such as 'tidak sesuai', 'sangat bagus', 'barang rusak', 'cepat sampai', and 'harga murah' often communicate clearer sentiment than either word considered separately. Examples of unigram, bigram, and domain-specific TF-IDF features associated with sentiment are presented in Table 7.

Table 7. Examples of Features Associated with Sentiment

Feature Type	Example Terms	Possible Sentiment Orientation
Unigram	bagus, murah, kecewa, rusak	Positive or negative
Bigram	sangat bagus, tidak sesuai, barang rusak, cepat sampai	Stronger contextual sentiment
Domain-specific terms	packing, seller, pengiriman, kualitas	Context-dependent

TF-IDF generates sparse vectors with many dimensions, a form that can be processed efficiently by conventional algorithms such as SVM and Naive Bayes. Previous classification research similarly reports that TF-IDF remains effective for supervised tasks when class polarity is strongly indicated by lexical patterns (Sebastiani, 2002; Kowsari et al., 2019; Li et al., 2022).

Model Performance Comparison

Comparative testing used 1,061 reviews: 550 negative and 511 positive. These counts show that stratification retained a close class balance in the held-out subset. The evaluated configurations paired TF-IDF with Multinomial Naive Bayes, Support Vector Machine, and Random Forest.

According to Table 8, every classifier exceeded 0.90 accuracy. The result demonstrates that TF-IDF represented sentiment-related vocabulary in Indonesian product reviews effectively. It also supports earlier findings that established classifiers remain useful when an appropriate feature representation is applied, particularly for domain-focused tasks with datasets of limited or moderate size (Kowsari et al., 2019; Li et al., 2022; Wankhade et al., 2022).

Table 8. Comparative Performance of the Classification Models

Model	Accuracy	Precision	Recall	F1-Score	Macro-F1	Weighted-F1
TF-IDF + Multinomial Naive Bayes	0.9538	0.9551	0.9530	0.9537	0.9537	0.9538
TF-IDF + Support Vector Machine	0.9595	0.9598	0.9591	0.9594	0.9594	0.9595
TF-IDF + Random Forest	0.9378	0.9378	0.9384	0.9378	0.9378	0.9378

Support Vector Machine ranked first among the evaluated methods, obtaining 0.9595 accuracy, 0.9594 macro-F1, and 0.9595 weighted-F1. Its advantage is consistent with the structure of TF-IDF data, which are sparse and high dimensional. A linear maximum-margin classifier can operate effectively in this setting by identifying a decision boundary that separates the two sentiment classes across the feature space.

Multinomial Naive Bayes also performed competitively, with 0.9538 accuracy and 0.9537 macro-F1. The small difference from SVM indicates that the dataset contains strong lexical cues, including 'bagus', 'cepat', 'aman', 'kecewa', 'rusak', and 'tidak sesuai'. This pattern agrees with previous studies: Naive Bayes can classify text effectively when word-occurrence patterns distinguish the classes, whereas SVM often gains an advantage in sparse, high-dimensional spaces (Sebastiani, 2002; Kowsari et al., 2019; Li et al., 2022).

Random Forest recorded the lowest of the three results, although its accuracy and macro-F1 remained high at 0.9378. Its relative disadvantage may stem from the sparse and high-dimensional characteristics of TF-IDF vectors, which tend to favor linear margin-based methods over tree ensembles. The finding should therefore not be interpreted as evidence that Random Forest is ineffective in general; rather, it appears less well matched to the lexical representation used in this experiment.

Figure 3 indicates that the absolute differences among models are modest, yet SVM leads consistently in accuracy, macro-F1, and weighted-F1. This consistency matters because a preferred classifier should combine high overall correctness with comparable treatment of both sentiments. Since macro-F1 weighs the two classes equally, SVM's score suggests that its advantage was not produced by favoring only one class.

No formal paired significance test, such as a paired t-test or Wilcoxon signed-rank test, was performed; consequently, the analysis does not report a p-value. Reliability was instead examined through the agreement of hold-out testing, stratified five-fold cross-validation, standard deviations, multiple performance measures, and confusion-matrix patterns. A future experiment could apply statistical tests to determine whether the observed differences are significant beyond sampling variation.

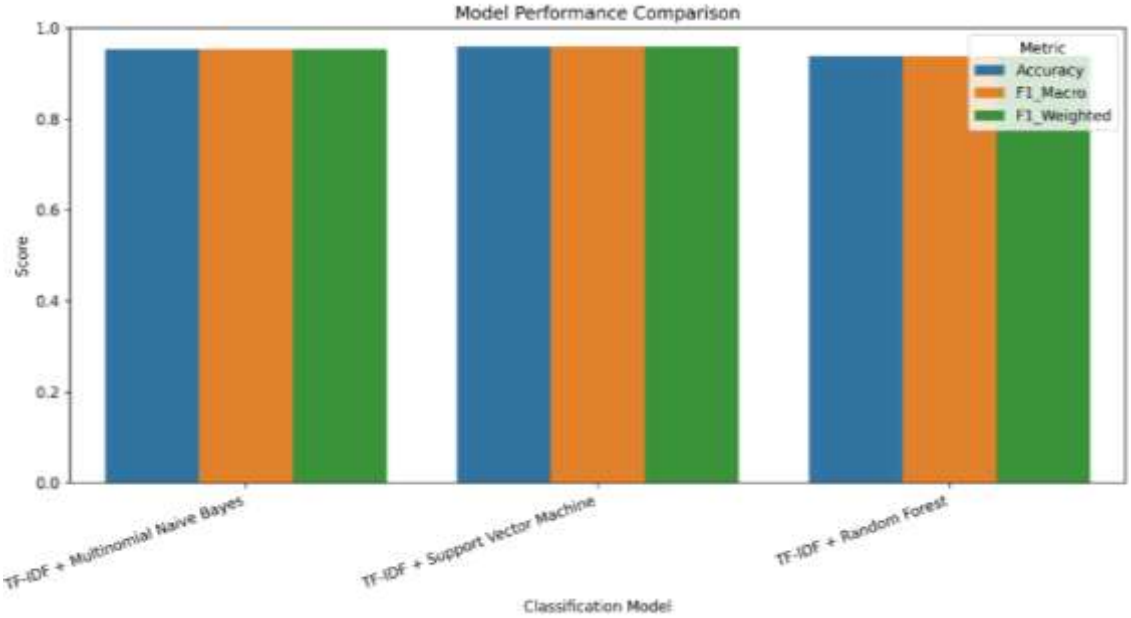

Figure 3. Performance Comparison across Classifiers

Cross-Validation Analysis

Stratified five-fold cross-validation was used to determine whether classifier performance remained stable beyond one training-test partition. The corresponding results are summarized in Table 9.

Table 9. Stratified Five-Fold Cross-Validation Performance

Model	Mean Accuracy	Std. Accuracy	Mean Precision Macro	Mean Recall Macro	Mean F1-Macro	Std. F1-Macro	Mean F1-Weighted
TF-IDF + Multinomial Naive Bayes	0.9533	0.0064	0.9546	0.9525	0.9531	0.0064	0.9532
TF-IDF + Support Vector Machine	0.9606	0.0051	0.9614	0.9600	0.9605	0.0051	0.9606
TF-IDF + Random Forest	0.9401	0.0049	0.9399	0.9403	0.9400	0.0049	0.9401

The cross-validation results are consistent with the hold-out findings. Support Vector Machine achieved the largest mean accuracy, 0.9606, and mean macro-F1, 0.9605. Its standard deviation of 0.0051 indicates limited variation across the five folds, suggesting that the strong result was not attributable to one unusually favorable partition.

Multinomial Naive Bayes likewise produced stable outcomes, with mean accuracy of 0.9533 and mean macro-F1 of 0.9531. The relatively small standard deviation shows that its performance changed little across folds, supporting its role as a dependable probabilistic baseline for Indonesian product-review sentiment analysis.

Random Forest again ranked third, yielding mean accuracy of 0.9401 and mean macro-F1 of 0.9400. Although these values remain substantial, they are below the corresponding SVM and Naive Bayes results. This pattern reinforces the interpretation that a tree ensemble is less suited than a linear classifier to the sparse, high-dimensional TF-IDF space used here.

Taken together, hold-out testing and cross-validation identify TF-IDF with Support Vector Machine as the most robust configuration. The close agreement between the two validation approaches indicates that this combination offers a stable and reproducible baseline for Indonesian e-commerce sentiment classification.

The study's novelty is its controlled and reproducible benchmark for Indonesian marketplace sentiment using PRDECT-ID. Rather than concentrating on broad-domain corpora, social media messages, or complex neural architectures, the research establishes a transparent conventional-machine-learning reference for product reviews. Controlled model comparison is integrated with unigram-bigram sensitivity testing, stratified cross-validation, confusion-matrix interpretation, coefficient-based feature inspection, and examination of classification errors. This combination explains model behavior in greater depth than reporting aggregate scores alone.

N-Gram Sensitivity Analysis

Beyond comparing algorithms, the experiment assessed how the selected n-gram representation influenced prediction. Unigrams, bigrams, and their combined configuration were tested to determine whether individual terms, two-word sequences, or both together most effectively represented sentiment in Indonesian e-commerce reviews.

Table 10. Classification Performance under Different N-Gram Settings

N-Gram Setting	Model	Accuracy	Precision Macro	Recall Macro	F1-Macro	F1-Weighted
Unigram	TF-IDF + Multinomial Naive Bayes	0.9227	0.9226	0.9227	0.9226	0.9227
Unigram	TF-IDF + Support Vector Machine	0.9444	0.9451	0.9438	0.9442	0.9444
Unigram	TF-IDF + Random Forest	0.9227	0.9227	0.9232	0.9227	0.9227
Bigram	TF-IDF + Multinomial Naive Bayes	0.9105	0.9140	0.9088	0.9099	0.9102
Bigram	TF-IDF + Support Vector Machine	0.9010	0.9039	0.8995	0.9005	0.9008
Bigram	TF-IDF + Random Forest	0.8831	0.8880	0.8811	0.8822	0.8826
Unigram + Bigram	TF-IDF + Multinomial Naive Bayes	0.9538	0.9551	0.9530	0.9537	0.9538
Unigram + Bigram	TF-IDF + Support Vector Machine	0.9595	0.9598	0.9591	0.9594	0.9595

Unigram + Bigram	TF-IDF + Random Forest	0.9378	0.9378	0.9384	0.9378	0.9378

Table 10 demonstrates that combining unigrams and bigrams produced the strongest result for each classifier. Support Vector Machine obtained the highest overall scores under this setting, with 0.9595 accuracy and 0.9594 macro-F1. The outcome indicates that a representation containing both isolated words and short word sequences captures Indonesian review sentiment more completely.

Using only unigrams reduced performance relative to the combined setting. SVM, for example, obtained 0.9442 macro-F1 from unigrams, which increased to 0.9594 when bigrams were added. The gain implies that short phrases supply contextual information unavailable from individual tokens alone.

The bigram-only representation was the weakest across all three algorithms, most likely because it created a sparse feature set. Numerous word pairs occur in only a few reviews and are therefore unreliable when unigrams are excluded. Bigrams are thus beneficial as supplementary contextual features but are less effective as the sole representation.

This result also has a linguistic explanation. Indonesian reviews frequently convey polarity through combinations such as 'tidak sesuai', 'barang rusak', 'sangat bagus', 'cepat sampai', 'tidak berfungsi', and 'tidak mengecewakan'. The meaning of these expressions depends on the interaction between their words, making combined unigram-bigram TF-IDF more appropriate than either feature type used independently.

Confusion Matrix Analysis

Confusion matrices were inspected to analyze how Multinomial Naive Bayes, Support Vector Machine, and Random Forest distributed correct predictions and errors between the negative and positive classes. The numerical confusion-matrix results for the three evaluated classifiers are summarized in Table 11.

Table 11. Confusion Matrices for the Evaluated Classifiers

a. TF-IDF + Multinomial Naive Bayes

Actual / Predicted	Negative	Positive
Negative	536	14
Positive	35	476

b. TF-IDF + Support Vector Machine

Actual / Predicted	Negative	Positive
Negative	533	17
Positive	26	485

c. TF-IDF + Random Forest

Actual / Predicted	Negative	Positive
Negative	507	43
Positive	23	488

The matrices reveal distinct error tendencies. Support Vector Machine delivered the strongest overall outcome by correctly labeling 1,018 of the 1,061 test reviews. It recognized 533 negative and 485 positive records, while only 17 negative reviews were predicted as positive and 26 positive reviews as negative. These counts indicate the most even performance across the two sentiment categories.

Multinomial Naive Bayes correctly classified 1,012 reviews but produced an asymmetric error pattern. It assigned 35 positive observations to the negative class, whereas only 14 negative observations were labeled positive. The model therefore appeared more conservative when predicting positive sentiment, potentially because negative lexical evidence dominated certain mixed or ambiguous texts.

Random Forest made 995 correct predictions and displayed the opposite tendency. Forty-three negative reviews were labeled positive, compared with 23 positive reviews labeled negative. Its greater tendency to predict the positive class may be problematic in practice because negative feedback, which is often critical for product and service improvement, can be overlooked.

The comparative matrix patterns reinforce the aggregate metrics. SVM combined the highest accuracy and macro-F1 with the most balanced distribution of errors. Naive Bayes identified negative reviews slightly more effectively but missed a larger number of positive reviews, whereas Random Forest produced the most false-positive sentiment assignments for genuinely negative texts. On this evidence, SVM is the most dependable classifier for the current task.

The corresponding visual representations of the confusion matrices are presented in Figure 4 for TF-IDF with Multinomial Naive Bayes, Figure 5 for TF-IDF with Random Forest, and Figure 6 for TF-IDF with Support Vector Machine.

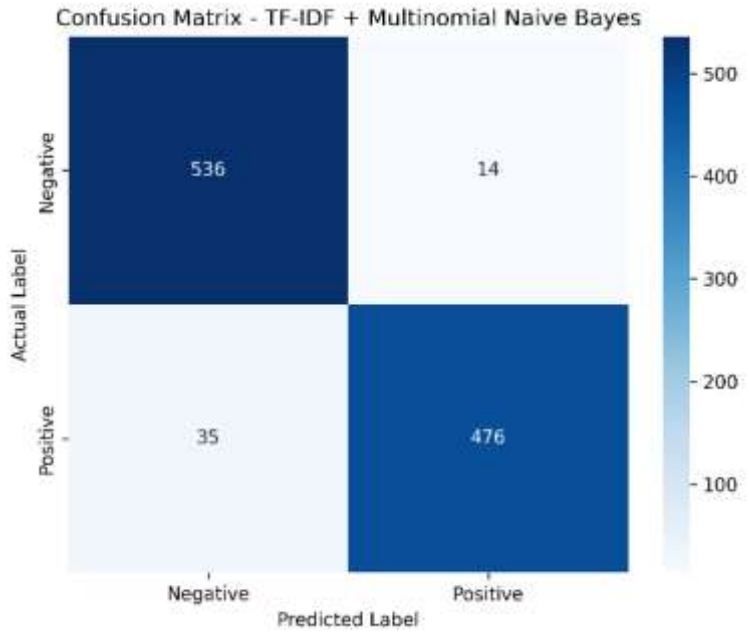

Figure 4. Confusion Matrix for TF-IDF with Multinomial Naive Bayes

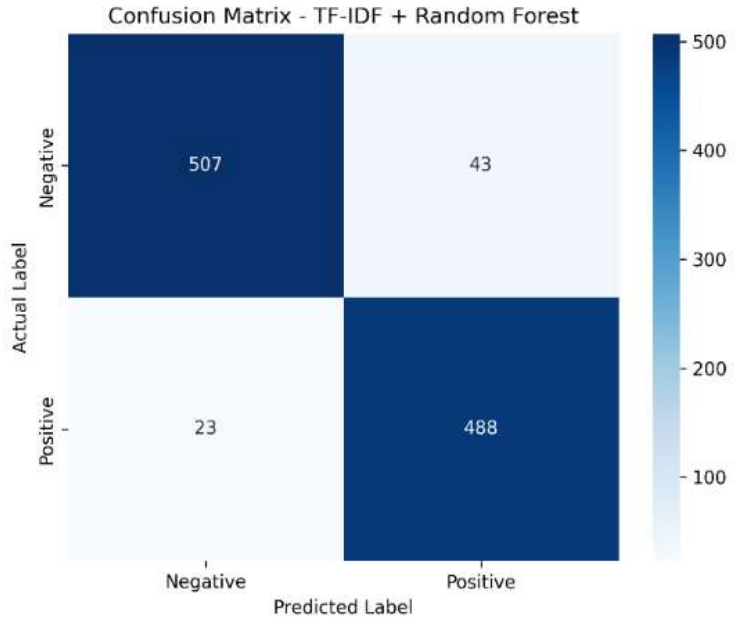

Figure 5. Confusion Matrix for TF-IDF with Random Forest

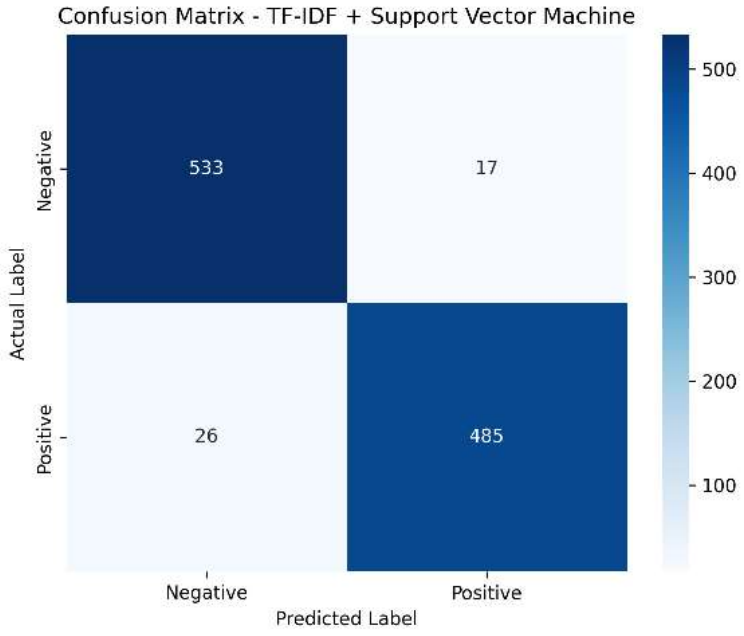

Figure 6. Confusion Matrix for TF-IDF with Support Vector Machine

Feature Interpretation of the SVM Model

To improve interpretability, the coefficients of the linear SVM were examined to identify terms that contributed most strongly to each class. Positive coefficient values favored the positive label, whereas negative values shifted predictions toward the

negative label. The highest-impact positive and negative features identified from the linear SVM coefficients are presented in Table 12.

Table 12. Highest-Impact Positive and Negative SVM Features

Positive-Oriented Terms	Coefficient	Negative-Oriented Terms	Coefficient
mantap	3.3974	tidak	-3.8269
bagus	3.1197	kurang	-2.3731
sesuai	2.2713	kecewa	-2.2645
aman	2.2080	tidak sesuai	-2.1231
cepat	2.1998	jelek	-1.8135
oke	2.1632	lama	-1.4045
baik	1.8918	rusak	-1.3944
alhamdulillah	1.5678	tidak bagus	-1.3885
keren	1.4747	tidak berfungsi	-1.3012
rapi	1.4240	parah	-1.2193

High positive coefficients were associated with direct expressions of satisfaction, including 'mantap', 'bagus', 'sesuai', 'aman', 'cepat', and 'baik'. These words commonly describe favorable product performance, responsive sellers, secure packaging, and satisfactory delivery.

Terms with large negative coefficients primarily consisted of negation and complaint vocabulary, including 'tidak', 'kurang', 'kecewa', 'tidak sesuai', 'jelek', 'lama', and 'rusak'. The prominence of multiword features such as 'tidak sesuai', 'tidak bagus', and 'tidak berfungsi' further supports the inclusion of bigrams alongside unigrams.

The coefficient analysis also validates the decision to retain negation during preprocessing. Removing words such as 'tidak', 'kurang', or 'bukan' as ordinary stopwords would eliminate important evidence of polarity. Selective rather than indiscriminate stopword filtering was therefore necessary.

Error Analysis

Although SVM produced the best aggregate performance, it still misclassified several reviews. These cases were examined to identify linguistic and contextual conditions associated with incorrect predictions. Representative misclassified reviews and their possible causes are summarized in Table 13.

Table 13. Representative Misclassified Reviews and Possible Causes

No.	Clean Review	Actual Label	Predicted Label	Possible Cause
1	barang sesuai deskripsi berfungsi baik cuma tidak puas pelayanan penjual pesanan diproses h pick up h dikomplen baru bergerak	Negative	Positive	Mixed sentiment; positive product terms dominate service complaint

2	pernah beli sebelumnya abis jd skrg beli makasih ka	Positive	Negative	Informal expression and weak explicit positive sentiment
3	bukan jam harga mahal didiskon jam sesuai harga beli kualitasnya	Positive	Negative	Negation and ambiguous comparison
4	mantaaff	Positive	Negative	Non-standard spelling not normalized
5	produk bocor kemasan krg bgs dr dus luar aja basah sampe dlm isi botol habis stengah mungkin jadi masukkan buat penjual kali kemasan lbh aman	Negative	Positive	Negative review contains constructive and neutral terms
6	barang cepat panas rusak	Negative	Positive	Contextual ambiguity of the word “cepat”
7	aplikasinya tidak kelihatan kalo pintu buka enggak automationnya lumayan lengkapnlah	Positive	Negative	Mixed complaint and positive evaluation
8	swlalu beli official shop tidak pernah kecewa makasi penjual	Positive	Negative	Negated negative phrase interpreted as negative
9	respon cepat kemasan rapih kecewa lubang teflonnya ngelupas	Negative	Positive	Positive service terms conflict with product complaint
10	pengiriman cepat tas tak mengecewakan	Positive	Negative	Negation phrase not fully captured

Five recurring sources of error were identified. First, reviews containing mixed sentiment combine favorable and unfavorable statements, making a single polarity label difficult to infer. A reviewer may praise the product while criticizing the seller's service; in such a case, TF-IDF can assign substantial weight to positive product terms even though the overall judgment is negative.

Second, unconventional spelling and informal variants remain difficult to normalize. The token 'mantaaff', for example, was misclassified because it was not mapped to the standard form 'mantap'. Expanding the normalization dictionary could improve coverage of common Indonesian spelling variations.

Third, positive constructions formed by negating negative words may confuse a lexical classifier. Expressions such as 'tidak pernah kecewa' and 'tak mengecewakan' are favorable, yet they contain terms ordinarily associated with dissatisfaction. TF-IDF records these word occurrences but cannot always represent their compositional interpretation.

Fourth, the polarity of certain words changes with context. 'Cepat' generally conveys a favorable meaning in 'pengiriman cepat', but it describes an undesirable condition in 'cepat panas'. Correct sentiment interpretation therefore depends on semantic context in addition to frequency-based lexical evidence.

Fifth, very short or indirect reviews may express sentiment without providing strong explicit indicators. When lexical evidence is weak, the classifier has limited information on which to base its decision, a common issue in short-text categorization.

Overall, the principal limitations of TF-IDF with SVM involve mixed opinions, nonstandard spelling, complex negation, and context-sensitive expressions. Future research could address these issues through broader normalization resources, sentiment lexicons, richer phrase features, formal significance testing, or contextual Indonesian language models based on BERT.

The findings also have practical implications for Indonesian e-commerce analytics. The validated TF-IDF-SVM configuration offers a transparent benchmark against which deep learning or transformer systems can be compared. In operational settings, it could support automated monitoring that flags dissatisfaction, identifies negative product feedback, and helps prioritize service improvements. Its modest computational requirements, interpretability, and reproducibility make it suitable for processing large collections of Indonesian marketplace reviews.

CONCLUSION

This research benchmarked three conventional machine learning algorithms for sentiment classification of Indonesian e-commerce product reviews in PRDECT-ID. A common experimental pipeline combined text preprocessing, TF-IDF feature construction, and controlled evaluation of Multinomial Naive Bayes, Support Vector Machine, and Random Forest. The assessment incorporated accuracy, precision, recall, F1-score, macro-F1, weighted-F1, confusion matrices, n-gram sensitivity testing, and stratified five-fold cross-validation. All three classifiers achieved accuracy above 0.90, confirming that TF-IDF provides an effective lexical representation for this dataset. Support Vector Machine delivered the strongest hold-out result, with 0.9595 accuracy, 0.9594 macro-F1, and 0.9595 weighted-F1. Its stability was also supported by cross-validation, which produced mean accuracy of 0.9606 and mean macro-F1 of 0.9605. Multinomial Naive Bayes followed closely and remains a computationally efficient probabilistic baseline when sentiment is strongly encoded in word patterns. Random Forest ranked third, although its performance was still high; the difference may reflect the closer suitability of linear margin-based methods to sparse, high-dimensional TF-IDF vectors. Overall, TF-IDF with SVM provides the most reliable and reproducible baseline among the tested configurations.

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

https://doi.org/10.5555/944919.944974

Harmandini, K. P., & L, K. M. (2024). Analysis of {TF-IDF} and {TF-RF} Feature Extraction on Product Review Sentiment. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, *8*(2), 929–937. https://doi.org/10.33395/sinkron.v8i2.13376

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, *13*(2), 1–41. https://doi.org/10.1145/3495162

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review. *Journal of King Saud University - Computer and Information Sciences*, *36*(4), 102048. https://doi.org/10.1016/j.jksuci.2024.102048

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning-Based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, *54*(3), 1–40. https://doi.org/10.1145/3439726

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135. https://doi.org/10.1561/1500000011

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing \& Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307.

https://doi.org/10.1162/COLI_a_00049

Uysal, A. K., & Gunal, S. (2014). The Impact of Preprocessing on Text Classification. *Information Processing \& Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), e1253. https://doi.org/10.1002/widm.1253