Bit-Tech (Binary Digital - Technology)

Vol.8, No.2, August 2025
Available online at: http://jurnal.kdi.or.id/index.php/bt

Automatic Classification of Public Complaints Using Naive Bayes

Rico Andrean Hardiansyah[1)*], Jamaludin Indra[2)], Dwi Sulistya Kusumaningrum[3)], Tohirin Al Mudzakir[4)]

[1)2)3)4)] *Teknik Informatika, Fakultas Ilmu Komputer, Universitas Buana Perjuangan Karawang, Indonesia*

[1)]if21.ricohardiansyah@mhs.ubpkarawang.ac.id

[2)]jamaludin.indra@ubpkarawang.ac.id

[3)]dwi.sulistya@ubpkarawang.ac.id

[4)]tohirin@ubpkarawang.ac.id

Article history:

Received 12 Aug 2025;
Revised 14 Aug 2025;
Accepted 19 Aug 2025;
Available online 10 Dec 2025

Keywords:

Naive Bayes
Public Complaint Classification
SMOTE
Text Mining
TF-IDF

Abstract

Public complaint services are essential for improving government service quality by providing a direct channel for citizens to report issues. In Karawang Regency, the Tanggap Karawang (TANGKAR) platform serves this function; however, the manual classification of complaints causes delays and potential misrouting, especially due to the highly imbalanced distribution of complaint categories. This study develops an automatic classification model for public complaints in eight categories economy, education, health, social, infrastructure, security, environment, and transportation by integrating Term Frequency–Inverse Document Frequency (TF–IDF), Multinomial Naive Bayes, and Synthetic Minority Oversampling Technique (SMOTE). This integration addresses domain-specific class imbalance challenges, combining the computational efficiency of Naive Bayes, the feature representation strength of TF–IDF, and the improved minority class recognition from SMOTE. A dataset of 800 complaint records from TANGKAR underwent preprocessing, including cleaning, case folding, normalization, tokenizing, stemming, and stopword removal. TF–IDF with unigram and bigram features was used for feature extraction, followed by classification under two scenarios: original unbalanced data and balanced data via SMOTE. Evaluation metrics included accuracy, precision, recall, F1-score, and confusion matrix. The model achieved 85.09% accuracy without SMOTE and 83.40% with SMOTE, with notable improvement in detecting minority categories after balancing. Although overall accuracy slightly decreased, SMOTE enhanced equitable prediction across all categories. This approach advances current public complaint classification methods by adapting to the linguistic diversity and uneven category distribution in actual e-government data, supporting faster and more accurate decision-making in public complaint management systems.

I. INTRODUCTION

Public complaint services are one of the important instruments in efforts to improve the quality of public services. The existence of a well-managed complaint channel allows the government to obtain direct information from the public about problems that occur in the field, while accelerating the handling process [1]. In Karawang Regency, the Regional Government through the Communication and Information Service developed Tanggap Karawang (TANGKAR) as an e-government platform that facilitates the submission of public aspirations and complaints. This platform accepts various types of complaints, ranging from infrastructure, health, environmental, social, security, education, economy, to transportation. However, the complaint classification process at TANGKAR is still carried out manually, which has caused delays in the distribution of reports to relevant agencies.

These problems can be overcome through the application of text mining technology, especially using classification algorithms such as Naive Bayes, which are known to be simple, fast, and have competitive performance on high-dimensional data [2]. Research in various domains has shown the effectiveness of Naive

[*] Corresponding author

doi: 10.32877/bt.v8i2.2992

Bayes, both on balanced and unbalanced data. The application of Naive Bayes combined with the Synthetic Minority Oversampling Technique (SMOTE) has been shown to improve precision and f-measure in sentiment analysis, particularly for detecting minority classes, although the overall accuracy improvement is not always significant [3].

In the field of education, the application of SMOTE to the classification of question topics based on Naive Bayes results in higher accuracy than without the use of the oversampling technique [4]. Meanwhile, in the health sector, the use of SMOTE-N in the classification of hepatitis diseases based on Naive Bayes has been shown to be effective in improving model performance, especially in minority classes that previously had low detection rates [5].

The increase in accuracy due to the use of SMOTE was also seen in a study that combined Naive Bayes with genetic algorithms in the classification of customer churn, where accuracy increased from 47.10% to 78.20% [6]. The effectiveness of SMOTE in improving the performance of Convolutional Neural Networks (CNN) on unbalanced datasets has also been proven through testing on various classification scenarios [7]. In addition, the application of the Synthetic Minority Oversampling Technique (SMOTE) technique in text sentiment analysis has been proven to be able to significantly improve accuracy, precision, and recall on unbalanced data [8].

However, most of the research did not focus on the domain of public complaints, which have the unique characteristic of a highly unequal distribution of classes. For example, in TANGKAR data, the infrastructure category has a much higher number of complaints than the economic or education category. This condition can cause the model to be biased against the majority class, so that the accuracy for the minority category is low. Previous research has indeed built a Naive Bayes-based public complaint classification system, but it has not addressed the problem of class imbalance, and even the application of SMOTE to similar data has been reported to significantly reduce accuracy and F1-score [9]. A high accuracy of 92% in the classification of community reports using Naive Bayes has also been achieved, but without the application of data balancing techniques [10]. This indicates that, while prior works have proven Naive Bayes' general effectiveness, they either overlook the severe imbalance present in public complaint datasets or fail to adapt balancing techniques in a way that preserves overall model performance.

Based on these gaps, this study offers a methodological adaptation that directly addresses the imbalance challenge in public complaint classification. By integrating TF–IDF for robust textual feature representation, Naive Bayes for efficient probabilistic classification, and SMOTE for minority class oversampling, the proposed approach is tailored to the linguistic diversity and skewed category distribution of real-world e-government data.

The objectives of this study are: (1) to build an automatic classification model that is able to group complaints into eight categories; (2) implement SMOTE to increase representation of minority categories; (3) compare the performance of the model on the original data and the data that has been balanced; and (4) conducting evaluations using accuracy, precision, recall, F1-score, and confusion matrix metrics. Each of these objectives is designed to systematically address the identified shortcomings ensuring that the model not only achieves high aggregate accuracy but also maintains equitable performance across all complaint categories. This research is expected to make a theoretical contribution to the development of text classification methods and practical benefits in accelerating and improving the accuracy of handling public complaints in local government environments.

II. Literature Review

The Naive Bayes algorithm is one of the simple but effective classification methods, especially in the case of natural language processing (NLP) and text classification. Its main advantages lie in its ability to handle high-dimensional data at low computational costs, as well as its probabilistic properties that facilitate the interpretation of results [11]. In various studies, Naive Bayes has been shown to have competitive performance when balanced with good text preprocessing techniques, such as tokenizing, stemming, and word weighting using Term Frequency–Inverse Document Frequency (TF–IDF). However, performance can decrease significantly when the training data has an unbalanced class distribution, as the model tends to be biased towards the majority class [12].

The problem of class imbalance in text data is a common challenge in machine learning. In these situations, the model tends to predict the class that has the largest proportion in the dataset, making it difficult to identify categories with a small amount of data (minority class). One of the widely used approaches to address this problem is the Synthetic Minority Oversampling Technique (SMOTE). This method works by creating a new synthetic sample on a minority class based on interpolation between existing data points, resulting in a more balanced distribution between the classes [13].

Various studies have shown that SMOTE is able to improve the performance of text classification models on unbalanced datasets. For example, the implementation of SMOTE in the classification of user reviews has been shown to increase the value of precision and recall, resulting in better detection of minority classes even though the overall accuracy sometimes decreases slightly [14]. The combination of SMOTE with text representation methods such as TF–IDF has been shown to provide consistent results on various types of algorithms, including Naive Bayes, Support Vector Machine (SVM), and Random Forest, especially in improving F1-scores in minority classes [15]. When comparing these works, it is evident that improvements are often achieved at the cost of slight accuracy reduction, suggesting a trade-off between balanced class performance and aggregate metrics a factor that is particularly relevant for real-world public service applications where fairness across categories is critical.

In addition to the standard SMOTE, several method modifications have been developed to address its drawbacks, such as the potential to produce synthetic samples that are not representative or too similar to the original data. One of them is Improved Random-SMOTE which utilizes the standard deviation feature to produce synthetic samples with better variation, while reducing the risk of overfitting [16]. This method is relevant for small to medium-sized datasets, including public complaint datasets, where the amount of data in certain categories is very limited. However, despite these algorithmic advancements, no prior work has demonstrated their targeted application to government-managed complaint platforms with multi-class imbalances and diverse linguistic expressions.

The application of SMOTE is not only limited to conventional algorithms, but can also be used to improve the performance of deep learning-based models. The integration of SMOTE prior to Convolutional Neural Networks (CNN) training on text data was shown to improve the model's ability to recognize patterns from minority classes, noting the need to adjust feature dimensions and model complexity for optimal results [17]. Such findings reinforce the cross-method applicability of SMOTE, yet they do not address the operational constraints of e-government systems where interpretability and low computational cost advantages of Naive Bayes remain essential.

On the other hand, the development of the Naive Bayes method is ongoing, including its application to spatial and temporal data that require the expansion of feature space to maintain classification accuracy. This method is also used in privacy-preserving classification scenarios, which proves the flexibility of Naive Bayes to adapt to a wide range of application needs [18].

Although the literature shows that SMOTE and Naive Bayes are effective in a variety of domains, studies that specifically examine their application to public complaints are still very limited. The public complaint dataset has unique characteristics: (1) a large number of categories with an unbalanced distribution, (2) a data format in the form of free text with high vocabulary variation, and (3) a direct implication to public services in the event of misclassificationSome studies, such as those on complaint classification in university service systems or municipal e-government portals, partially address this domain but either ignore the imbalance problem or fail to integrate balancing methods effectively often resulting in biased models toward majority classes. The majority of previous studies used common datasets such as product reviews, news, or social media, so they have not explored in depth the impact of balancing on per-class metrics in the domain of public complaints. This study differentiates itself by applying Naive Bayes with TF–IDF weighting and SMOTE specifically to actual multi-class public complaint data from the TANGKAR platform, enabling a direct evaluation of performance trade-offs between overall accuracy and balanced category recognition in a real operational context.

III. Methods

This research stage is designed in a structured manner to ensure that each process runs systematically and measurably. The overall flow of the method can be seen in Figure 1, which visualizes the sequence of steps from the collection of raw data to the visualization of the classification results. This process includes eight main interrelated stages, namely: (1) data collection as the foundation of the research, (2) text preprocessing to prepare raw data into a processable form, (3) extraction of features to represent text in numerical form, (4) division of data into training and test sets, (5) modeling using classification algorithms, (6) evaluation of results to measure model performance, (7) visualization to display the dominant word pattern in the data, and (8) preparation of conclusions based on the analysis obtained.

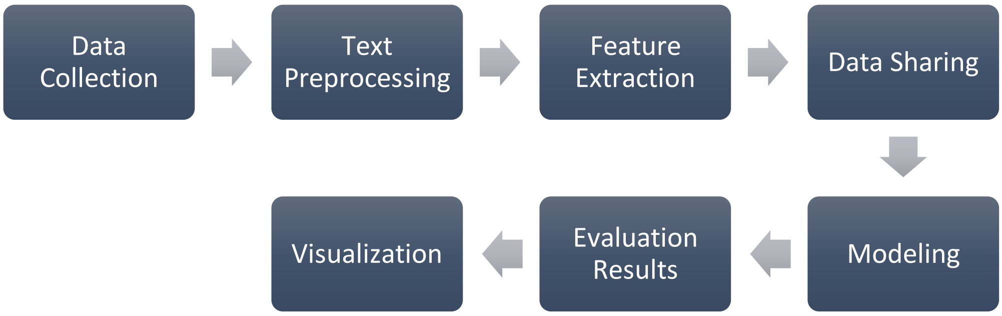

Fig 1. Research Stages

Based on the research flow illustrated in Figure 1, the following is explained in detail each stage that is passed in this study.

A. Data Collection

This research uses data from the Karawang Tanggap (Tangkar) platform, which is managed by the Karawang Regency Communication and Information Office. The platform contains public complaints recorded throughout 2024 through the official Tanggap Karawang website, with a total of around 2,700 reports. Each complaint entry has four main attributes, namely the date of the complaint, the category of the complaint, the content of the complaint, and the organizational unit responsible for following up on the report. In this study, as many as 800 complaint data have been classified into eight categories.

B. Text Preprocessing

Text preprocessing plays a crucial role in reducing noise and ensuring a consistent representation of text for further analysis. The process begins with data cleaning, which involves eliminating unnecessary word elements such as characters outside the A-Z alphabet, links, and special symbols like hashtags and @ symbols in usernames [19]. Next, case folding is applied, where all the letters in the text are converted to lowercase. This step helps address inconsistencies caused by capital letters in different documents, making text search and processing more uniform [20].

Following this, non-standard language and writing normalization is performed. Abbreviations and slang forms are translated into their standard equivalents using domain-specific normalization dictionaries (e.g., 'klo' becomes 'if' and 'yg' becomes 'yang') [21]. This step helps standardize non-standard words, typographical errors, or abbreviations . Tokenizing follows, which involves breaking the text into individual word tokens. For languages like Indonesian, a tokenizer sensitive to punctuation and compound words is used; however, for simpler cases, whitespace tokenization suffices after cleaning.

Stemming is the next step, which reduces words to their root form by removing various suffixes. This process is done without regard to grammatical context, simplifying the complexity in text analysis. Additionally, stopword removal is carried out to eliminate words that occur frequently but do not contribute significantly to the analysis, such as 'and', 'that', and 'in'. Finally, the output of the preprocessing stage is saved either as a clean string or a list of tokens, which will serve as input for feature extraction in the next steps of the text analysis process.

C. Feature Extraction

The feature extraction stage in this study was carried out using the Term Frequency–Inverse Document Frequency (TF–IDF) method to convert the pre-processed text into a numerical representation that can be processed by the classification algorithm. TF–IDF calculates the weight of a word based on the frequency of its occurrence in a document compared to the frequency of its occurrence throughout the corpus, so words that appear frequently in many documents will get a low weight, while words that appear infrequently but are relevant to a particular document will have a high weight. In its implementation, TF–IDF is implemented with parameter settings that consider a single ngram (unigram) and a combination of two words (bigram) to capture the meaning of a more specific phrase. The minimum document frequency limit is set to ignore words that only appear once, while the maximum document frequency is used to remove words that are too common. The result of this process is a feature matrix with dimensions corresponding to the number of unique words selected, where each value in the matrix represents the TF–IDF weight of that word in a document. This representation then becomes an input for the Multinomial Naive Bayes model for the training and testing process. The mathematical formulation of TF–IDF is presented below (1).

$$tf_{t.d} = \frac{The\ number\ of\ occurrences\ of\ term\ t\ in\ document\ d}{The\ total\ number\ of\ words\ in\ document\ d} \tag{1}$$

Term Frequency (TF) is used to calculate the number of occurrences of a word in a particular document. The formula for calculating the value of Inverse Document Frequency (IDF) is presented below (2):

$$idf_d = log\ \frac{N}{n_t} \tag{2}$$

Inverse Document Frequency (IDF) is used to determine the importance of a word by considering how many documents contain that word. The symbol N denotes the total number of documents in the dataset (3).

$$tfidf_{\ t.d} = log_{\ t.d}\ \mathrm{X}\ idf_d \tag{3}$$

Notation:

t : The term or keyword being analyzed
d : The document in which the term appears
t.d : The TF-IDF value of term ttt in document ddd
Tf : The number of occurrences of term ttt in document ddd
Idf : The inverse document frequency, which measures how rare the term ttt is across the entire set of documents

D. Data Sharing

Data sharing is carried out in a stratified split manner to maintain the proportion of complaint categories in training data and testing data. In this study, two scenarios were used. The first scenario uses the original data without a balancing process, while the second scenario applies the Synthetic Minority Oversampling Technique (SMOTE) to the training data to address the imbalance in the amount of data between categories. The implementation of SMOTE is carried out only on training data to prevent information leakage to test data. This approach allows comparison of model performance under the original data distribution conditions and at the balanced data conditions.

E. Modeling

At this stage, information analysis is carried out based on a predetermined method, namely Naïve Bayes. This step is carried out for the process of classifying public complaints based on the previously analyzed values. The expected result is the separation of complaint categories into eight main classifications, namely economy, education, health, social, infrastructure, security, environment and transportation. This classification process uses a Naïve Bayes algorithm, which applies a probabilistic approach to determine the most appropriate category based on the text of the complaint given. The Naive Bayes algorithm consists of two main stages, namely the training process and testing. In the training stage, the algorithm analyzes the data by calculating the probability distribution of each feature against the target class. Then, at the test stage, the patterns that have been obtained are used to determine the class of the new data (4).

$$P(c_j \mid w_i) \frac{P(w_i \mid c_j) \times P(c_j)}{P(w_i)} \tag{4}$$

Notation:

$P(cj|wi)$: The probability that a document belongs to category cj, given that the word wi appears.
$P(wi|cj)$: The probability that the word wi appears given that the document belongs to category cj.
$P(cj)$: The prior probability of category cj occurring across all documents.
$P(wi)$: The probability of the occurrence of word wi across all documents regardless of category.

F. Evaluation Results

Model performance testing was carried out to determine the accuracy achieved by the Naïve Bayes algorithm in the process of classifying public complaint data into eight categories, namely economy, education, health, social, infrastructure, security, environment, and transportation. The model's performance assessment is carried out by comparing the resulting classification results with the original label on the test data. Evaluation is performed using a confusion matrix to measure the performance of the model classification based on the comparison between the actual label and the predicted results, which are visualized in the form of a matrix

G. Visualization

The visualization stage is carried out to provide a clearer picture of the characteristics of the data and modeling results. The first visualization shows the distribution of the amount of data in each complaint category before and after the implementation of SMOTE. This comparison is used to demonstrate the effectiveness of data balancing techniques in reducing class distribution inequality. Furthermore, a word cloud was created for each complaint category by displaying the words that had the highest TF–IDF weight, making it easier to identify the dominant keywords that represented the main issue in each category. The results of the model's predictions are also visualized in the form of a heatmap-based confusion matrix to comprehensively describe the number of true and false predictions in each category. In addition, ROC and precision–recall curves are also displayed to evaluate the performance of the model in terms of predictive probability, especially in distinguishing between minority and majority classes. This visualization as a whole serves not only as an analysis tool, but also as a validation medium for the relevance of the extracted features and the accuracy of the classifications generated by the model.

Overall, this research method is designed to ensure that each stage runs systematically from data collection to visualization of results. The approach used allows for effective processing of public complaint data through text cleansing and normalization, numerical representation using TF–IDF, and classification based on Naive Bayes algorithms. The use of two data sharing scenarios, both without and with SMOTE, provides a comprehensive evaluation of the effect of data balancing on model performance. The evaluation and visualization stage that is carried out not only measures the level of accuracy, but also provides in-depth insights regarding the distribution of predictions and the relevance of the features used. With this design, the research is expected to be able to produce a classification model that is accurate, adaptive to data variations, and applicable to support the management and follow-up of public complaints in a faster and measurable manner.

IV. RESULTS

A. Dataset

The dataset used in this study is a collection of community complaint reports obtained from the Karawang Response (TANGKAR) platform. Each entry contains information about the source of the complaint, the content of the report, and the categories that have been defined. An example of a data snippet can be seen in Table 1, which shows four sample complaints from various categories, including infrastructure, transportation, security, and environment. This variation becomes an important basis for the classification process, as each category has different language and keyword characteristics, requiring proper pre-processing and feature extraction stages to produce an accurate prediction model.

TABLE 1
KARAWANG RESPONSE COMPLAINT DATA

Source	Contents of the Complaint	Category
Webiste	Reporting back regarding street lighting 2 weeks ago, then a few days after repairing the light went out again and never turned on again, that's how it is, sir, because it's very uncomfortable.	Infrastructure
Website	Cars parked carelessly, creating a danger to other drivers. Along the West Karawang highway, dozens of large cars are parked freely.	Transportation
Website	After school, they don't go straight back to their home website.	Security
Website	There is illegal garbage, please clean it up, and please give me CCTV so that you know who is throwing garbage carelessly, it has piled up the garbage again, thank you.	Environment

As seen in Table 1, each complaint entry reflects a diverse and specific problem according to the category that has been determined. This diversity of topics suggests that complaint data has significant linguistic complexity and vocabulary variations, ranging from technical descriptions related to infrastructure to social and environmental behavioral complaints. This complexity makes the classification process require a careful text processing approach so that the model is able to accurately recognize the language patterns in each category. As such, this dataset is not only a key resource in model training and testing, but also plays an important role in evaluating the algorithm's ability to automatically distinguish different types of public complaints.

B. Preprocessing Text

Pre-processing is the initial process in text analysis that serves to clean and prepare raw data before entering the feature extraction and model building stages. In this study, preprocessing was carried out on the complaint content column (text) with the stages of text cleansing, case folding, normalization, tokenization, stemming, and stopword removal. The initial stage of text pre-processing consists of three main processes. Cleaning is done to remove irrelevant characters such as punctuation, numbers, links, and mentions, so that the text becomes clean and free of elements that do not provide significant meaning. Furthermore, Case Folding converts the entire letter to lowercase to avoid word shape differences due to capitalization, so that word matching becomes more consistent. Finally, normalization is used to improve non-standard words, abbreviations, or typos into word forms according to Indonesian rules, in order to ensure uniformity and accuracy in word representation. The final results after going through these three processes are shown in Table 2.

TABLE 2
RESULTS AFTER CLEANING, CASE FOLDING, AND NORMALIZATION

Before	After
lampu jalan umum yg mati di daerah bengle majalaya samping perum sadjati garden city klo malem sangat gelap	lampu jalan umum yang mati di daerah bengle majalaya samping perum sadjati garden city kalau malam sangat gelap

The next stage involves three advanced processes, namely Tokenizing, Stemming, and Stopword Removal. In the Tokenizing process, text is broken down into pieces of words or tokens to facilitate the analysis of sentence structure. Furthermore, Stemming is used to return a word to its basic form by removing the suffix, although in this example there is no change in the word because all the words are already in the base form. Finally, Stopword

Removal removes common words that appear frequently but do not contribute significantly to the analysis, such as the words "yang" and "di". The end result of these three stages results in a more concise and focused representation of the text that will be used in the feature extraction process. As shown in table 3.

TABLE 3
RESULTS AFTER CLEANING, CASE FOLDING, AND NORMALIZATION

Before	After
lampu jalan umum yang mati di daerah bengle majalaya samping perum sadjati garden city kalau malam sangat gelap	['lampu', 'jalan', 'mati', 'daerah', 'bengle', 'majalaya', 'samping', 'perum', 'sadjati', 'garden', 'city', 'malam', 'gelap']

With the completion of all pre-processing stages including Cleaning, Case Folding, Normalization, Tokenizing, Stemming, and Stopword Removal, the complaint text data has been successfully transformed from a raw form full of variations to a clean, consistent, and focused word representation on relevant information. This transformation not only reduces noise and redundancy in the data, but also ensures that every word left has a significant contribution to the analysis process. These final results serve as a strong foundation for the feature extraction and modeling stages, as the quality of the input data directly affects the performance of the classification model that will be built in this study.

C. TF-IDF Word Weighting

At the text classification stage, the use of word weights is very helpful in extracting the meaning of the document. TF-IDF is used to assign important values to words based on their frequency in documents and the degree to which they exist in other sets of documents. Term Frequency (TF) indicates the frequency with which a word appears in a document, while Inverse Document Frequency (IDF) measures the importance of the word based on how rarely it appears throughout the document. The result of the multiplication of the two produces a TF-IDF value that reflects the level of importance of a word to the document. Table 4 below shows the top words with TF-IDF values generated from public complaint data.

TABLE 4
WORD WEIGHTING

Term	TF	IDF	TF-IDF
road	303	2.29	692.94
karawang	111	3.11	345.75
garbage	77	3.65	280.78
Parking	76	3.68	279.85
lampu	63	3.65	229.73

Based on Table 4. The word "road" shows the highest Term Frequency value of 303 and the TF-IDF value of 692.94, which is the most dominant word in the corpus of public complaint data. It shows that issues related to road infrastructure are one of the most frequently submitted complaint topics by the community. In addition, the word "karawang" obtained a TF-IDF weight of 345.75.

D. Naive Bayes Modeling

At this stage, the modeling process is applied to the pre-processing data using the Naive Bayes algorithm, with a total of 800 complaint data. The data used is actual data without balancing or oversampling. The dataset is divided into two parts, namely 80% for training data and 20% for test data, with a total of 161 test data. Based on the distribution of categories shown in Figure 2, it can be seen that the infrastructure category has the highest amount of data, followed by social and environmental, while the economic and education categories have the lowest amount of data. This imbalance is one of the factors that can affect the performance of the model in the initial testing stage.

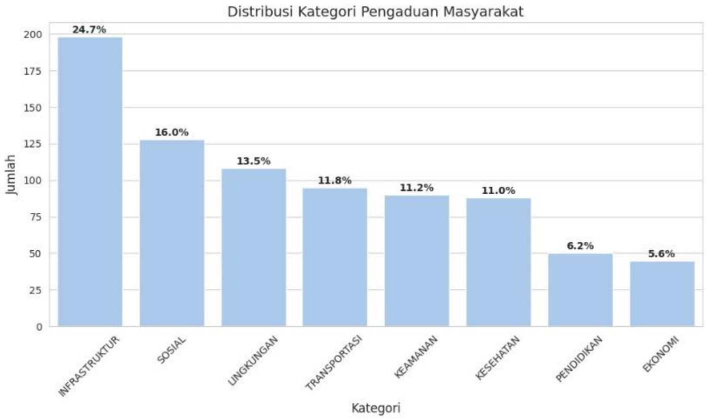

Fig 2. Data Distribution

After that, the oversampling process uses the Synthetic Minority Oversampling Technique (SMOTE), the distribution of data in each complaint category becomes balanced. Each of them has a relatively similar amount of data, which is 138 data per category.

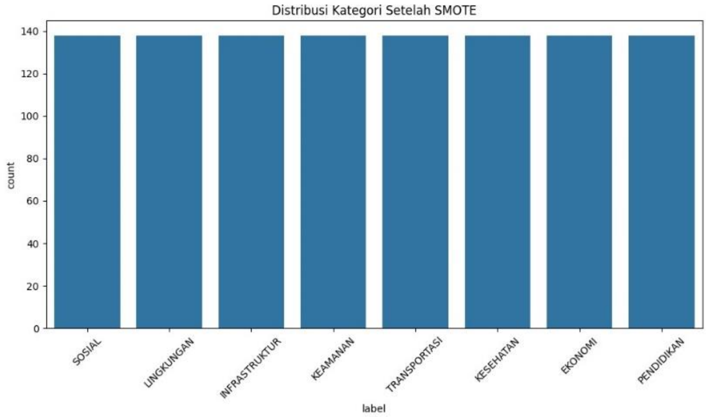

Fig 3. Data Distribution After SMOTE

The distribution of data after the oversampling process is shown in Figure 3. At this stage, the dataset is divided in a 70:30 ratio, of which 70 percent of the data is used for the model training process and the other 30 percent, as many as 241 data, is used for testing. The classification is carried out using the Naive Bayes algorithm.

TABLE 5
ACCURACY VALUES

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	85,09%	85,92%	85,26%	84,67%
Naive Bayes With SMOTE	83,40%	83,09%	83,43%	82,38%

Table 5 presents the results of the application of the Naïve Bayes algorithm to original data and data that has gone through the oversampling process using the Synthetic Minority Oversampling Technique (SMOTE). Modeling using original data without class balancing, the model produced quite good performance with an accuracy of 85.09%, accuracy of 85.92%, recall of 85.26%, and an f1-score of 84.67%. As a technique to overcome the imbalance of distribution between categories, an oversampling process is carried out using SMOTE. After the data was balanced and the model was retrained, the accuracy results were 83.40%, accuracy 83.09%, recall 83.43%, and f1-score 82.38%. Despite a slight decrease in scores on some evaluation metrics, the implementation of SMOTE contributed to improving the distribution of classes so that the model could recognize minority categories more evenly.

Based on the results of the model evaluation on the original data and the data that has been balanced with SMOTE, the next step is to see the model's performance directly in the test data classification process. Table 6 presents some of the results of the model's classification of the test data, which contain the content of the complaint, the actual label, and the prediction label generated by the Naïve Bayes algorithm.

TABLE 6
CLASSIFICATION RESULTS FROM TEST DATA

Contents of the Complaint	Label	Predictions
The smell of garbage from BT15 is getting worse	Environment	Environment
A Tale of Thieves and Thieves	Security	Security
The driver allegedly hit a truck that was parked illegally on the side of the road. Dark road conditions and the existence of trucks that often park carelessly	Transportation	Transportation
Individuals on behalf of the Environmental Service, have often come	Security	Health
I want to improve my second child's birth certificate	Social	Social
There is a teenager using marijuana-type drugs	Security	Security

In general, the model was able to classify most of the test data correctly, such as in complaints related to "garbage smell from BT15 getting worse" which was identified correctly as the Environmental category, as well as "purwasari prone to brawls and begals" which were predicted to be appropriate in the Safety category. However, there are also cases of misclassification, for example complaints of "individuals on behalf of the environmental agency, have come so often" which should be included in the Safety category but are predicted to be Health. These findings suggest that although the model's accuracy is relatively high, there is still room for improvement, particularly in distinguishing categories that have similarities in context or vocabulary.

E. Evaluation

The evaluation of the performance of the classification model aims to assess the extent to which the model can identify and group public complaint data into eight predetermined classes, namely economy, infrastructure, security, health, environment, education, social, and transportation. The evaluation of the model used in this study utilizes the confusion matrix as an evaluation tool that is able to show how well the model is in classifying, by displaying the number of correct and false predictions for each class.

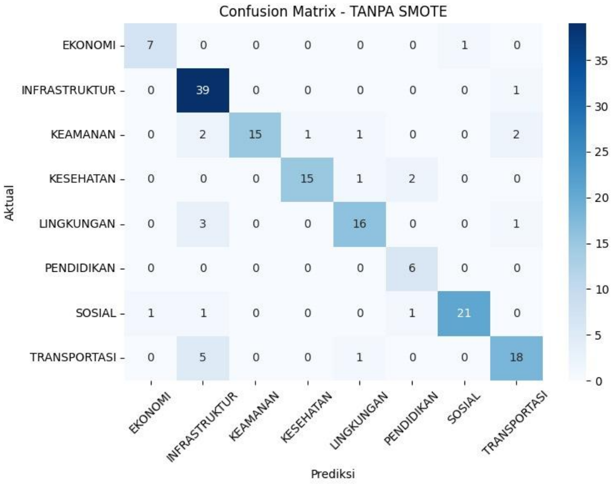

Fig 4. Confusion Matrix Naive Bayes

Based on Figure 4 of the results of the evaluation through the confusion matrix, the model has a good level of accuracy in classifying several categories of public complaints. The Infrastructure category obtained the correct classification results, namely 39 data out of a total of 40. The Social and Transportation categories also performed well with 21 and 18 data being correctly classified respectively. Categories such as Safety, Health, and Environment Although most of the data was correctly predicted, there were still misclassifications to other categories. This indicates that there is a difference in accuracy between categories that is influenced by the distribution of the amount of data. The Economics and Education categories, which have a smaller amount of data, show lower accuracy and tend to be classified into more dominant classes.

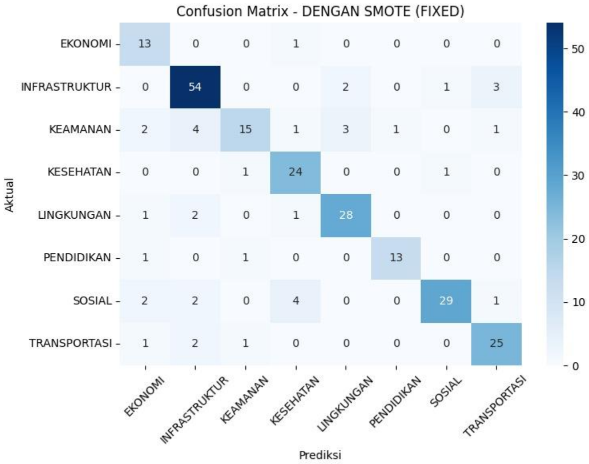

Fig 5 Confusion Matrix Naive Bayes with SMOTE

Based on Figure 5 confusion matrix, the application of the SMOTE method to training data succeeded in increasing the distribution of classes, so that the Naïve Bayes model could recognize more categories of community complaints. This is shown by the True Positive (TP) values that are quite high in most classes, such as Infrastructure (TP = 54), Environment (TP = 28), Social (TP = 29), and Transportation (TP = 25). The Economics and Health categories also showed good results, with (TP = 13) and (TP = 24 respectively), as well as a low number of misclassifications. Showing that the model has a fairly stable predictive ability in these classes. Misclassification still occurs especially in the Security category, where the model often misclassifies data into classes Infrastructure, Environment, and Social.

V. Discussion

The results showed that the application of the Naive Bayes algorithm to the original data resulted in an accuracy of 85.09%, precision of 85.92%, recall of 85.26%, and an f1-Score of 84.67%. After oversampling using the Synthetic Minority Oversampling Technique (SMOTE), the accuracy decreased to 83.40%, accuracy 83.09%, recall 83.43%, and f1-Score 82.38%. However, the results on the confusion matrix show an increase in detection in minority categories, such as Economy and Health, which previously had a low number of true predictions. This improvement in minority class recognition is particularly relevant in the context of public service platforms, as accurate classification of less-reported issues can lead to faster routing to the appropriate agencies, potentially reducing policy response times and improving citizen satisfaction through more equitable service delivery. These findings are in line with studies on the classification of rebate texts that show that SMOTE is effective in improving minority class representation and detection in categories with limited data [22].

The application of SMOTE to the Double-Layered Convolutional Neural Networks (CNN) architecture for medical diagnosis has been proven to improve detection accuracy in rare disease classes, although it requires special attention to feature dimension setting and regularization techniques to prevent overfitting [23]. In the context of the classification of public complaints, the application of Naive Bayes without a balancing technique to the complaints of students of the Faculty of Engineering, University of Muhammadiyah Makassar managed to achieve an accuracy of 91%, showing that this model is able to provide quite good performance despite facing class imbalances [24], but has a bias against the majority class. This supports the use of balancing techniques such as SMOTE to overcome uneven distributions. However, other studies have underlined that SMOTE has limitations, such as the potential for overfitting due to the high similarity between synthetic data and the original data, as well as the possibility of producing samples that do not fully represent natural distributions [25]. In the domain of public complaints, this limitation may be even more pronounced because synthetic samples could unintentionally distort the original semantic context of citizen reports creating data points that appear linguistically valid but misrepresent the intent, urgency, or nuance of the actual complaint. This risk underscores the importance of post-balancing validation and domain expert review before model deployment in real operational systems.

In addition, some studies state that Naive Bayes' performance on unbalanced data is highly dependent on the quality of the features used. Research on early detection of cervical cancer, for example, suggests that the combination of Naive Bayes with SMOTE can improve accuracy in minority classes, but its effectiveness is greatly influenced by precise feature extraction techniques [26]. Similar results were also seen in traditional and national song classification research, where the implementation of SMOTE succeeded in improving the representation of minority classes and improving the overall performance of the model [27].

The limitations of this study include the relatively small size of the dataset (800 data), significant distribution differences between categories, and the use of only one oversampling method, SMOTE, without comparing it with other balancing techniques such as Borderline-SMOTE or Adaptive Synthetic Sampling (ADASYN).

Moreover, the reliance on text preprocessing and TF–IDF as the sole feature extraction method may limit the model's ability to capture deeper semantic relationships in complaint narratives, which could be addressed in future research using contextual embeddings. For further research, it is recommended to increase the size of the dataset, especially the minority category, explore alternative balancing methods or a combination of balancing and ensemble techniques, compare Naive Bayes with other algorithms such as Support Vector Machine or Random Forest, and consider embedding-based text representations such as FastText or IndoBERT to improve the quality of features prior to classification.

VI. Conclusions

This research contributes significantly to the development of a public complaint classification method by integrating Term Frequency–Inverse Document Frequency (TF–IDF), the Naive Bayes algorithm, and the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in multi-class data. The novelty lies in the tailored combination of these techniques, which not only ensures computational efficiency but also adapts them to the unique linguistic characteristics of public complaint data. The proposed approach enhances operational decision-making by visualizing data distribution, dominant words, and classification results, allowing public service managers to identify urgent issues in real time. The research also demonstrates that TF–IDF and Naive Bayes can successfully classify complaints into distinct categories while SMOTE effectively improves minority class representation, leading to more balanced predictions.

The findings highlight a trade-off between aggregate accuracy and class coverage, with SMOTE slightly affecting overall performance but significantly enhancing minority class detection. This paper offers valuable insights for addressing class imbalance in public service complaint systems and underscores the importance of fairness in service delivery. Future work should focus on expanding the dataset for better generalization, exploring alternative balancing techniques like Borderline-SMOTE or ADASYN, and investigating deep learning approaches such as IndoBERT to capture richer contextual features. These advancements are expected to improve predictive accuracy and interpretability, fostering the adoption of such systems in government complaint-handling workflows.

References

[1] R. Mutaqin, "Sistem Informasi Pengaduan Masyarakat Desa Sumberanyar Kecamatan Paiton Berbasis Android," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 4, pp. 1960–1972, Dec. 2021, doi: 10.35957/jatisi.v8i4.1199.

[2] A. Wibisono, S. Dadi Rizkiono, and A. Wantoro, "Filtering Spam Email Menggunakan Metode Naive Bayes," *TELEFORTECH J. Telemat. Inf. Technol.*, vol. 1, no. 1, pp. 9–17, Jul. 2020, doi: 10.33365/tft.v1i1.685.

[3] M. Hadwan, M. Al-Sarem, F. Saeed, and M. A. Al-Hagery, "An Improved Sentiment Classification Approach for Measuring User Satisfaction toward Governmental Services' Mobile Apps Using Machine Learning Methods with Feature Engineering and SMOTE Technique," *Appl. Sci.*, vol. 12, no. 11, pp. 1–25, May 2022, doi: 10.3390/app12115547.

[4] Orvalamarva, Oktariani Nurul Pratiwi, and Faqih Hamami, "Application of SMOTE Method on Topic Based Question Classification Using Naïve Bayes Algorithm," *Indones. J. Comput. Sci.*, vol. 13, no. 4, pp. 284–301, Jul. 2024, doi: 10.33022/ijcs.v13i4.4179.

[5] S. Fathmah, D. Kartini, F. Abadi, I. Budiman, and M. I. Mazdadi, "Implementation of PPCA Imputation, SMOTE-N Class Balancing in Hepatitis Classification Using Naïve Bayes," *JUITA J. Inform.*, vol. 12, no. 2, p. 169, 2024, doi: 10.30595/juita.v12i2.21528.

[6] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020, doi: 10.52465/joscex.v1i1.5.

[7] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Appl. Sci.*, vol. 13, no. 6, pp. 1–34,Mar. 2023, doi: 10.3390/app13064006.

[8] N. Fajriyah, N. T. Lapatta, D. W. Nugraha, and R. Laila, "Implementasi Svm Dan Smote Pada Analisis Sentimen Media Sosial X Terhadap Pelantikan Agus Harimurti Yudhoyono," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 2, pp. 1359–1370, 2025, doi: 10.29100/jipi.v10i2.6246.

[9] I. G. B. A. Budaya and I. K. P. Suniantara, "Comparison of Sentiment Analysis Algorithms with SMOTE Oversampling and TF-IDF Implementation on Google Reviews for Public Health Centers," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 1077–1086, Jul. 2024, doi: 10.57152/malcom.v4i3.1459.

[10] K. Wabang, O. D. Nurhayati, and Farikhin, "Application of The Naïve Bayes Classifier Algorithm to

Classify Community Complaints," *J. RESTI*, vol. 6, no. 5, pp. 872–876, 2022, doi: 10.29207/resti.v6i5.4498.

[11] S. S. Prasetiyowati and Y. Sibaroni, "Unlocking the potential of Naive Bayes for spatio temporal classification: a novel approach to feature expansion," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00958-x.

[12] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00943-4.

[13] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci. Rep.*, vol. 15, no. 1, pp. 1–20, 2025, doi: 10.1038/s41598-025-05791-7.

[14] B. Javid and H. Mashayekhi, "Classification of imbalanced user reviews using a generative approach," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 64, 2025, doi: 10.1007/s13278-025-01477-0.

[15] C. H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00782-9.

[16] Y. Zhang, L. Deng, and B. Wei, "Imbalanced Data Classification Based on Improved Random-SMOTE and Feature Standard Deviation," *Mathematics*, vol. 12, no. 11, pp. 1–17, 2024, doi: 10.3390/math12111709.

[17] T. Al-Shehari *et al.*, "Comparative evaluation of data imbalance addressing techniques for CNN-based insider threat detection," *Sci. Rep.*, vol. 14, no. 1, pp. 1–18, 2024, doi: 10.1038/s41598-024-73510-9.

[18] D.-H. Vu, T.-S. Vu, and T.-D. Luong, "An efficient and practical approach for privacy-preserving Naive Bayes classification," *J. Inf. Secur. Appl.*, vol. 68, p. 103215, 2022, doi: https://doi.org/10.1016/j.jisa.2022.103215.

[19] A. Z. Amrullah, A. Sofyan Anas, and M. A. J. Hidayat, "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," *J. BITe*, vol. 2, no. 1, pp. 40–44, 2020, doi: 10.30812/bite.v2i1.804.

[20] T. J. Firdaus, J. Indra, S. A. P. Lestari, and H. Hikmayanti, "Sentiment Analysis of the Sambara Application Using the Support Vector Machine Algorithm," *J. Tek. Inform.*, vol. 5, no. 4, pp. 1183–1192, 2024, doi: 10.52436/1.jutif.2024.5.4.2673.

[21] R. Kondo *et al.*, "Text Normalization for Japanese Sentiment Analysis," in *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, 2025, pp. 149–157. doi: 10.18653/v1/2025.wnut-1.16.

[22] T. Simbolon, A. P. Wibawa, I. A. E. Zaeni, and A. R. Ismail, "Text classification of traditional and national songs using naïve bayes algorithm," *Sci. Inf. Technol. Lett.*, vol. 3, no. 2, pp. 59–72, 2022, doi: 10.31763/sitech.v3i2.1215.

[23] M. M. Musthafa, I. Manimozhi, T. R. Mahesh, and S. Guluwadi, "Optimizing double-layered convolutional neural networks for efficient lung cancer classification through hyperparameter optimization and advanced image pre-processing techniques," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, pp. 1–21, 2024, doi: 10.1186/s12911-024-02553-9.

[24] Sunarti, Ridwang, and M. A. M. Hayat, "Klasifikasi Pengaduan Pelayanan Fakultas Teknik Universitas Muhammadiyah Makassar menggunakan Natural Language Processing," *Arus J. Sains dan Teknol.*, vol. 2, no. 2, pp. 572–579, 2024, doi: 10.57250/ajst.v2i2.667.

[25] I. M. Alkhawaldeh, I. Albalkhi, and A. J. Naswhan, "Challenges and limitations of synthetic minority oversampling techniques in machine learning," *World J. Methodol.*, vol. 13, no. 5, pp. 373–378, 2023, doi: 10.5662/wjm.v13.i5.373.

[26] N. S. Rahmi, N. W. S. Wardhani, M. B. Mitakda, R. S. Fauztina, and I. Salsabila, "SMOTE Classification and Random Oversampling Naive Bayes in Imbalanced Data : (Case Study of Early Detection of Cervical Cancer in Indonesia)," in *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, 2022, pp. 1–6. doi: 10.1109/ICITDA55840.2022.9971421.

[27] A. Wibowo, A. F. N. Masruriyah, and S. Rahmawati, "Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes for Imbalanced Datasets," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 197–207, 2025, doi: 10.35882/jeeemi.v7i1.596.