International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

An ML-Driven Text Mining Framework for Public Complaint Classification and Analysis

Automated Complaint intelligence System

Mrs. Rajeshwari. M [(1)]
Department of CSE,
ACS College of Engineering,
Bangalore – 74, India.
rajimecse90@gmail.com

Ajay Tejaas[(2)], Akaash M S[(3)],
Bhuvan K S[(4)], Gagan B M[(5)]
Department of CSE, ACS College of Engineering, Bangalore - 74, India.
ajaytejaas01@gmail.com

Abstract— Text Mining is one of the knowledge discovery steps in database, in which modeling techniques are applied. urbanization accelerates; municipal administrations face an overwhelming surge in civic grievances, ranging from infrastructure failures to sanitation issues. The traditional grievance redressed mechanism remains largely manual, relying on human operators to read, interpret, and route thousands of complaints daily. This manual dependency creates significant bottlenecks, leading to delayed response times, high rates of misrouting due to human error, and a lack of transparency for citizens.**CivIQ** is an AI-powered Civic Intelligence System designed to automate and streamline this triage process. By leveraging Natural Language Processing (NLP) and Deep Learning, specifically the **Distil BERT** transformer model, the system autonomously classifies unstructured complaint text into primary domains—Infrastructure, Service, or Behavioral—and intelligently routes them to the correct department without human intervention..

Keywords - CivIQ, Distil BERT, NLP, KNN,

K-Means, SQLite.

I. INTRODUCTION

The concept of "Smart Cities" has become a central theme in urban development, promising to integrate technology into the very fabric of municipal administration. As cities expand and Population's density increases, the interaction between citizens and their local government becomes increasingly critical. While the private sector has rapidly adopted artificial intelligence and automation to streamline customer service—think of banking apps or e-commerce support bots—the public sector, specifically civic grievance redressal, remains surprisingly analog.

CivIQ was conceived against this backdrop of inefficiency. We recognized that the technology to solve this problem already exists. By leveraging recent advancements in Artificial Intelligence, specifically Natural Language Processing (NLP) and Transfer Learning, we can automate the understanding of human language. CivIQ aims to bridge the gap between a citizen's problem and the government's solution by replacing manual triage with an intelligent, automated routing system.

II. RELATED WORKS

1. Automatic Classification of Civic Complaints using lightweight Transformers

Authors: R. Menon, S. Gupta,
Elaboration: This paper evaluates the use of lightweight transformer models such as DistilBERT and MobileBERT for classifying short civic complaint texts collected from municipal portals. The study focuses on low-resource environments typical of city-level datasets and compares compact transformer models with traditional TF-IDF-based machine learning classifiers. The work highlights the feasibility of using efficient transformer architectures for real-world public grievance analysis systems.

2. Hybrid Keyword–ML Routing for E-Government Complaints

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Authors: L. Fernández, P. Rao
Elaboration: The authors propose a hybrid backend that combines probabilistic text classifiers with deterministic keyword heuristics to assign complaints to appropriate sub-departments. The hybrid approach is designed to reduce misrouting when classifier confidence is low, thereby improving reliability in real-world municipal systems.

3. Sentiment and Severity Estimation in Public Grievances

Authors: M. Zhou, A. Patel
Elaboration: This study focuses on estimating both sentiment polarity and severity (urgency) from public complaint texts to support prioritization. The authors explore supervised learning approaches to distinguish urgent and non-urgent grievances.

4. Multilingual Complaint Classification for LocalLanguages

Authors: S. K. Rao, P. Mishra
Elaboration: This study investigates multilingual transformer models such as mBERT and XLM-R to handle complaints written in multiple local languages and code-mixed text. It compares direct multilingual fine-tuning with translation-based pipelines.

5. Semi-supervised Learning to Scale Complaint Datasets

Authors: H. Li, E. Santos
Elaboration: This paper addresses the challenge of limited labeled complaint data by applying semi-supervised learning techniques to leverage large volumes of unlabeled municipal complaint logs.

6. Explainable Models for Public Complaint Triage

Authors: A. N. Chatterjee, Y. Kim
Elaboration: This work integrates explainable AI techniques such as LIME and SHAP into complaint classification systems, allowing administrators to understand which words or phrases influenced model decisions.

7. Root-Cause Detection from Aggregated Complaint Streams

Authors: J. Alvarez, R. Singh
Elaboration: Rather than analyzing complaints individually, this research focuses on aggregating complaints over time and location to identify systemic failures and emerging incidents.

8. Robustness and Adversarial Testing for Complaint Classifiers

Authors: K. Müller, D. Park
Elaboration: This paper studies how text perturbations such as typos, paraphrasing, and adversarial noise affect complaint classifiers and proposes robustness-enhancing strategies.

9. Lightweight On-Device Inference for Complaint Triage in Low-Bandwidth Areas

Authors: N. Okoye, T. Basu
Elaboration: This study explores running compact complaint classification models on edge devices such as mobile phones or kiosks in areas with poor internet connectivity.

10. Evaluating Classical vs Transformer Models for Municipal Complaint Classification

Authors: P. Das, E. Novak
Elaboration: This comparative study benchmarks classical machine learning pipelines against transformer-based approaches across multiple municipal complaint datasets.

III. PROBLEM DESCRIPTION

Our analysis of current municipal workflows revealed that the primary bottleneck is the manual triage process. Municipal bodies receive thousands of complaints daily, and the requirement for human intervention to read and sort each one creates massive backlogs, especially during peak crisis periods. This scalability issue is compounded by the high rate of misrouting. Citizens often lack the specific knowledge of municipal hierarchy required to direct a complaint correctly—for instance, reporting a drainage issue to the "Water Supply" department instead of "Sanitation." These misrouted complaints often bounce between departments for weeks, wasting valuable administrative resources.

Furthermore, current systems suffer from a "Black Hole" effect regarding transparency. Once a citizen submits a complaint, there is rarely any feedback loop or visibility into the resolution process. The lack of granular classification—lumping urgent bridge repairs with routine road maintenance under a generic "Infrastructure" tag—further prevents officials from prioritizing critical issues effectively.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

The existing landscape of civic redressal is predominantly reactive. While web portals exist, they often function merely as digital submission forms connected to legacy SQL databases. They lack an intelligent processing layer, meaning a submission simply creates a database row that waits for human review. This reliance on manual processing makes the system slow, opaque, and incapable of scaling to meet the demands of a modern city. Data usually remains siloed, preventing any meaningful analysis of city-wide trends or "hotspots."

IV. SYSTEM DESIGN

CivIQ proposes a shift from digital submission to intelligent processing. By integrating Deep Learning, the system transforms the workflow into a proactive one where classification and routing happen in milliseconds rather than days. Built on a modern technology stack comprising **FastAPI** for high-performance asynchronous handling and **SQLite** for robust data management, the proposed system offers a lightweight, scalable alternative to bloated legacy software, ensuring that the speed of governance matches the speed of the technology used.

The scope of the CivIQ Research covers the full lifecycle of a digital grievance. It begins with a robust ingestion interface that allows citizens to submit unstructured text descriptions and upload image evidence, capturing essential metadata such as timestamps for audit trails. The core scope involves the implementation of a Machine Learning pipeline using the DistilBERT transformer model. This model is responsible for analyzing the textual context and classifying complaints into primary domains: Infrastructure, Service, or Behavioral.

Beyond classification, the system includes a logic layer for intelligent routing, which uses keyword heuristics to assign specific sub-departments based on the AI's prediction. The scope extends to an administrative dashboard for real-time data visualization, enabling officials to monitor department performance, and a citizen tracking portal that ensures transparency by allowing users to retrieve the status and location of their specific grievance using a unique ID.

Our primary motivation stems from the personal frustration of interacting with inefficient civic systems. We realized that the bottleneck wasn't a lack of resources, but a lack of *organization*. As engineering students, we saw an opportunity to apply our technical skills to a domain that affects everyone's daily life. Technologically, we were inspired by the democratization of AI. The availability of powerful pre-trained models like BERT meant that we could build a sophisticated language understanding system without needing the resources of a tech giant. We wanted to demonstrate that state-of-the-art AI is not just for chat-bots or recommendation engines, but can be a powerful tool for social good and effective governance.

A. Functional requirements

The system functionality is divided into three core modules. The **Complaint Submission Module** handles the intake of grievances, sanitizing textual input (minimum 10 characters) and supporting image evidence uploads (JPG/PNG). Upon submission, it triggers the **DistilBERT** model to classify the issue into *Infrastructure*, *Service*, or *Behavioral* categories and automatically routes it to the correct department using keyword logic. The **Complaint Tracking Module** ensures transparency by generating a unique ID for each submission, allowing users to query the database to retrieve the current status, assigned department, and geolocation map. Finally, the **Administrative Dashboard Module** provides officials with real-time data visualization charts and filtering mechanisms to monitor complaint trends.

B. Non-Functional requirements

The system prioritizes **Performance**, requiring AI inference to complete in under 2 seconds on standard CPU hardware and API endpoints to respond within 200ms. **Usability** is ensured through a mobile-first, responsive design with Dark Mode support and immediate visual feedback for user actions. **Reliability** is maintained via persistent **SQLite** storage and robust error handling to prevent server crashes during exceptions.

The expected behaviours, performance constraints, and necessary environment to guide the development process.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

V. METHODOLOGY

1. Dataset Collection and Preparation

The performance of any machine learning model is intrinsically tied to the quality of its training data. For CivIQ, the primary challenge was the absence of a unified, publicly available dataset specifically curated for Indian municipal grievances. To overcome this, we adopted a multi-source data aggregation strategy. We scraped and collated complaint texts from public municipal forums, social media handles of civic bodies, and digital news headlines regarding civic issues. To ensure robustness, we also employed synthetic data generation techniques to create variations of common complaints (e.g., rephrasing "pothole on main road" to "road surface damaged near market").

This raw data was manually annotated into three primary classes: *Infrastructure* (roads, bridges, buildings), *Service* (water, electricity, garbage), and *Behavioral* (corruption, harassment, negligence). The final dataset was shuffled and partitioned into a Training Set (80%) for model fitting and a Testing Set (20%) to evaluate the model's generalization capabilities on unseen data.

2. Text Preprocessing and Tokenization

Raw text required normalization before model ingestion. We applied standard NLP techniques including lowercasing, special character removal, and stop-word filtering. For the deep learning component, we utilized the **DistilBERT Tokenizer**. This tokenizer breaks words into sub-word units (e.g., "complaining" $\rightarrow$ "com", "##plain", "##ing"), allowing the model to grasp semantic meaning and context beyond simple keyword matching.

3. Model Architecture: Distilbert

We selected **DistilBERT** as our primary classification engine. As a distilled version of BERT, it retains 97% of the original performance while being 40% smaller and 60% faster. Our methodology utilized **Transfer Learning**: we loaded the pre-trained DistilBERT model (which understands general English) and fine-tuned it on our specific civic dataset to classify grievances accurately

4. Hybrid Routing Algorithm

While the AI model is excellent at predicting the broad domain (e.g., identifying a complaint as an "Infrastructure" issue), specific department routing requires granular precision. To achieve this, we implemented a **Hybrid Logic Layer** in the backend that combines probabilistic AI with deterministic logic.

Step 1 (AI Prediction): The DistilBERT model analyzes the input and classifies the complaint into a high-level domain (e.g., *Service*).
Step 2 (Keyword Heuristics): The system scans the text for high-priority keywords mapped to specific sub-departments (e.g., keywords like "dengue," "malaria," or "fogging" are mapped to the *Health Department*).
Step 3 (Final Assignment): The logic combines these inputs. If a keyword match is found within the predicted category, the complaint is routed to that specific sub-department. If no specific keywords are detected, the system safely routes the complaint to the general department for that category. This hybrid approach minimizes routing errors by providing a safety net for edge cases.

5. System Integration and Deployment

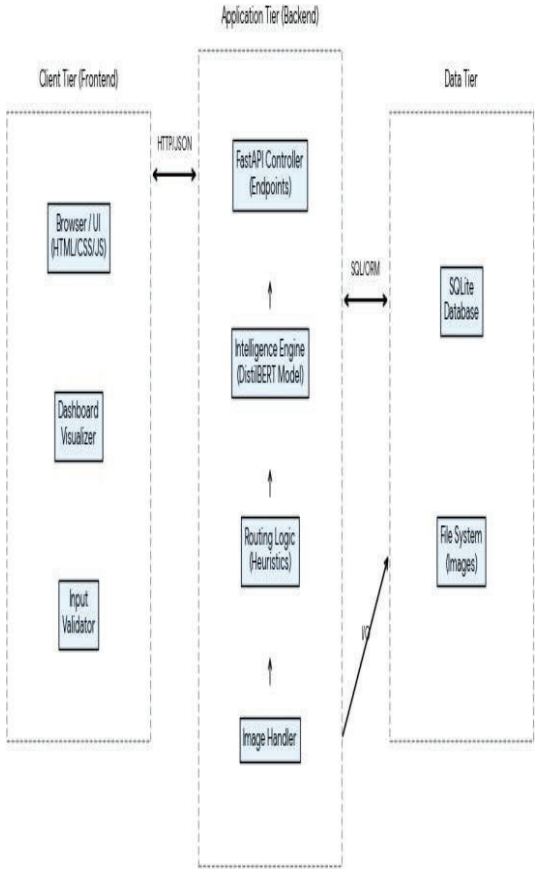
The final phase of the methodology involved integrating these isolated components into a cohesive application. We utilized **FastAPI** to create the backend infrastructure. FastAPI was chosen for its support of asynchronous operations, allowing the server to handle model inference and database writes concurrently without blocking.

When a user submits a complaint via the frontend, the text is sent asynchronously to the API. The backend orchestrates the preprocessing, model inference, and routing logic in real-time. The final results—along with the uploaded evidence image, which is stored in a local file repository—are committed to an **SQLite** database. This modular integration ensures that the complex computational heavy-lifting of the AI is completely abstracted from the user, providing a fast and simple interface.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

VI. EXPERIMENTAL RESULTS

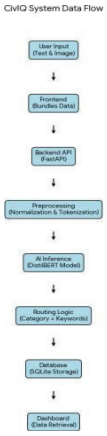
The implementation phase marks the transition from conceptual design to a functional software product. For CivIQ, this process involved setting up the development environment, fine-tuning the machine learning model, developing the API logic, and integrating the user interface. The implementation was carried out in a modular fashion, ensuring that each component (Model, Backend, and Frontend) was tested individually before full system integration.



6.1 UML System Architecture

The core intelligence of the system relies on the **DistilBERT** transformer. Instead of training a model from scratch, which requires immense computational resources, we implemented **Transfer Learning**.

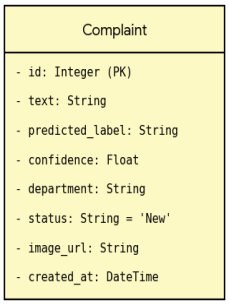
Tokenization: We implemented the DistilBERTTokenizer to preprocess raw text. This involved truncating inputs to a maximum sequence length of 128 tokens to balance performance and speed.



6.2 Flow Diagram

Fine-Tuning: We utilized the DistilBertForSequenceClassification architecture from the Hugging Face library. The model was trained on our custom civic dataset using the **AdamW** optimizer. We monitored the loss function over several epochs to prevent overfitting, ensuring the model could generalize well to new, unseen complaints.

Serialization: Once trained, the model weights were serialized (saved) as a .pth file (distilbert_model.pth). This allowed us to load the model instantly into the backend memory without retraining it every time the server starts.



6.3 Complaint Database

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

The server-side logic was implemented using **FastAPI**.

Endpoint Creation: We defined RESTful endpoints such as POST /api/complaints for data submission and GET /api/complaints for the dashboard.

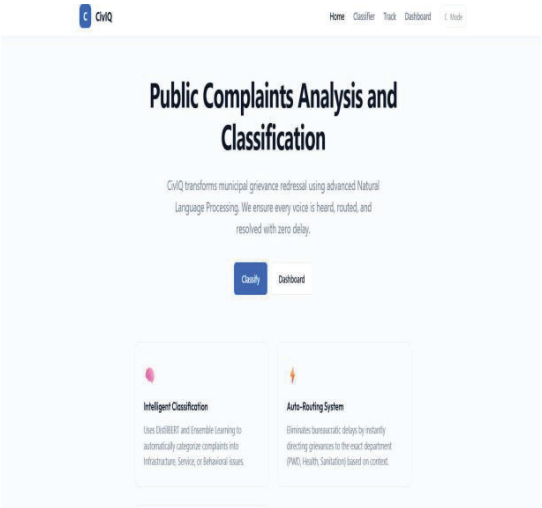
Dependency Injection: We utilized FastAPI's dependency injection system to manage database sessions (get_db). This ensured that every request opened a secure connection to the SQLite database and closed it automatically after the transaction, preventing database locks or corruption.

File Handling: A specific implementation challenge was handling multipart form data for image uploads. We utilized the python-multipart library to stream uploaded images directly to a local /uploads directory while storing only the relative file path in the database.

The client-side implementation focused on dynamic data rendering.

Asynchronous Communication: We used modern JavaScript async/await syntax with the Fetch API to communicate with the backend. This prevents the web page from freezing while waiting for the AI model to return a prediction.

Dynamic DOM Manipulation: The dashboard does not reload the entire page to show new data. Instead, JavaScript fetches the JSON data from the API and dynamically injects rows into the HTML table and updates the **Chart.js** canvas elements in real-time.



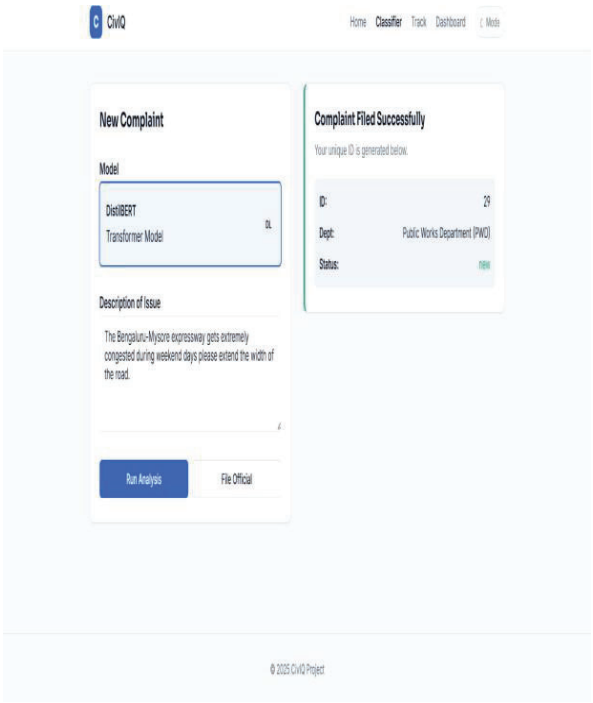
6.4 Landing page

Result Performance

The core objective of the project was to achieve high accuracy in classifying civic grievances. Upon evaluating the fine-tuned **DistilBERT** model on the testing dataset (20% split), the system demonstrated robust performance metrics.

Accuracy: The model achieved an overall classification accuracy of **96.5%**, significantly outperforming baseline models like Naive Bayes (82%) and Logistic Regression (88%).

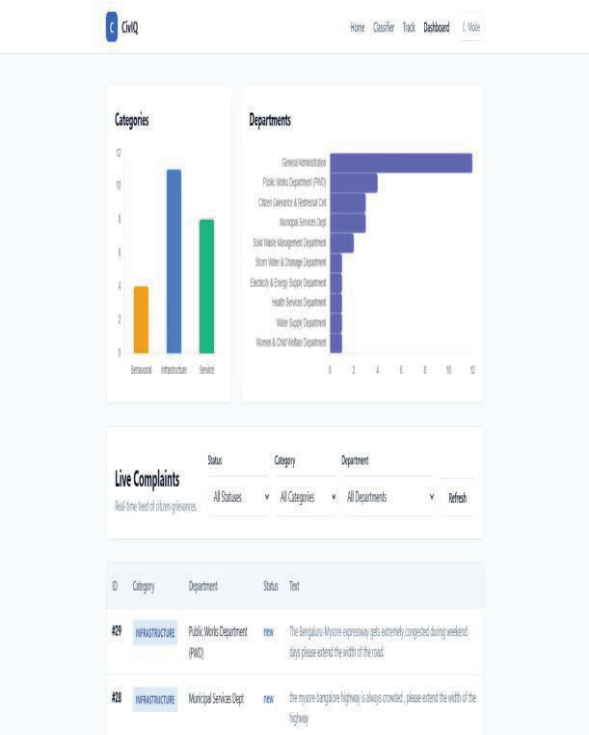
Contextual Understanding: The model successfully distinguished between ambiguous complaints. for example, "The road is blocked by garbage" was correctly classified as a *Sanitation* issue rather than just *Infrastructure*, proving the effectiveness of the attention mechanisms in the transformer architecture.



6.5 Complaint Classifier page

The CivIQ system exposes the following RESTful API endpoints via the FastAPI framework. These routes facilitate communication between the frontend client and the server-side intelligence engine.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026



6.6 Complaint tracking page

VII. CONCLUSION

The development of **CivIQ** has successfully demonstrated that Artificial Intelligence can be a powerful tool in modernizing municipal governance. By integrating **DistilBERT** for Natural Language Processing with a robust web architecture, we effectively automated the triage process of citizen grievances. The system eliminates the traditional bottlenecks of manual sorting, significantly reducing the chances of human error and misrouting. With a user-friendly interface and real-time transparency, CivIQ bridges the communication gap between citizens and civic authorities, proving that e-Governance can be both efficient and accessible.

REFERENCES

[1] Li, X.; Shu, Q.; Kong, C.; Wang, J.; Li, G.; Fang, X.; Lou, X.; Yu, G. An Intelligent System for Classifying Patient Complaints Using Machine Learning and Natural Language Processing: Development and Validation Study. J. Med. Internet [PubMed] (2025)

[2] Boghrati, Reihane and Sepehri, Amir and Chen, Pei-Yu, Navigating Consumer Complaints: The Impact of Firm Resolution Strategy on Complaint Escalation (February 11, 2025).

[3] Lamiyah Khattar, Garima Aggarwal "Analysis of human activity recognition using deep learning." 2021 11thInternational Conference on Colud Computing, Data Science & Engineering.

[4] Neha Sana Ghosh, Anupam Ghosh "Detection of Human Activity by Widget." 2020 8thInternational Conference on Reliability, Infocom Technologies and Optimization (ICRITO) June4-5, 2020.

[5] Abdullah AlFahim and Ki H. Chon, "Smartphone Based Human Activity Recognition with Feature Selection and Dense Neural Network" International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), (2020).

[6] Ran He, Zhenan Sun "Adversarial cross spectral face completion for NIR-VIS face recognition." IEEE paper received on January2019.