

COMPARATIVE COMPLAINT TEXT CLASSIFICATION USING SVM, CNN, AND BERT WITH PRIORITY SCORING FOR NYC 311 SERVICE REQUESTS

Putri Sabilla Nur Ryzky¹, Riska Dhenabayu²

Digital Business / Faculty of Economics and Business
State University of Surabaya
putri.23499@mhs.unesa.ac.id, riskadhenabayu@unesa.ac.id

Abstract

Public complaint management systems like NYC 311 generate millions of service requests annually, posing significant challenges regarding manual categorization and resource prioritization. This study has two objectives: (1) to compare the classification performance of Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Bidirectional Encoder Representations from Transformers (BERT) on a large-scale city complaint dataset, and (2) to develop a complaint prioritization framework that combines normalized complaint frequency with average resolution time. The NYC 311 Service Request dataset, comprising 1,203,418 cleaned records, was utilized, with a sample of 20,000 records selected via stratified sampling. Pre-processing steps included converting text to lowercase, removing numbers and special characters, and normalizing whitespace. All models were evaluated without class weighting using a shared test set of 4,000 records. The results indicate that BERT achieved the highest macro F1-score (0.9167), outperforming SVM (0.8861) and CNN (0.8900), while all three models recorded similar aggregate accuracy rates (95.60%–95.83%). Both SVM and CNN yielded a recall of zero for the "Noise – Street/Sidewalk" category, whereas BERT was able to partially recognize it demonstrating that contextual language understanding is essential for distinguishing complaint categories with semantic ambiguity. Priority score analysis identified "Illegal Parking" (0.6007) and "HEAT/HOT WATER" (0.5437) as the highest-priority complaint types, while geographic analysis revealed that Brooklyn and the Bronx bear a disproportionately higher burden of complaints compared to other boroughs. These findings indicate that transformer-based classification combined with composite priority assessment can support data-driven municipal complaint management that accounts for geographical differences.

Keywords: Natural Language Processing; Text Classification; Bert; Convolutional Neural Network; Support Vector Machine; NYC 311; Complaint Prioritization; Priority Scoring

Abstrak

Sistem pengelolaan pengaduan publik seperti NYC 311 menghasilkan jutaan permintaan layanan setiap tahunnya, yang menimbulkan tantangan signifikan terkait kategorisasi manual dan penentuan prioritas sumber daya. Studi ini memiliki dua tujuan: (1) membandingkan kinerja klasifikasi Support Vector Machine (SVM), Convolutional Neural Network (CNN), dan Bidirectional Encoder Representations from Transformers (BERT) pada dataset pengaduan kota berskala besar, dan (2) mengembangkan kerangka kerja penentuan prioritas pengaduan yang menggabungkan frekuensi pengaduan dinormalisasi dengan waktu penyelesaian rata-rata. Dataset Permintaan Layanan NYC 311, mencakup 1.203.418 data bersih, digunakan dalam penelitian ini, dengan sampel sebanyak 20.000 data yang dipilih melalui Teknik stratified sampling. Tahap pra-pemrosesan meliputi perubahan teks menjadi huruf kecil, penghapusan angka dan karakter khusus, serta normalisasi spasi. Semua model dievaluasi tanpa pembobotan kelas menggunakan set pengujian bersama yang terdiri dari 4.000 data. Hasil penelitian menunjukkan BERT mencapai skor F1 makro tertinggi (0,9167), mengungguli SVM (0,8861) dan CNN (0,8900), sementara ketiga model mencatat tingkat akurasi agregat yang serupa (95,60%–95,83%). Baik SVM maupun CNN menghasilkan nilai recall nol untuk kategori "Kebisingan – Jalan/Trotoar", sedangkan BERT mampu mengenalinya sebagian; hal ini menunjukkan bahwa pemahaman bahasa kontekstual sangat penting untuk membedakan kategori pengaduan yang memiliki ambiguitas semantik. Analisis skor prioritas mengidentifikasi "Parkir Liar" (0,6007) dan "Pemanas/Air Panas" (0,5437) sebagai jenis pengaduan dengan prioritas tertinggi, sementara analisis geografis mengungkapkan

bahwa wilayah Brooklyn dan Bronx menanggung beban pengaduan yang secara proporsional lebih tinggi dibandingkan wilayah kota lainnya. Temuan ini mengindikasikan klasifikasi berbasis transformer yang dikombinasikan dengan penilaian prioritas gabungan dapat mendukung pengelolaan pengaduan kota berbasis data yang memperhitungkan perbedaan geografis.

Kata kunci: Pemrosesan Bahasa Alami; Klasifikasi Teks; Bert; Jaringan Saraf Konvolusional; Support Vector Machine; Nyc 311; Penentuan Prioritas Keluhan; Penilaian Prioritas

INTRODUCTION

Urban governments in major metropolitan cities face an increasingly complex challenge in managing large volumes of citizen complaints submitted through public reporting systems. The NYC 311 service request system alone processes millions of complaints annually, spanning categories including noise disturbances, illegal parking, heating failures, sanitation issues, and infrastructure conditions. The sheer volume of incoming requests makes manual categorization operationally infeasible, creating an urgent need for automated systems capable of classifying complaints accurately and routing them to appropriate handling units in real time (Isgbekele E et al., 2025; Pio et al., 2024)

Machine learning-based text classification has emerged as a primary approach for automating complaint categorization across various domains. Support Vector Machines (SVM) utilizing Term Frequency-Inverse Document Frequency (TF-IDF) representations have demonstrated robust performance in high-dimensional text classification tasks, including those involving crowdsourced complaint datasets (Abdullah D & Abdulazeez A, 2021; Madyatmadja et al., 2023). However, TF-IDF-based approaches represent text as static frequency vectors and fail to capture word order or contextual meaning, thereby limiting their ability to distinguish between complaint categories that share similar vocabulary but differ in context. Deep learning approaches, such as Convolutional Neural Networks (CNNs), partially overcome these limitations by learning distributed token representations and capturing local sequential patterns within the text (BOZYİĞİT et al., 2022; Shamsi & Abdallah, 2022; Tang et al., 2021). Studies applying CNNs and attention-based architectures to sentiment and review classification indicate that deep learning models can generalize across heterogeneous, user-generated text domains, including complaint and feedback systems (Bigne et al., 2021; Chen et al., 2021). Nevertheless, CNN embeddings are learned from scratch using training data and capture only local patterns at the phrase

level, lacking access to the broader contextual understanding required to resolve the semantic ambiguity often found in complaint categories.

The emergence of large-scale pre-trained language models specifically BERT (Bidirectional Encoder Representations from Transformers) has significantly advanced the state of the art in text classification. Unlike SVMs and CNNs, BERT generates contextual token representations based on the full bidirectional sentence context, enabling semantic disambiguation for categories with similar vocabularies a capability highly relevant to complaint classification (Devlin et al., 2018; Jurafsky & Martin, 2026; Prottasha et al., 2022). Fine-grained classification tasks in specialized domains, such as financial lexicon analysis and civil aviation complaint categorization, have demonstrated that transformer models outperform classical baseline methods when contextual understanding is required (Cai et al., 2025; Consoli et al., 2022). In multi-class complaint classification, recent studies employing BERT and transformer-based models consistently report superior performance, particularly regarding minority categories and semantically ambiguous ones (Alarifi et al., 2023; Jain et al., 2025). Despite these advantages, the additional computational costs and complexity associated with fine-tuning BERT compared to SVMs and CNNs must be justified by tangible and operationally meaningful performance gains; this is an aspect that remains under-explored in the complaint classification literature concerning multi-class scenarios with imbalanced data distributions.

Despite various advancements, significant gaps remain in the literature. First, most existing complaint classification studies evaluate model performance solely based on aggregate accuracy, failing to examine per-class metrics that could reveal failures in minority categories. A classifier achieving high overall accuracy but yielding zero recall for specific complaint categories would result in the systematic misrouting or mishandling of all complaints of that type a consequence directly impacting service equity that aggregate accuracy

fails to capture (Alarifi et al., 2023; Lee et al., 2019). Second, existing studies treat classification as a standalone task, without extending the analysis to generate actionable prioritization insights that could directly inform municipal resource allocation decisions. Third, geographic analysis regarding the distribution of complaints across administrative sub-units crucial for differentiated service responses in cities with heterogeneous spatial characteristics remains under-explored in the complaint classification literature (Madyatmadja et al., 2023; Saraiva et al., 2022). Fourth, the literature lacks comparative studies evaluating SVM, CNN, and BERT under identical experimental conditions using the same municipal complaint dataset and consistent preprocessing, sampling, and evaluation protocols; consequently, it is difficult to draw conclusions regarding which approach is most suitable for this specific application domain.

This study aims to address this gap by pursuing two primary objectives: (1) systematically comparing the classification performance of SVM, CNN, and BERT models on a large-scale municipal complaint dataset under controlled experimental conditions, and (2) developing and evaluating a data-driven complaint prioritization framework that integrates classification results with complaint frequency and resolution time to support resource allocation for public services. To achieve these objectives, the study focuses on the ten most frequently occurring complaint categories in the NYC 311 dataset; collectively, these categories encompass the majority of service requests and represent the complaint types that are most operationally significant for municipal government management.

This study makes four specific contributions to the complaint classification literature. First, it presents a systematic comparative evaluation of SVM, CNN, and BERT for multi-class complaint text classification, utilizing identical datasets, preprocessing pipelines, sampling procedures, and evaluation protocols; this enables a controlled assessment of the trade-offs between classical, deep learning, and pre-trained transformer approaches. Second, the study demonstrates that aggregate accuracy is insufficient as a sole evaluation criterion in imbalanced multi-class contexts. While SVM, CNN, and BERT achieved nearly identical overall accuracy (95.60%–95.83%), SVM and CNN failed completely on the "Noise – Street/Sidewalk" category, whereas BERT achieved partial success a

distinction with direct operational consequences that is revealed only through macro-F1 scores and per-class recall. Third, the study proposes a composite "Priority Score" framework that combines normalized complaint frequency and average resolution time to generate a data-driven ranking of complaint categories for service resource allocation. Fourth, the study presents a borough-level geographic analysis identifying spatial disparities in complaint burdens across New York City's five boroughs, with implications for needs-based service deployment (Alarifi et al., 2023; Pio et al., 2024; Saraiva et al., 2022).

RESEARCH METHODS

Research Design

This study employs a quantitative approach using machine learning methods applied to text data from a public complaint management system. The research simultaneously addresses two objectives: (1) complaint text classification, which categorizes each service request into one of ten predefined complaint categories using three machine learning models; and (2) complaint prioritization, which generates a data-driven ranking of complaint categories to support public service resource allocation decisions. The overall research workflow comprises ten sequential stages: dataset collection, data cleaning, text preprocessing, class selection, stratified sampling, dataset splitting, model training, model evaluation, response time analysis, and priority score analysis (Alarifi et al., 2023; Madyatmadja et al., 2023).

Dataset Description

The dataset used in this study is the NYC 311 Service Requests dataset, which consists of publicly available records of non-emergency complaints submitted by residents of New York City, NY, USA, to the city government via the 311 reporting system. The initial dataset comprised 1,218,109 records. Following a data cleaning process in which records with missing values in the "Complaint Type" or "Descriptor" columns were removed using the `dropna()` operation the number of usable records was reduced to 1,203,418. The five key columns utilized in this study are: Problem (formerly Complaint Type), Problem Detail (formerly Descriptor), Created Date, Closed Date, and Borough (Madyatmadja et al., 2023).

Only the ten most frequently occurring complaint categories were retained for this

classification task: Illegal Parking, Heating/Hot Water, Noise – Residential, Blocked Vehicle Access, Road Conditions, Snow or Ice, Unsanitary Conditions, Noise – Street/Sidewalk, Plumbing Systems, and Traffic Light Conditions. This selection criterion was based on frequency; collectively, these top ten categories represent the most operationally significant complaint types within the dataset and provide sufficient linguistic and topical diversity for comparative model evaluation. Table 1 summarizes the full dataset characteristics and the data-splitting configuration.

Table 1. Dataset Characteristics

Description	Detail
Data Source	NYC 311 Service Requests
Initial Records	Before cleaning
After Cleaning	After missing value removal
Sampling Method	Stratified Sampling
Training Set	80% of sample
Test Set	20% of sample
Number of Classes	Top 10 complaint types
Split Strategy	Stratified train-test split

Text Preprocessing

The dataset-level cleaning procedure was applied consistently before model-specific text representation. After cleaning, SVM used TF-IDF vectorization, CNN used word-level tokenization and sequence padding, while BERT used the tokenizer associated with the pretrained bert-base-uncased mode. Missing values were removed first, followed by lowercase conversion, numeric character removal, special character removal, and whitespace normalization. No stemming or lemmatization was applied, as complaint descriptions consist predominantly of short noun phrases where morphological normalization provides marginal benefit and risks distorting domain-specific terminology. The workflow comprises five sequential steps summarized in Table 2, accompanied by concrete examples of the "before" and "after" states for each transformation. (BOZYİĞİT et al., 2022; Tang et al., 2021).

Table 2. Text Preprocessing Steps

Step	Process	Before
1	Missing value removal	Records with null Complaint Type or Descriptor
2	Lowercase conversion	"LOUD MUSIC"

3	Number removal	"noise from apt 123"
4	Special character removal	"noise!!!"
5	Whitespace normalization	"noise from street"

In the first step, records with missing or null values in the Complaint Type or Descriptor fields were removed. In the second step, all text was converted to lowercase to eliminate case-based token duplication (e.g., "LOUD" and "loud" treated as the same token). In the third step, numeric characters were removed, as complaint quantities or apartment numbers carry no semantic relevance to complaint categorization. In the fourth step, special characters including punctuation and symbols were removed to reduce noise in the token vocabulary. In the fifth step, multiple consecutive whitespace characters resulting from prior removals were normalized to single spaces (BOZYİĞİT et al., 2022; Tang et al., 2021).

No stemming or lemmatization was applied in this pipeline. Complaint descriptions in the NYC 311 dataset consist predominantly of short noun phrases and domain-specific terms for which morphological normalization provides marginal benefit and risks distorting terminology that is semantically meaningful in context (e.g., "plumbing" vs. "plumb") would eliminate a significant difference in meaning (BOZYİĞİT et al., 2022). Following these general preprocessing steps, SVM employs TF-IDF vectorization, CNN uses word-level tokenization and sequence padding, while BERT utilizes the WordPiece tokenizer associated with the 'bert-base-uncased' checkpoint.

Data Sampling and Splitting

Given the large scale of the cleaned dataset (1,203,418 records), stratified sampling was employed to extract a computationally manageable subset of 20,000 records while preserving the class distribution of the original dataset. Stratified sampling ensures that the proportion of each complaint category in the sample mirrors its proportion in the entire dataset, thereby preventing the under-representation of minority classes which could otherwise introduce bias during the model training and evaluation stages (Madyatmadja et al., 2023).

To validate the chosen sample size, a sensitivity analysis was conducted using the same preprocessing procedures, TF-IDF configuration, stratified data splitting, and LinearSVC model employed in the primary SVM experiment. Three sample sizes 5,000, 10,000, and 20,000 records were evaluated. As shown in Table 3, classification performance remained stable across all three sizes, with accuracy ranging from 95.60% to 95.65% and weighted F1-scores ranging from 0.9369 to 0.9376. Since increasing the sample size from 10,000 to 20,000 did not yield a significant performance improvement, the 20,000-record sample was retained as a practical balance between data representativeness and computational feasibility.

Table 3. Sample-Size Sensitivity Analysis of SVM

Sample	Train/ Test	Accuracy (%)	Weighted F1
5,000	4,000/ 1,000	95,60%	0.9369
10,000	8,000/ 2,000	95,65%	0.9376
20,000	16,000/ 4,000	95,60%	0.9372

For SVM and CNN, the sample of 20,000 data points was split using the stratified train-test split method with an 80:20 ratio (utilizing the `stratify` parameter and `random\_state = 42`), resulting in 16,000 training samples and 4,000 testing samples. For BERT, a three-stage stratified split was applied: the 20,000 data points were first divided into 80% for training-validation (16,000) and 20% for testing (4,000); subsequently, the training-validation set was further split at a 90:10 ratio, yielding 14,400 training samples, 1,600 validation samples, and 4,000 testing samples. The validation set was used to monitor training convergence and was not utilized for the final evaluation. The final evaluation for all three models was conducted using the same testing set (set aside earlier) containing 4,000 data points, ensuring a consistent basis for evaluation and direct comparability.

Classification Models

Three classification models were trained and evaluated under identical experimental conditions to ensure a fair and controlled comparison of their respective architectures. No class weighting was applied to the models, allowing their performance to be compared under uniform training conditions and enabling the contribution of

each model's representation approach to be distinguished from techniques designed to address class imbalance. These models represent three distinct approaches to text classification: a classical machine learning approach utilizing sparse, frequency-based features (SVM), a deep learning approach based on local distributional representations (CNN) and a pre-trained transformer approach leveraging bidirectional contextual representations (BERT).

The SVM model was implemented using LinearSVC from the Scikit-learn library with TF-IDF feature representation. The vectorizer was configured with a maximum vocabulary limit of 5,000 features and English stop-word removal. LinearSVC was selected due to its computational efficiency in processing high-dimensional, sparse text data and its proven effectiveness in multi-class text classification using TF-IDF features (Abdullah D & Abdulazeez A, 2021; James et al., 2017).

Table 4. CNN Architecture

Layer	Type	Function
1	Embedding	Vocab=5,000; dim=128
2	Conv1D	Filters=128; kernel=5; ReLU
3	GlobalMaxPooling1D	-
4	Dense	64 units; ReL
5	Dropout	Rate = 0.5
6	Dense (output)	10 units; soft

BERT is implemented using the `bert-base-uncased` pre-trained checkpoint from the Hugging Face Transformers library. The model comprises 12 transformer encoder layers, 768 hidden units, 12 self-attention heads, and approximately 110 million parameters. Text is tokenized using the WordPiece tokenizer with a maximum sequence length of 128 tokens, allowing out-of-vocabulary terms to be represented through subword decomposition. Fine-tuning is performed by adding a classification head to the [CLS] representation at the final layer and updating all model parameters using 14,400 training samples via the Hugging Face Trainer API. BERT employs bidirectional self-attention to capture contextual information from both directions, thereby generating contextual representations capable of capturing semantic relationships that transcend the sparse features of TF-IDF or the local patterns learned by CNNs

(Devlin et al., 2018; Jain et al., 2025; Prottasha et al., 2022).

Model Configuration and Experimental Setting

Table 5 presents a consolidated summary of all model configuration parameters and experimental settings for the three models. To enhance reproducibility, a fixed random seed of 42 was consistently applied across all experiments covering stratified sampling, train-test splitting, and model initialization where applicable. This same random seed was maintained throughout the comparative experiments to ensure that observed performance differences reflected model architecture rather than variations resulting from random initialization (Jurafsky & Martin, 2026).

Tabel 5. Model Configuration Summary

Parameter	SVM	CNN	BERT
Text representation	TF-IDF	Embedding	WordPiece
Vocabulary size	5,000	5,000	30,522 (fixed)
Max sequence length	-	100 tokens	128 tokens
Algorithm/Basis	LinearSVC	Sequential CNN	Bert-base-uncased
Optimizer	-	Adam	AdamW
Learning rate	-	Default (0.001)	2×10^{-5}
Epoch	-	10	4
Batch size	-	32	16
Weight decay	-	-	0.01
Dropout rate	-	0.5	-
Class weighting	None	None	None
Random seed	42	42	42
Data split	80:20	80:20	72:8:20

Evaluation Metrics

The three models were evaluated using the same held-out test set, consisting of 4,000 records. Seven metrics were used for evaluation: accuracy, precision, recall, F1-score, weighted F1-score, macro F1-score, and the confusion matrix. Table 6 summarizes each metric and its computational scope.

Table 6. Evaluation Metrics

Metric	Formula
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
Precision	$TP / (TP + FP)$

Recall	$TP / (TP + FN)$
F1-Score	$2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$
Weighted F1	Frequency-weighted average of per-class F1-Score
Macro F1	Unweighted average of per-class F1-Score
Confusion Matrix	Matrix of predicted vs. actual class labels

While accuracy measures the proportion of correct classifications, this metric can be misleading on imbalanced datasets because strong performance on the majority class can dominate the results. Therefore, precision and recall are also evaluated for each class via a classification report. Precision indicates the proportion of correctly predicted instances for a class, whereas recall measures the proportion of actual instances of a class that were successfully identified. The F1-score combines precision and recall using their harmonic mean, making it highly useful for evaluating performance when class distributions are imbalanced. In this study, the weighted F1-score calculated as the average of individual F1-scores weighted by class frequency is used as the primary aggregate metric because it reflects overall performance while accounting for class imbalance. The macro F1-score calculated as the unweighted average of per-class F1-scores is also employed as a key metric because it assigns equal importance to each class, making it more sensitive to poor performance or total failure within minority classes. This is particularly relevant for the "Noise – Street/Sidewalk" category, where low recall might have only a limited impact on the weighted F1-score due to the relatively small number of samples (Lee et al., 2019; Madyatmadja et al., 2023).

A confusion matrix serves as an additional diagnostic tool to examine misclassification patterns between individual classes that might not be apparent through aggregate metrics. This is especially important for categories with semantic links such as "Noise – Residential" and "Noise – Street/Sidewalk" which are prone to being confused with one another due to lexical and semantic similarities. Consequently, a model might achieve high accuracy and weighted F1-scores yet fail to identify instances of a minority class, as evidenced by a recall value of 0.00. Such failures are reflected more clearly in per-class results, the macro F1-score, and the confusion matrix than in accuracy alone (Alarifi et al., 2023; Cai et al., 2025).

Response Time and Priority Score Analysis

In addition to classification performance, this study applies a complaint prioritization framework. The resolution time for each complaint is calculated as the difference between the Closed Date and the Created Date, expressed in hours and defined as Resolution Hours. The average Resolution Hours for each complaint category are then calculated and normalized to derive a Response Time Score (Isgbekele E et al., 2025; Pio et al., 2024).

Subsequently, a composite Priority Score is calculated for each of the 10 complaint categories by combining the two normalized indicators using the following formula:

$$\text{Priority Score} = (0.6 \times \text{Frequency Score}) + (0.4 \times \text{Response Time Score})$$

The Frequency Score is derived by dividing the number of complaints in each category by the highest number of complaints across all categories, yielding a value ranging from 0 to 1. Similarly, the Response Time Score is calculated by normalizing the average resolution time for each category against the maximum average resolution time. A higher weight assigned to frequency (0.6) reflects its role as a primary indicator of service load, whereas the response time component (0.4) accounts for urgency by prioritizing categories that require longer average resolution times. These weights were determined empirically and represent a limitation of the study; therefore, future research could validate them using a stakeholder-based approach, such as the Analytic Hierarchy Process (AHP) (Alarifi et al., 2023; Pio et al., 2024).

Borough Analysis

The final analytical stage examines the geographic distribution of complaints across New York City's five administrative boroughs: Brooklyn, Queens, Bronx, Manhattan, and Staten Island. For each of the top-10 complaint categories, the total number of records per borough was computed from the full cleaned dataset (1,203,418 records) prior to stratified sampling. This ensures that the borough distribution reflects the actual geographic prevalence of complaints at city scale rather than the sampled subset used for model training. The analysis focuses particularly on the three highest-priority complaint types Illegal Parking, HEAT/HOT WATER, and Noise - Residential to identify boroughs carrying disproportionately high service

burdens and to derive policy implications for differentiated resource allocation (Madyatmadja et al., 2023; Saraiva et al., 2022).

RESULTS AND DISCUSSION

Dataset Overview and Exploratory Analysis

Before model training, an exploratory analysis was conducted to characterize the distribution of complaint types in the NYC 311 dataset. As shown in Figure 1, Illegal Parking is the most frequently reported complaint (~184,000 records), followed by HEAT/HOT WATER (~163,000) and Noise - Residential (~110,000). The distributional imbalance where Illegal Parking is roughly 7.5× more frequent than Traffic Signal Condition is a critical characteristic of the dataset, as it directly determines the bias pressure on each classification model. All three models were evaluated under identical conditions without class weighting, ensuring a fair comparison of how each architecture handles this imbalance given only the textual features and the distributional properties of the data. (Madyatmadja et al., 2023).

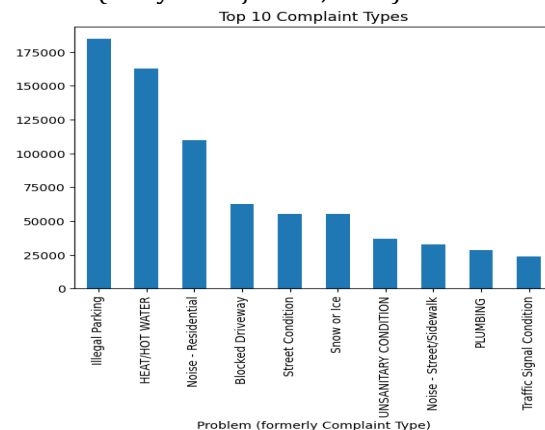


Figure 1. Distribution of Top 10 Complaint Types in the NYC 311 Dataset

SVM Classification Results

The SVM model implemented using LinearSVC with TF-IDF representation (max_features = 5,000, English stopwords removed, random_state = 42), achieved an accuracy of 0.9560 and a weighted F1-score of 0.9372. Table 7 presents the full per-class classification report, showing the eight out of ten complaint categories achieved perfect or near-perfect precision, recall, and F1-score. This reflects the strong discriminative capability of TF-IDF for categories with distinctive vocabularies. However, the Noise - Street/Sidewalk class yielded precision, recall, and F1-score values of 0.00, meaning all 173 test samples from this

category were misclassified. Consistent with this, the Noise – Residential class showed high recall (1.00) but lower precision (0.77), confirming that the classifier misclassified all Noise – Street/Sidewalk samples into the Noise – Residential category. The macro F1-score of 0.8861 reflects the penalty imposed by this complete class failure when each class is weighted equally as opposed to by frequency the aggregate score drops substantially, revealing the limitation invisible in accuracy alone. (Abdullah D & Abdulazeez A, 2021; BOZYİĞİT et al., 2022).

Table 7. Per-class Classification Report - SVM

Class	P	R	F1
Blocked Driveaway	1.00	1.00	1.00
HEAT/HOT WATER	1.00	1.00	1.00
Illegal Parking	1.00	1.00	1.00
Noise – Residential	0.77	1.00	0.87
Noise – Street/Sidewalk	0.00	0.00	0.00
PLUMBING	1.00	1.00	1.00
Snow or Ice	1.00	1.00	1.00
Street Condition	1.00	1.00	1.00
Traffic Signal Condition	0.99	1.00	1.00
UNSANITARY CONDITION	1.00	1.00	1.00
Weighted Avg	0.92	0.96	0.94
Accuracy			0.96

CNN Classification Results

The CNN model, implemented using a sequential architecture (Embedding → Conv1D → GlobalMaxPooling1D → Dropout → Dense; Adam optimizer, 10 epochs, batch size 32, random_seed = 42), achieved an accuracy of 0.9583 and a weighted F1-score of 0.9400. Table 8 presents a comprehensive per-class report. Notably, despite employing a fundamentally different architecture learning distributed token embeddings and local phrase patterns rather than sparse TF-IDF vectors the CNN exhibited an identical failure in the Noise – Street/Sidewalk category: precision, recall, and F1-score values of 0.00, with all 167 test samples from this class misclassified into the Noise – Residential category (recall = 1.00, precision = 0.78). The macro F1-score of 0.8900 was slightly higher than that of the SVM (0.8861) but reflected the same structural failure. The convergence of the SVM and CNN on

this identical failure pattern is a key finding it indicates that the classification failure is not attributable to a specific algorithm alone but is a systematic consequence of the inability to distinguish text between these two noise categories using standard representation approaches. Neither frequency-based representations (TF-IDF) nor local distributional representations (CNN embeddings) were capable of capturing sufficient contextual signals to differentiate these classes (Shama et al., 2024; Tang et al., 2021; Zhang et al., 2021).

Table 8. Per-class Classification Report - CNN

Class	P	R	F1
Blocked Driveaway	1.00	1.00	1.00
HEAT/HOT WATER	1.00	1.00	1.00
Illegal Parking	1.00	1.00	1.00
Noise – Residential	0.77	1.00	0.87
Noise – Street/Sidewalk	0.00	0.00	0.00
PLUMBING	1.00	1.00	1.00
Snow or Ice	1.00	1.00	1.00
Street Condition	1.00	1.00	1.00
Traffic Signal Condition	0.99	0.99	0.99
UNSANITARY CONDITION	1.00	1.00	1.00
Weighted Avg	0.92	0.90	0.89
Accuracy			0.96

BERT Classification Results

The BERT model, fine-tuned using bert-base-uncased (4 epochs, learning rate 2e-5, batch size 16, weight decay 0.01, random_seed = 42), achieved an accuracy of 0.9560 and a weighted F1-score of 0.9496. Table 9 presents a comprehensive per-class report. BERT was the only model among the three to yield a non-zero recall for the Noise – Street/Sidewalk class: precision = 0.49, recall = 0.23, and F1-score = 0.31. Out of 173 test samples in this class, BERT correctly identified 40, whereas SVM and CNN failed to identify a single one correctly. For the Noise – Residential class, BERT recorded precision = 0.80, recall = 0.93, and F1-score = 0.86 reflecting a more balanced trade-off compared to SVM and CNN; the latter two models achieved a recall of 1.00 only by misclassifying all Noise – Street/Sidewalk samples into that category. The macro F1-score of 0.9167 was significantly higher than those of SVM (0.8861) and CNN

(0.8900), providing quantitative evidence as requested by the reviewers that BERT demonstrates superior coverage across complaint classes when each class is weighted equally (Devlin et al., 2018; Jain et al., 2025; Prottasha et al., 2022).

Table 9. Per-class Classification Report - BERT

Class	P	R	F1
Blocked Driveway	1.00	1.00	1.00
HEAT/HOT WATER	1.00	1.00	1.00
Illegal Parking	0.99	1.00	0.99
Noise - Residential	0.80	0.92	0.86
Noise - Street/Sidewalk	0.48	0.27	0.36
PLUMBING	1.00	1.00	1.00
Snow or Ice	1.00	1.00	1.00
Street Condition	1.00	1.00	1.00
Traffic Signal Condition	1.00	1.00	1.00
UNSANITARY CONDITION	1.00	1.00	1.00
Weighted Avg	0.95	0.92	0.95
Accuracy			0.96

Model Performance Comparison

Table 10. Classification Model Performance Comparison

Model	Accuracy (%)	Weighted F1-Score
SVM	95.60	0.93
CNN	95.83	0.94
BERT	95.60	0.94

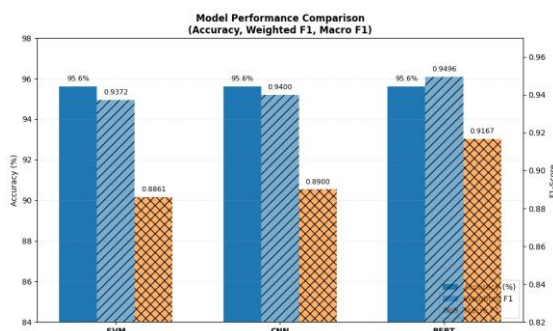


Figure 2. Model Performance Comparison – Accuracy, Weighted F1, and Macro F1

The macro F1 score provides the clearest quantitative distinction: BERT (0.9167) outperforms CNN (0.8900) and SVM (0.8861) by

0.027 and 0.031 points, respectively. Since the macro F1 calculates the unweighted average of per-class F1 scores, this metric explicitly penalizes the complete failure observed in SVM and CNN regarding the Noise - Street/Sidewalk category, while acknowledging BERT's partial success in that class. Weighted F1 also demonstrates BERT's superiority (0.9496 versus 0.9400 and 0.9372), although the margin is smaller because the Noise - Street/Sidewalk class being the smallest carries less weight in the frequency-weighted average. The combination of macro F1 and per-class recall reinforces the case for BERT beyond dispute: it is the only model that achieves non-zero coverage for every complaint category in the evaluation set (Alarifi et al., 2023; Devlin et al., 2018).

Classification Error Analysis

Figure 3 presents the confusion matrix for all three models in a single panel to facilitate direct comparison. For both the SVM and CNN models, the dominant misclassification pattern involved the misassignment of all sample from the Noise - Street/Sidewalk category to the Noise - Residential category, specifically all 173 test samples of Noise - Street/Sidewalk samples were still misclassified as Noise - Residential, while the remaining 40 were correctly identified, indicating a partial yet significant improvement. The other nine classes exhibited negligible misclassification rates across all three models, suggesting that the classification challenge is concentrated specifically on a single pair of categories rather than representing a general limitation of the models.

Confusion Matrix Comparison of SVM, CNN, and BERT

		CNN									
True Label	Blocked Driveway	311	0	0	0	0	0	0	0	0	0
	HEAT/HOT WATER	0	891	0	0	0	0	0	0	0	0
	Illegal Parking	0	0	991	0	0	0	0	0	0	0
	Noise - Residential	0	0	0	584	0	0	0	0	0	0
	Noise - Street/Sidewalk	0	0	0	164	0	0	0	0	0	0
	PLUMBING	0	0	0	0	0	152	0	0	0	0
	Snow or Ice	0	0	0	0	0	0	293	0	0	0
	Street Condition	0	0	0	0	0	0	0	295	1	0
	Traffic Signal Condition	0	0	0	0	0	0	0	0	123	0
	UNSANITARY CONDITION	0	0	0	0	0	0	0	0	0	195
	Predicted Label	Blocked Driveway	HEAT/HOT WATER	Illegal Parking	Noise - Residential	Noise - Street/Sidewalk	PLUMBING	Snow or Ice	Street Condition	Traffic Signal Condition	UNSANITARY CONDITION



Figure 3. Confusion Matrices – SVM, CNN, BERT

The fact that both SVM and CNN models failed on this specific pair of classes despite their architectural differences suggests that the misclassifications stem from data-level characteristics rather than model-level deficiencies. Residents reporting noise disturbances from roadways or sidewalks often use language that is lexically indistinguishable from reports concerning residential arease (e.g., "loud music", "shouting", or "construction noise") without explicitly stating the source location in their free-text descriptions. This linguistic overlap presents a challenge that is difficult to resolve relying solely on available textual features. While BERT's bidirectional attention mechanism subtle contextual distribution patterns across the entire token sequence, the resulting improvements remain limited. This

indicates a need for additional non-textual features specifically GPS coordinates or location indicators derived from addresses to affectively distinguish between these categories.

Response Time Analysis

Resolution time is calculated as the arithmetic difference between the Closed Date and the Created Date for each complaint record, expressed in hours. The complaint categories with the longest average resolution times Tattooing (~1,441 hours), Mobile Food Vendors (~1,441 hours), and Poison Ivy (~1,441 hours) approach the 60-day threshold; this likely reflects an administrative closure artifact where the system automatically closes tickets upon reaching the maximum retention limit, rather than reflecting actual resolution. These categories fall outside the scope of the ten complaint types examined in this study and should be treated as data quality artifacts. Among the ten categories included, the HEATING/HOT WATER category has the longest average resolution time, followed by PLUMBING and UNSANITARY CONDITIONS. These categories involve coordination among various parties including building owners, building inspectors, and regulatory enforcement agencies which structurally extends the closure time beyond the duration required for an immediate operational response (Isgbekele E et al., 2025; Pio et al., 2024).

Priority Score Analysis

$$\text{Priority Score} = (0.6 \times \text{Frequency Score}) + (0.4 \times \text{Response Time Score})$$

Table 11. Priority Score Components and Final Rankings for Top 10 Complaint Types

Rank	Complaint Type	Freq. Score	Resp. Score	Priority Score
1	Illegal Parking	1.0000	0.0011	0.6007
2	HEAT/HOT WATER	0.8571	0.0929	0.5437
3	Noise – Residential	0.5954	0.0033	0.3574
4	UNSANITARY CONDITION	0.4075	0.0020	0.2454
5	Street Condition	0.3482	0.0001	0.2091
6	Blocked Driveway	0.3398	0.0008	0.2040
7	Snow or Ice	0.3337	0.0005	0.2004
8	PLUMBING	0.3172	0.0019	0.1906
9	Noise – Street/Sidewalk	0.1770	0.0011	0.1065

10	Traffic Signal Condition	0.1540	0.0001	0.0926
----	--------------------------	--------	--------	--------

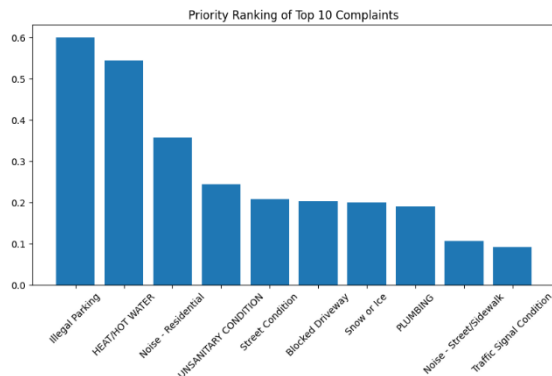


Figure 4. Priority Score Ranking of Top 10 Complaint Types

Table 11 presents the decomposed priority scores. Near-zero Response_Score values across all categories confirm that the frequency component almost entirely drives the rankings raising a question about whether the 0.4 weight on response time sufficiently differentiates time-sensitive categories. Illegal Parking leads (Priority Score = 0.6007), reflecting its dominant complaint volume. HEAT/HOT WATER ranks second (0.5437) due to high frequency combined with an elevated response score consistent with multi-step coordination and direct health implications. UNSANITARY CONDITION ranks fourth despite lower absolute frequency because its response score is proportionally higher. The scoring weights (0.6/0.4) are empirically set; stakeholder validation is advisable before operational deployment (Alarifi et al., 2023; Periyasamy et al., 2026; Pio et al., 2024).

Table 11 presents a breakdown of priority scores, including the frequency and response time components, as well as the final composite score. Response_Score values close to zero across all ten categories confirm that the frequency component is the primary determinant of the final ranking under the current weighting scheme. The Illegal Parking category ranks highest with a Priority Score of 0.6007 (Freq_Score = 1.0000), reflecting its status as the complaint category with the highest volume in the dataset. The Heating/Hot Water category ranks second (0.5437), driven by a combination of high complaint frequency (Freq_Score = 0.8571) and the highest Response_Score (0.0929) among the ten categories; this aligns with the need for phased coordination for heating system repairs and the direct health and safety impacts of unaddressed

heating system failures during winter (Alarifi et al., 2023; Periyasamy et al., 2026).

Poor sanitation conditions ranked fourth (0.2454) despite having a lower absolute complaint frequency than the residential noise category because their proportional response time score was higher, reflecting the escalating public health risks associated with unresolved sanitation issues. This demonstrates that the composite formula captures a dimension of urgency that would not be revealed by rankings based solely on frequency. However, the dominance of the frequency component evidenced by Response_Score values approaching zero across all categories raises methodological concerns; the 0.4 weight assigned to response time is insufficient to significantly alter the ranking order (which is dominated by frequency) within the current dataset. This weighting scheme (0.6/0.4) was determined empirically without stakeholder validation and requires recalibration using participatory policy methods, such as the Analytic Hierarchy Process (AHP), before the framework is applied to operational decision-making (Alarifi et al., 2023; Pio et al., 2024).

Borough Analysis

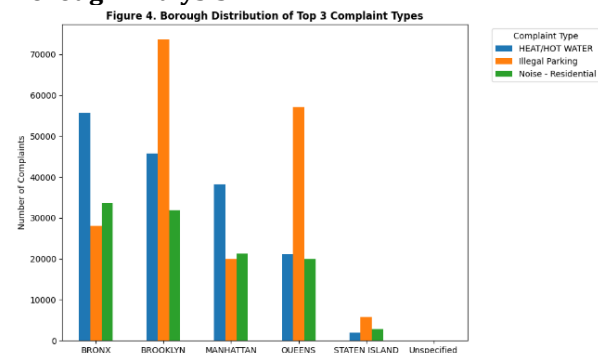


Figure 5. Borough Distribution of Top 3 Complaint Types

Figure 5 reveals geographically differentiated complaint profiles. Illegal Parking is concentrated in Brooklyn (73,613 – 39.9%) and Queens (57,117 – 31.0%), consistent with high residential density and multi-family dwellings without private parking. HEAT/HOT WATER is highest in the Bronx (55,593), reflecting aging rental housing stock and structural economic inequality. Noise – Residential peaks in the Bronx (33,744) and Brooklyn (31,867). These patterns indicate that borough-uniform resource allocation systematically under-serves Brooklyn and the Bronx (Hien et al., 2022; Madyatmadja et al., 2023; Saraiva et al., 2022).

These findings indicate that the burden of complaints is geographically uneven across administrative boroughs and correlates with structural differences in housing availability, land use, and socioeconomic conditions, rather than merely reflecting random reporting variations. The Bronx and Brooklyn consistently record the highest volumes across various complaint categories; consequently, a uniform resource allocation strategy across all boroughs would systematically result in underserved communities in these areas. These patterns support a needs-based approach to service resource distribution, wherein borough-specific complaint profiles generated through the classification and prioritization systems proposed in this study serve as the basis for allocating resources for enforcement, inspections, and maintenance (Madyatmadja et al., 2023; Saraiva et al., 2022).

Discussion

This study makes four contributions to the existing literature on complaint classification. First, a comparison of three models under identical experimental conditions demonstrates that aggregate accuracy is not a reliable primary selection criterion in the context of imbalanced multi-class classification. While the three models achieved accuracy levels ranging from 0.9560 to 0.9583, their capabilities in handling the minority class "Noise – Street/Sidewalk" differed significantly. SVM and CNN yielded a recall of zero, whereas BERT achieved a recall of 0.23 and a macro F1 score of 0.9167, compared to 0.8861 and 0.8900 for SVM and CNN, respectively. These findings align with research by (Madyatmadja et al., 2023) who also found that aggregate metrics obscure inter-class differences in crowdsourced complaint classification and by (Lee et al., 2019), who recommended class-level reporting as a necessary complement to accuracy metrics in medical complaint classification.

Secondly, the shared failure of SVM and CNN models on the "Noise – Street/Sidewalk" class reinforces the interpretation that these errors reflect data-level limitations rather than specific model deficiencies. These findings extend the observations of (BOZYİĞİT et al., 2022) regarding lexical ambiguity between adjacent complaint categories and align with the views of (Cai et al., 2025) and (Saraiva et al., 2022) that spatial features are essential for addressing classification challenges involving geographic context. The

performance improvement achieved by BERT driven by bidirectional contextual attention mechanisms rather than explicit spatial information is consistent with results reported by (Jain et al., 2025) and (Prottasha et al., 2022) for transformer models in semantically ambiguous classification tasks, although the improvement observed in this study remains partial.

Third, the prioritization assessment framework extends the study's analytical contribution, moving beyond mere classification to support operational decision-making. The scores outlined in Table 11 indicate that response time contributes minimally to the ranking under the current weighting scheme; this raises a design question for future implementation: can the formula's intended sensitivity to urgency be maintained when the frequency factor so strongly dominates? This concern aligns with the challenges of translating multi-criteria assessments into policy-relevant tools, as noted (Alarifi et al., 2023) and (Pio et al., 2024). Fourth, analysis at the borough level reveals that the complaint burden exhibits a geographic distribution pattern correlated with known socioeconomic and housing inequalities across New York City a finding with direct implications for equitable service allocation that goes beyond the insights provided by classification accuracy alone.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study demonstrates the potential of machine learning to automate the classification and prioritization of citizen complaints at the metropolitan scale using the NYC 311 Service Requests dataset. Based on the experimental results, four main conclusions can be drawn.

First, BERT achieved the strongest overall performance, with an accuracy of 95.60%, weighted F1-score of 0.9496, and macro F1-score of 0.9167. It was also the only model to correctly identify a portion of the minority Noise – Street/Sidewalk class, achieving a recall of 0.23 and an F1-score of 0.31. In contrast, SVM and CNN achieved comparable accuracy levels but produced zero recall and F1-score for this class. These findings indicate that accuracy alone is insufficient for model selection in imbalanced multi-class classification. Weighted F1 and macro F1, together with per-class recall, provide a more informative assessment of overall and minority-class

performance (Alarifi et al., 2023; Devlin et al., 2018; Lee et al., 2019).

Second, the consistent difficulty in distinguishing Noise – Street/Sidewalk from Noise – Residential across the models suggests that the problem is largely related to the characteristics of the available data. Citizens may describe these complaint types using similar language without explicitly indicating the location or source of the noise. Although BERT's contextual representation enabled partial improvement, a substantial proportion of the Noise – Street/Sidewalk samples remained misclassified. This indicates that textual information alone may not be sufficient to fully distinguish semantically related complaint categories. Future research could therefore incorporate supplementary spatial and contextual features, such as geographic coordinates, borough information, or temporal characteristics, to improve classification performance (BOZYİĞİT et al., 2022; Cai et al., 2025; Saraiva et al., 2022)

Third, the Priority Score framework extended complaint ranking beyond frequency alone by incorporating both complaint volume and response time. Illegal Parking received the highest Priority Score (0.6007), primarily because of its dominant complaint volume, while HEAT/HOT WATER ranked second (0.5437), reflecting the combination of high frequency and relatively longer response time. The ranking demonstrates that the proposed framework can incorporate an urgency dimension that is not captured by frequency-based prioritization alone. However, the relatively small contribution of the normalized response-time scores means that the current weighting scheme remains largely frequency-driven. Since the 0.6 and 0.4 weights were determined empirically without stakeholder validation, future studies should assess and refine these weights using approaches such as the Analytic Hierarchy Process (AHP) and through consultation with municipal stakeholders (Alarifi et al., 2023; Periyasamy et al., 2026; Pio et al., 2024)

Fourth, the borough-level analysis indicates that complaint burdens are geographically uneven. Brooklyn recorded the highest Illegal Parking volume (73,613 complaints; 39.9% of the category total), while the Bronx recorded the highest HEAT/HOT WATER volume (55,593 complaints). Both boroughs also showed high volumes of Noise – Residential complaints. These patterns suggest that a uniform resource allocation strategy may not adequately reflect

differences in service demand across boroughs and may disproportionately affect areas with higher complaint burdens. This finding highlights the importance of geographically informed resource allocation for improving service equity and operational efficiency (Hien et al., 2022; Madyatmadja et al., 2023; Saraiva et al., 2022)

Several limitations should be considered when interpreting these findings. First, the main experiments relied on a single stratified 80:20 train-test split rather than k-fold cross-validation. Although the sample-size sensitivity analysis showed stable SVM performance across 5,000, 10,000, and 20,000 records, the results for all three models remain dependent on the selected sample and split configuration. Future research should employ k-fold cross-validation or repeated hold-out evaluation to obtain more robust performance estimates. Second, the experiments used 20,000 records from a cleaned dataset containing 1,203,418 records, representing approximately 1.7% of the available data. While the sensitivity analysis supports the stability of SVM performance across different sample sizes, the generalizability of CNN and BERT to the full dataset remains to be verified. Third, the models relied exclusively on the Problem Detail text field. Although this design isolates the contribution of textual representations, excluding spatial, temporal, and demographic information limits the models' ability to resolve context-dependent categories such as Noise – Street/Sidewalk. Finally, the Priority Score weighting scheme was empirically defined without direct validation from municipal stakeholders. Consequently, its effectiveness for real-world resource allocation requires further validation using stakeholder input and operational testing.

Suggestion

Based on the findings and limitations of this study, five directions are recommended for future research. First, future studies should integrate non-textual features, particularly GPS coordinates, submission time, and borough-level socioeconomic indicators, alongside complaint text. These additional features may help resolve spatially ambiguous categories, such as Noise – Street/Sidewalk and Noise – Residential, which cannot always be reliably distinguished from textual information alone (Cai et al., 2025; Saraiva et al., 2022)

Second, the weighting scheme of the Priority Score should be validated through

participatory approaches, such as the Analytic Hierarchy Process (AHP) or consultation with relevant municipal stakeholders. This would ensure that the relative contributions of complaint frequency and response time better reflect actual service priorities and strengthen the framework's applicability as a decision-support tool (Alarifi et al., 2023; Pio et al., 2024).

Third, future research could evaluate larger or domain-adapted transformer models, including RoBERTa and DeBERTa, as well as instruction-tuned large language models in zero-shot or few-shot settings. Such models may provide stronger contextual understanding of domain-specific complaint language and offer alternatives to conventional supervised fine-tuning (Almuqati et al., 2026; Jain et al., 2025).

Fourth, future experiments should employ k-fold cross-validation or repeated hold-out evaluation to obtain more robust performance estimates, particularly for minority classes that are sensitive to test-set composition. The classification scope could also be expanded beyond the top 10 complaint categories to evaluate the scalability of the proposed approach across the broader range of NYC 311 complaint types (Lee et al., 2019; Madyatmadja et al., 2023)

Fifth, future studies should investigate the temporal dynamics of complaint patterns, including seasonal changes, weather-related increases, and event-driven surges, together with their interaction with borough-level characteristics. Combining temporal, spatial, and textual information could support more adaptive and anticipatory resource allocation, moving municipal services from reactive complaint handling toward proactive planning (Isgbekele E et al., 2025; Tesfagergish et al., 2022)

REFERENCES

- Abdullah D, & Abdulazeez A. (2021). Machine Learning Applications based on SVM Classification: A Review. *Qubahan Academic Journal*, 1(2), 81-90. <https://doi.org/10.48161/Issn.2709-8206>
- Alarifi, G., Rahman, F., & Hossain, S. (2023). Prediction and Analysis of Customer Complaints Using Machine Learning Techniques. *International Journal of E-Business Research*, 19(1). <https://doi.org/10.4018/IJEER.319716>
- Almuqati, A., Albarrak, A., Mudasser, Q., & Wahsheh, H. (2026). A data-driven approach to answering customer service e-mail inquiries in the Arabic language. *International Journal of Advanced Technology and Engineering Exploration*, 13(134), 123-143. <https://doi.org/10.19101/IJATEE.2025.121220769>
- Bigne, E., Ruiz, C., Cuenca, A., Perez, C., & Garcia, A. (2021). What drives the helpfulness of online reviews? A deep learning study of sentiment analysis, pictorial content and reviewer expertise for mature destinations. *Journal of Destination Marketing and Management*, 20. <https://doi.org/10.1016/j.jdmm.2021.100570>
- BOZYİĞİT, F., DOĞAN, O., & KILINÇ, D. (2022). Categorization of Customer Complaints in Food Industry Using Machine Learning Approaches. *Journal of Intelligent Systems: Theory and Applications*, 5(1), 85-91. <https://doi.org/10.38016/jista.954098>
- Cai, H., Shao, X., Zhou, P., & Li, H. (2025). Multi-Label Classification of Complaint Texts: Civil Aviation Service Quality Case Study. *Electronics (Switzerland)*, 14(3). <https://doi.org/10.3390/electronics14030434>
- Chen, W. K., Riantama, D., & Chen, L. S. (2021). Using a text mining approach to hear voices of customers from social media toward the fast-food restaurant industry. *Sustainability (Switzerland)*, 13(1), 1-17. <https://doi.org/10.3390/su13010268>
- Consoli, S., Barbaglia, L., & Manzan, S. (2022). Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems*, 247. <https://doi.org/10.1016/j.knosys.2022.108781>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://doi.org/10.48550/arXiv.1810.04805>
- Hien, N. T., Su, Y. L., Sann, R., & Thanh, L. T. P. (2022). Analysis of Online Customer Complaint Behavior in Vietnam's Hotel Industry. *Sustainability (Switzerland)*, 14(7). <https://doi.org/10.3390/su14073770>
- Isgbekele E, Valerie O, & Michael O. (2025). View of A Systematic Review of AI-Based Customer Feedback Summarization Techniques. *NIPES Journal of Science and Technology Research*, 7, 607-615. <https://doi.org/https://doi.org/10.37933/nipes/7.4.2025.SI71>



- Jain, U., Mishra, P., Dash, A., & Pandey, A. (2025). Multi-label multi-class text classification-enhanced attention in transformers with knowledge distillation. In *Journal of Applied Research and Technology* (Vol. 23, Number 1). <https://doi.org/10.22201/icat.24486736e.2025.23.1.2484>
- James, Gareth., Witten, Daniela., Hastie, Trevor., & Tibshirani, Robert. (2017). *An introduction to statistical learning: with applications in R*. Springer : Springer Science+Business Media.
- Jurafsky, D., & Martin, J. H. (2026). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft*.
- Lee, S. H., Levin, D., Finley, P. D., & Heilig, C. M. (2019). Chief complaint classification with recurrent neural networks. *Journal of Biomedical Informatics*, 93, 103158. <https://doi.org/https://doi.org/10.1016/j.jbi.2019.103158>
- Madyatmadja, E. D., Sianipar, C. P. M., Wijaya, C., & Sembiring, D. J. M. (2023). Classifying Crowdsourced Citizen Complaints through Data Mining: Accuracy Testing of k-Nearest Neighbors, Random Forest, Support Vector Machine, and AdaBoost. *Informatics*, 10(4). <https://doi.org/10.3390/informatics10040084>
- Periyasamy, R., Govindaraji, M., Nasurulla, I., Srinivasan, V., & Devi, K. R. (2026). Optimizing call center agent efficiency through deep learning-based classifications using SMFCCAE. *International Journal of Reconfigurable and Embedded Systems*, 15(1), 31–41. <https://doi.org/10.11591/ijres.v15.i1.pp31-41>
- Pio, P. G. C., Sigahi, T., Rampasso, I. S., Satolo, E. G., Serafim, M. P., Quelhas, O. L. G., Filho, W. L., & Anholon, R. (2024). Complaint management: comparison between traditional and digital banks and the benefits of using management systems for improvement. *International Journal of Productivity and Performance Management*, 73(4), 1050–1070. <https://doi.org/10.1108/IJPPM-08-2022-0430>
- Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., & Baz, M. (2022). Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning. *Sensors*, 22(11). <https://doi.org/10.3390/s22114157>
- Saraiva, M., Matijošaitienė, I., Mishra, S., & Amante, A. (2022). Crime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics. *ISPRS International Journal of Geo-Information*, 11(7). <https://doi.org/10.3390/ijgi11070400>
- Shama, F., Aziz, A., & Deya, L. B. M. (2024). CitySolution: A complaining task distributive mobile application for smart city corporation using deep learning. *SoftwareX*, 27. <https://doi.org/10.1016/j.softx.2024.101829>
- Shamsi, A. A. Al, & Abdallah, S. (2022). Sentiment Analysis of Emirati Dialects. *Big Data and Cognitive Computing*, 6(2). <https://doi.org/10.3390/bdcc6020057>
- Tang, X., Mou, H., Liu, J., & Du, X. (2021). Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-021-91189-0>
- Tesfagergish, S. G., Kapočūtė-Dzikienė, J., & Damaševičius, R. (2022). Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning. *Applied Sciences (Switzerland)*, 12(17). <https://doi.org/10.3390/app12178662>
- Zhang, A., Lipton, Z. C., Li, M. U., & Smola, A. J. (2021). *Dive into Deep Learning*.

