

Public Complaint Text Classification in the Wargaku Application Using TF-IDF and Support Vector Machine

Alfina Dian Febyani^{1*}, I Kadek Dwi Nuryana¹ , Nisa Dwi Angresti² 

¹State University of Surabaya, Surabaya, Indonesia

alfinadian.21001@mhs.unesa.ac.id, dwinuryana@unesa.ac.id

²Universitas Andalas, Padang, Indonesia

nisadwi@it.unand.ac.id

Abstract. The Wargaku application is utilized by Surabaya residents to submit complaints regarding population administration services. As the number of complaints continues to increase, manual complaint categorization becomes inefficient and prone to misclassification. This study proposes an automated public complaint text classification system using Natural Language Processing (NLP) and machine learning techniques. The dataset consists of 2,303 public complaints distributed across 18 service categories. During data preparation, complaint texts were preprocessed and transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) method. To address the class imbalance problem, Synthetic Minority Over-sampling Technique (SMOTE) and RandomOverSampler were applied before model training. Three machine learning algorithms, namely Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN), were evaluated using accuracy, precision, recall, and F1-score. Hyperparameter tuning was performed to optimize model performance under three train-test split scenarios. Experimental results show that the SVM model achieved the best performance using the 85:15 split, with an accuracy of 93.96%, a training F1-score of 96.08%, and a testing F1-score of 94.15%. The best-performing model was deployed as a Streamlit-based web application for automatic complaint classification. These findings demonstrate that combining TF-IDF, class balancing techniques, and SVM provides an effective and practical solution for improving the efficiency and consistency of digital public complaint services.

Keywords: *Natural language processing; Public complaints; Streamlit application; Support vector machine; Text classification; TF-IDF*

1. Introduction

The rapid development of information technology has transformed public service delivery by enabling governments to provide more accessible, responsive, and citizen-oriented services ([Purwita et al., 2025](#)). Digital complaint platforms have become an important channel for citizens to communicate service problems and provide feedback to public institutions. In Surabaya, Indonesia, the WargaKu application facilitates the submission of complaints, suggestions, and aspirations related to public services. Previous studies have highlighted the role of WargaKu in supporting public service delivery, transparency, responsiveness, and citizen participation ([Intani et al., 2022](#); [May & Fanida, 2022](#)). However, the effectiveness of digital complaint platforms also depends on the ability of government institutions to process and route incoming reports efficiently.

The dataset examined in this study consists of 2,303 complaints submitted to the Surabaya Population and Civil Registration Office (Dispendukcapil) during 2022–2024 and distributed across 18 service categories. The distribution is highly imbalanced, with KTP-EL accounting for 631 complaints, followed by Birth Certificate (323) and Family Card (246), whereas several categories contain fewer than ten observations. In the current process, users manually select the complaint category when submitting a report, which may result in inappropriate categorization and increase the workload of service officers ([Adilita & Rahmawati, 2024](#)). Similar problems have been reported in other government complaint platforms. In the JakLapor service in Jakarta, incoming reports were manually classified, and inappropriate category selection motivated the development of automated text classification models ([Intani et al., 2022](#)).

Natural Language Processing (NLP) and machine learning provide an alternative approach for automatically categorizing complaint texts. Text representations such as Term Frequency–Inverse Document Frequency (TF-IDF) can transform unstructured complaints into numerical features, while Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN) provide alternative classification approaches ([Breiman, 2001](#); [Joachims, 1998](#)). Research on citizen reports in Indonesia has demonstrated the applicability of contextual text analytics and machine learning for Indonesian-language complaint classification ([Madyatmadja et al., 2022](#)). Other studies have compared multiple machine learning algorithms for citizen complaints, including K-Nearest Neighbors, Random Forest, SVM, and AdaBoost, demonstrating that classification performance varies across algorithms and datasets ([Madyatmadja et al., 2022](#)). ([Hardiansyah et al., 2025](#)) similarly applied TF-IDF and Naïve Bayes to classify university service complaints and reported an accuracy of 91%, while ([Karsana et al., 2023](#)) demonstrated differences in text-classification performance across machine learning approaches.

Class imbalance presents an additional challenge in the present study because complaint categories are represented by substantially different numbers of observations. Imbalanced data can cause classification models to favor majority classes and consequently reduce their ability to recognize minority categories ([Chen et al., 2024](#)). The Synthetic Minority Over-sampling Technique (SMOTE) addresses this problem by generating synthetic observations for minority classes ([Chawla et al., 2002](#)). Recent international research has also demonstrated the relevance of balancing techniques in complaint classification. ([Li et al., 2025](#)), for example, developed an NLP-based system for classifying patient complaints using 1,465 records for training and validation and 376 records for external testing; the study applied SMOTE to address imbalanced categories and evaluated Logistic Regression, Multinomial Naïve Bayes, and SVM. These findings reinforce the importance of explicitly addressing class imbalance when developing automated complaint classification systems.

Although previous studies have demonstrated the feasibility of machine learning for complaint classification, several gaps remain. Existing studies differ in their application contexts, numbers of complaint categories, dataset distributions, feature representations, and evaluation procedures. Moreover, limited research has systematically compared multiple machine learning models while simultaneously examining class-balancing strategies, hyperparameter optimization, and different train-test configurations using a highly imbalanced government complaint dataset. The practical deployment of the resulting model in a government service environment also remains relatively underexplored. Therefore, further empirical evaluation is needed using real operational complaint data with multiple service categories and substantially uneven class distributions.

To address this gap, this study develops an automated NLP-based classification system for WargaKu complaints using SVM, RF, and NN. TF-IDF is applied for feature extraction, while SMOTE and RandomOverSampler are compared to address class imbalance. Hyperparameter tuning is conducted under three train-test split scenarios (75:25, 80:20, and 85:15), and model performance is evaluated using accuracy, precision, recall, and F1-score. Finally, the best-performing model is implemented in a Streamlit-based application to support automated complaint categorization and routing at the Surabaya Population and Civil Registration Office.

This study contributes by providing an empirical comparison of machine learning models using a real-world, highly imbalanced, multiclass government complaint dataset and by connecting model evaluation with practical deployment. The findings provide methodological evidence for automated complaint classification and practical insights into the application of NLP-based machine learning to improve the efficiency and consistency of public complaint management.

2. Methods

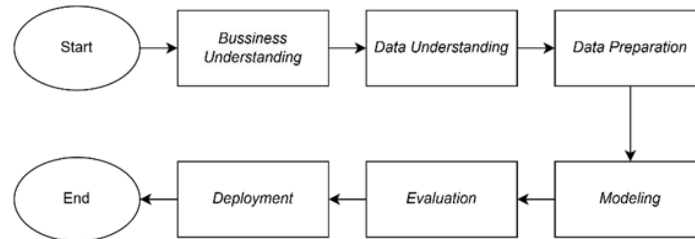


Figure 1 CRISP-DM Research Framework

This study adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, a widely recognized methodology for developing data mining and machine learning solutions. CRISP-DM consists of six iterative phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The iterative nature of the framework allows improvements to be incorporated throughout the research process, ensuring that the developed model aligns with the research objectives and practical requirements.

2.1. Business Understanding

The primary challenge addressed in this study is the increasing number of public complaints submitted through the Wargaku application, which currently requires users to manually select complaint categories. This manual process frequently results in incorrect classification, delays complaint handling, and increases the workload of officers at the Surabaya Population and Civil Registration Office (Dispendukcapil). Therefore, this study aims to develop an automated text classification system capable of predicting complaint categories directly from the complaint text, thereby improving the efficiency, consistency, and accuracy of complaint management.

2.2. Data Understanding

The dataset consists of public complaints submitted through the Wargaku application between 2022 and 2024. A total of 2,303 complaint records were collected and categorized into 18 service classes. The dataset exhibits an imbalanced distribution, where several categories, such as KTP-EL, Birth Certificate, and KK, contain substantially more complaints than categories such as KITAS, Statement of Absolute Responsibility (SPTJM), and Internship Information. Understanding this distribution is essential because class imbalance may bias machine learning models toward majority classes and reduce classification performance for minority categories.

2.3. Data Preparation

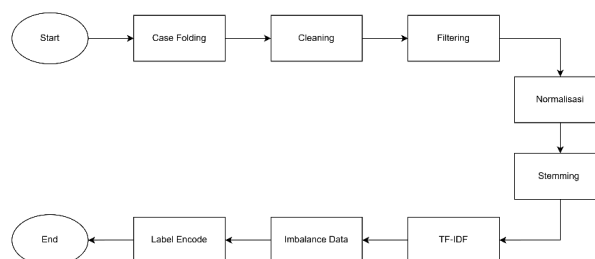


Figure 2 Text Preprocessing Flowchart

The TF-IDF feature extraction method was selected because it effectively represents textual documents by assigning higher weights to informative terms while reducing the influence of frequently occurring words. Compared with CountVectorizer, TF-IDF generally provides better discrimination for text classification because it considers both term frequency and inverse document frequency, making it more suitable for sparse textual data ([Gunawan & Santoso, 2021](#); [Karsana et al., 2023](#)).

The maximum number of features was limited to 300 to reduce computational complexity while preserving the most informative terms. Preliminary experiments showed that increasing the number of features beyond this threshold produced only marginal performance improvements while increasing training time and model complexity.

Furthermore, stemming was performed using the Porter Stemmer implementation available in the Sastrawi library to normalize words into their root forms, thereby reducing vocabulary size and improving feature consistency during classification ([Hidayat, 2022](#)).

2.4. Modeling

This study formulates the problem as a multiclass text classification task in which each complaint is assigned to one of 18 public service categories. Three supervised machine learning algorithms were evaluated: Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN). These algorithms were selected because they represent different classification paradigms and have demonstrated competitive performance in previous text classification studies ([Gunawan & Santoso, 2021](#)).

To optimize model performance, hyperparameter tuning was conducted using GridSearchCV with three-fold cross-validation. The evaluated hyperparameters differed according to each algorithm and were selected based on their influence on classification performance. Model evaluation was conducted using three train-test split scenarios: 75:25, 80:20, and 85:15. These scenarios were intentionally selected to investigate the stability of each classifier under different proportions of training data. Increasing the proportion of training data allows the models to learn richer textual patterns, whereas maintaining an adequate testing set ensures reliable and unbiased performance evaluation. Comparing multiple train-test splits also provides additional evidence regarding the robustness and consistency of each classifier rather than relying on a single experimental setting.

2.5. Evaluation

The classification models were evaluated using accuracy, precision, recall, and F1-score on both training and testing datasets. In addition to overall performance metrics, category-level performance was analyzed using a confusion matrix and classification report to identify categories that remained difficult to classify. This comprehensive evaluation enables assessment of both the predictive performance and generalization capability of each algorithm.

2.6. Deployment

The best-performing classification model was exported in .pkl format using the Joblib library and deployed as a web-based application using Streamlit. The application provides a simple graphical user interface (GUI) that allows users to enter complaint text and instantly obtain the predicted complaint category. This deployment demonstrates the practical applicability of the proposed approach in supporting complaint management within the Surabaya Population and Civil Registration Office (Dispendukcapil).

3. Results and Discussion

3.1. Text Data Preprocessing

The preprocessing stage is carried out to clean and prepare the text data so that it is ready for the modeling process. This process consists of several key steps, as illustrated in the following figure:

3.1.1. Case Folding

Case Folding is the process of converting all characters in a text into lowercase letters to ensure consistency during text processing. This step is essential because most text classification models treat "KTP" and "ktp" as different tokens. By standardizing all characters to lowercase, we reduce redundancy and improve the quality of feature extraction.

Table 1 Example of Case Folding

Before	After
selamat siang bapak ibu DISPENDUKCAPIL saya mau mengurus akta kematian om saya melalui online via website klampid namun tidak bisa mohon informasinya terima kasih	selamat siang bapak ibu dispendukcapil saya mau mengurus akta kematian om saya melalui online via website klampid namun tidak bisa mohon informasinya terima kasih

3.1.2. Cleaning

Cleaning is the process of removing non-alphabetic characters, numbers, punctuation marks, and special symbols that are irrelevant to text analysis. These elements often do not contribute meaningful information for classification tasks and may introduce noise into the model. By eliminating them, the text becomes cleaner and more suitable for further processing such as tokenization or vectorization.

Table 2 Example of Text Cleaning

Before	After
selamat siang bapak/ibu DISPENDUKCAPIL saya mau mengurus akta kematian om saya melalui online via website klampid https://klampid-dispendukcapil.surabaya.go.id/ namun tidak bisa. mohon informasinya terima kasih	selamat siang bapak ibu DISPENDUKCAPIL saya mau mengurus akta kematian om saya melalui online via website klampid namun tidak bisa mohon informasinya terima kasih

3.1.3. Filtering

Filtering involves removing common or frequently occurring words, known as stopwords, that do not carry significant meaning in the context of classification. Examples include words like “and,” “the,” “of,” and “from.” These words generally do not contribute to distinguishing between categories and may dilute the effectiveness of feature extraction. By removing them, the model can focus on more meaningful terms that are relevant to the classification task.

Table 3 Example of Stopword Removal

Before	After
selamat siang bapak ibu dispendukcapil saya mau mengurus akta kematian om saya melalui online via website klampid namun tidak bisa mohon informasinya terima kasih	selamat siang bapak ibu dispendukcapil saya mau urus akta kematian om saya lewat online via website klampid namun tidak bisa mohon informasi terima kasih

3.1.4. Normalization

Normalization is the process of converting non-standard or informal words into their standard forms using a normalization dictionary. This step is particularly important in processing text from user-generated content, where misspellings, abbreviations, or colloquial expressions are common. By

standardizing the vocabulary, the model can better recognize and learn consistent patterns across the dataset, leading to more accurate classification.

Table 4 Example of Text Normalization

Before	After
selamat siang bapak/ibu	selamat siang bapak/ibu
DISPENDUKCAPIL saya mau	Dispendukcapil, saya ingin
mengurus akta kematian om saya	mengurus akta kematian paman
melalui online via website klampid	saya melalui daring melalui situs
https://klampid-dispendukcapil.surabaya.go.id/	web Klampid (https://klampid-dispendukcapil.surabaya.go.id/),
namun tidak bisa. mohon	namun tidak dapat diakses. Mohon
informasinya terima kasih	informasinya, terima kasih.

3.1.5. Stemming

Stemming is the process of reducing words to their root or base form. For example, the word “mengurus” is stemmed to “urus.” This step helps in grouping different forms of a word under a single representation, which simplifies the vocabulary and improves the performance of the classification model. Stemming is particularly useful in languages with rich morphology, such as Indonesian, where affixes can create multiple forms of the same root word.

Table 5 Example of Stemming

Before	After
selamat siang bapak ibu	selamat siang bapak ibu
dispendukcapil saya mau mengurus	dispendukcapil saya mau urus akta
akta kematian om saya melalui	mati om saya lewat online via
online via website klampid namun	website klampid namun tidak bisa
tidak bisa mohon informasinya	mohon informasi terima kasih
terima kasih	

3.1.6. TF-IDF (Term Frequency-Inverse Document Frequency)

This stage transforms the preprocessed text into a numerical representation in the form of vectors with a maximum of 300 features, based on the frequency of word occurrence in the document and its importance relative to the entire corpus. The representation is generated using the TF-IDF (Term Frequency–Inverse Document Frequency) method, which emphasizes words that appear frequently in a particular document but rarely in others, making them more informative for the classification process.

The final result of the preprocessing stage is a numerical feature matrix representing each public complaint document. This matrix is then used as input for the classification algorithms.

Table 6 TF-IDF Weighting

Document	Certificate	Birth		Citizen	Area
D1	0.00000	0.00000		0.00000	0.00000
D2	0.00000	0.00000		0.00000	0.00000
D3	0.00000	0.00000		0.00000	0.00000
D4	0.00000	0.00000		0.00000	0.00000
D5	0.83348	0.00000		0.55256	0.00000

The table above shows the TF-IDF weight values for several keywords from four complaint documents. These weights indicate the importance of a word within a document relative to the entire corpus. The higher the value, the more significant the word is in that particular document.



Figure 3 Word Cloud Visualization

The word cloud is used to illustrate the most frequently occurring words in the complaint data corpus. Words such as “ktp,” “akta,” “lahir,” and “keluarga” dominate, reflecting the types of population administration services most commonly complained about by the public.

3.1.7. Handling Imbalance Data

To address class imbalance, the number of data samples in each category was balanced using SMOTE (Synthetic Minority Over-sampling Technique) and RandomOverSampler. These methods were applied to ensure that each category contained an equal number of samples, which helps improve model performance and prevents bias toward majority classes during training.

Table 7 Class Distribution Before and After Oversampling

Category Name	Original Count	Oversampled Data
KTP-EL	631	631
Akta Kelahiran	323	631
KK	246	631
Klampid	218	631
Akta Kematian	124	631
Pelayanan Petugas Loket Dispendukcapil Siola	118	631
Konsolidasi NIK	113	631
KIA	94	631
Pindah Masuk Kota Surabaya	91	631
Pindah Keluar Kota Surabaya	88	631
Akta Perkawinan	77	631
Pindah Antar Kecamatan	73	631
Pindah Antar Kelurahan	38	631
Legalisir Kependudukan	28	631
Akta Perceraian	27	631
Surat Pertanggungjawaban Mutlak (SPTJM)	9	631
KITAS	4	631
Informasi Magang	1	631

3.1.7. Label Encoding

This process converts each complaint category into a unique numeric label to be used as the classification target. It ensures compatibility with machine learning models that require numerical input.

Table 8 Label Encoding of Complaint Categories

Category Name	Encoded Value
KTP-EL	0
Akta Kelahiran	1
KK	2
Klampid	3
Akta Kematian	4
Pelayanan Petugas Loker Dispendukcapil Siola	5
Konsolidasi NIK	6
KIA	7
Pindah Masuk Kota Surabaya	8
Pindah Keluar Kota Surabaya	9
Akta Perkawinan	10
Pindah Antar Kecamatan	11
Pindah Antar Kelurahan	12
Legalisasi Kependudukan	13
Akta Perceraian	14
Surat Pertanggungjawaban Mutlak (SPTJM)	15
KITAS	16
Informasi Magang	17

3.2. Modeling

This study utilizes three machine learning algorithms to perform classification on public complaint texts sourced from the Wargaku application: Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN). The implementation details of each model are described below:

3.2.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a commonly used algorithm in machine learning for solving classification and regression problems. It functions by identifying the optimal hyperplane that separates data points from different categories with the widest possible margin, which refers to the distance between the hyperplane and the closest data points from each class (Taufiqurrahman et al., 2021). In this study, the SVM was tuned using parameters C and gamma, with optimization performed through 3-fold Grid Search Cross Validation (GridSearchCV). Input data was converted into numerical form using TF-IDF (Term Frequency–Inverse Document Frequency) and reduced to 300 features. The evaluation results demonstrated that the SVM achieved an accuracy of 93.96% and obtained the highest F1-score of 94.15%, with a small gap from the training F1-score (~0.0193), indicating strong consistency and accuracy.

3.2.2. Random Forest (RF)

Random Forest (RF) is an ensemble learning technique that constructs multiple decision trees on randomly sampled subsets of data (bagging) and selects features at random for each tree (Adrian et al., 2021). In this study, the model was built using 300 TF-IDF features, and data balancing was performed with SMOTE and RandomOverSampler techniques. Important hyperparameters such as max_depth, min_samples_split, and n_estimators were tuned to optimize model performance. While RF achieved the highest classification accuracy (95.15%), a considerable F1-score gap between training and testing results suggested a risk of overfitting.

3.2.3. Neural Network (NN)

The Neural Network (NN) is a machine learning algorithm inspired by the human brain's biological neural structure. It consists of multiple layers of interconnected nodes (neurons), including

an input layer, one or more hidden layers, and an output layer (Dharmawan et al., 2022). This model employed the ReLU activation function and a regularization parameter (alpha) of 0.002. Similar to the other models, TF-IDF features were used as input. Although its performance was slightly below SVM, NN still produced a high F1-score of 93.88%, indicating competitive classification capabilities.

After preprocessing, the dataset was prepared for modeling. The three models were compared to determine the best-performing algorithm for classifying complaints into 18 categories. The dataset was divided into training and testing sets using three ratio splits: 75:25, 80:20, and 85:15. The evaluation process was carried out using accuracy and F1-score metrics. Additionally, hyperparameter tuning was applied to improve each model's performance.

Based on the experimental results, the SVM model achieved the most optimal performance when using an 85:15 training-to-testing split, with an accuracy of 93.96%, a training F1-score of 96.08%, and a testing F1-score of 94.15%. These results confirm that SVM provides reliable and stable performance for text classification tasks related to public complaints.

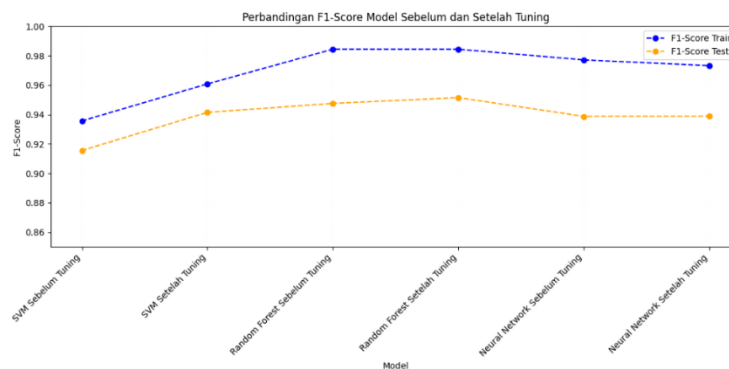


Figure 4 Comparison of F1-Scores Before and After Hyperparameter Tuning (85:15 Split)

Figure 4 presents a comparison of F1-scores before and after hyperparameter tuning using the 85:15 train-test ratio. The Support Vector Machine (SVM) model demonstrated an improvement in F1-score from 0.9155 to 0.9415 after tuning, representing the best performance across all test scenarios. Additionally, the F1-score gap between training and testing data for SVM was the smallest (± 0.0193), indicating excellent generalization ability and a low risk of overfitting.

Table 9 Model Evaluation Results

Data Split	Model	Accuracy (After Tuning)	F1-score (Train)	F1-score (Test)
75:25	Support Vector Machine	0.929577	0.961303	0.931639
	Random Forest	0.939437	0.986077	0.940011
	Neural Network	0.931338	0.975088	0.932161
80:20	Support Vector Machine	0.934419	0.960659	0.936521
	Random Forest	0.946743	0.985085	0.947084
	Neural Network	0.937060	0.976814	0.937338
85:15	Support Vector Machine	0.939554	0.960819	0.941477
	Random Forest	0.951291	0.984434	0.951503
	Neural Network	0.938380	0.973262	0.938833

The experimental results indicate that the Support Vector Machine (SVM) consistently achieved the most balanced performance across all train-test split scenarios. The highest testing performance was obtained using the 85:15 split, resulting in an accuracy of 93.96% and an F1-score of 94.15%. In addition, the difference between the training and testing F1-scores was approximately 0.0193, indicating good generalization capability and a relatively low risk of overfitting.

The superior performance of SVM can be attributed to the characteristics of the TF-IDF feature representation. TF-IDF produces high-dimensional and sparse feature vectors, which are particularly suitable for margin-based classifiers such as SVM. By maximizing the decision margin between classes, SVM effectively identifies discriminative textual patterns even when the feature space contains a large number of dimensions. Similar findings have been reported by [\(Gunawan & Santoso, 2021\)](#) who demonstrated that SVM performs consistently well for Indonesian text classification tasks involving high-dimensional textual features. Likewise, [\(Karsana et al. 2023\)](#) reported that SVM achieved competitive performance compared with Artificial Neural Network (ANN) and Naïve Bayes in multiclass text classification, confirming its effectiveness in handling sparse textual data. Therefore, the findings of this study further support previous research that identifies SVM as one of the most reliable algorithms for multiclass text classification.

Although Random Forest achieved the highest training accuracy, its larger gap between training and testing performance indicates a tendency toward overfitting. Decision tree-based ensemble methods are prone to overfitting when trained on high-dimensional sparse textual features, particularly after oversampling techniques generate additional synthetic samples. Under these conditions, Random Forest may learn dataset-specific patterns instead of generalized decision boundaries, resulting in excellent training performance but lower generalization capability when classifying previously unseen complaint texts.

Meanwhile, the Neural Network model achieved competitive performance but remained slightly below SVM. This outcome is expected because Neural Network models generally require substantially larger datasets to effectively learn semantic representations. Although SMOTE and RandomOverSampler increased the number of training samples, the original dataset remained relatively limited compared with datasets commonly used for deep learning applications. Consequently, the Neural Network could not fully exploit its representation learning capability, resulting in slightly lower classification performance than SVM.

The experimental results also demonstrate that increasing the proportion of training data positively influenced model performance. Among the three evaluation scenarios, the 85:15 train-test split consistently produced the highest testing performance across all classifiers. This finding suggests that providing more training samples enables the models to learn richer lexical patterns while the remaining testing data are still sufficient to provide reliable performance evaluation.

Table 10 F1-Score by Category (SVM Model)

Category	F1-Score
0	0.94
1	0.97
2	0.96
3	0.95
4	1.00
5	0.97
6	0.98
7	0.88
8	0.92
9	0.74
10	0.92
11	1.00
12	0.92
13	0.97
14	0.96
15	0.94

16	0.97
17	0.96

Based on the classification report presented in Table 10, the SVM model achieved F1-scores above 0.90 for the majority of complaint categories, with Category 4 and Category 11 achieving a perfect F1-score of 1.00. These results indicate that the proposed preprocessing pipeline and SVM classifier effectively captured the linguistic characteristics of most complaint types, resulting in highly accurate multiclass classification.

The relatively lower performance observed in Categories 7 and 9 is likely attributable to overlapping vocabulary among complaint categories and the limited contextual information contained in short complaint texts. Furthermore, although SMOTE and RandomOverSampler successfully balanced the class distribution, synthetically generated samples cannot fully replace the semantic diversity found in naturally occurring complaint data. These findings further demonstrate that the combination of TF-IDF feature extraction, data balancing techniques, and the Support Vector Machine algorithm provides an effective approach for multiclass public complaint classification. The proposed model not only achieved high predictive performance but also offers practical applicability for supporting automated complaint categorization within the Wargaku application, thereby improving the efficiency and consistency of public service delivery.

3.3. Comparison of Actual and Predicted Labels

In this study, an analysis was conducted to compare the predictions generated by the best-performing Support Vector Machine (SVM) model against the actual labels in the test dataset. The goal was to evaluate how accurately the model could classify complaint types according to their true categories. The evaluation was carried out directly on sample complaint texts representing various service types under the Surabaya Population and Civil Registration Office (Dispendukcapil).

Table 11 Comparison of SVM Predictions and Actual Labels

Complaint	Prediction	Actual
mohon arahannya bagaimana cara untuk mengganti ttd yg salah ada di e ktp	KTP-EL	KTP-EL
ubah alamat KK gmn caranya?di KK ditulis rt.09 padahal seharusnya rt.08	KK (Kartu Keluarga)	KK (Kartu Keluarga)
maaf mau bertanya apakah terkena denda jika mengurus akta Kematian terlambat? Januari baru urus kematiannya bulan Juli..klo terkena denda brp terima kasih	Akta Kematian	Akta Kematian
mohon informasi kapan saya bisa terima akta kelahiran An : Syamsul Arifin	Akta Kelahiran	Akta Kelahiran
selamat pagi. mohon informasi untuk pengurusan surat pindah masuk ke kota SBY apaa saja ya syarat yang di perlukan ? ini saya sudah cabut berkas dari Lengkong (Nganjuk) untuk masuk ke Surabaya..	Pindah Masuk Kota Surabaya	Pindah Masuk Kota Surabaya
mohon informasi nya . terima kasih		

3.4. System Implementation

The system development method used in this study is the Waterfall model, which consists of sequential stages starting from requirements analysis to testing. This approach was chosen due to its systematic flow and suitability for small-scale system development, such as this complaint classification application.

The deployment phase in this study utilizes Streamlit. The Streamlit application can be accessed through web browsers such as Google Chrome, Opera, Mozilla Firefox, or Microsoft Edge, and is

accessible via both laptops and desktop computers. The classification model implemented is Support Vector Machine (SVM), which had undergone training and hyperparameter tuning to achieve the most stable and accurate performance, particularly in terms of F1-score on the test data. The model is saved in .pkl format using the Pickle library, allowing it to be easily loaded without retraining. The system is designed to accept user input in the form of complaint texts, process them, and return the corresponding service category such as KTP-EL, KK, Birth Certificate, and others.



Figure 5 Streamlit Application Wireframe

The system design process begins with the creation of a wireframe as shown in the figure, serving as the initial structure for the application's user interface. The wireframe outlines the basic visual layout of the text-based complaint category prediction application. The design includes essential elements such as a complaint input field, a prediction button, and an output display area to show the classification result.

Figure 6 Complaint Classification Output Interface

Based on the figure, the application is capable of automatically predicting complaint categories such as "KTP-EL," "KK," or "Birth Certificate." With integrated automatic preprocessing, the system can accept raw text and still provide accurate classification results. The implementation results show that the system can deliver instant complaint category predictions, making it easier for Dispendukcapil Surabaya staff to handle complaints effectively.

4. Conclusion

This study proposed an automated public complaint text classification system for the Wargaku application using Natural Language Processing and compared the performance of Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN). Experimental results demonstrate that SVM achieved the highest classification performance, obtaining an accuracy of 93.96% and an F1-score of 94.15% using the 85:15 train-test split after hyperparameter tuning. The developed

Streamlit-based application further demonstrates the practical applicability of the proposed model for assisting complaint categorization within the Surabaya Population and Civil Registration Office.

Despite these promising results, several limitations should be acknowledged. First, the dataset was collected exclusively from the Wargaku application, limiting the generalizability of the findings to other government institutions. Second, the study employed TF-IDF as the feature representation, which primarily captures lexical information rather than contextual semantics. Third, the classification task was formulated as a single-label problem, whereas real-world complaints may contain multiple service-related topics.

Future research should investigate transformer-based language models such as IndoBERT or multilingual BERT to capture richer contextual information. In addition, evaluating the proposed approach using larger and more diverse datasets from multiple public service institutions, as well as exploring multi-label classification strategies, may further improve model robustness and practical applicability.

Acknowledgments

The authors would like to thank the Department of Population and Civil Registration (Dispendukcapil) of Surabaya for providing the data used in this research. The authors also express their appreciation to the supervising lecturers for their guidance and valuable input during the preparation of this article. This research received no funding from any institution.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

Conceptualization, [Alfina Dian Febyani and I Kadek Dwi Nuryana]; Methodology, [Alfina Dian Febyani and Nisa Dwi Angresti]; Software, [Alfina Dian Febyani]; Data curation, [Alfina Dian Febyani and Nisa Dwi Angresti]; Formal analysis, [Alfina Dian Febyani]; Investigation, [Alfina Dian Febyani]; Visualization, [Alfina Dian Febyani and Nisa Dwi Angresti]; Supervision, [I Kadek Dwi Nuryana]; Writing – original draft, [Alfina Dian Febyani and Nisa Dwi Angresti]; Writing – review & editing, [I Kadek Dwi Nuryana]. All authors have read and agreed to the published version of the manuscript.

Ethics Statement

This study used secondary administrative complaint data provided by the Surabaya Population and Civil Registration Office (Dispendukcapil). The study did not involve direct recruitment or interaction with human participants. Ethical approval was not applicable to the data analysis conducted in this study.

AI Use Declaration

The authors used ChatGPT (OpenAI) as a generative AI-assisted tool during the preparation of this manuscript, primarily for language editing, grammatical refinement, and improving the clarity and organization of the manuscript. The authors reviewed and verified all AI-assisted content and take full responsibility for the accuracy, originality, interpretation, and final content of the published work.

References

- Adilita, V. R., & Rahmawati, I. D. (2024). Efektivitas Penggunaan Aplikasi WargaKu dalam Meningkatkan Pelayanan Publik di Kota Surabaya. *LANCAH: Jurnal Inovasi Dan Tren*, 2(2b). <https://doi.org/10.35870/ljit.v2i2b.2855>

- Adrian, M. R., Putra, M. P., Rafialdy, M. H., & Rakhmawati N.A. (2021). *Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB*. 7.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions. *Artificial Intelligence Review*, 57(6), 137. <https://doi.org/10.1007/s10462-024-10759-6>
- Dharmawan, S., Viny,), Mawardi, C., Novario,), & Perdana, J. (2022). *Jurnal Ilmu Komputer dan Sistem Informasi Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT)*.
- Gunawan, K. I., & Santoso, J. (2021). Multilabel Text Classification Menggunakan SVM dan Doc2Vec Classification Pada Dokumen Berita Bahasa Indonesia. *Journal of Information System, Graphics, Hospitality and Technology*, 3(01), 29–38. <https://doi.org/10.37823/insight.v3i01.126>
- Hardiansyah, R. A., Indra, J., Kusumaningrum, D. S., & Mudzakir, T. Al. (2025). Automatic Classification of Public Complaints Using Naive Bayes. *Bit-Tech*, 8(2), 1548–1559. <https://doi.org/10.32877/bt.v8i2.2992>
- Intani, S. M., Nasution, B. I., Aminanto, M. E., Nugraha, Y., Muchtar, N., & Kanggrawan, J. I. (2022). Automating Public Complaint Classification Through JakLapor Channel: A Case Study of Jakarta, Indonesia. *IEEE ISC2 2022 - 8th IEEE International Smart Cities Conference*, 1–6. <https://doi.org/10.1109/ISC255366.2022.9922346>
- Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Machine Learning: ECML-98, Lecture Notes in Computer Science*, 1398, 137–142. <https://doi.org/10.1007/BFb0026683>
- Karsana, M. M. N., Muslim, K., & Astuti, W. (2023). Single-label and multi-label text classification using ANN and comparison with Naïve Bayes and SVM. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 7(2), 857. <https://doi.org/10.30865/mib.v7i2.6024>
- Li, X., Shu, Q., Kong, C., Wang, J., Li, G., Fang, X., Lou, X., & Yu, G. (2025). An Intelligent System for Classifying Patient Complaints Using Machine Learning and Natural Language Processing: Development and Validation Study. *Journal of Medical Internet Research*, 27, e55721. <https://doi.org/10.2196/55721>
- Madyatmadja, E. D., Yahya, B. N., & Wijaya, C. (2022). Contextual Text Analytics Framework for Citizen Report Classification: A Case Study Using the Indonesian Language. *IEEE Access*, 10, 31432–31444. <https://doi.org/10.1109/ACCESS.2022.3158940>
- May, I. P. A., & Fanida, E. H. (2022). ANALISIS EFEKTIVITAS APLIKASI WARGAKU SURABAYA DALAM MENUNJANG PELAYANAN PUBLIK MASYARAKAT KOTA SURABAYA. *Publika*, 1553–1568. <https://doi.org/10.26740/publika.v11n1.p1553-1568>
- Purwita, A. W., Izzati, B. M., Cinthya, M., Elfaiz, E. A., & Abdillah, R. (2025). Analisis Transformasi Digital Pelayanan Kesehatan Publik melalui Implementasi Aplikasi SiKuat di Puskesmas Kota Sidoarjo. *KERNEL: Jurnal Riset Inovasi Bidang Informatika Dan Pendidikan Informatika*, 6(1), 81–88. <https://doi.org/10.31284/j.kernel.2025.v6i1.7784>
- Taufiqurrahman, F., Al Faraby, S., & Purbolaksono, M. D. (2021). *Klasifikasi Teks Multi Label pada Hadis Terjemahan Bahasa Indonesia Menggunakan Chi-Square dan SVM*.