Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16280
e-ISSN : 2541-2019
p-ISSN : 2541-044X

# Customer Complaint Classification at PT Pos Indonesia Manokwari Using Naive Bayes and Random Forest

**Rizhmaria Ester Vieta Saphira[1)*], Christian Dwi Suhendra[2)], Lilis Indrayani[3)]**
**[1,2,3)]University Of Papua**
[1)]rizhmarias@gmail.com, [2)]c.suhendra@unipa.ac.id, [3)]lilisindrayani8@gmail.com,

**Submitted** :May 26, 2026 | **Accepted** : June 14, 2026 | **Published** : July 5, 2026

**Abstract:** Customer complaints represent an important source of information for evaluating service quality and improving organizational performance. However, the increasing volume of complaints received by PT Pos Indonesia Manokwari makes manual complaint classification inefficient and time-consuming. This study aims to compare the performance of Naive Bayes and Random Forest algorithms for customer complaint classification using the Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction method. The dataset consisted of 1,490 customer complaint records collected from the Customer Complaint Handling (CCH) system and categorized into twelve complaint classes. The research process included data cleaning, case folding, stopword removal, TF-IDF transformation, dataset splitting, model training, and performance evaluation. The classification models were evaluated using accuracy, precision, recall, F1-weighted score, F1-macro score, and 5-fold cross-validation. The experimental results showed that Random Forest achieved better performance than Naive Bayes. Random Forest obtained an accuracy of 87.92%, precision of 85.22%, recall of 87.92% an F1-weighted score of 86.30%, and an F1-macro score of 70.85%, while Naive Bayes achieved an accuracy of 84.90%, an F1-weighted score of 84.00%, and an F1-macro score of 48.41%. The cross-validation results produced an average accuracy of 71.81%. Although Random Forest achieved the highest hold-out accuracy, the cross-validation results indicate performance variation across different data partitions, which may be caused by class imbalance among complaint categories. These findings demonstrate that Random Forest is more effective for multiclass customer complaint classification and can support the development of automated complaint management systems at PT Pos Indonesia Manokwari.

**Keywords:** Complaint Classification; Machine Learning; Naive Bayes; Random Forest; TF-IDF

## INTRODUCTION

The rapid advancement of information technology has significantly transformed business operations across various sectors, including postal and logistics services. Organizations increasingly rely on digital systems to manage large volumes of data efficiently and effectively. Among the various types of organizational data, customer complaints constitute an important source of information that can be utilized to evaluate service quality, identify operational problems, and support strategic decision-making processes(Q. Li et al., 2022; Mao et al., 2024).

PT Pos Indonesia is one of the largest postal and logistics service providers in Indonesia. Along with the growth of e-commerce and shipping activities, the number of customer complaints submitted to the company has also increased. These complaints cover various issues, including delivery delays, undelivered shipments, returned packages, delivery status inquiries, failed deliveries, and other service-related problems. The large volume and diversity of complaint categories make manual classification increasingly inefficient and time-consuming(Q. Li et al., 2022; Akuma et al., 2022; Mao et al., 2024).

Text classification, a branch of text mining and machine learning, has been widely adopted to automate the categorization of textual data (Q. Li et al., 2022). Automated complaint classification enables organizations to identify complaint categories more quickly and accurately, thereby improving response times and customer satisfaction (Y. Mao et al., 2024). Among the most commonly used classification algorithms are Naive Bayes and

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Random Forest. Naive Bayes is a probabilistic classifier known for its computational efficiency and effectiveness in handling high-dimensional text data, while Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

Several previous studies have compared the performance of Naive Bayes and Random Forest in text classification tasks. Research findings generally indicate that Random Forest achieves higher classification accuracy, while Naive Bayes offers advantages in computational simplicity and faster training times. In addition, the TF-IDF feature extraction method has been widely used to represent textual information effectively by assigning weights based on term importance within documents.

Although numerous studies have investigated text classification techniques, most previous research has primarily focused on improving overall classification accuracy. Limited studies have examined multiclass customer complaint classification problems involving imbalanced class distributions, where minority complaint categories are often more difficult to classify accurately. In addition, relatively few studies have compared classification performance using both F1-weighted and F1-macro metrics, which provide a more comprehensive evaluation of model effectiveness across majority and minority classes. The key challenge is not only achieving high overall accuracy, but also maintaining classification performance across minority complaint categories. Furthermore, studies focusing on customer complaint classification in the postal and logistics sector, particularly using complaint data from PT Pos Indonesia Manokwari, remain limited.

Therefore, this study aims to compare the performance of Naive Bayes and Random Forest algorithms for customer complaint classification using TF-IDF feature extraction. The contributions of this study are threefold. First, it provides a comparative evaluation of two widely used machine learning algorithms on customer complaint data. Second, it demonstrates the effectiveness of TF-IDF in representing complaint text features. Third, it provides practical recommendations for developing automated complaint management systems at PT Pos Indonesia Manokwari.

## LITERATURE REVIEW

Text classification is a widely used technique in text mining and machine learning for automatically categorizing textual data into predefined classes. According to(Q. Li et al., 2022), text classification has evolved from traditional machine learning approaches to more advanced deep learning methods and remains one of the most important applications of natural language processing. The increasing volume of textual data generated through digital platforms has encouraged organizations to adopt automated classification techniques to improve efficiency and decision-making processes. Recent studies by (Ahmadi et al., 2025) further demonstrate that text classification techniques continue to play an important role in practical applications such as cybersecurity, spam detection, and email categorization.

One of the most commonly used algorithms in text classification is Naive Bayes.(Aprianti et al., 2025) stated that Naive Bayes performs well in handling high-dimensional text data because of its probabilistic approach and computational efficiency. Similarly, (Halim et al., 2025) reported that Naive Bayes provides satisfactory classification performance with relatively low computational cost, making it suitable for large-scale text processing tasks. Recent studies by (Helmayanti, Hamami, & Fa'rifah, 2023; Khoerunnisa, Shiddieq, & Nurhayati, 2025 )also demonstrated that the combination of TF-IDF and Naive Bayes can achieve reliable classification performance in various text mining applications, including sentiment analysis and news classification. Despite these advantages, the algorithm may experience performance degradation when dealing with complex feature relationships and imbalanced datasets.

Another widely used algorithm is Random Forest, an ensemble learning method that combines multiple decision trees to improve prediction performance. According to (Fitriyani et al., 2026)Random Forest generally achieves higher classification accuracy than Naive Bayes due to its ability to capture complex interactions among features and reduce overfitting. Likewise,(Santoso, Nugroho, & Asyfiya, 2025)found that Random Forest consistently outperformed several traditional machine learning algorithms in sentiment analysis tasks because of its robust ensemble structure. (Breiman, 2001), who originally introduced Random Forest, explained that the algorithm improves predictive performance through the aggregation of multiple decision trees. Furthermore, (Riswanto, Hidayat, & Nasiri, 2026)highlighted that Random Forest has been successfully implemented across various scientific domains because of its robustness and generalization capability.

Feature extraction is another important stage in text classification. (S. Akuma et al., 2022)explained that Term Frequency–Inverse Document Frequency (TF-IDF) is one of the most widely used feature extraction techniques because it effectively represents the importance of words within documents. Furthermore, (C. A. Nurhaliza Agustina, R. Novita, Mustakim, & N. E. Rozanda, 2024) demonstrated that TF-IDF significantly improves classification performance by generating meaningful numerical representations of textual information. Similar findings were reported by (Helmayanti et al., 2023; Khoerunnisa et al., 2025), who found that TF-IDF contributes significantly to improving classification accuracy when combined with machine learning algorithms such as Naive Bayes.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16280
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Several studies have directly compared the performance of Naive Bayes and Random Forest.(Nugroho, Hayati, & Jabir, 2025) reported that Random Forest achieved higher classification accuracy than Naive Bayes in public sentiment classification tasks. Similarly,(Meinita & Anshori, 2025) found that Random Forest produced more accurate results in customer service chatbot intent classification, while Naive Bayes offered faster training and lower computational complexity. Comparable results were also reported by (Fahrezi Putra Ichsansyah & Korespondensi, 2026; Halim et al., 2025; Prakoso Indaryono, 2024), who observed that Random Forest generally achieved better classification performance than Naive Bayes when handling complex and multiclass datasets.These findings indicate that Random Forest is generally more effective for complex text classification problems, whereas Naive Bayes remains advantageous in terms of computational efficiency.

In the context of customer complaint analysis, machine learning techniques have been increasingly applied to automate complaint categorization and improve service management. According to (Y. Mao et al., 2024)customer feedback and complaint data contain valuable information The increasing availability of customer-generated textual data has encouraged organizations to adopt automated complaint classification systems to improve response quality and service efficiency. Previous studies have shown that machine learning-based complaint classification can support decision-making and customer relationship management processes(Meinita & Anshori, 2025; Nugroho et al., 2025). that can be utilized to evaluate service quality and support organizational decision-making. However, studies focusing specifically on customer complaint classification in postal and logistics services remain limited, particularly those involving multiclass complaint datasets.

Based on the reviewed literature, both Naive Bayes and Random Forest have demonstrated promising performance in text classification tasks. However, limited research has examined their effectiveness in classifying customer complaints within the postal and logistics sector, especially using complaint data from PT Pos Indonesia Manokwari. Therefore, this study compares the performance of Naive Bayes and Random Forest using TF-IDF feature extraction to identify the most suitable classification model for automated customer complaint management.

## METHOD

### Research Design

This study employed a quantitative approach using a comparative experimental method to evaluate the performance of Naive Bayes and Random Forest algorithms for customer complaint classification. Quantitative research is appropriate for measuring model performance through objective evaluation metrics and statistical analysis. The comparative experimental method was used to compare the effectiveness of the two classification algorithms on the same dataset and identify the most suitable model for customer complaint classification (Q. Li et al., 2022)

The study followed a text mining framework consisting of data collection, preprocessing, feature extraction, model training, and performance evaluation stages.

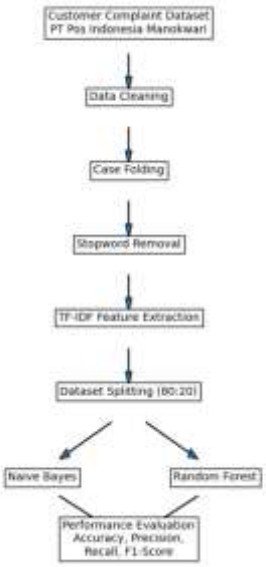



Figure 1. Research Framework

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16280
e-ISSN : 2541-2019
p-ISSN : 2541-044X

**Dataset**

The dataset used in this study was obtained from the Customer Complaint Handling (CCH) system of PT Pos Indonesia Manokwari. The dataset consisted of customer complaints recorded between January 2025 and March 2026. These complaints represented various operational issues experienced by customers, including delivery delays, undelivered shipments, shipment returns, shipment information requests, failed deliveries, and other customer service issues.

Before the classification process, the dataset underwent data cleaning procedures to remove duplicate records, incomplete entries, and irrelevant attributes. According to(S. Akuma et al., 2022), data cleaning is an essential stage in text mining because low-quality textual data can introduce noise and negatively affect classification performance. After preprocessing, the final dataset contained 1,490 complaint records distributed across twelve complaint categories.

The distribution of complaint categories indicates that the dataset suffers from a severe class imbalance problem. The largest class, Undelivered Shipment (Redelivery), contains 412 records (27.65%), whereas the smallest class, Wrong Routing (Salah Salur), contains only 5 records (0.34%). This results in a majority-to-minority ratio of approximately 82:1. Such an imbalanced distribution may bias classification models toward majority classes and reduce their ability to accurately classify minority complaint categories. Consequently, evaluation metrics such as F1-Macro are important because they provide a balanced assessment of classification performance across all classes. This class imbalance may also contribute to lower F1-Macro scores compared with Accuracy and F1-Weighted scores.

Table 1. Dataset Characteristics

| **Dataset Characteristics** | Value |
|---|---|
| Initial Dataset Size | 1.500 |
| Number of Attributes | 25 |
| Final Dataset Size | 1.490 |
| Number of Complaint Classes | 12 |

**Data Preprocessing**

Text preprocessing was conducted to improve data quality and prepare complaint texts for machine learning classification. According to (S. Akuma et al., 2022), preprocessing reduces textual noise and improves the effectiveness of feature extraction methods. The preprocessing stage consisted of data cleaning, case folding, and stopword removal. Data cleaning was performed to remove duplicate records, URLs, special characters, symbols, and irrelevant information from complaint texts. Subsequently, all text documents were converted into lowercase letters through case folding to ensure consistency during text processing. Common Indonesian stopwords such as dan, yang, di, and ke were removed because they do not contribute meaningful information to the classification process and may increase feature dimensionality. Stemming was not applied in this study because several complaint terms represent operational keywords used in postal and logistics services that may lose contextual meaning after stemming. As a result, the preprocessing stage preserved important domain-specific terminology while reducing textual noise. For example, the sentence “My package has not been delivered and the shipment status has not changed since February 10, 2026” was transformed into “package delivered shipment status changed february” after preprocessing.

**Feature Engineering and TF-IDF Transformation**

Feature engineering was performed to transform textual complaint data into numerical representations suitable for machine learning algorithms. In this study, the Term Frequency–Inverse Document Frequency (TF-IDF) method was employed. (C. A. Nurhaliza Agustina et al., 2024) reported that TF-IDF is one of the most effective feature extraction techniques for text classification because it represents the importance of terms within documents while reducing the influence of common words.

The TF-IDF configuration used an n-gram range of (1,2), allowing the model to capture both individual words (unigrams) and two-word phrases (bigrams). Unigrams help identify important keywords within complaint texts, while bigrams capture complaint-related phrases that may provide additional contextual information. Furthermore, the maximum number of features was limited to 3,000 to reduce noise and computational complexity while retaining the most informative terms. The sublinear_tf parameter was also applied to scale term frequencies logarithmically, thereby reducing the influence of extremely frequent terms. After the feature extraction process, a total of 1,397 features were generated and used as input for the classification models.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16280
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2. TF-IDF Parameters

| Parameter | Value |
|---|---|
| N-Gram | (1,2) |
| Max Features | 3000 |
| Sublinear TF | True |
| Final Features | 1397 |

**Dataset Splitting**

After feature extraction, the dataset was divided into training and testing subsets using the hold-out method with an 80:20 ratio. According to(Q. Li et al., 2022), this splitting strategy is commonly used in text classification studies because it provides a balance between model training and model evaluation. From the total of 1,490 complaint records, 1,192 records (80%) were allocated to the training set and 298 records (20%) were allocated to the testing set.

Table 3. Complaint Class Distribution

| Complaint Class | Total | Percentage(%) |
|---|---|---|
| Belum Terima (Antar Ulang) | 412 | 27.65 |
| Keterlambatan | 374 | 25.10 |
| Permintaan Berita Acara | 361 | 24.23 |
| Pengembalian/Retur | 158 | 10.60 |
| Informasi Kiriman | 82 | 5.50 |
| Gagal Antar | 33 | 2.21 |
| Lainnya | 22 | 1.48 |
| Salah Serah | 17 | 1.14 |
| Belum Terima (Status Delivered) | 12 | 0.81 |
| Pra Collecting | 8 | 0.54 |
| Salah Update Status | 6 | 0.40 |
| Salah Salur | 5 | 0.34 |
| Total | 1490 | 100.00 |

The distribution of complaint categories indicates that the dataset suffers from a severe class imbalance problem. The largest class, Undelivered Shipment (Redelivery), contains 412 records (27.65%), whereas the smallest class, Wrong Routing, contains only 5 records (0.34%). This results in a majority-to-minority ratio of approximately 82:1. Such an imbalanced distribution may bias classification models toward majority classes and reduce their ability to accurately classify minority complaint categories. Consequently, evaluation metrics such as F1-Macro are important because they provide a balanced assessment of classification performance across all classes.

**Classification Algorithms**

**Naive Bayes**

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem. The algorithm assumes that all features are conditionally independent given the target class. According to(Halim et al., 2025), Naive Bayes remains one of the most popular algorithms for text classification because of its computational efficiency and ability to handle high-dimensional textual data. Previous studies have also demonstrated that Naive Bayes achieves satisfactory performance in various text classification tasks, particularly when combined with TF-IDF feature extraction (Helmayanti et al., 2023; Khoerunnisa et al., 2025).

The Naive Bayes classifier was implemented using the MultinomialNB algorithm with an alpha smoothing parameter of 0.1. MultinomialNB was selected because it is widely used for text classification tasks involving TF-

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16280
e-ISSN : 2541-2019
p-ISSN : 2541-044X

IDF features. The alpha parameter was applied to reduce the effect of zero probabilities during classification and improve model robustness.

**Random Forest**

Random Forest is an ensemble learning algorithm that combines multiple decision trees to generate a final prediction through majority voting. According to Breiman (2001), Random Forest is an effective approach for improving classification accuracy while reducing overfitting.

The Random Forest classifier was implemented using 300 decision trees (n_estimators = 300), with no restriction on maximum tree depth (max_depth = None) and random_state = 42 to ensure reproducibility of the experimental results. The ensemble structure enables the model to capture complex feature relationships and improve classification performance.

**Model Evaluation**

The classification models were evaluated using Accuracy, Precision, Recall, F1-Weighted, and F1-Macro metrics. According to (Q. Li et al., 2022), these metrics provide comprehensive information regarding model performance and are widely used in multiclass text classification studies. Because the dataset contains multiple complaint categories with an imbalanced class distribution, F1-Macro was considered an important evaluation metric as it assigns equal importance to all classes regardless of their size.

To evaluate model stability and generalization capability, 5-Fold Cross Validation was conducted. In addition to overall accuracy, particular attention was given to F1-Macro because it provides a more balanced evaluation of classification performance across both majority and minority complaint categories.

## RESULT

**Complaint Category Distribution**

The dataset consisted of 1,490 customer complaint records distributed across 12 complaint categories. The largest categories were *Undelivered Shipment (Redelivery)*, *Delivery Delay*, and *Official Report Request*. These categories accounted for the majority of customer complaints recorded in the dataset.

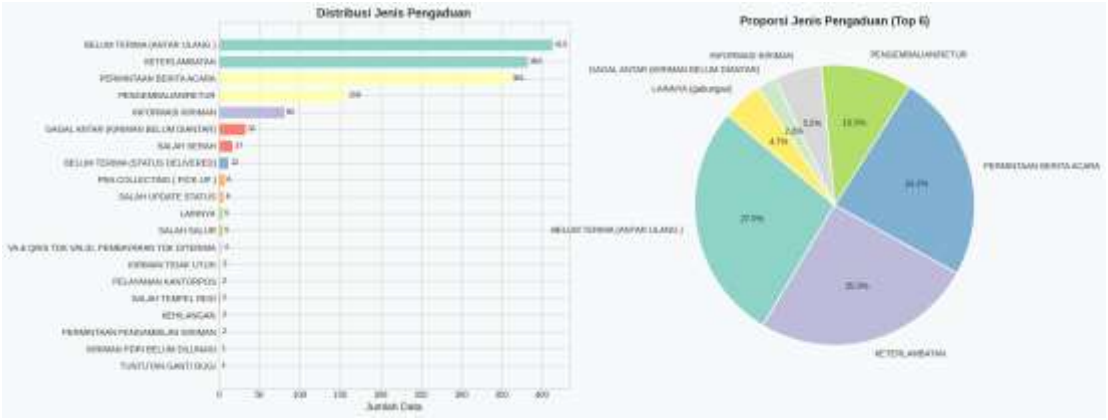



Figure 2. Complaint Category Distribution

The results indicate that delivery-related complaints dominate the dataset, suggesting that shipment and delivery processes remain the primary source of customer service issues.

**Shipment Service and Complaint Trend Analysis**

The distribution of complaints based on shipment service type and monthly complaint trends was analyzed to identify service categories that generated the highest number of customer complaints.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16280
e-ISSN : 2541-2019
p-ISSN : 2541-044X

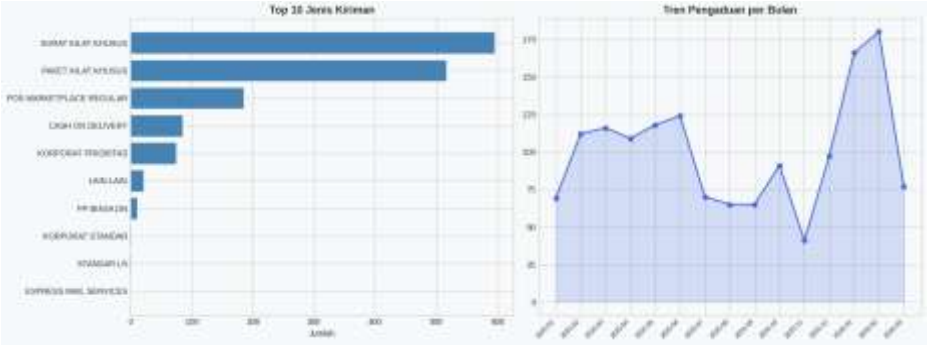



Figure 3. Shipment Service and Monthly Complaint Trend

The results show that *Surat Kilat Khusus* and *Paket Kilat Khusus* generated the highest number of complaints. Furthermore, the monthly complaint trend reached its peak in February 2026, indicating increased operational challenges during that period.

**Text Preprocessing Results**

Text preprocessing was conducted through data cleaning, case folding, and stopword removal. These processes successfully eliminated noise such as URLs, special characters, duplicate information, and irrelevant terms, resulting in cleaner textual data suitable for feature extraction.

The preprocessing results demonstrate that complaint texts became more structured and informative after removing irrelevant textual components.

**Word Cloud Analysis**

Word cloud visualization was generated to identify the most frequently occurring terms within customer complaints.

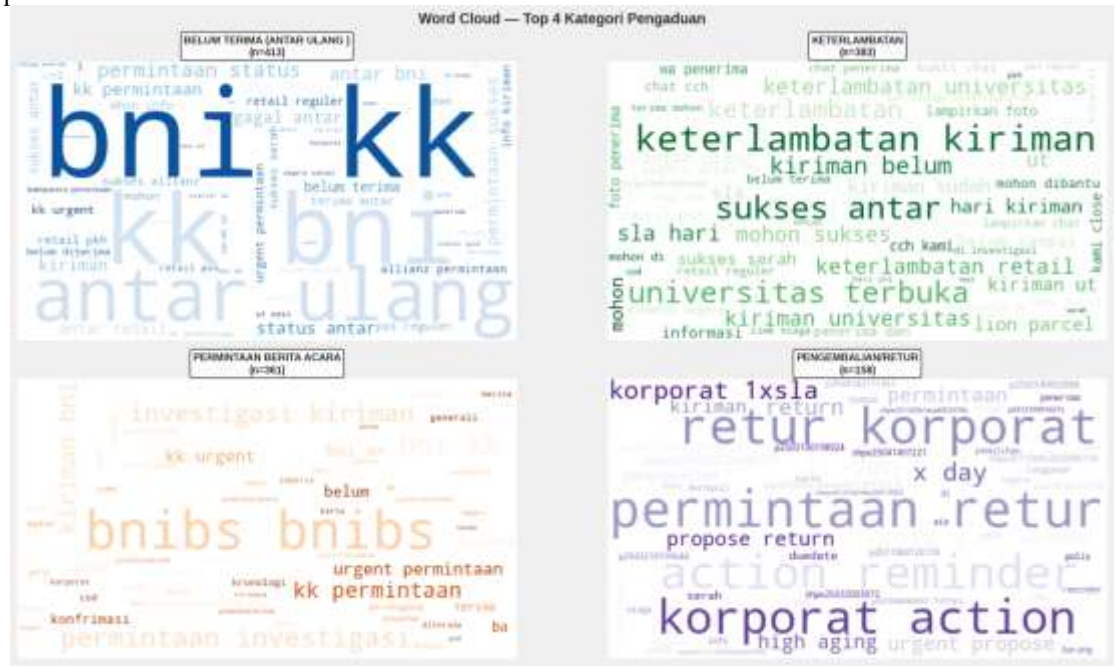



Figure 5. Word Cloud Visualization

The visualization reveals that each complaint category contains distinctive keywords. In the *Undelivered Shipment (Redelivery)* category, terms such as *delivery*, *redelivery*, and *status* appeared most frequently. Meanwhile, the *Delivery Delay* category was dominated by terms related to shipment delays and delivery schedules.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16280
e-ISSN : 2541-2019
p-ISSN : 2541-044X

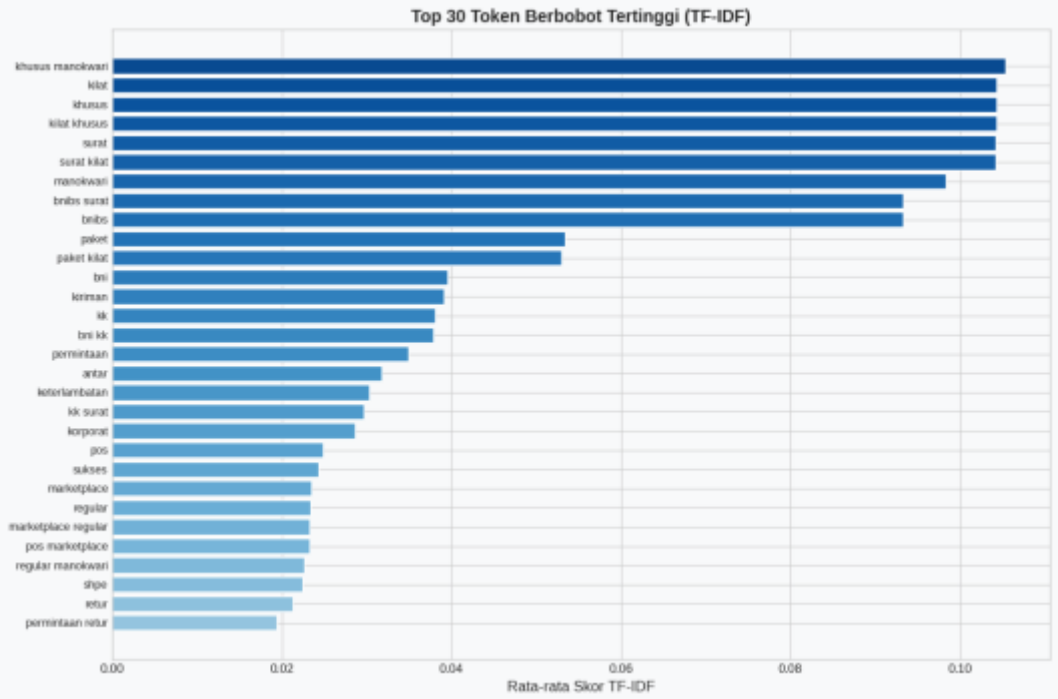
**TF-IDF Feature Extraction Results**

The TF-IDF method was applied to transform complaint texts into numerical feature vectors. According to (Agustina et al., 2024), TF-IDF is effective in representing term importance within textual datasets and improving classification performance

Table 4. TF-IDF Information

| Parameter | Value |
|---|---|
| Shape Training Matrix | (1192,1397) |
| Shape Testing Matrix | (298,1397) |
| Number of Tokens | 1.397 |
| Matrix Density | 0,99% |

Figure 6. Top TF-IDF Features




The TF-IDF transformation generated 1,397 textual features that were subsequently used as input for the classification models.

**Classification Performance**

The performance of the Naive Bayes and Random Forest classifiers was evaluated using Accuracy, Precision, Recall, F1-Weighted Score, and F1-Macro Score.

Table 5. Classification Performance Comparison

| Metric | Naive Bayes | Random Forest |
|---|---|---|
| Accuracy | 84.90% | 87.92% |
| Precision | 82.75% | 85.22% |
| Recall | 84.90% | 87.92% |
| F1-Weighted | 84.00% | 86.30% |
| F1-Macro | 48.41% | 70.85% |

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16280
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The results indicate that Random Forest achieved higher performance than Naive Bayes across all evaluation metrics. The Random Forest model obtained the highest accuracy of 87.92%, while Naive Bayes achieved an accuracy of 84.90%.

**Confusion Matrix Results**

Figure 7. Confusion Matrix

The confusion matrix demonstrates that most complaint categories were correctly classified by the Random Forest model. The *Official Report Request* category exhibited the highest classification performance, while a small number of minority-class complaints were misclassified.

**Cross Validation Results**

To evaluate model stability and generalization capability, 5-Fold Cross Validation was conducted.

Table 6. Cross Validation Results

| Fold | Accuracy (%) |
|---|---|
| Fold 1 | 57,05 |
| Fold 2 | 78,52 |
| Fold 3 | 76,17 |
| Fold 4 | 78,52 |
| Fold 5 | 68,79 |

The cross-validation results indicate that the proposed classification model showed performance variation when evaluated using different data partitions. The average accuracy of 71.81% demonstrates the model's ability to generalize across multiple validation subsets.

The performance gap between hold-out testing accuracy (87.92%) and the average cross-validation accuracy (71.81%) suggests that the model is affected by uneven class distribution across complaint categories. The severe class imbalance present in the dataset may cause the model to perform well on specific testing partitions while showing reduced consistency across different validation folds**.**

## DISCUSSION

This study evaluated the performance of Naive Bayes and Random Forest algorithms for customer complaint classification at PT Pos Indonesia Manokwari using TF-IDF feature extraction. The experimental results demonstrated that Random Forest outperformed Naive Bayes across all evaluation metrics. Random Forest achieved an accuracy of 87.92%, a precision of 85.22%, a recall of 87.92%, an F1-Weighted score of 86.30%, and an F1-Macro score of 70.85%. These results indicate that Random Forest was more effective in handling multiclass customer complaint classification because its ensemble learning mechanism can capture more complex feature relationships and improve generalization performance**.** The results are consistent with previous studies that reported the superiority of Random Forest over traditional machine learning algorithms in text classification tasks.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

In contrast, Naive Bayes achieved an accuracy of 84.90%, an F1-Weighted score of 84.00%, and an F1-Macro score of 48.41%. Although the overall accuracy was relatively high, the substantially lower F1-Macro score indicates difficulties in classifying minority complaint categories. Because the dataset contained a severe class imbalance, F1-Macro provides a more meaningful evaluation than accuracy alone since it assigns equal importance to all classes regardless of their size (Kyaw et al., 2024). The lower F1-Macro score suggests that Naive Bayes was less effective in recognizing minority complaint categories, which may be attributed to its assumption of feature independence when processing textual data.

The results also demonstrate the effectiveness of TF-IDF feature extraction for representing customer complaint texts. The TF-IDF transformation generated 1,397 textual features that captured important information related to complaint categories and operational issues. The use of unigram and bigram features enabled the model to capture both individual keywords and complaint-related phrases, thereby improving classification performance. In addition, the application of sublinear TF helped reduce the dominance of highly frequent terms, allowing more informative features to contribute to the classification process. However, TF-IDF relies primarily on term frequency information and does not fully capture semantic relationships between words, which may limit its effectiveness when dealing with more complex complaint texts. This limitation may contribute to classification errors, particularly for complaint categories with similar vocabularies or limited training samples.

The dataset exhibited a severe class imbalance problem, with the largest complaint category containing 412 records and the smallest category containing only 5 records, resulting in a majority-to-minority ratio of approximately 82:1. This imbalance likely influenced model performance, particularly on minority complaint classes. Furthermore, the performance gap between hold-out testing accuracy (87.92%) and the average cross-validation accuracy (71.81%) suggests that the model is affected by uneven class distribution across different data partitions. Although Random Forest achieved the highest overall performance, the variation observed during cross-validation indicates that class imbalance remains a challenge for achieving consistent generalization across all complaint categories. Therefore, future studies should consider applying data balancing techniques to improve classification performance on minority classes and enhance model stability.

**Limitation of Study**

This study has several limitations. First, the dataset was collected only from PT Pos Indonesia Manokwari, which may limit the generalizability of the findings to other branches or logistics companies. Second, the dataset exhibited an imbalanced class distribution, which may affect the classification performance of minority complaint categories. The minority classes contain very few samples, which limits the reliability of performance evaluation for those categories. Third, this study only compared two machine learning algorithms, namely Naive Bayes and Random Forest, without evaluating more advanced approaches such as Gradient Boosting, XGBoost, or deep learning models.

## CONCLUSION

This study compared the performance of Naive Bayes and Random Forest algorithms for customer complaint classification at PT Pos Indonesia Manokwari using TF-IDF feature extraction. A total of 1,490 customer complaint records distributed across twelve complaint categories were processed through data cleaning, case folding, stopword removal, and feature extraction stages before classification.

The experimental results demonstrated that Random Forest outperformed Naive Bayes across all evaluation metrics, achieving higher Accuracy, Precision, Recall, F1-Weighted, and F1-Macro scores. However, further improvement is required to address class imbalance and improve model generalization across different data partitions, as indicated by the difference between hold-out testing and cross-validation results. These findings indicate that Random Forest has strong potential for supporting automated customer complaint classification and improving complaint management efficiency at PT Pos Indonesia Manokwari.

The main contribution of this study lies in the comparative evaluation of two widely used machine learning algorithms using real-world customer complaint data. The results provide practical insights for developing intelligent customer service systems and enhancing operational decision-making processes within postal and logistics services.

Future research may utilize larger and more diverse datasets collected from multiple branches of PT Pos Indonesia. In addition, data balancing techniques and advanced machine learning approaches, such as deep learning and transformer-based models, should be explored to further improve classification performance and model generalization.

## REFERENCES

Ahmadi, M., Khajavi, M., Varmaghani, A., Ala, A., Danesh, K., & Javaheri, D. (2025). *Leveraging Large Language Models for Cybersecurity: Enhancing SMS Spam Detection with Robust and Context-Aware Text Classification*. Retrieved from http://arxiv.org/abs/2502.11014

*name of corresponding author

Aprianti, Y., Lia Hananto, A., Shofiah Hilabi, S., Informasi, S., & Buana Perjuangan Karawang, U. (2025). *Klasifikasi Sentimen Komentar Pengguna pada Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes*. *9*, 2025. https://doi.org/10.47002/metik.v9i1.1023

Breiman, L. (2001). *Random Forests* (Vol. 45).

C. A. Nurhaliza Agustina, R. Novita, Mustakim, & N. E. Rozanda. (2024). The Implementation of TF-IDF and Word2Vec on Booster Vaccine Sentiment Analysis Using Support Vector Machine Algorithm. *Procedia Computer Science*, *234*, 156–163.

Fahrezi Putra Ichsansyah, R., & Korespondensi, P. (2026). *PERBANDINGAN ALGORITMA NAÏVE BAYES DAN RANDOM FOREST DALAM KLASIFIKASI SENTIMEN ULASAN PENGGUNA KREDIVO DI PLAY STORE COMPARISON OF NAÏVE BAYES AND RANDOM FOREST ALGORITHMS IN SENTIMENT CLASSIFICATION OF KREDIVO USER REVIEWS ON THE PLAY STORE*. *13*(2), 297–308.

Fitriyani, A., Ibrahim, I., Studi Teknik Informatika, P., Belitung No, J., Sumur Bandung, K., Bandung, K., & Barat, J. (2026). *Sistemasi: Jurnal Sistem Informasi Perbandingan Kinerja Algoritma Naïve Bayes, Random Forest, dan Support Vector Machine dalam Analisis Sentimen Aplikasi Weverse Performance Comparison of Naïve Bayes, Random Forest, and Support Vector Machine Algorithms in Sentiment Analysis of the Weverse Application*. Retrieved from http://sistemasi.ftik.unisi.ac.id

Halim, A., Zidan, F., Handayani, I., & Anggara, A. (2025). *Sentiment analysis of the 2024 election using the naïve bayes method using data x*. *14*(2), 225–234. Retrieved from www.ejournal.isha.or.id/index.php/Mandiri

Helmayanti, S. A., Hamami, F., & Fa'rifah, R. Y. (2023). PENERAPAN ALGORITMA TF-IDF DAN NAÏVE BAYES UNTUK ANALISIS SENTIMEN BERBASIS ASPEK ULASAN APLIKASI FLIP PADA GOOGLE PLAY STORE. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, *4*(3), 1822–1834. https://doi.org/10.35870/jimik.v4i3.415

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., & Alegre, E. (2023). *Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach*. Retrieved from https://talosintelligence.com/reputation_center/email_rep

Khoerunnisa, S., Shiddieq, D. F., & Nurhayati, D. (2025). Penerapan Algoritma Naive Bayes dengan Teknik TF-IDF dan Cross Validation untuk Analisis Sentimen Terhadap Starlink. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(2), 566–577. https://doi.org/10.57152/malcom.v5i2.1852

Kyaw, P. H., Yoon, Y. I., & Kim, J. H. (2024). A Systematic Review of Deep Learning Techniques for Imbalanced Classification Problems. Electronics, 13(19), 3823. https://doi.org/10.3390/electronics13193823

Meinita, R., & Anshori, I. F. (2025). Perbandingan Algoritma Naïve Bayes dan Random Forest untuk Klasifikasi Intent Chatbot Layanan Pelanggan. *Jurnal Algoritme*, *6*(1), 186–198. https://doi.org/10.35957/algoritme.v5i3.12639

Nugroho, A. K., Hayati, L. N., & Jabir, S. R. (2025). Analisis Perbandingan Metode Naive Bayes dan Random Forest pada Klasifikasi Sentimen Publik terhadap Aplikasi Identitas Kependudukan Digital (IKD). *Jurnal Algoritma*, *22*(2). https://doi.org/10.33364/algoritma/v.22-2.2729

Prakoso Indaryono, N. A. (2024). ANALISA PERBANDINGAN ALGORITMA RANDOM FOREST DAN NAÏVE BAYES UNTUK KLASIFIKASI CURAH HUJAN BERDASARKAN IKLIM DI INDONESIA. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *9*(1), 158–167. https://doi.org/10.29100/jipi.v9i1.4421

Q. Li, et al. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, *13*(2), 31.

Riswanto, R., Hidayat, T., & Nasiri, A. (2026). PENERAPAN ALGORITMA RANDOM FOREST DALAM BERBAGAI BIDANG KEILMUAN : SYSTIMATIC LITERATUR REVIEW. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *11*(1), 285–294. https://doi.org/10.29100/jipi.v11i1.7316

S. Akuma, T. Lubem, & I. T. Adom. (2022). Comparing Bag of Words and TF-IDF with Different Models for Hate Speech Detection from Live Tweets. *International Journal of Information Technology*, *14*(7), 3629–3635.

Santoso, B. A., Nugroho, I., & Asyfiya, D. U. (2025). *TIN: Terapan Informatika Nusantara Perbandingan Algoritma Naïve Bayes, Support Vector Machine, dan Random Forest Untuk Analisis Sentimen Komentar Politik Youtube*. *6*(4). https://doi.org/10.47065/tin.v6i4.8326

Y. Mao, Q. Liu, & Y. Zhang. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, *36*(4), 102048.

*name of corresponding author