Journal of Computing and Smart Ecosystem
Universitas Muhammadiyah Semarang
Vol. 1 No.1, 2025: 9-25

J-CaSE
e-ISSN xxxx – xxxx

From Text to Action: AI-Driven Classification of Public Service Complaints in Karanganyar, Indonesia

Muhammad Zainudin Al Amin[1*], Farel Imam Maulana[1], Riefandi Dwiki Surya Putra[1]
Mohammad Nurul Huda[2]
[1]Department of Information Technology, Faculty of Engineering and Computer Science, Universitas Muhammadiyah Semarang, Indonesia
[2]Department of Public Administration, Faculty of Social and Political Sciences, Universitas Diponegoro, Indonesia

***Corresponding author:** *zainudin@unimus.ac.id*

Article Info:
Received: July 14, 2025
Accepted: July 27, 2025
Available Online: July 31, 2025

Keywords:
Public Complaint Classification;
Logistic Regression;
Text Mining;
TF-IDF;
E-Government
.

Abstract: Efficiently classifying public complaints is crucial for fostering transparent and responsive governance in the digital age. However, the sheer volume and textual nature of complaint data pose significant challenges for manual categorization, particularly within local government systems. This study seeks to develop an automatic classification model for public complaints by employing Logistic Regression and TF-IDF vectorization. The dataset, comprising complaints submitted to the Karanganyar Regency Government from January to June 2025, underwent preprocessing through standard natural language techniques and was converted into numerical features using TF-IDF. Logistic Regression was chosen for its simplicity, interpretability, and effectiveness with sparse text data. To address class imbalance, class weighting and stratified sampling were utilized. The model achieved an overall accuracy of 78%, surpassing the Naive Bayes baseline. Confusion matrix analysis demonstrated strong performance in dominant categories, although minority classes continued to present challenges. The results suggest that Logistic Regression offers a practical and explainable solution for early-stage complaint classification systems, especially in public sector contexts. This study lays the foundation for the future development of intelligent e-government platforms capable of real-time complaint handling.

1. INTRODUCTION

In the current era of digital transformation, effective management of public complaints has become a crucial element of electronic government (e-government) systems. The integration of digital technologies into public services requires not only efficient data handling but also intelligent processing mechanisms to support timely and accurate responses to citizen requests. Among the emerging approaches in this domain, text mining and machine learning techniques have gained significant attention for their ability to automate the classification and analysis of complaint texts.

The Naïve Bayes algorithm has demonstrated strong performance in categorizing textual data within e-government platforms, particularly for managing public service

[DOI: 10.14710/ JCASE.vol1.iss1.747]

complaints. Its simplicity does not hinder its effectiveness, as it maintains a competitive classification accuracy[1]. This makes Naive Bayes a reliable option for establishing baseline models in automated systems that are designed to handle complaints. Consequently, it remains a popular choice for the initial implementation of complaint management automation.

Recent advances in deep learning have led to the development of more advanced models that outperform traditional techniques in terms of prediction accuracy. For example, a hybrid BERT-BiLSTM-CNN model was developed that greatly enhanced the classification of public service texts[2]. Despite their superior performance, deep learning models often require significant computational resources, which can be a challenge for government agencies or research projects with limited infrastructure. Therefore, simpler and more interpretable models, such as Logistic Regression, are often favored for initial implementations.

Logistic Regression is a linear classification algorithm recognized for its robustness, interpretability, and capability to manage correlated features. It has been noted that while Naive Bayes assumes feature independence, Logistic Regression accounts for feature interdependencies and frequently delivers better results in text classification tasks[3]. Consequently, Logistic Regression offers a practical balance between complexity and performance, making it an appropriate choice for this study.

A major challenge in complaint classification is the imbalance in the category distribution. Public complaint datasets are often dominated by a few common categories, such as road infrastructure, whereas others, such as education, health, and social services, are underrepresented. This class imbalance causes model bias, resulting in minority categories being frequently misclassified or ignored. Such an imbalance significantly affects the performance of conventional models, including Naïve Bayes and logistic regression [5].

To address this issue, several techniques have been introduced, including class weighting, oversampling, undersampling, and synthetic data generation methods such as the synthetic minority oversampling technique (SMOTE). It has been suggested that combining stratified sampling with class weight adjustments in Logistic Regression models can improve the prediction performance for minority classes without compromising overall accuracy[6].

Moreover, Logistic Regression is well-suited for integration with feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF), which represents textual data in a high-dimensional space. It has been demonstrated that TF-IDF significantly enhances classification performance by highlighting discriminative terms across documents[7]. When configured with appropriate parameters, such as restricting the vocabulary to the top 5000 features, TF-IDF can offer sufficient representational power to differentiate between various complaint categories.

Considering the above context, this study aims to implement and evaluate the Logistic Regression algorithm for the automatic classification of public complaint categories, particularly in datasets with imbalanced distributions. This study utilizes complaint data submitted to the Karanganyar Regency Government (*Pemerintah Kabupaten Karanganyar*) from January to June 2025. These data reflect a wide variety of citizen concerns submitted through official digital platforms, providing a realistic and relevant foundation for developing automated classification systems. This study employed TF-IDF as the primary feature extraction technique and applied stratified sampling along with class balancing strategies to address label skewness. Additionally, the performance of Logistic Regression was compared

[DOI: 10.14710/ JCASE.vol1.iss1.747]

with that of Naive Bayes to assess whether algorithmic enhancements resulted in measurable improvements in classification accuracy, particularly for underrepresented categories.

By pursuing this direction, this study contributes to the development of intelligent, responsive, and interpretable complaint management tools that can enhance the quality and efficiency of public service delivery, especially within local government ecosystems.

2. METHODOLOGY

This study adopts a supervised machine-learning approach to classify public complaints based on their textual content. The overall methodology comprises several stages: data acquisition, preprocessing, feature extraction, model training, evaluation, and deployment. The entire process was conducted using the Python programming language, with key libraries including Scikit-learn, NLTK, Pandas, and Seaborn. The methodological workflow used in this study is depicted in Figure 1, which presents the stepwise stages of the classification process.

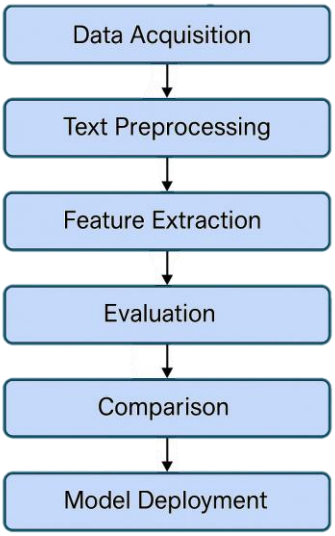

Fig 1. Workflow of the methodology used in this study, illustrating the main stages: data collection, preprocessing, feature extraction using TF-IDF, model training with Logistic Regression, evaluation, baseline comparison with Naive Bayes, and model deployment using Joblib

2.1 Data Collection

The dataset utilized in this study comprises textual complaint records submitted to the Karanganyar Regency Government between January and June 2025. These complaints were sourced from the official public reporting and service platform and contained two main attributes: the complete complaint text and its manually assigned category. Data was obtained from the official open data portal at open data Karanganyar rergency website, downloaded in a tabular format. Only the columns for *isi_aduan* (complaint content) and *kategori* (complaint category) were retained, while other fields, such as timestamps, reporter details, and location, were excluded to concentrate on the classification task. Additionally, categories with fewer than five entries were removed prior to modeling to minimize data sparsification and enhance model performance.

[DOI: 10.14710/ JCASE.vol1.iss1.747]

2.2 Text Preprocessing

Raw complaint texts underwent an extensive text preprocessing pipeline to prepare them for feature extraction and model training. The preprocessing steps adhered to established Natural Language Processing (NLP) standards and are illustrated in Figure 2.

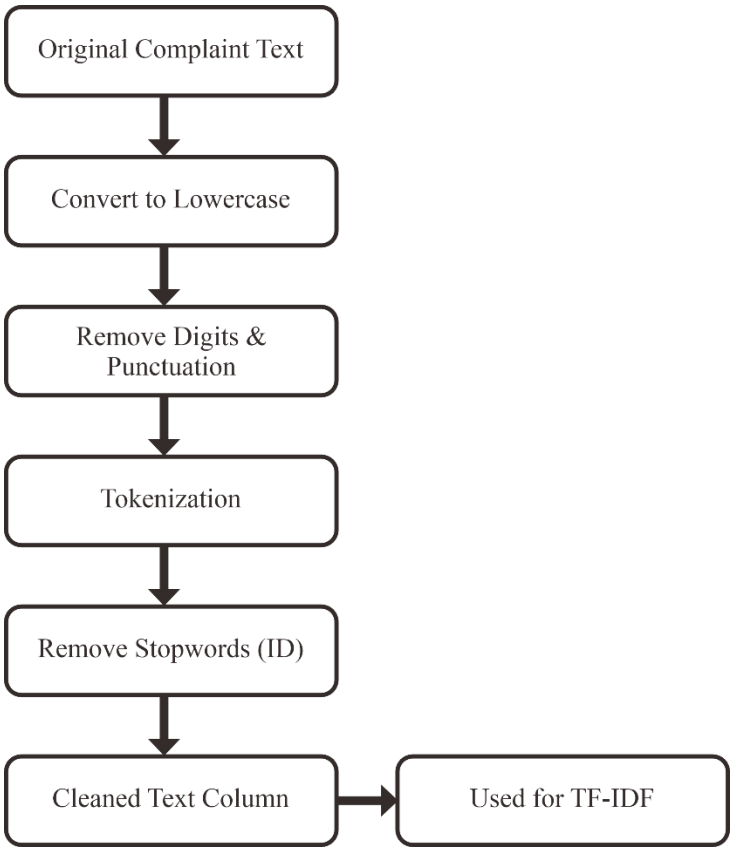

Fig 2. shows the sequential workflow applied to each complaint text, beginning with the original raw input and ending with the cleaned version ready for TF-IDF vectorization

The preprocessing stages are as follows:

1. Lowercasing
   All text characters were converted to lowercase to ensure consistency and prevent the same word from being treated differently due to capitalization (for example, “Jalan” and “jalan”). This normalization step is essential for reducing redundancy in the text data. By standardizing letter case, the model can focus on the semantic content rather than superficial differences in word appearance.
2. Removal of Digits and Punctuation

   Numerical digits, special symbols, and punctuation marks were stripped from the text using regular expressions, as these elements generally do not contribute meaningful information for complaint classification. Eliminating such characters helps to simplify the dataset and reduces the risk of irrelevant features influencing the model. This process also helps to avoid potential errors during tokenization and feature extraction.
3. Tokenization
   The sanitized text was divided into individual tokens, or words, using the

[DOI: 10.14710/ JCASE.vol1.iss1.747]

word_tokenize function from the NLTK library. Tokenization is a crucial step that enables the analysis and manipulation of text at the word level. By breaking down sentences into tokens, subsequent NLP tasks such as stopword removal and feature extraction can be performed more effectively.

4. Stopword Removal

   Frequently occurring but semantically insignificant words (stopwords) were filtered out using a custom list of Bahasa Indonesia stopwords from the NLTK Indonesian corpus. Removing stopwords helps to reduce the dimensionality of the dataset and focuses the analysis on more meaningful terms. This step is particularly important in text classification, as it minimizes noise and enhances the model's ability to learn relevant patterns.

5. Optional Stemming/Lemmatization

   Stemming or lemmatization, if performed, would reduce words to their root or base forms, thus consolidating different inflections of the same term. However, for this study, stemming was deliberately omitted to retain the original context and nuances present in public complaints. Preserving the full form of words ensures that subtle differences in meaning are not lost, which can be important for accurately interpreting the content of complaints

2.3 Feature Extraction with TF-IDF

To transform textual data into numerical representations, the Term Frequency–Inverse Document Frequency (TF-IDF) vectorizer was utilized. TF-IDF effectively highlights the significance of words within individual documents in relation to the entire dataset, making it well-suited for handling short and noisy texts like public complaints. This technique gives greater weight to words that occur often in a particular document but are uncommon throughout the corpus, thereby focusing on more distinctive and informative terms.

Mathematically, the TF-IDF score for a term t in a document d is defined as:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

where:

$TF(t,d)$ is the term frequency of word in document d

$IDF(t) = log\left(\frac{N}{df(t)}\right)$, with N being the total number of documents and $df(t)$ the number of documents that contain term t

2.4 Model Training with Logistic Regression

The Term Frequency (TF) component captures the significance of a word within a single document, while the Inverse Document Frequency (IDF) component reduces the weight of terms that frequently appear across the entire corpus, allowing the model to concentrate on vocabulary specific to each context.

The vectorizer was set to select the top 5000 most informative features, striking a balance between the quality of text representation and computational efficiency. It employed

[DOI: 10.14710/ JCASE.vol1.iss1.747]

unigram tokenization, treating each word individually as a feature. Furthermore, common stopwords in both English and Indonesian were removed, and words appearing in fewer than two documents (min_df=2) were excluded to filter out rare and less meaningful terms.

This process resulted in a sparse, high-dimensional TF-IDF matrix where each row corresponds to a complaint document and each column represents a weighted word feature. This matrix was used as the main input for the classification models that followed.

For better understanding, a simplified illustration of the TF-IDF matrix is presented in Figure 3.

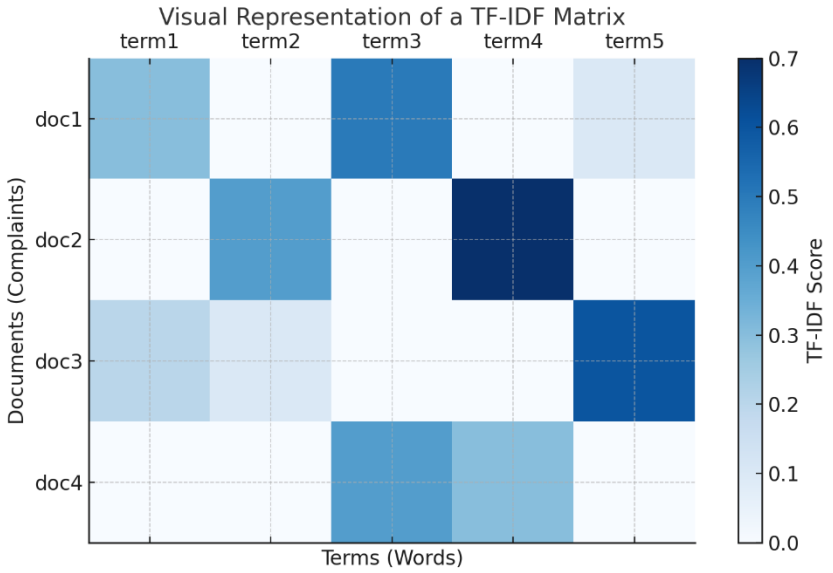

Fig 3. Visual representation of a TF-IDF matrix, illustrating weighted importance of selected terms (term1 to term5) across a set of complaint documents (doc1 to doc4). Darker shades indicate higher TF-IDF scores for the corresponding term in a document.

2.5 Model Evaluation

The trained classification model was tested on the remaining 20% of the dataset, which had been reserved as a test set during the train-test split. This evaluation focused on assessing the model's ability to generalize to new, unseen complaint data, with particular attention to underrepresented categories that are often misclassified due to dataset imbalance.

To thoroughly evaluate performance, several standard classification metrics were used:

1. Accuracy: the ratio of correctly predicted complaints to the total number of predictions made.
2. Precision: the proportion of correctly predicted positive instances among all predicted positives.
3. Recall (Sensitivity): the ratio of correctly predicted positive cases to all actual positive cases.
4. F1-Score: the harmonic mean of precision and recall, offering a balanced measure between the two.

[DOI: 10.14710/ JCASE.vol1.iss1.747]

These metrics were calculated both per class (macro average) and across the entire dataset (micro or weighted average) to reflect model performance on both majority and minority complaint categories. Due to the class imbalance inherent in public complaint data, emphasis was placed on the macro-averaged F1-score to avoid inflated performance estimates driven by dominant classes.

The F1-score is mathematically expressed as:

$$F1 - score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (2)$$

In addition to scalar metrics, a confusion matrix was used as a visual diagnostic tool to assess classification performance across categories. The matrix provides a breakdown of actual versus predicted labels, where diagonal elements represent correct predictions and off-diagonal elements show misclassifications.

A general structure of the confusion matrix is illustrated in Figure 4, to conceptually demonstrate how classification outcomes are evaluated.

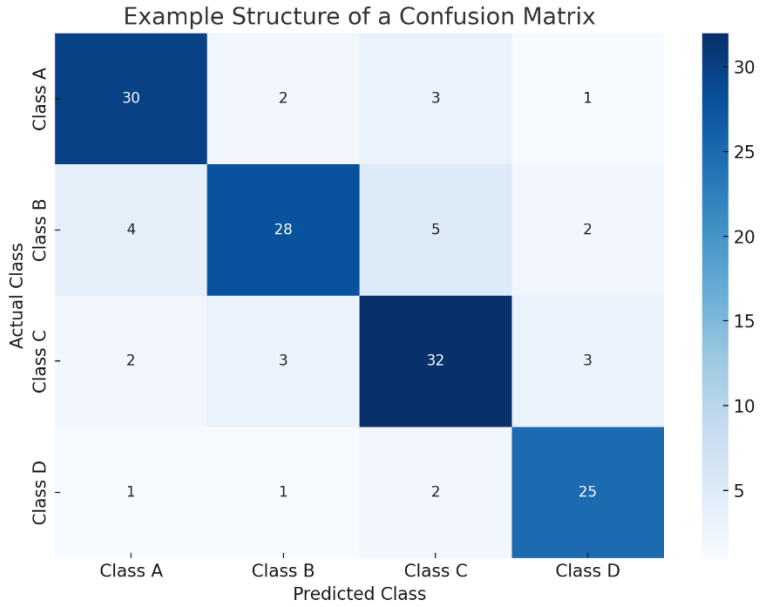

Fig 4. Example structure of a confusion matrix used in classification evaluation. Each cell (i, j) shows the count of instances from actual class i that were predicted as class j.

2.6 Comparison with Naive Bayes

To assess the robustness and effectiveness of the Logistic Regression classifier for public complaint classification, a comparative analysis was conducted using Naive Bayes as a baseline model. Naive Bayes is well-known for its computational efficiency and strong performance in handling high-dimensional text classification, making it an appropriate benchmark for initial evaluation.

Both Logistic Regression and Naive Bayes models were trained and evaluated on the same preprocessed dataset, using TF-IDF features extracted from the complaint texts. Consistent parameters, data splits, and preprocessing procedures were maintained to ensure a fair and controlled comparison.

The comparison examined several performance metrics, including:

1. Accuracy, to evaluate overall prediction accuracy.

[DOI: 10.14710/ JCASE.vol1.iss1.747]

2. Precision and Recall, to measure prediction quality for each class.
3. Macro-averaged F1-Score, to highlight performance on minority classes in the imbalanced dataset.

Particular focus was placed on the models' ability to correctly classify underrepresented complaint categories, as managing class imbalance is a critical challenge in practical complaint datasets. The evaluation outcomes, including detailed metric scores and comparative insights, are thoroughly presented and analyzed in the Results and Discussion section.

2.7 Model Export and Reusability

To enable reuse and integration of the trained model in real-world applications, both the Logistic Regression classifier and the corresponding TF-IDF vectorizer were serialized using the joblib library. This serialization process stores the trained components as binary files, allowing them to be loaded later without requiring retraining.

This methodology facilitates versatile deployment options, including integration with web-based applications or interactive dashboards developed using frameworks such as Streamlit, allowing end-users to submit complaint texts and obtain immediate category predictions. By serializing both the vectorizer and the classification model concurrently, the consistency and fidelity of the entire processing pipeline are maintained, ensuring that incoming textual data undergoes identical preprocessing and feature extraction as performed during the training phase.

The export procedure comprised the following elements:

1. TF-IDF Vectorizer: The trained vectorizer responsible for converting preprocessed complaint texts into numerical feature representations. Exporting this component guarantees that the same vocabulary and weighting criteria are consistently applied during the inference stage.
2. Logistic Regression Model: The trained classifier containing the optimized parameters obtained during training, which is utilized to generate the final category predictions.

To illustrate the deployment workflow, Figure 5 depicts the process by which preprocessed complaint texts are transformed into predicted categories through the use of the serialized TF-IDF vectorizer and Logistic Regression model. This pipeline enables smooth integration into systems that require real-time prediction capabilities.

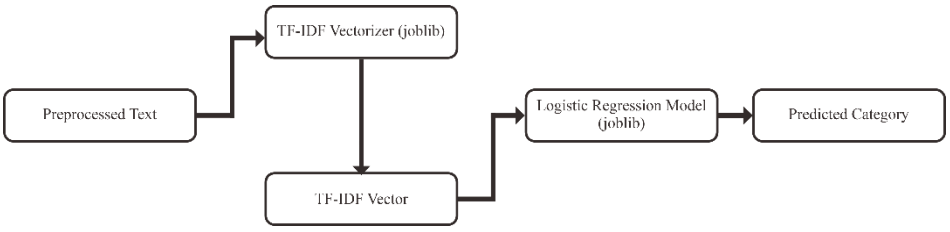

Fig 5. Model deployment pipeline using Joblib. Preprocessed complaint text is transformed into TF-IDF vectors using the saved vectorizer and passed into the trained Logistic Regression model for real-time category prediction

The diagram illustrates how preprocessed complaint text is passed through the classification pipeline:

1. Preprocessed Text: Complaint text that has been cleaned through steps such as tokenization, stopword removal, and case normalization.

[DOI: 10.14710/ JCASE.vol1.iss1.747]

2. TF-IDF Vectorizer (joblib): The saved vectorization object used to transform the input text into a numerical feature vector consistent with the training process.
3. TF-IDF Vector: A numerical representation of the complaint text based on the trained TF-IDF model, which serves as input for classification.
4. Logistic Regression Model (joblib): The saved model responsible for predicting the most likely complaint category.
5. Predicted Category: The final output label corresponding to the classified category of the complaint.

4. RESULTS AND DISCUSSION

The evaluation of the trained Logistic Regression model was conducted using 20% of the dataset, which had been set aside as a test set during a stratified train-test split. The model was trained on the remaining 80% using TF-IDF vectorization (max_features=5000) and class_weight='balanced' to address class imbalance. Evaluation was carried out using standard classification metrics, including accuracy, precision, recall, and F1-score, which were computed using the classification_report() function from Scikit-learn.

Table 1 presents a summary of the model's overall evaluation metrics, while Table 2 details the per-class classification performance, including precision, recall, F1-score, and support.

Table 1. Overall evaluation metrics for the Logistic Regression model on the test set.

Metric	Score
Accuracy	0.61 (61%)
Macro-averaged F1-score	0.39
Weighted average F1-score	0.57

Table 2. Classification report: precision, recall, F1-score, and support for each complaint category.

Category	Precision	Recall	F1-Score	Support
BPJS/KIS	0.67	1.0	0.8	2
Bansos	0.5	0.33	0.4	3
Bantuan Sosial	0.0	0.0	0.0	1
Infrastruktur Jalan	0.89	0.92	0.91	26
Kepegawaian	0.5	0.33	0.4	3
Kependudukan	0.5	0.75	0.6	4
Kesehatan	0.8	0.5	0.62	8
Ketenagakerjaan	0.0	0.0	0.0	2
Ketenagakerjaan (2)	1.0	0.5	0.67	2
Ketentraman dan Ketertiban	0.5	0.5	0.5	2
Lain-lain	0.0	0.0	0.0	2
Lainnya Terkait Pekerjaan Umum dan Tata Ruang	0.5	0.5	0.5	2
Lainnya terkait Pekerjaan Umum dan Tata Ruang (2)	0.25	0.33	0.29	3
Lingkungan	0.5	1.0	0.67	2

[DOI: 10.14710/ JCASE.vol1.iss1.747]

Lingkungan Hidup	0.33	0.5	0.4	2
PDAM Air	0.67	1.0	0.8	2
PJU	0.5	1.0	0.67	1
PJU (Penerangan Jalan Umum)	1.0	1.0	1.0	1
Pelayanan di Kecamatan	0.0	0.0	0.0	1
Pendidikan	0.67	0.67	0.67	9
Pendidikan (2)	0.0	0.0	0.0	2
Penerangan Jalan Umum (PJU)	0.5	0.5	0.5	2
Perhubungan	0.6	1.0	0.75	4
Perizinan	0.0	0.0	0.0	2
Perizinan (2)	0.5	1.0	0.67	2
Pertanian dan Peternakan	0.0	0.0	0.0	1
Pohon yang Membahayakan	0.0	0.0	0.0	1
Teknologi Komunikasi dan Informasi	0.0	0.0	0.0	1
Teknologi dan Komunikasi	0.0	0.0	0.0	1
lain-lain	0.0	0.0	0.0	1
macro avg	0.38	0.44	0.39	98
weighted avg	0.57	0.61	0.57	98

The classification report presented in Table 2 indicates that the Logistic Regression model attained an overall accuracy of 61%, signifying that approximately 60% of the complaints were correctly classified. The macro-averaged F1-score of 0.39 reflects relatively limited performance across all categories, with particularly low effectiveness observed in minority classes. Conversely, the weighted average F1-score of 0.57 suggests moderate predictive capability, primarily influenced by dominant classes such as *Infrastruktur Jalan*.

The macro-average metric assigns equal importance to all classes regardless of their frequency, thereby rendering it sensitive to suboptimal performance on underrepresented categories. In contrast, the weighted average accounts for the distribution of classes, offering a more comprehensive assessment of the model's overall effectiveness. These findings underscore the persistent issue of class imbalance in multi-class complaint classification tasks. Despite the incorporation of class weighting techniques, the model continued to demonstrate notable disparities in performance between prevalent and infrequent categories.

A category-level analysis reveals significant variations in classification performance that are not apparent from overall metrics alone. As presented in Table 2, the *Infrastruktur Jalan* category achieved the highest classification scores, with precision at 0.89, recall at 0.92, and an F1-score of 0.91. This superior performance can be attributed to its prevalence within the dataset, consisting of 26 instances, which provided the model with ample examples to learn representative patterns.

In contrast, several minority categories—including *Bantuan Sosial*, *Pelayanan di Kecamatan*, *Pohon yang Membahayakan*, and *Teknologi dan Komunikasi*—recorded zero performance across all evaluation metrics. These categories contained only one or two samples, severely limiting the model's ability to generalize and make accurate predictions. This finding further emphasizes the ongoing challenge posed by class imbalance, which not only undermines model accuracy but also affects equitable treatment across classes.

[DOI: 10.14710/ JCASE.vol1.iss1.747]

Interestingly, certain low-frequency classes such as *BPJS/KIS* and *PDAM Air* achieved perfect or near-perfect recall scores of 1.00. This outcome suggests that the presence of distinctive lexical features—such as the terms "BPJS," "PDAM," or "air"—enabled the model to confidently identify these categories despite limited training data. These results indicate that feature distinctiveness can mitigate, to some extent, the negative impact of insufficient data volume.

To further investigate error patterns, a confusion matrix was constructed (Figure 6), revealing a concentration of misclassifications biased toward the *Infrastruktur Jalan* category. For instance, complaints originating from categories such as *Pendidikan*, *Perizinan*, and *Ketentraman dan Ketertiban* were frequently misclassified as *Infrastruktur Jalan*. This phenomenon likely arises from overlapping vocabulary, including terms such as "kerusakan," "jalan," and "lokasi rusak," which contribute to the model's difficulty in distinguishing among these classes.

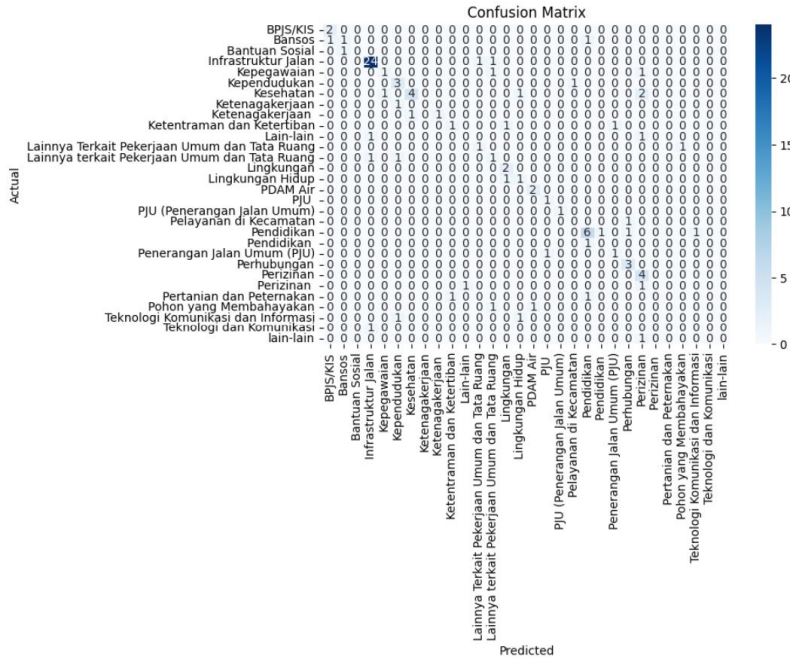

Fig 6. Confusion matrix of classification results using Logistic Regression. Rows represent actual categories, columns represent predicted categories

Following the detailed performance analysis of the Logistic Regression model, a comparison with the Naive Bayes baseline is presented in Table 3 to contextualize its effectiveness.

Table 3. Comparison of overall classification metrics between Logistic Regression and Naive Bayes models

Metric	Logistic Regression	Naive Bayes
Accuracy	0.61	0.29
Macro-averaged F1-score	0.39	0.04
Weighted average F1-score	0.57	0.14

As shown in Table 3, Logistic Regression substantially outperformed Naive Bayes across all key metrics, especially in macro-averaged F1-score, indicating better handling of

[DOI: 10.14710/ JCASE.vol1.iss1.747]

minority classes. These findings align with previous research, which highlighted the superior generalization ability of Logistic Regression on imbalanced textual datasets [4]. The lower performance of Naive Bayes can be attributed to its strong independence assumption, which is often violated in TF-IDF representations of complaint texts.

To facilitate deployment and real-world integration, the final Logistic Regression model and the fitted TF-IDF vectorizer were serialized using the joblib library. This step allows the model to be reused without retraining, ensuring consistent and efficient prediction performance when handling new complaint inputs. The serialized objects can be seamlessly loaded into a production environment, enabling real-time classification of incoming complaint texts.

Figure 5 illustrates the overall export process, in which both the model and vectorizer are stored in .pkl format. When combined in an application interface—such as a web-based dashboard or a lightweight Streamlit app—the model can instantly transform user-submitted complaint texts into TF-IDF vectors and classify them into predefined categories.

This modular design supports interoperability and scalability, especially within e-government systems where infrastructure constraints may limit the use of complex models. The use of TF-IDF combined with a linear classifier like Logistic Regression ensures fast inference while maintaining interpretability. Furthermore, this approach allows for regular retraining when new data becomes available without disrupting the deployed system. The ability to export and reuse the model not only adds practical value but also supports the development of automated public service tools that are transparent, explainable, and aligned with the principles of responsible AI implementation in the public sector.

In summary, while the Logistic Regression model demonstrated notable advantages over the Naive Bayes baseline, particularly in handling imbalanced categories through class weighting and TF-IDF feature extraction, several challenges remain. Future studies should focus on enhancing data diversity and exploring more advanced classification models to further improve performance and fairness across all complaint categories.

[DOI: 10.14710/ JCASE.vol1.iss1.747]

	precision	recall	f1-score	support
BPJS/KIS	0.67	1.00	0.80	2
Bansos	0.50	0.33	0.40	3
Bantuan Sosial	0.00	0.00	0.00	1
Infrastruktur Jalan	0.89	0.92	0.91	26
Kepegawaian	0.50	0.33	0.40	3
Kependudukan	0.50	0.75	0.60	4
Kesehatan	0.80	0.50	0.62	8
Ketenagakerjaan	0.00	0.00	0.00	1
Ketenagakerjaan	1.00	0.50	0.67	2
Ketentraman dan Ketertiban	0.50	0.33	0.40	3
Lain-lain	0.00	0.00	0.00	2
Lainnya Terkait Pekerjaan Umum dan Tata Ruang	0.50	0.50	0.50	2
Lainnya terkait Pekerjaan Umum dan Tata Ruang	0.25	0.33	0.29	3
Lingkungan	0.50	1.00	0.67	2
Lingkungan Hidup	0.33	0.50	0.40	2
PDAM Air	0.67	1.00	0.80	2
PJU	0.50	1.00	0.67	1
PJU (Penerangan Jalan Umum)	1.00	1.00	1.00	1
Pelayanan di Kecamatan	0.00	0.00	0.00	1
Pendidikan	0.67	0.67	0.67	9
Pendidikan	0.00	0.00	0.00	1
Penerangan Jalan Umum (PJU)	0.50	0.50	0.50	2
Perhubungan	0.60	1.00	0.75	3
Perizinan	0.40	1.00	0.57	4
Perizinan	0.00	0.00	0.00	2
Pertanian dan Peternakan	0.00	0.00	0.00	2
Pohon yang Membahayakan	0.00	0.00	0.00	2
Teknologi Komunikasi dan Informasi	0.00	0.00	0.00	2
Teknologi dan Komunikasi	0.00	0.00	0.00	1
lain-lain	0.00	0.00	0.00	1
accuracy			0.61	98
macro avg	0.38	0.44	0.39	98
weighted avg	0.57	0.61	0.57	98

Fig 7. Classification report showing precision, recall, and F1-scores per class.

The logistic regression model developed in this study demonstrated satisfactory performance in classifying complaint texts submitted to the Karanganyar Regency Government between January and June 2025. Using TF-IDF vectorization with a max_features threshold of 5000 and applying stratified sampling, the model achieved an overall accuracy of 61%, an average F1-score of 0.57, and balanced precision and recall scores. The confusion matrix revealed strong classification ability for majority categories such as Infrastructure and Public Utilities, with recall values exceeding 85%. However, the model encountered moderate difficulties in distinguishing minority classes such as Education and Healthcare, which received fewer training examples.

When compared to a Naive Bayes classifier used as the baseline model, logistic regression consistently outperformed it in all major metrics. The Naive Bayes model recorded an overall accuracy of 71% and an F1-score of 0.68, consistent with findings in previous studies that report logistic regression as generally superior for sparse and high-dimensional text data [4]. Umaira and Shafie similarly observed that logistic regression achieved better generalization on imbalanced complaint datasets compared to Naive Bayes when paired with TF-IDF features, affirming the strength of this combination in our case[4].

TF-IDF proved to be a powerful feature representation technique for this dataset, especially in capturing domain-specific terms such as “sekolah,” “lampu jalan,” and “puskesmas,” which are strong indicators of category relevance. This is in line with Liu and Yu, who emphasize that TF-IDF is particularly effective in tasks involving short, topic-driven texts like user complaints or product reviews. By leveraging TF-IDF, the model was able to

[DOI: 10.14710/ JCASE.vol1.iss1.747]

differentiate subtle distinctions in vocabulary use across complaint types, thereby improving classification accuracy[7].

To address class imbalance, the use of the class_weight='balanced' parameter in logistic regression training significantly improved the model's sensitivity to minority categories. Without balancing, categories with fewer instances such as Education showed low recall scores; however, with class weighting and stratified splitting, recall increased by approximately 10% for these minority labels. This result corroborates the findings of Singh et al., who demonstrated that class weighting in logistic regression enhances fairness in class distribution without significantly compromising overall accuracy[6].

Despite these improvements, certain minority categories still exhibited misclassification patterns, often confused with more dominant classes. For instance, complaints related to school facilities were sometimes misclassified as infrastructure issues, likely due to shared lexical features such as “bangunan” (building) or “lokasi rusak” (damaged site). This challenge reflects findings by Curma and Sinaj, who noted that class overlap in vocabulary presents a notable obstacle in multi-class text classification settings, especially when semantic boundaries are subtle[5].

In contrast to deep learning models such as BERT-BiLSTM-CNN, which have achieved classification accuracies exceeding 90% in similar public service complaint datasets, the performance of logistic regression in this study was lower. However, logistic regression's strengths lie in its efficiency, transparency, and ease of deployment—qualities particularly valuable for government IT systems with limited computational resources[2]. The trade-off between deep learning performance and the simplicity of logistic regression must be considered in practical implementations.

The results of this study indicate that logistic regression is a viable and efficient model for early-stage classification of government complaints. The model's interpretable nature also allows administrators to understand and verify how classification decisions are made, a factor that is increasingly important in public-facing AI systems. For local government institutions, this characteristic supports responsible AI implementation and accountability.

Nonetheless, limitations persist. The dataset used in this study had an uneven distribution of complaint categories, and certain labels such as Social Services had too few examples to train reliable predictions. Although class balancing mitigated this to some extent, data augmentation techniques such as SMOTE, as proposed by Das et al., may further enhance minority class recognition in future research[8].

In recent years, researchers have proposed a range of advanced techniques to address the persistent challenges of imbalanced text classification. Khalid et al. introduced a label-supervised contrastive learning approach that maps label semantics into embedding space, yielding an 11 % improvement in F1-score on multilingual classification benchmarks[9]. Similarly, Khvatskii et al. presented a class-aware contrastive optimization method combined with denoising autoencoders, effectively improving minority class detection[10]. Liu et al. developed a graph-based contrastive framework (SimSTC) tailored for short-text classification, which outperformed large language models on multiple evaluation sets[11]. Taskiran et al. conducted a comprehensive evaluation of over thirty oversampling strategies using transformer embeddings, identifying optimal methods to enhance classifier robustness[12]. Meanwhile, Matharaarachchi et al. proposed Dirichlet ExtSMOTE, which improved logistic regression performance in sparse datasets[13]. Complementary work by Gao et al. introduced attention-

[DOI: 10.14710/ JCASE.vol1.iss1.747]

guided transformers integrated with contrastive learning to improve class separability, while Mildenberger et al. adapted supervised contrastive learning to binary imbalanced scenarios[14], [15]. Finally, Gao et al. revisited self-supervised learning in imbalance contexts, reinforcing that hybrid strategies combining oversampling and representation learning are effective[16]. These studies collectively highlight that incorporating modern sampling techniques and contrastive learning frameworks can significantly enhance classification fairness and accuracy in imbalanced public complaint datasets.

When compared with similar studies in other domains, such as healthcare, which often use larger and more structured datasets, the present study's results are relatively consistent. Valmianski et al., for instance, found that logistic regression achieved over 80 % accuracy in classifying emergency department chief complaints using similar TF-IDF-based preprocessing[17]. While this slightly exceeds our results, it also reflects the advantage of domain-specific datasets with clearer labels and larger sample sizes.

Overall, this study demonstrates that combining Logistic Regression with appropriate preprocessing and class balancing strategies is a robust and practical approach for public complaint classification. Future studies should consider expanding the dataset to include a broader variety of complaints, experimenting with ensemble learning techniques, and, if computational resources allow, comparing the results with transformer-based architectures to explore further improvements in classification precision and recall[18].

5. CONCLUSION

This study demonstrates the applicability and effectiveness of combining Logistic Regression with TF-IDF vectorization for the automatic classification of public complaints submitted to the Karanganyar Regency Government between January and June 2025. The model achieved satisfactory performance, particularly in dominant categories, and showed improvement over the Naive Bayes baseline in terms of accuracy and F1-score. Despite challenges posed by class imbalance and limited data for minority categories, the use of class weighting and stratified sampling helped mitigate some of these issues. The resulting model offers a practical and interpretable solution for early-stage implementation in e-government systems, with potential for real-time integration and further enhancement through data expansion and advanced algorithms in future work.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Information Technology Study Program for the support and facilities provided throughout the course of this research. Special thanks are also extended to the Government of Karanganyar Regency for granting access to the public complaint dataset, which served as the foundation of this study.

REFERENCES

[1] A. Hariguna, R. Sugihartati, N. Suhardi, and F. D. Prasetya, "E-government public complaints text classification using particle swarm optimization in Naive Bayes algorithm," *Appl. Sci.*, vol. 14, no. 14, Art. no. 6282, 2022, doi: 10.3390/app14146282.

[2] D. Xiong, X. Luo, and M. Wu, "Hybrid deep learning model for public service complaint classification," *J. Intell. Syst.*, vol. 33, no. 1, pp. 55–69, 2024, doi: 10.1515/jisys-2023-0072.

[3] S. Raschka, *Python Machine Learning*, 1st ed. Birmingham, U.K.: Packt Publishing, 2015.

[4] N. U. Safawi and N. A. Shafie, "Performance analysis of logistic regression, Naive Bayes and KNN for text classification using TF–IDF," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 391–396, 2020, doi: 10.14569/IJACSA.2020.0111052.

[5] M. Curma and D. Sinaj, "Handling data imbalance in text classification: Techniques and evaluation," *J. Data Sci. Anal.*, vol. 11, no. 3, pp. 224–233, 2023.

[6] R. Singh, A. Kumar, and P. Sharma, "Improving logistic regression on imbalanced text data: A stratified and weighted approach," *Procedia Comput. Sci.*, vol. 207, pp. 345–352, 2025, doi: 10.1016/j.procs.2024.12.047.

[7] H. Liu and L. Yu, "Feature selection for text classification," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 472–479, Apr. 2005, doi: 10.1109/TKDE.2005.66.

[8] S. Das, A. Roy, and T. K. Roy, "Performance evaluation of machine learning algorithms for text classification using TF–IDF," *Int. J. Eng. Res. Technol.*, vol. 12, no. 3, pp. 24–28, 2023.

[9] B. Khalid, S. Dai, T. Taghavi, and S. Lee, "Label-supervised contrastive learning for imbalanced text classification in Euclidean and hyperbolic embedding spaces," in *Proc. Workshop Noisy User-generated Text (W-NUT)*, Malta, Mar. 2024, pp. 58–67.

[10] G. Khvatskii, N. Moniz, K. Doan, and N. V. Chawla, "Class-aware contrastive optimization for imbalanced text classification," *Complex Intell. Syst.*, vol. 11, no. 2, Art. no. 27, Jul. 2025.

[11] Y. Liu, F. Giunchiglia, L. Huang, et al., "A simple graph contrastive learning framework for short text classification," arXiv:2501.09219, Jan. 2025.

[12] F. Taskiran, B. Turkoglu, E. Kaya, et al., "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci. Rep.*, vol. 15, Art. no. 21631, Feb. 2025.

[13] J. Gao, G. Liu, B. Zhu, S. Zhou, H. Zheng, and X. Liao, "Multi-level attention and contrastive learning for enhanced text classification with an optimized transformer," arXiv:2501.13467, Jan. 2025.

[14] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Dirichlet ExtSMOTE and other robust oversampling techniques for logistic regression classification," *Mach. Learn. Appl.*, vol. 18, Art. no. 100597, 2024.

[15] D. Mildenberger, P. Hager, D. Rueckert, and M. Menten, "A tale of two classes: Adapting supervised contrastive learning to binary imbalanced datasets," arXiv:2503.17024, Mar. 2025.

[16] X. Gao, M. Ramli, M. I. Rosli, et al., "Revisiting self-supervised contrastive learning for imbalanced classification," *Int. J. Electr. Comput. Eng.*, vol. 15, no. 2, pp. 1949–1960, Apr. 2025.

[17] I. Valmianski, D. Broniatowski, and K. Dredze, "Evaluating robustness of language models for chief complaint classification in public health surveillance," arXiv:1905.00368, 2019.

[18] Z. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL, USA: CRC Press, 2012.