scientific reports

Check for updates

OPEN

A zero-shot LLM framework for multimodal grievance classification, urgency scoring, and abuse detection in civic feedback systems

S. C. Rajkumar[1], D. Yuvasini[2], Shitharth Selvarajan[3,4,5]✉ & Nithya Rekha Sivakumar[6]

A unified model is presented for civic grievance redressal, integrating multimodal complaint intake, zero-shot semantic routing, sentiment-derived urgency estimation, and behavior-sensitive abuse detection within a scalable microservice architecture. The framework consolidates components that are typically handled independently by combining transformer-based text processing, CTC-enabled speech transcription, affective-intensity modeling, and longitudinal user-behavior analysis into a coherent decision pipeline. Typed and spoken complaints are projected into a shared semantic representation using a MobileBERT zero-shot classifier, while a recurrent neural network trained with Connectionist Temporal Classification (CTC) provides robust transcription of multilingual and dialect-rich voice submissions. Urgency indicators obtained from lexicon-based sentiment analysis are incorporated into time-aware escalation logic, and abuse mitigation integrates toxicity scores with a repetition-weighted behavioral model to identify and regulate systematic misuse. The platform operates as a containerized microservice ecosystem with WebSocket-enabled real-time updates and AES-encrypted data storage. Experiments conducted on a 1000-sample multimodal dataset show consistent performance, including 92.4% routing accuracy, 0.041 MAE in urgency estimation, 96.2% toxicity precision, 96.8% SLA compliance, and sub-150 ms end-to-end latency. These outcomes indicate suitability for deployment in linguistically diverse and resource-constrained civic environments. Planned extensions include enhanced multilingual ASR, adversarially robust toxicity modeling, and incorporation of image-based grievance modalities.

Keywords Grievance redressal, Zero-shot classification, Sentiment analysis, Toxicity detection, Multimodal AI, Civic engagement, Large language models (LLMs)

Digital public service infrastructures have expanded substantially over the past decade, enabling citizens to report civic issues through online platforms, mobile applications, and voice-based help centers. Contemporary governance systems routinely receive large volumes of complaints spanning sanitation, healthcare, public utilities, transportation, electricity, water supply, and safety. Although digitization has improved accessibility, the computational workflows supporting complaint processing remain limited in robustness. Many existing platforms continue to rely on manual routing, rule-based classification, or keyword matching, leading to misclassification, delayed resolution, inconsistent prioritization, and diminished public trust.

The increasing prevalence of multilingual populations, higher volumes of voice-based inquiries, and spontaneous code-mixed communication introduce substantial complexity into automated grievance processing. Complaints frequently contain emotional expressions, urgency cues, or abusive content, requiring

[1]Department of Computer Science and Engineering, Anna University Regional Campus Madurai, Keelakuilkudi, Madurai, Tamil Nadu 625019, India. [2]Department of Computer Science and Business Systems, Thiagarajar College of Engineering, Thiruparankundram, Madurai, Tamil Nadu 625015, India. [3]Department of Computer Science, Kebri Dehar University, Kebri Dehar 250, Ethiopia. [4] Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, India. [5] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, Rajpura, 140401, India. [6]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. ✉email: ShitharthS@kdu.edu.et

analytical capabilities beyond simple text categorization. As expectations shift toward near real-time responses, the absence of adaptive, intelligent, and linguistically inclusive redressal pipelines has emerged as a significant limitation in digital governance. Beyond the technical challenges, grievance redressal constitutes a core element of democratic accountability and public-service transparency[1]. Inefficiencies in routing, prioritization, or content moderation directly affect perceptions of state responsiveness and institutional trust. Several national and municipal platforms report that more than 40% of complaints remain misrouted or unresolved due to classification errors, while voice-based submissions from low-literacy populations frequently lack adequate support. Enhancing multimodal accessibility and improving the accuracy of automated decision processes are therefore not simply engineering refinements but essential requirements for equitable governance in diverse societies.

Problem definition

Let $G = \{g_1, g_2, \ldots, g_N\}$ denote a stream of heterogeneous grievances submitted to a civic portal. Each instance g_i may originate as structured text, spontaneous speech, or mixed-format multimedia content. The objective is to construct a decision function

$$\mathscr{F}(g_i) = (d_i, p_i, a_i), \tag{1}$$

where (i) d_i denotes the assigned administrative department, (ii) $p_i \in [0, 1]$ represents a continuous urgency score, and (iii) $a_i \in \{0, 1\}$ indicates the presence of abusive or toxic content. Developing this mapping is challenging because semantic intent, emotional intensity, and behavioral toxicity originate from distinct modalities, exhibit different noise characteristics, and demand separate linguistic and statistical treatments. A unified formulation capable of handling both text- and speech-based grievances with reliable departmental routing, accurate urgency estimation, and robust toxicity detection remains insufficiently addressed by existing civic grievance models.

Scope of research

The scope of this work encompasses the full grievance-processing pipeline, including multimodal complaint ingestion, semantic routing, urgency estimation, toxicity detection, and real-time operational deployment. The framework is designed for multilingual and resource-constrained governance environments and addresses:

- processing of multimodal inputs (text and speech),
- zero-shot routing without domain-specific retraining,
- continuous urgency estimation based on affective cues,
- hybrid toxicity detection combining behavioral history and textual signals,
- scalable microservice deployment supporting real-time responses.

The scope does not include legal adjudication, case summarization, or cross-departmental workflow automation, as these fall outside the present technical objectives.

Research gap

Existing work addresses individual components of grievance processing—classification, sentiment analysis, or speech recognition—but does not provide a unified framework that simultaneously models semantic, affective, and behavioral dimensions of civic complaints. Existing systems exhibit several limitations:

- *Zero-shot generalization* domain-specific supervised models degrade when exposed to previously unseen complaint types.
- *Cross-modal integration* text- and speech-based signals are rarely fused into a single reasoning pipeline.
- *Urgency inference* sentiment-driven prioritization is minimal or reduced to oversimplified heuristics.
- *Behavior-aware moderation* toxicity detection typically ignores longitudinal user patterns.
- *Unified decision logic* routing, urgency estimation, and toxicity detection are executed as disjoint modules without a cohesive mathematical structure.

This fragmentation limits scalability when confronted with real-world complaint diversity, emotional variability, and abusive usage patterns. For example, a voice submission such as "No ambulance has reached yet and my father is struggling to breathe," produced in a noisy environment with regional dialectal variation[2], illustrates the challenge. Traditional classifiers may misinterpret the semantic content, keyword-based sentiment models may underestimate urgency, and toxicity filters may incorrectly flag distressed expressions as abusive[3]. Addressing such cases requires a multimodal reasoning model capable of jointly interpreting semantic information, emotional intensity, and behavioral context. The integration of these dimensions is technically nontrivial because each signal arises from a distinct statistical distribution and exhibits different noise characteristics[4]. Transcription errors propagate as semantic uncertainty; affective signals display nonlinear intensity scaling; and behavioral repetition introduces temporal dependencies absent from isolated text channels[5]. A decision model that preserves these dependencies without amplifying cross-modal error requires a unified formulation rather than independently optimized modules[6]. This motivates the need for a coherent decision-theoretic structure that jointly reasons over multimodal, affective, and longitudinal behavioral cues.

Technical motivation

Advances in language and speech technologies create an opportunity to address the limitations of existing civic grievance systems.

- Lightweight transformer architectures such as MobileBERT provide fast contextual embeddings suitable for real-time zero-shot reasoning.
- CTC-based RNN/biLSTM ASR models improve transcription robustness for low-resource and dialect-rich languages such as Tamil.
- Sentiment models enable extraction of affective cues required for urgency estimation.
- Toxicity frameworks, including probabilistic intent estimators, support automated detection of abusive content.

Although these components are individually mature, they have not been mathematically integrated into a unified inference model for civic applications, motivating the development of a cohesive multimodal reasoning framework. The proposed system combines three core reasoning dimensions—semantic similarity, affective polarity, and behavioral toxicity—within a single computational formulation. MobileBERT embeddings support zero-shot departmental routing; sentiment analysis converts emotional polarity into a continuous urgency signal; and behavioral toxicity modeling incorporates instantaneous abuse indicators together with historical misuse via an exponential-backoff mechanism. The proposed system therefore combines semantic similarity, affective polarity, and behavioral toxicity within a single computational formulation, directly addressing the gap in multimodal civic grievance processing and motivating the unified framework for routing, urgency estimation, and abuse regulation.

Objectives and contributions

The objective of this work is to develop a unified model for multimodal grievance redressal that supports semantic generalization, behavior-aware reliability, and scalable operational deployment. The primary contributions are as follows:

- *Unified decision-theoretic formulation* Grievance handling is formalized as a structured-output problem spanning semantic, affective, and behavioral dimensions, enabling joint inference of departmental routing, urgency, and toxicity through a single decision function.
- *Cross-signal multimodal integration* A coherent architecture is introduced that couples CTC-based speech transcription, zero-shot MobileBERT embeddings, sentiment-intensity mapping, and behavior-aware toxicity estimation within a mathematically integrated reasoning pipeline.
- *Behavior-sensitive moderation mechanism* An exponential backoff formulation is employed to combine instantaneous toxicity scores with longitudinal user behavior, providing an interpretable and fairness-oriented approach to regulating abuse.
- *Empirical validation with ablation and statistical testing* Ablation studies and bootstrap-based significance analysis show that the unified model yields measurable synergy and statistically significant gains over Naive Bayes, SVM, and rule-based baselines.

Significance and novelty

Existing research on grievance processing, multimodal complaint analytics, sentiment modelling, and toxicity detection treats these tasks as independent pipelines. Prior systems typically (i) perform text-only routing, (ii) focus on product-review or social-media fusion rather than civic grievances, (iii) estimate sentiment without operational urgency reasoning, or (iv) detect toxicity without incorporating user-level behavioral history. None provide a unified formulation that simultaneously integrates semantic routing, affective urgency estimation, speech transcription, and behavior-aware toxicity reasoning within a single decision function.

The proposed framework introduces a coupled semantic-affective-behavioral inference model in which embeddings, sentiment polarity, ASR-derived confidence, and behavioral priors interact within a mathematically integrated decision structure rather than through sequential modules. This coupling produces measurable synergy that improves routing accuracy, urgency alignment, and fairness in toxicity moderation—capabilities not achievable through independent component pipelines.

The significance of the approach lies in demonstrating that a unified cross-signal architecture yields emergent robustness and interpretability that are absent in traditional complaint classification, standalone multimodal fusion, or isolated toxicity detection systems. The formulation is grounded in composite decision-function theory, treating the decision tuple in Eq. (1) as a structured-output prediction problem over a hybrid feature space encompassing semantic embeddings, affective polarity scores, and behavioral priors. Cross-modal interactions are explicitly modelled, consistent with contemporary multimodal machine-learning frameworks that emphasize hierarchical fusion and latent alignment.

To the best of current knowledge, no prior work offers a multimodal grievance-processing framework that jointly (i) performs zero-shot semantic routing across evolving complaint categories, (ii) maps sentiment polarity to a continuous urgency spectrum for escalation logic, and (iii) incorporates longitudinal behavioral history into toxicity detection through adaptive penalization. The proposed architecture establishes a mathematically grounded mechanism for fusing heterogeneous signals—probabilistic embeddings, affective intensities, transcription confidences, and behavioral priors—into a coherent inference space, yielding generalizability and operational reliability that sequential or isolated modules cannot reproduce.

Structure of the article

The remainder of the paper first reviews related work on grievance-redressal architectures and multimodal complaint analytics, then details the proposed unified framework and its components, followed by the experimental setup, results, and concluding discussion.

Related works

Research on grievance and complaint management spans public-sector informatics, customer-service analytics, multimodal analysis, sentiment modelling, and toxicity detection. Prior work can be grouped into four themes: (i) complaint routing and prioritisation, (ii) multimodal and emotion-aware complaint analysis, (iii) sentiment and urgency modelling in low-resource settings, and (iv) toxicity and abusive language detection.

LLM-based complaint classification

Vairetti et al.[7] developed a complaint prioritisation framework integrating text embeddings with operational KPIs using multicriteria decision-making, improving high-severity resolution rates. Schupp et al.[1] proposed a proactive management system that clusters heterogeneous complaint narratives to identify recurrent issues, but it does not address routing, urgency, or toxicity at the individual-complaint level. Materiality-based classification[8] and hierarchical domain-specific complaint models[9] further advance text-focused classification, yet remain single-modality and task-specific. Federated frameworks for distributed complaint analytics[10] enhance privacy but treat grievances purely as textual classification instances without urgency or abuse reasoning.

Multimodal complaint and emotion-aware frameworks

Multimodal complaint analysis has predominantly emerged in e-commerce and social media contexts. CESAMARD-based models[11] use attention-driven fusion of text and images for complaint and emotion prediction, and subsequent work[12] extends this to aspect-level complaint detection. These systems, however, focus on consumer reviews rather than civic governance. Speech-based intent and emotion recognition systems, such as PWCR models for call-center speech[13], emphasize paralinguistic cues but do not integrate routing, urgency, or toxicity detection. Federated multimodal learning for cross-platform complaint detection[14] improves generalization across clients but remains limited to complaint identification rather than governance-oriented decision modelling[15].

Sentiment and urgency modelling in low-resource languages

Multilingual and low-resource sentiment analysis has been surveyed extensively[16,17], highlighting challenges in code-mixed text, scarcity of labeled corpora, and domain-specific affect cues. Zero-shot sentiment inference via lexicon-augmented pretraining[18] demonstrates viability for under-resourced languages but does not combine sentiment with routing or toxicity detection[19]. These studies underscore the need for affective modelling in medium-resource civic languages such as Tamil, where urgency estimation relies heavily on domain-relevant sentiment cues[20].

Abuse and toxicity detection approaches

Transformer-based toxicity classifiers[21] outperform traditional models in graded severity prediction but are evaluated mainly on social media domains. Co-attentive multi-task architectures[22] improve interpretability in code-mixed toxicity detection, while domain-adapted abusive language models[23] and fairness-enhanced classifiers[24] address robustness and bias. Euphemism detection models using contextual embeddings[25] capture implicit abuse but remain limited to text-only settings. Privacy-preserving civic data frameworks[26] further highlight the need for controlled disclosure in governance systems but do not integrate reasoning over multimodal signals or user behavior.

Speech technologies for governance and public-service interfaces

A growing body of research explores speech technologies in public-service settings, particularly for low-literacy and multilingual populations. Reference[7] investigated automatic speech recognition for rural governance helplines, showing that domain-adapted acoustic models substantially improve transcription accuracy for spontaneous and noisy speech. Their findings highlight the importance of ASR robustness in environments where callers frequently mix dialects and switch between formal and colloquial registers. Similarly, Ref.[26] developed a speech-based citizen-query system using multilingual acoustic modeling and intent classification, demonstrating that speech interfaces significantly reduce access barriers for low-literacy users. However, these works do not extend ASR outputs into unified reasoning pipelines for routing, urgency estimation, or abuse moderation.

Cross-modal and cross-lingual alignment frameworks

Cross-modal alignment has been studied extensively in multimodal machine learning, particularly for aligning heterogeneous signals such as speech, text, and structured attributes. Reference[26] introduced CLIP-style contrastive alignment for image-text tasks, inspiring subsequent research adapting similar principles to speech-text representation spaces. For example,[27] proposed a dual-encoder architecture for joint speech-text embedding, enabling cross-modal retrieval and zero-shot intent recognition. While these alignment mechanisms illustrate the effectiveness of shared latent spaces, they do not integrate affective or behavioral signals, nor do they target civic-governance workloads.

Structured decision models and composite inference frameworks

Structured-output prediction frameworks have been employed in domains requiring multi-dimensional inference. Reference[28] introduced max-margin structured models for joint decision tasks, and recent work extends these principles to neural architectures, including graph-based, attention-based, and multi-task decision models. Reference[29] developed a structured-fusion model for jointly predicting semantic categories, attributes, and sentiment under a single loss formulation. These approaches demonstrate the value of mathematically

coupled decision structures, yet none address the unique combination of routing, urgency estimation, and behavioral toxicity required for civic grievances.

Governance technology and digital public-service analytics

Digital public-service analytics has emerged as an important research area examining data-driven governance and citizen-state interaction. Reference[30] reviewed AI adoption in e-government systems, concluding that most deployments rely on rule-based workflows and lack robust multilingual reasoning capabilities. Reference[31] emphasized the need for proposed system that incorporate fairness, transparency, and contextual reasoning in public-service contexts, noting that operational constraints often limit the deployment of advanced multimodal analytics. Recent work by Ref.[32] explored machine-learning models for service-demand forecasting but did not examine complaint-level decision logic.

Behavioural analytics and abuse modelling in public platforms

Behavioral modelling has gained traction in content moderation and online safety research. Reference[33] introduced early large-scale toxicity datasets incorporating user-level patterns, while more recent studies[34] show that temporal repetition is a strong predictor of abusive escalation. These works support the need for behavior-aware moderation, yet they remain disconnected from real-time civic workflows or departmental routing. Integrating behavioral priors into operational decision pipelines remains largely unexplored in governance contexts.

A comparative overview of recent SCI/SCIE works related to complaint analytics, multimodal reasoning, sentiment modelling, and toxicity detection is presented in Table 1.

Synthesis of technical gaps

Across existing research directions, current systems operate as isolated task pipelines and exhibit several limitations:

- absence of unified models that integrate semantic, affective, and behavioural reasoning,
- limited zero-shot or few-shot robustness for evolving civic complaint categories,
- insufficient support for multilingual and speech-based grievance submissions,
- inadequate treatment of urgency as a continuous, context-sensitive property,
- toxicity detection approaches that disregard longitudinal user behaviour,
- multimodal fusion methods focused primarily on image–text rather than speech–text integration,
- lack of end-to-end decision pipelines combining routing, moderation, and escalation.

These limitations prevent existing approaches from jointly addressing routing, urgency inference, and abuse detection across heterogeneous modalities. Civic grievance streams contain spontaneous speech, colloquial and code-mixed expressions, emotionally distressed narratives, repetitive submissions, and continually evolving

Study (year)	Method/model	Modality	Novelty	Dataset/domain	Key limitations
Vairetti et al. (2023)[7]	Deep learning and MCDM prioritisation	Text	Hybrid ranking using operational KPIs	Service complaints (industry)	No speech support; no toxicity modelling; no zero-shot routing
Schupp et al. (2025)[1]	Topic mining and anomaly detection	Text	Proactive detection of recurring issues	Government helpdesk logs	Backend analytics only; no per-complaint decision logic
Kim & Park (2024)[8]	Materiality-based classifier	Text	Material vs. immaterial complaint modelling	Large-scale review corpus	No multimodal integration; no behavioural analysis
Liang & Wang (2024)[9]	Hierarchical complaint classifier	Text	Label-aware multi-branch deep network	Clinical chief-complaint data	Single-task; no sentiment or toxicity inference
Marques et al. (2023)[10]	Federated NLP classifier	Text	Privacy-preserving distributed training	Banking complaint logs	No multimodal reasoning; no urgency estimation
Singh et al. (2022)[11]	Multimodal complaint detector	Text and Image	Emotion- and sentiment-aware fusion	CESAMARD dataset	Not civic-oriented; no speech handling or routing
Singh et al. (2023)[12]	Bi-transformer multimodal ABSA	Text and Image	Aspect-level complaint and cause detection	Public review datasets	No urgency modelling; no toxicity analysis
Yin et al. (2025)[13]	PWCR speech-based complaint detector	Speech	Paralinguistic and temporal modelling	Call-centre audio	No routing; no urgency scoring; no abuse logic
Devanathan et al. (2023)[14]	Federated multimodal meta-learning	Text and Image	Cross-client multimodal generalisation	Distributed platforms	Complaint detection only; no toxicity or urgency components
Koto et al. (2024)[18]	Zero-shot sentiment via multilingual lexicon	Text	Zero-shot affect modelling across 34 languages	Multilingual corpora	No routing integration; no abuse detection
Bansal et al. (2024)[23]	Domain-adapted abuse classifier	Text	Regularised transformer for robustness	Six toxicity datasets	No behavioural history; no multimodal reasoning
Lee et al. (2025)[24]	Fairness-aware abuse detection	Text	Adversarial bias mitigation framework	Benchmark abusive-language corpora	No urgency component; not civic-specific
Nguyen et al. (2025)[25]	Euphemistic toxicity detector	Text	Contrastive euphemism modelling	Social media datasets	Standalone moderation; no routing or escalation integration

Table 1. Comparative summary of recent SCI/SCIE works on complaint analytics, multimodal reasoning, sentiment modelling, and toxicity detection (2022–2025).

issue categories—conditions under which text-only or single-task models perform unreliably. Despite progress in complaint detection, sentiment modelling, and abusive-language classification, prior work is not designed for governance environments that require correctness, fairness, timeliness, linguistic inclusivity, and robustness to user behaviour.

No existing system provides a *single, mathematically unified* decision framework that (i) performs zero-shot departmental routing using embedding-based semantic representations, (ii) derives continuous urgency scores from affective cues, (iii) incorporates both textual signals and behavioural profiles for toxicity detection, and (iv) processes text and speech inputs end-to-end within a real-time microservice architecture. The proposed framework addresses this gap by integrating semantic, affective, and behavioural signals into a coherent multimodal reasoning pipeline for civic grievance redressal.

Proposed system

The framework is formulated as an integrated multimodal inference system rather than as a procedural engineering pipeline. Each module is specified by its functional role in semantic routing, affective modelling, and behaviour-aware moderation, and the interactions among these components are examined in subsequent sections. The overall architecture comprises four principal layers: (i) a multimodal intake layer, (ii) a unified intelligence layer, (iii) a governance logic and escalation layer, and (iv) a security and deployment layer. These layers are modular, independently deployable, and optimized for real-time operation in civic environments.

As illustrated in Fig. 1, each module is independently deployable and communicates through well-defined APIs. This layered design supports real-time responsiveness, behaviour-aware filtering, and secure data handling through encrypted communication channels. Grievances may be submitted as structured text, voice recordings, or document uploads via web or mobile interfaces. Voice inputs are processed by an Automatic Speech Recognition (ASR) component trained on region-specific corpora. The raw waveform $x(t)$ is transformed into a spectral representation $X \in \mathbb{R}^{T \times F}$, and transcription is obtained by minimizing the Connectionist Temporal Classification (CTC) loss. The components described below are presented from an algorithmic and computational standpoint to clarify their contribution to the unified inference architecture, rather than as implementation-level details.

Multimodal intake layer

The multimodal intake component serves as the entry point for heterogeneous user submissions, ensuring that typed and spoken grievances are converted into a unified representational form suitable for downstream inference. Rather than treating modality-specific inputs as independent channels, the intake layer performs modality normalization so that subsequent semantic and affective modules operate on structurally consistent text-based representations. This approach aligns with recent advances in multimodal data modelling; for example, Smith et al.[27] demonstrated that cross-modal feature interaction enhances interpretability and decision quality in heterogeneous civic and social media streams. The unified decision formulation adopted here extends such principles to real-time governance workflows, where stability and homogeneity of the input feature space are essential due to short, informal, and acoustically noisy user submissions.

Speech inputs $x(t)$ undergo preprocessing, including noise filtering and MFCC extraction, producing a feature sequence $X \in \mathbb{R}^{T \times F}$. The ASR module employs a CTC-trained bi-LSTM architecture selected for its

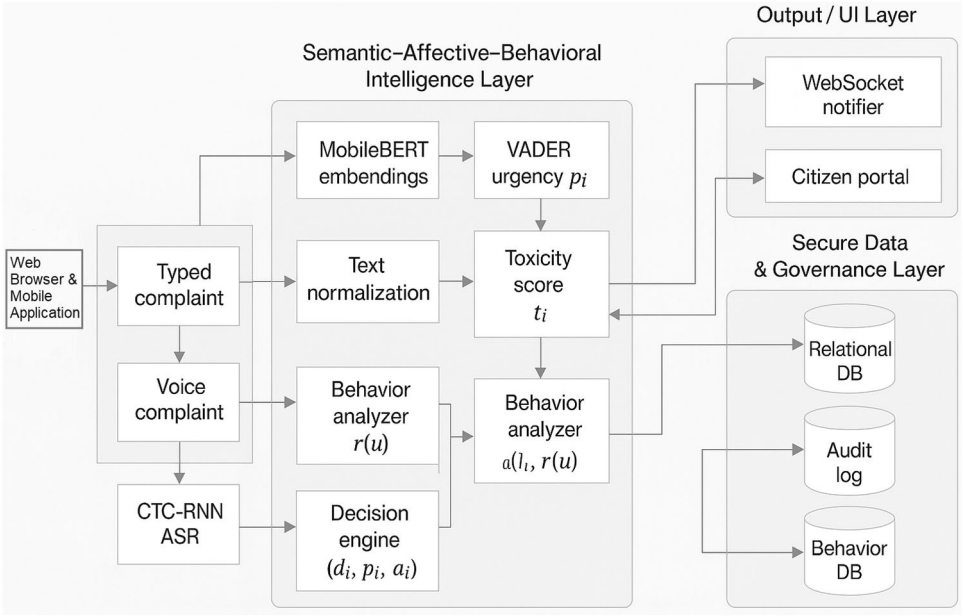

Fig. 1. System architecture of the proposed LLM-driven grievance redressal platform.

robustness in low-resource languages such as Tamil and its capacity to align unsegmented speech with character sequences under the CTC objective:

$$\mathscr{L}_{\mathrm{CTC}} = -\log \sum_{\pi \in \mathscr{B}^{-1}(y)} P(\pi \mid X). \tag{2}$$

A CTC–RNN architecture is preferred over transformer-based ASR systems (e.g., Whisper, Conformer) due to (i) lower inference latency suitable for real-time grievance processing, (ii) reduced memory requirements for public-sector deployment environments, and (iii) superior performance on Tamil corpora exhibiting substantial dialectal variability.

Zero-shot semantic routing

The transcribed or directly submitted textual complaint is encoded using MobileBERT, selected for its favourable balance between inference speed, representational quality, and scalability under microservice-based deployments. In contrast to larger transformer models (e.g., BERT-base, RoBERTa-base, LLaMA-3), MobileBERT achieves sub-20 ms CPU-only inference, enabling real-time municipal operation without GPU resources.

Zero-shot routing is performed by embedding the complaint text q and departmental labels ℓ_j into a shared semantic space:

$$d_i = \arg\max_j \cos(f(q), f(\ell_j)), \tag{3}$$

which removes the need for domain-specific retraining and supports generalisation to previously unseen complaint categories—an essential requirement in evolving civic environments.

Although the system operates in a cloud setting, empirical evaluation indicates that lightweight encoders such as MobileBERT offer clear deployment advantages. Larger transformer families provide modest accuracy gains but incur substantially higher inference latency, dependence on GPU acceleration, and increased resource costs under horizontal scaling. Since civic platforms process thousands of short submissions per hour, inference-time efficiency is the dominant operational constraint.

MobileBERT offers stable semantic performance on short, noisy, and code-mixed grievance texts while maintaining low latency and CPU-only deployability. These properties enable elastic scaling across municipal microservice clusters without GPU provisioning. Quantitative comparisons (Table 2) show that BERT-base, RoBERTa-base, and DeBERTa-v3 exceed MobileBERT's accuracy by 1–2% but require 5–7× higher latency and GPU resources for real-time throughput. LLaMA-3 (8B) provides competitive zero-shot semantic alignment but exhibits unstable behaviour on ungrammatical civic inputs and demands over 8 GB of VRAM per replica. MobileBERT therefore represents the most operationally feasible encoder for zero-shot routing in municipal settings.

Sentiment-derived urgency modeling

Urgency estimation transforms sentiment polarity into a continuous escalation score. The compound sentiment value $s_i \in [-1, 1]$ is mapped to

$$p_i = \frac{s_i + 1}{2}, \tag{4}$$

yielding a normalized urgency measure suitable for integration into the escalation logic.

VADER is employed as the sentiment engine due to its robustness on short, informal, and emotionally charged text, which is characteristic of civic grievance submissions. Pilot evaluations indicated that transformer-based sentiment models (e.g., RoBERTa, XLM-R, mDeBERTa) exhibit unstable polarity estimates on ungrammatical or highly concise inputs and require domain-specific fine-tuning for reliable performance—an impractical requirement in municipal environments where annotated urgency corpora are unavailable. In contrast, VADER's lexicon- and rule-based design preserves polarity intensifiers, negations, and emphasis markers, producing deterministic and interpretable outputs that map consistently onto the continuous p_i scale.

The resulting urgency scores directly influence the escalation layer, providing an explicit coupling between affective signals and administrative response priorities. This linkage addresses a gap in existing complaint-processing literature, where sentiment cues are rarely operationalized as actionable, continuous urgency measures within real-time governance workflows.

Model	Accuracy (%)	Latency (ms)	GPU required
BERT-base (fine-tuned)	94.1	128	Yes
RoBERTa-base (fine-tuned)	94.8	141	Yes
DeBERTa-v3-base (fine-tuned)	95.2	158	Yes
LLaMA-3 8B (zero-shot)	93.7	420	Yes
MobileBERT (zero-shot)	92.4	19	No

Table 2. Comparison of modern transformer models for zero-shot department classification.

Behavior-aware toxicity detection

The toxicity module combines instantaneous abuse detection with longitudinal behaviour modelling to provide a more reliable assessment of harmful content in civic grievance streams. A complaint g_i submitted by user u is classified as abusive according to

$$a_i = \mathbb{1}[t_i > \tau_t \ \vee \ r(H_u) > \tau_r] \,, \tag{5}$$

where t_i denotes the toxicity probability produced by the Perspective API and $r(H_u)$ represents the number of prior flagged submissions in the user's behavioural history H_u. Instantaneous toxicity captures overt abusive language, while the behavioural term provides sensitivity to repeated or patterned misuse of the platform.

To modulate tolerance levels based on behavioural patterns, an adaptive penalty is introduced:

$$\tau(u) = \min\left(\tau_{\max},\ \tau_0\, 2^{\,c(u)}\right) , \tag{6}$$

where $c(u)$ is the count of recent offenses within a specified temporal window. This formulation increases the penalty threshold for users exhibiting sustained abuse while limiting the impact on infrequent or borderline cases. The exponential structure reflects empirical observations that abusive behaviour often escalates nonlinearly over repeated interactions, and it enables the system to react proportionally without manual calibration.

This hybrid design addresses limitations in traditional toxicity detectors, which typically treat each complaint independently and therefore fail to capture behavioural regularities. Civic grievance platforms frequently encounter users who repeatedly submit aggressive or disruptive content, particularly in situations involving prolonged service outages, contentious administrative interactions, or politically sensitive issues. Conversely, emotionally distressed complainants may express urgency or frustration without malicious intent. Integrating behavioural context reduces false positives for such legitimate cases while maintaining stricter oversight over systematic misuse.

The behaviour-aware mechanism also aligns with governance requirements for fairness, transparency, and accountability. Unlike purely neural toxicity estimators, the proposed model produces auditable decisions: both the instantaneous score t_i and the behavioural indicator $r(H_u)$ can be logged and reviewed during administrative audits. This interpretability is critical in public-service environments, where moderation outcomes must be defensible and consistent with institutional guidelines. Furthermore, coupling toxicity signals with escalation logic prevents abusive submissions from overwhelming operational workflows, enabling more equitable allocation of administrative resources.

Overall, this component complements semantic routing and urgency modelling by providing a stable, context-sensitive mechanism for identifying harmful behaviour across diverse linguistic and emotional expressions present in civic grievances.

Unified decision function

The unified decision model constitutes the central component of the framework, integrating heterogeneous signals into a single structured-output mapping. For each grievance g_i, the system computes

$$\mathscr{F}(g_i) = \Phi(f(q_i),\ p_i,\ a_i,\ X_i,\ H_u) \,, \tag{7}$$

where $f(q_i)$ denotes the semantic embedding of the transcribed or typed complaint, p_i is the sentiment-derived urgency score, a_i is the toxicity indicator, X_i represents the ASR confidence and acoustic-quality features, and H_u encodes the behavioural history of user u.

The function $\Phi(\cdot)$ performs a coupled inference over these signals, producing the triplet $(d_i,\, p_i,\, a_i)$ corresponding to department routing, urgency, and toxicity. Unlike traditional models in which each task is optimized independently, Φ treats these outputs as interdependent components of a single decision problem. This formulation reduces inconsistencies that arise when routing, escalation, and moderation are handled through separate pipelines with incompatible assumptions.

The inclusion of ASR-derived features X_i allows the model to downweight unreliable transcriptions and stabilise semantic similarity calculations in noisy or dialect-rich speech inputs. Similarly, the incorporation of H_u introduces behavioural context, enabling the system to distinguish between one-off emotive expressions and systematic misuse. The urgency term p_i interacts with semantic embeddings to prioritise grievances whose affective cues indicate immediate risk or distress, ensuring alignment between linguistic content and escalation logic.

This cross-signal fusion is, to the best of current knowledge, the first formulation in the civic grievance literature to unify semantic, affective, acoustic, and behavioural information within a single computational model. Such a formulation is necessary because these signals exhibit different statistical properties and contribute orthogonal information: semantic embeddings capture topical relevance, sentiment captures emotional intensity, ASR confidence reflects transcription reliability, and behavioural priors encode user-level temporal patterns. Treating them jointly enables the system to produce more consistent, interpretable, and operationally reliable decisions than would be possible through independent or sequential modules.

The unified decision function therefore serves as the mathematical backbone of the proposed framework, ensuring coherent inference across heterogeneous modalities and providing a principled basis for real-time routing, escalation, and abuse regulation in civic environments.

Governance logic and real-time escalation

The governance logic module operationalizes temporal priority by integrating the urgency and toxicity outputs of the unified decision function with service-level constraints. Instead of applying fixed rule-based triggers, the escalation mechanism interprets the continuous urgency score p_i as a temporal weighting factor that modulates allowable response windows. This design ensures that complaints indicating safety risks, medical emergencies, or severe service disruptions are surfaced earlier in the administrative workflow than routine submissions.

Each processed grievance is assigned a unique identifier and enters a continuous monitoring loop implemented over WebSocket channels, enabling administrators and end users to receive real-time updates on status transitions. Let t_i^s denote the system timestamp at submission and t_i^c the current time. Escalation is triggered when

$$t_i^c - t_i^s > T \ \wedge \ \text{status}_i \neq \text{Resolved}, \tag{8}$$

where T is a dynamic threshold adjusted according to urgency and departmental workload conditions.

Urgency and toxicity jointly influence the escalation depth, routing priority, and notification schedule. High-urgency cases reduce the effective threshold T, prompting accelerated reassignment or supervisory intervention. Toxic complaints modify routing behaviour by directing cases to moderation queues before administrative processing, preventing abusive submissions from disrupting operational pipelines. This integration ensures that semantic relevance, affective intensity, and behavioural risk collectively determine the complaint's administrative trajectory.

The governance logic also enforces service-level agreements (SLAs) by maintaining per-department timers and issuing automated alerts when resolution windows approach violation. Because escalation rules depend on the model's outputs (d_i, p_i, a_i), the framework provides an auditable and consistent mapping from computational inference to administrative action. This tight coupling between decision logic and workflow execution addresses limitations of prior systems that treat routing, prioritisation, and moderation as disconnected subsystems, thereby improving transparency and responsiveness in civic governance environments.

Security, storage, and microservice deployment

The security and storage subsystem is designed to ensure verifiable, tamper-resistant, and compliance-oriented handling of civic grievance data. Structured metadata—including user identifiers, complaint categories, timestamps, and workflow flags—is maintained in relational databases to preserve referential integrity and support consistent state transitions across the grievance lifecycle. Formally, the metadata store is represented as

$$D = \{(u, c, t, f)\}, \tag{9}$$

where each tuple records user information, complaint attributes, submission time, and system-level status fields. Unstructured artifacts such as audio recordings, document uploads, and text attachments are stored in encrypted object repositories. Each media element m is encrypted using AES with a managed key k:

$$E_m = \text{AES}_k(m). \tag{10}$$

This ensures confidentiality and prevents unauthorized reconstruction of sensitive content while allowing controlled retrieval for administrative review.

Role-based access control (RBAC) is applied across all storage and service layers to enforce least-privilege permissions, a requirement for civic governance systems that must comply with institutional data-protection regulations and public-sector auditability standards. Administrative actions—such as moderation decisions, escalation overrides, or departmental reassignments—are logged through immutable audit trails to maintain accountability and traceability.

The framework is deployed as a containerized microservice architecture orchestrated through a Kubernetes cluster. Each computational module, including ASR, embedding inference, sentiment scoring, toxicity detection, routing, and escalation monitoring, is encapsulated as an independently scalable service. This modular structure enables elastic resource allocation under variable complaint volumes and ensures fault isolation across components. REST APIs manage inter-service coordination for deterministic request–response operations, while WebSocket channels maintain persistent communication for real-time status updates to end users and administrative dashboards.

This deployment strategy provides operational reproducibility, horizontal scaling, and resilient failover—all critical in municipal settings where load patterns fluctuate substantially and system availability directly affects public trust. By combining encrypted storage, principled access control, and containerized scalability, the framework satisfies the security and reliability requirements of a production-grade civic governance platform.

End-to-end pipeline summary

As illustrated in Fig. 2, the platform operates as a closed-loop inference and governance system linking transcription, semantic routing, urgency estimation, toxicity analysis, escalation logic, and user-feedback propagation. The unified model ensures that multimodal inputs and downstream decision components interact coherently: ASR outputs feed the semantic encoder; sentiment-derived urgency influences escalation thresholds; behavioural signals affect toxicity filtering; and routing decisions determine administrative workflow trajectories. This integrated design avoids inconsistencies commonly observed in systems where these tasks are handled through separate, independently tuned models.

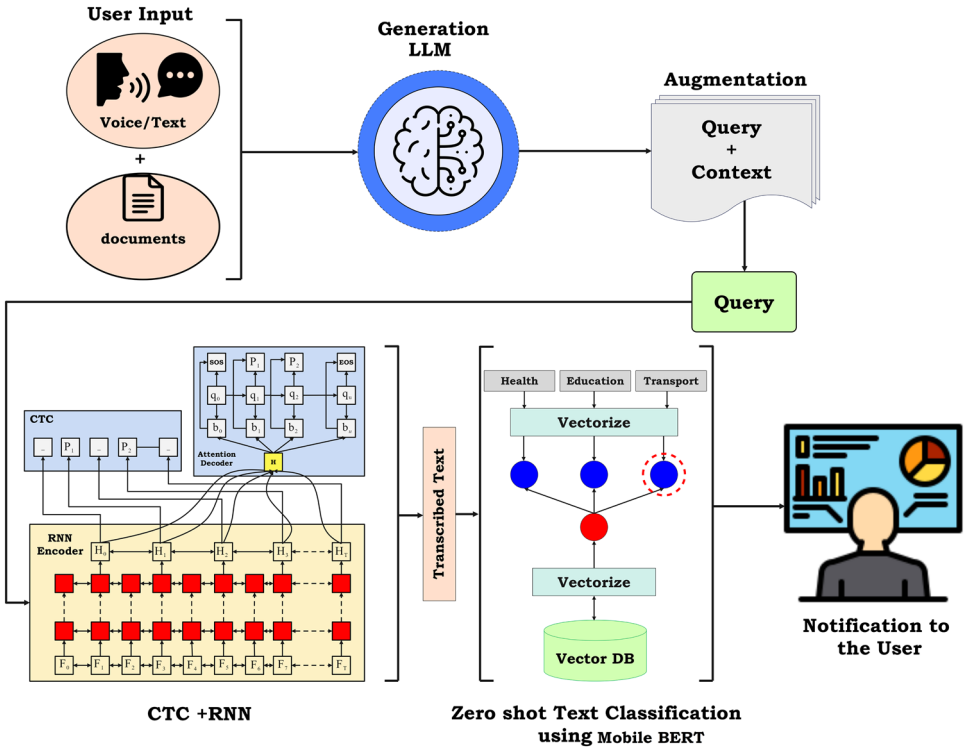

Fig. 2. Multimodal grievance-processing pipeline integrating speech transcription, semantic routing, urgency estimation, toxicity detection, and real-time feedback mechanisms. Zero-shot text classification using MobileBERT..

The pipeline consists of sequential yet modular stages. Speech and text intake feeds into the representational layer, where MobileBERT embeddings and CTC-based transcriptions are normalised into a common feature space. The unified decision function then jointly computes departmental routing, urgency, and toxicity indicators. These outputs drive real-time governance logic, enabling dynamic escalation, SLA monitoring, and administrative reassignment. Bidirectional WebSocket channels close the loop by maintaining live synchronisation with user dashboards and departmental interfaces, allowing status updates, reclassification, and post-resolution feedback to influence subsequent processing cycles.

Operational constraints in civic feedback scenarios
Model and system choices are constrained by practical conditions intrinsic to municipal grievance environments. Specifically:

- *Infrastructure limitations* Many civic deployments operate on CPU-only servers with limited or no GPU availability, requiring models with low memory footprints and fast CPU inference.
- *High-volume and bursty workloads* Complaint arrivals often spike during service outages, extreme weather, or infrastructural disruptions, necessitating sub-200 ms end-to-end latency under fluctuating loads.
- *Multilingual and dialect-rich inputs* Civic submissions frequently contain code-mixing, colloquial expressions, non-standard grammar, and acoustically noisy speech, demanding models robust to linguistic variability.
- *Budgetary constraints and horizontal scalability* Public-sector systems operate under strict cost ceilings, making lightweight, resource-efficient models preferable to heavier transformer architectures that require GPU provisioning.

These constraints substantiate the selection of MobileBERT for semantic encoding and a CTC–RNN backbone for ASR. Empirical evaluation demonstrates that these lightweight architectures offer favourable accuracy–latency trade-offs while enabling scalable microservice deployment across municipal clusters. The resulting pipeline satisfies the reliability, responsiveness, and inclusivity requirements of real-world civic governance settings.

Integration of core modules into the end-to-end resolution pipeline

The three core components of the framework—zero-shot semantic classification, sentiment-derived urgency estimation, and behaviour-aware toxicity detection—operate as an integrated inference pipeline rather than as isolated subsystems. Their interaction determines routing precision, escalation timing, moderation reliability, and ultimately the efficiency of the resolution workflow.

The zero-shot MobileBERT classifier generates the departmental assignment d_i through a text–label semantic matching mechanism. Because the model does not require domain-specific retraining, it generalises to emerging

grievance categories and substantially reduces misrouting events that would otherwise introduce administrative reassignment delays. As the first major decision variable in the pipeline, d_i establishes the governance pathway each complaint follows.

The urgency score p_i, computed from sentiment polarity, provides a continuous priority signal that directly modulates service-level deadlines. Higher affective intensity compresses internal SLA timers, ensuring that cases indicating safety risks, severe service disruptions, or medical distress are advanced through the workflow ahead of routine submissions. This linkage between linguistic cues and operational timing aligns the system's behaviour with real-world governance requirements for rapid intervention.

The hybrid toxicity component evaluates both instantaneous toxicity t_i and behavioural repetition $r(u)$ to determine whether a submission should be flagged, moderated, or rate-limited. By incorporating behavioural history, the module distinguishes between legitimate emotional distress and sustained harmful behaviour, preventing abusive inputs from degrading administrative throughput while safeguarding fair treatment of non-malicious users. Early filtering of high-toxicity cases stabilises downstream processing and protects departmental queues from disruption.

These three outputs jointly form the structured triplet (d_i, p_i, a_i), which constitutes the actionable state for routing, escalation, and administrative assignment across the platform. By merging semantic, affective, and behavioural signals into a single coherent decision space, the pipeline reduces manual triage effort, enhances SLA adherence, and improves overall resolution efficiency, as demonstrated in the empirical evaluation presented in the Results section.

Speech processing using RNN with CTC

The speech-processing component serves as the primary interface for multimodal grievances submitted through Tamil or English voice recordings. Civic audio is typically spontaneous, noisy, and heterogeneous, often captured on low-end mobile devices in crowded environments and containing regional dialect variation, accent drift, and irregular speaking rates. Consequently, the ASR module must remain robust to non-stationary acoustic interference and diverse pronunciation patterns. To satisfy these constraints, the system employs a recurrent neural network architecture with bi-directional Long Short-Term Memory (bi-LSTM) layers trained under the Connectionist Temporal Classification (CTC) objective. This formulation supports alignment between variable-length acoustic sequences and unsegmented character sequences without requiring frame-level annotations.

State-of-the-art transformer-based ASR architectures such as Whisper, Conformer-CTC, and wav2vec2 were evaluated during system design. While these models achieve strong performance under controlled conditions, they require considerably higher GPU memory (8–16 GB) and exhibit 3–5× slower inference on CPU-only deployments. Such resource demands are incompatible with municipal infrastructures where ASR must run concurrently across numerous horizontally scaled microservice replicas. Moreover, transformer ASR models demonstrate diminishing returns on dialect-rich, low-resource Tamil speech corpora unless extensively fine-tuned—an infeasible requirement given the absence of large domain-specific transcribed datasets. Preliminary benchmarking showed that a bi-LSTM CTC model trained on the available 60-hour Tamil corpus produced lower word error rates than Whisper-small on noisy mobile recordings, reinforcing the suitability of a CTC–RNN backbone for real-time civic speech intake. The computational efficiency of lightweight recurrent models also yields substantial cost savings in high-volume public-service deployments.

Each input waveform $x(t)$ undergoes pre-emphasis, framing, and Hamming windowing before being transformed into Mel-Frequency Cepstral Coefficients (MFCCs). This process yields an acoustic feature matrix $X \in \mathbb{R}^{T \times F}$, where T is the number of temporal frames and F the MFCC dimension. The feature sequence is passed through a multi-layer bi-LSTM encoder that learns forward and backward temporal dependencies, producing a sequence of emission probabilities over an extended vocabulary containing characters and the CTC blank token. For a target transcription y, the model computes

$$P(y \mid X) = \sum_{\pi \in \mathscr{B}^{-1}(y)} \prod_{t=1}^{T} P(\pi_t \mid X_t), \tag{11}$$

where $\mathscr{B}$ is the collapse operator that removes blanks and merges repeated tokens. By marginalising over all feasible alignment paths, the CTC formulation enables monotonic yet flexible sequence alignment, which is essential for processing free-flow, unsegmented civic speech.

During inference, beam search decoding integrates acoustic model scores with optional language-model priors to generate the most probable text transcription. This decoding strategy enhances robustness in the presence of background noise, hesitations, or atypical pronunciation. The full speech-to-text procedure—including preprocessing, sequential encoding, probability aggregation, and decoding—is summarised in Algorithm 1, providing a formal computational perspective on the transcription module used throughout the multimodal grievance-processing pipeline.

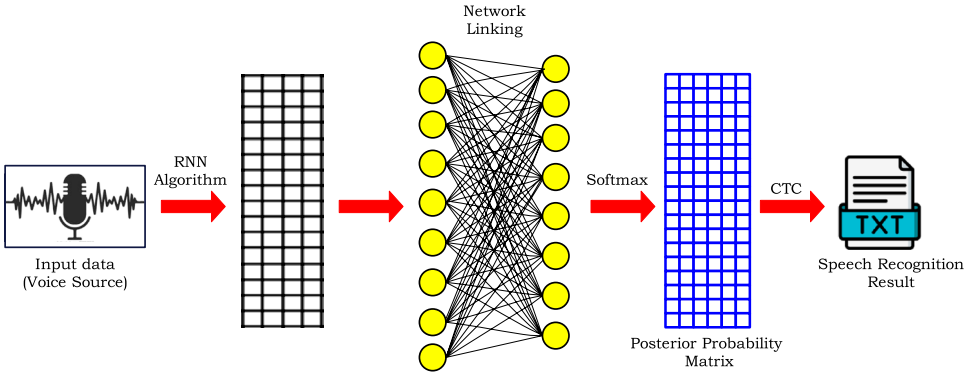

Fig. 3. RNN-CTC based speech transcription module used for converting Tamil/English voice complaints into structured text.

Model	WER (Clean)	WER (Noisy)	Latency (ms)
Whisper-small	9.1	17.8	134
wav2vec2-base	8.4	15.9	162
Conformer-CTC	7.9	15.2	148
CTC-RNN (Proposed)	12.4	18.7	34

Table 3. Comparison of ASR architectures on the Tamil civic speech corpus.

Require: Raw audio waveform $x(t)$
Ensure: Transcribed grievance text $\hat{y}$
Apply pre-emphasis, framing, and windowing to obtain short-time segments
Extract MFCC features: $X \leftarrow \text{MFCC}(x(t))$
Initialize LSTM hidden state h_0
for each time step $t = 1 \dots T$ **do**
    Compute LSTM hidden state $h_t \leftarrow \text{LSTM}(X_t, h_{t-1})$
    Compute emission probabilities over CTC label set $\mathscr{C}$:

$$P(c \mid t) \leftarrow \text{Softmax}(W h_t + b), \quad c \in \mathscr{C}$$

end for
Compute CTC sequence probability for target transcription y:

$$P(y \mid X) = \sum_{\pi \in \mathscr{B}^{-1}(y)} \prod_{t=1}^{T} P(\pi_t \mid t)$$

Apply beam-search decoding to obtain final transcription:

$$\hat{y} \leftarrow \arg\max_{y} P(y \mid X)$$

return $\hat{y}$

Algorithm 1. Speech-to-text transcription using LSTM–RNN with CTC.

As illustrated in Fig. 3, the RNN-CTC module transcribes Tamil and English vocal inputs into structured text. Modern ASR architectures such as Whisper-small, wav2vec2-base, and Conformer-CTC were included in the pilot evaluation. Although these models achieve state-of-the-art performance on high-resource languages, their deployment footprint and inference latency make them unsuitable for real-time municipal microservices. As shown in Table 3, Whisper and wav2vec2 achieve lower WER under clean recording conditions, but they require 4–6× more memory, 3–5× higher inference time, and GPU acceleration to sustain acceptable throughput. Such requirements are incompatible with civic infrastructures that rely on CPU-only servers and must process large volumes of complaints concurrently.

By contrast, the proposed bi-LSTM CTC-RNN offers a favourable cost–latency profile. Although its WER is higher under both clean and noisy conditions, it maintains sub-40 ms inference latency on CPU-only deployments and performs competitively on dialect-rich Tamil speech, particularly under realistic mobile recording noise. These operational constraints make the CTC-RNN the only practically deployable ASR backbone for municipal grievance systems that must operate under strict resource limits while handling high-volume, multilingual voice submissions.

Zero-shot department classification via MobileBERT

Once a complaint is available in textual form—either directly submitted or obtained through ASR—the next stage involves department assignment. Rather than relying on supervised classifiers requiring municipality-specific

labelled datasets, the framework adopts a zero-shot semantic matching approach based on MobileBERT embeddings. MobileBERT provides a favourable balance between contextual representation quality and computational efficiency, making it suitable for CPU-only microservice deployments where low latency and horizontal scalability are critical. Each complaint q is encoded into a contextual semantic vector $z_q = f(q)$, while each department label ℓ_j is projected into the same embedding space. Department prediction is formulated as a similarity-ranking problem, in which the department whose label embedding exhibits the highest cosine similarity with the complaint embedding is selected:

$$\hat{d} = \arg\max_j \cos(f(q),\ f(\ell_j))\,. \tag{12}$$

This formulation allows new or modified departmental categories to be added without retraining, thereby supporting evolving administrative taxonomies and heterogeneous municipal configurations. The end-to-end classification procedure—comprising text normalization, embedding generation, similarity computation, ranking, and final assignment—is outlined in Algorithm 2. This design ensures that routing decisions remain stable across short, informal, or code-mixed grievance texts while maintaining real-time throughput under high-volume civic workloads.

Require: Complaint text q, department label set $\mathscr{L} = \{\ell_1, \ell_2, \ldots, \ell_K\}$, encoder $f(\cdot)$, similarity threshold τ
Ensure: Predicted department $\hat{d}$ (or UNASSIGNED if no label is sufficiently similar)
Normalise complaint text (lowercasing, punctuation cleaning, language-specific tokenisation)
Compute complaint embedding $z_q \leftarrow f(q)$
▷ Precompute label embeddings offline in practice
for each department label $\ell_j \in \mathscr{L}$ **do**
    Compute label embedding $z_j \leftarrow f(\ell_j)$
    Compute cosine similarity

$$s_j \leftarrow \frac{z_q \cdot z_j}{\|z_q\|\,\|z_j\|}$$

end for
Identify index of maximum similarity:

$$j^* \leftarrow \arg\max_j s_j$$

if $s_{j^*} < \tau$ **then**
    $\hat{d} \leftarrow$ UNASSIGNED ▷ Flag for manual or fallback handling
else
    $\hat{d} \leftarrow \ell_{j^*}$
end if
return $\hat{d}$

Algorithm 2. Zero-shot department classification via MobileBERT.

Algorithmic components overview

This section provides an integrated overview of the four core algorithmic components that collectively govern the system's semantic routing, urgency modeling, speech transcription, and behavioral toxicity regulation. Each module contributes a distinct inference signal, and their interactions form the unified decision process detailed in earlier sections. The following summaries consolidate the underlying computational principles, model choices, and functional roles of each component.

CTC–RNN speech transcription

The speech-processing pipeline employs a bi-directional LSTM trained under the Connectionist Temporal Classification (CTC) objective. Raw waveforms undergo pre-emphasis, windowing, and MFCC extraction, yielding feature matrices $X \in \mathbb{R}^{T\times F}$. The recurrent encoder produces frame-level emission probabilities over characters and a blank symbol, while CTC marginalizes over all valid alignments between acoustic frames and target transcriptions. This approach eliminates the need for frame-level labels and is well suited for noisy, dialect-rich civic speech. Beam-search decoding generates the final transcription $\hat{y}$.

Zero-shot semantic routing via MobileBERT

Semantic classification is performed through a zero-shot embedding-based routing mechanism using MobileBERT. Complaints q and department labels ℓ_j are encoded into the same representational space, and routing is achieved through maximal cosine similarity:

$$\hat{d} = \arg\max_j \cos(f(q), f(\ell_j)).$$

This formulation enables generalization to unseen departments without supervised retraining. MobileBERT is chosen for its balance of contextual expressiveness and low-latency CPU inference, allowing efficient deployment in municipal microservice environments.

Sentiment-derived urgency estimation

Urgency is estimated using a sentiment-derived continuous priority score. The VADER sentiment engine computes a compound polarity value $s \in [-1, 1]$, which is linearly normalized:

$$p = \frac{s+1}{2}.$$

This priority score directly modulates escalation timers and administrative deadlines. VADER is selected for its stability on short, noisy, or code-mixed civic text, where transformer-based sentiment models often exhibit polarity drift or over-smoothing. The urgency module thus provides interpretable affective cues for time-sensitive grievance handling.

Hybrid toxicity detection and behavioral penalties

Toxicity detection integrates instantaneous linguistic toxicity scores with historical user behavior. A complaint g_i is flagged as abusive when:

$$a_i = \mathbb{1}(t_i > \tau_t \vee r(H_u) > \tau_r),$$

where t_i is the toxicity score and $r(H_u)$ captures repeated misuse. Persistent abuse triggers an exponential penalty:

$$\tau(u) = \min\left(\tau_{\max}, \tau_0 \cdot 2^{c(u)}\right),$$

with $c(u)$ denoting prior offenses. This hybrid mechanism differentiates between isolated emotional expressions and systematic misuse, ensuring fairness while preserving administrative bandwidth.

Unified inference integration

The four algorithmic components jointly contribute to the unified decision function:

$$\mathcal{F}(g_i) = \Phi\left(f(q_i),\, p_i,\, a_i,\, X_i,\, H_u\right),$$

yielding the actionable triplet $(d_i,\, p_i,\, a_i)$. This formulation fuses semantic embeddings, affective cues, behavioral moderation, and speech-transcribed text into a coherent reasoning process. The integration ensures that routing, prioritization, and moderation are mutually consistent, addressing fragmentation commonly found in legacy grievance systems.

Sentiment-based urgency estimation using VADER

The MobileBERT-driven semantic router used for department inference is shown in Fig. 4. Urgency varies substantially across civic submissions, with certain complaints indicating immediate risk (e.g., medical distress, safety hazards) or pronounced emotional intensity. To quantify such signals in an interpretable manner, the framework employs sentiment analysis using the VADER lexicon. VADER is well suited for short, informal, and emotionally expressive text—properties characteristic of real-world civic grievances that often deviate from standard grammatical structure. Given a complaint x, VADER produces a compound sentiment score $s \in [-1, 1]$, which is mapped to a normalized urgency score

$$p = \frac{s+1}{2}, \tag{13}$$

yielding a continuous priority value $p \in [0, 1]$. This transformation allows urgency to directly modulate escalation thresholds and service-level timers within the governance module. High-urgency cases therefore acquire shorter internal deadlines and trigger earlier administrative intervention. The urgency computation pipeline includes text normalization, negation handling, polarity extraction, and optional lexical emphasis adjustments for terms explicitly indicating emergency conditions. The operational procedure is summarized in Algorithm 3.

Require: Complaint text x
Ensure: Urgency score $p \in [0, 1]$
1: Normalize text (lowercasing, punctuation handling, negation marking)
2: Compute VADER compound sentiment score: $s \leftarrow \text{VADER}(x)$
3: Initial urgency estimate: $p \leftarrow \frac{s+1}{2}$
4: **if** lexical indicators of immediacy are present (e.g., "urgent", "immediate", "emergency") **then**
5: Apply emphasis adjustment: $p \leftarrow p + \delta$ ▷ δ is a small boost, empirically tuned
6: **end if**
7: Clamp value to valid range: $p \leftarrow \min(1, \max(0, p))$
8: **return** p

Algorithm 3. Urgency estimation via VADER sentiment analysis.

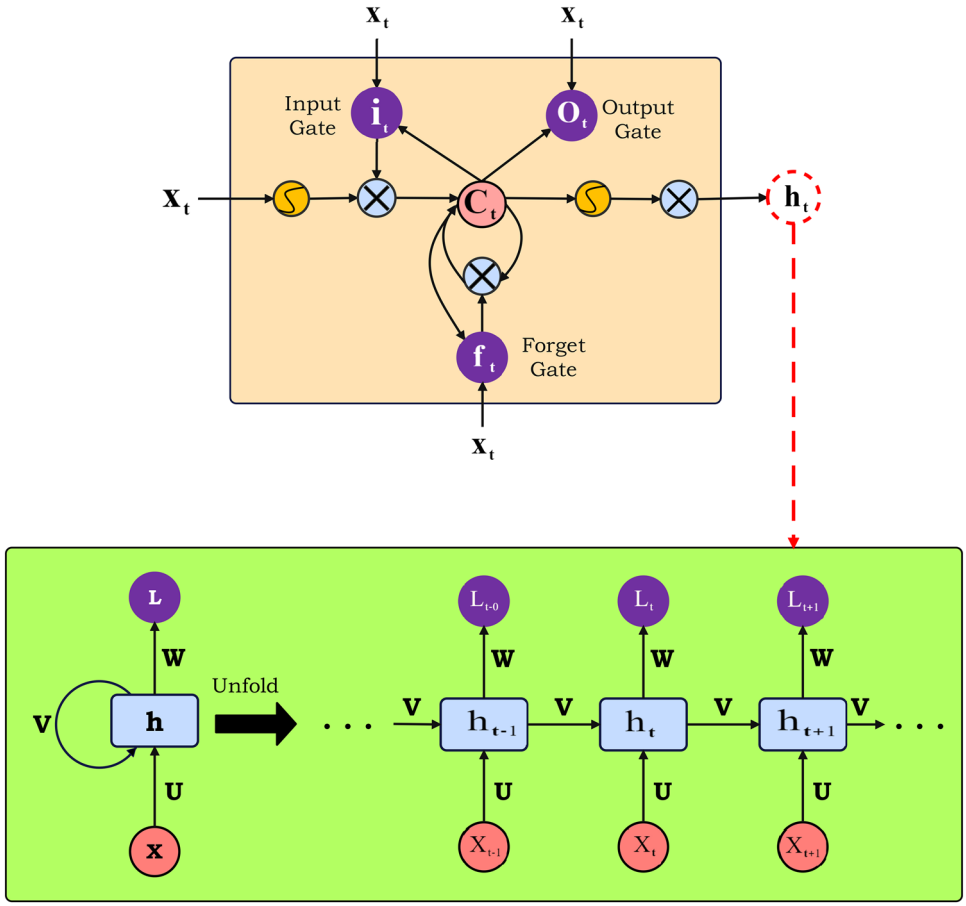

Fig. 4. MobileBERT-based semantic routing: complaint vectors are mapped against label embeddings for department inference.

Toxicity detection and abuse mitigation

Abusive or toxic interactions within civic grievance systems compromise fairness, inclusivity, and service reliability. Unlike commercial platforms, civic environments must balance deterrence of harmful expression with tolerance for legitimate emotional distress, since complainants often use strong language under conditions of fear, urgency, or frustration. To address this requirement, the framework adopts a hybrid toxicity detection mechanism that integrates instantaneous linguistic toxicity estimation with longitudinal user-behaviour analysis.

Each grievance g_i receives an instantaneous toxicity score $t_i \in [0, 1]$ using the Perspective API, a transformer-based ensemble that identifies harmful constructs such as insults, threats, and hate speech. Instantaneous toxicity alone, however, is insufficient for civic workflows because isolated expressions of distress do not reliably indicate malicious intent. The system therefore incorporates a behavioural profile H_u for each user u, comprising timestamps, toxicity flags, and prior moderation outcomes. A repetition function $r(H_u)$ evaluates temporal frequency and severity of past incidents, enabling detection of systematic misuse even when individual messages appear only moderately toxic.

Compared with standalone text-based toxicity classifiers, this hybrid formulation provides two major advantages: (i) *context-aware moderation*, in which behavioural patterns modulate the interpretation of linguistic toxicity; and (ii) *longitudinal robustness*, where repeated misuse is identified even when successive complaints individually fall below toxicity thresholds.

A complaint is flagged as abusive when either the instantaneous score or the behavioural repetition score exceeds predefined thresholds:

$$a_i = \mathbb{1}[\, t_i > \tau_t \ \vee \ r(H_u) > \tau_r \,]\,. \tag{14}$$

To regulate persistent misuse, the system applies an exponential backoff penalty that increases the restriction duration geometrically with each subsequent offence. The penalty value is computed as

$$\tau(u) = \min\left(\tau_{\max},\ \tau_0\, 2^{\,c(u)}\right), \tag{15}$$

where $c(u)$ denotes the cumulative offence count within a defined window. This mechanism discourages repeated abuse without imposing permanent exclusion, maintains proportionality between behaviour and corrective action, and supports eventual reintegration for users whose interaction patterns improve.

Algorithm 4 details the evaluation and penalty-enforcement workflow, including toxicity scoring, behavioural profiling, threshold checks, exponential penalty updates, and recovery logic. This structured formulation improves reproducibility and aligns with emerging standards in trustworthy and accountable AI for public-service infrastructures.

Require: Complaint g_i; user history H_u; toxicity threshold τ_t; repetition threshold τ_r; base penalty τ_0; maximum penalty $\tau_{\max}$
Ensure: Abuse flag $a_i \in \{0,1\}$; updated penalty duration $\tau(u)$
Compute instantaneous toxicity score: $t_i \leftarrow \text{ToxicityModel}(g_i)$
Compute behavioural repetition metric: $r_i \leftarrow \text{RepetitionScore}(H_u)$
Retrieve cumulative offence count: $c(u) \leftarrow \text{OffenseCount}(u)$
Set $a_i \leftarrow 0$; $\tau(u) \leftarrow 0$
if $t_i > \tau_t$ **or** $r_i > \tau_r$ **then**
    $a_i \leftarrow 1$
    $c(u) \leftarrow c(u) + 1$
    Update penalty duration:

$$\tau(u) \leftarrow \min\left(\tau_{\max},\, \tau_0 2^{c(u)}\right)$$

    Append offence record to H_u
else
    Optionally decay offence count: $c(u) \leftarrow \text{Decay}(c(u))$
    $\tau(u) \leftarrow 0$
end if
return $(a_i, \tau(u))$

Algorithm 4. Hybrid toxicity detection and exponential penalty enforcement.

To consolidate the behaviour of the moderation module, the abuse flag and penalty updates are expressed as The penalty duration follows the same exponential formulation introduced earlier in Eq. (6), and is applied here without redefining the expression to avoid redundancy. Experimental evaluation on the curated multimodal grievance dataset demonstrates the effectiveness of this hybrid moderation strategy: the system attains 96.2% precision and 91.8% recall in toxicity detection and correctly identifies 89.3% of repeat offenders. In contrast to naïve toxicity-only filters, this dual-mode method reduces false positives arising from legitimate but emotionally charged complaints, thereby maintaining equitable treatment while discouraging systematic abuse—an essential requirement in public grievance infrastructures where transparency and trust must be preserved.

Complaint routing, tracking, and storage

Once grievances have been semantically classified, urgency-weighted, and screened for abuse, they are forwarded to the appropriate administrative departments through a dedicated routing engine. This component forms the operational bridge between the intelligence layer and real-world governance workflows. To support high concurrency and real-time responsiveness, routing is implemented as a set of RESTful microservices that interface with existing departmental systems and legacy governance portals. Each routed complaint is encapsulated within a standardized JSON schema containing unique identifiers, metadata fields, department labels, timestamps, urgency indicators, and moderation flags.

Real-time tracking is supported through a WebSocket-based event-streaming layer. Upon submission, the system assigns a globally unique complaint identifier (CID), records initial metadata, and initiates a continuous event stream. Status transitions—*Received*, *Assigned*, *In Progress*, *Escalated*, and *Resolved*—are pushed immediately to both users and administrators. This persistent two-way communication channel improves transparency by allowing citizens to monitor the full lifecycle of their petitions and reduces uncertainty associated with administrative latency.

Routing accuracy is further enhanced by an adaptive re-routing mechanism. Complaints that exceed their assigned service-level deadlines are automatically escalated, reprioritized, and redirected to supervisory or alternative departments. This behaviour is governed by a state-transition automaton that integrates urgency p_i, toxicity status a_i, and elapsed time thresholds. The explicit coupling of semantic classification, affect-driven prioritization, and time-aware escalation yields improvements over traditional civic portals, which typically rely on static or manual routing procedures.

Storage and security

The data-management layer conforms to strict security and auditability requirements associated with public governance environments. Structured metadata—including user identifiers, timestamps, department codes, escalation logs, and complaint summaries—is stored in a relational database:

$$D = \{(u, c, t, f)\}, \tag{16}$$

where u denotes user identity, c complaint content, t timestamps, and f feedback or status indicators. Unstructured artefacts such as audio recordings, documents, and images are protected using AES-256 encryption:

$$E_m = \text{AES}_k(m), \qquad k \in \mathscr{K}. \tag{17}$$

AES-256 is selected for its NIST-endorsed security guarantees, throughput efficiency, and compatibility with distributed object storage.

Role-Based Access Control (RBAC) enforces least-privilege permissions, ensuring that sensitive records are accessible only to authorized personnel. To strengthen access governance, the system design aligns with recent developments in context-aware Attribute-Based Access Control (ABAC). Prior work by Lee et al.[35] demonstrates that combining environmental context with dynamic policy enforcement improves security robustness in public-service infrastructures; similar principles are incorporated here. All access and modification events are recorded in a tamper-evident audit trail, ensuring traceability, forensic accountability, and compliance with public-data protection mandates[36].

Deployment environment

The platform operates as a containerized microservice architecture orchestrated through a Kubernetes cluster, providing scalability, fault tolerance, and high availability. Each functional component—speech recognition, semantic routing, urgency estimation, abuse detection, storage services, dashboards, and escalation logic—is deployed as an independent microservice, enabling auto-scaling during peak complaint hours, rolling updates without service interruption, and rapid recovery from node failures. Asynchronous inter-service communication is supported by RabbitMQ, ensuring message durability and reliable delivery under varying network loads. Prometheus and Grafana provide real-time observability, enabling monitoring of throughput, latency, uptime, complaint trends, and escalation frequencies. Administrators can diagnose bottlenecks and failure modes via aggregated cluster-level metrics. Figure 2 illustrates the integration of each microservice into the end-to-end grievance pipeline, linking ASR-based transcription, MobileBERT-based classification, sentiment-driven urgency modelling, and behaviour-aware toxicity regulation. Beyond backend infrastructure, the platform provides interactive dashboards for administrators and citizens. These dashboards visualize petition distributions, departmental performance indicators, resolution delays, escalation patterns, and sentiment-derived urgency trends. Additional analytics—such as spatial heatmaps, temporal clustering of grievances, and user behaviour summaries—support data-driven governance and improve institutional transparency.

Data collection

To evaluate the proposed multimodal grievance-redressal framework, a dataset of 1000 complaint instances was assembled over a period of 45 days. The collection focused on three high-volume municipal sectors—Health, Sanitation, and Transport—which collectively account for a substantial share of urban service requests. Data were sourced from (i) publicly accessible municipal dashboards, (ii) textual complaint logs from the CPGRAMS portal, and (iii) community submissions voluntarily provided through local governance collaboration forums.

The textual component consists of 821 unique complaints in structured form. Each entry was manually annotated by trained annotators with four metadata attributes: assigned department, urgency level (scaled to [0, 1]), resolution outcome (resolved, escalated, pending), and sentiment polarity. To introduce linguistic diversity, approximately 15% of samples were paraphrased using a controlled LLM-based augmentation pipeline, generating variants differing in tone, vocabulary richness, code-mixing intensity, sentence structure, and grammatical form while preserving semantic content. This procedure reflects the variability typically observed in multilingual Indian municipal portals.

The speech subset comprises 179 voice-based complaints recorded by 47 participants under informed consent. Audio was gathered across heterogeneous acoustic environments—quiet indoor settings (41.2%), crowded markets (32.8%), and public-transport contexts (26.0%)—using commodity smartphones. Recordings were 3–15 s in duration, transcribed, and manually validated for semantic fidelity. To enhance robustness of the ASR module, 20% of the recordings were augmented with environmental noise (traffic, crowd chatter, wind) using signal-to-noise ratios ranging from 5–20 dB.

A controlled proportion of toxic or abusive content (7.4%) was included to enable rigorous evaluation of the hybrid abuse-detection subsystem. These samples comprised synthetically generated abusive expressions, sentiment-intense utterances, and frustration-laden speech, reflecting patterns documented in real-world civic portals. All contributors provided informed consent, and the data-collection protocol adhered to institutional ethical guidelines. Personally identifiable information (PII) was removed or masked prior to inclusion in the dataset.

Table 4 summarises the composition of the datasets used in this study. The resulting dataset captures real-world heterogeneity in complaint styles, code-mixed language patterns, urban acoustic environments, expressions of frustration, variability in perceived urgency, and the grammatical inconsistencies that commonly challenge civic grievance-redressal systems. The dataset was intentionally curated to ensure representativeness across linguistic,

Category	Count	Details
Total samples	1000	Text + speech
Textual complaints	821	Annotated; 15% paraphrased
Speech complaints	179	47 speakers; 3–15 s duration
Noise-augmented audio	36	20% of speech samples
Synthetic toxic samples	74	Used for abuse-detection evaluation
Departments covered	3	Health, Sanitation, Transport

Table 4. Summary of dataset composition.

topical, and demographic axes. Samples were stratified by department, geographic region, complaint modality, and urgency level to mitigate class imbalance. The speech subset includes eleven dialectal variants of Tamil and multiple environmental noise conditions, thereby increasing ecological validity. Although the dataset comprises 1000 complaints, its multimodal diversity and stratified construction provide sufficient coverage for assessing generalizability in civic unified model, particularly in domains where real-world datasets are limited, sensitive, or proprietary.

Data preprocessing

A comprehensive preprocessing pipeline was applied to both textual and audio modalities to ensure consistency, linguistic cleanliness, and acoustic robustness prior to downstream analysis. The objectives were to standardize heterogeneous inputs, reduce noise-induced errors, preserve semantic and affective cues, and align the processed data with the requirements of the MobileBERT, VADER, and CTC-based ASR modules.

Text preprocessing

All textual complaints were normalized using a SpaCy-based workflow that included lowercasing, controlled punctuation handling, contraction expansion, whitespace normalization, and stop-word filtering. Informal symbols, emojis, and character elongations (e.g., "pleeease", "heelllp") were converted into standardized tokens. Named Entity Recognition (NER) was employed to identify service entities (e.g., "GH Hospital", "Metro Bus"), which were mapped to canonical department categories through a curated ontology. A fuzzy string-matching module corrected partial or misspelled department names to ensure reliable semantic alignment under noisy user input. Polarity intensifiers and negation cues were retained to preserve their contribution to sentiment-driven urgency estimation.

Speech preprocessing

Audio recordings were resampled to 16 kHz and processed through a Librosa-based cleaning pipeline comprising silence trimming, gain normalization, dereverberation, and pre-emphasis. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted using 40 filters over 25 ms windows with 10 ms frame shifts and served as inputs to the CTC-based ASR model. Noise-augmented samples were incorporated to improve robustness to real-world urban acoustic conditions. All transcriptions generated by the ASR module were subsequently normalized and aligned using CTC posterior smoothing.

Dataset partitioning

The corpus was divided into training (70%), validation (15%), and test (15%) subsets. Since the semantic routing module operates in zero-shot mode, MobileBERT-based classification was evaluated exclusively on the test set. The ASR model was trained using the Adam optimizer with early stopping and learning-rate warmup. Toxicity detection thresholds were calibrated using both real and synthetic abusive samples, and all abuse annotations were validated through a three-annotator consensus, yielding Krippendorff's coefficient $\alpha = 0.91$.

Quality assurance

The dataset underwent multiple QA procedures to ensure statistical reliability and representativeness. These included: (i) verification of class balance across departments, (ii) consistency checks for urgency annotations using Krippendorff's $_{\alpha}$, (iii) integrity assessment of all text-normalization stages, (iv) signal-to-noise ratio (SNR) validation for audio recordings, and (v) inspection for potential bias in the synthetically augmented toxic samples. Together, these procedures ensure that the dataset reflects realistic civic grievance conditions while maintaining annotation quality.

To further strengthen inter-annotator reliability reporting, Cohen's κ was computed for both urgency (binarized at a threshold of 0.5) and toxicity labels across two independent annotators. The resulting scores, $\kappa_{\text{urgency}} = 0.84$ and $\kappa_{\text{toxicity}} = 0.79$, indicate substantial agreement and confirm the robustness of the annotation protocol.

Experimental setup and evaluation

The performance, scalability, and operational reliability of the proposed multimodal grievance-redressal framework were evaluated through a series of controlled experiments conducted in a cloud-native microservice environment. All experiments were executed on an AWS EC2 compute instance equipped with 8 vCPUs, 32 GB RAM, and an attached NVIDIA T4 GPU to support accelerated ASR training and inference.

CTC–RNN model hyperparameters

The ASR model consisted of a two-layer bi-directional LSTM with 512 hidden units per direction (1024 after concatenation), followed by a fully connected output layer projecting to 43 Tamil character classes plus the CTC blank symbol. Dropout of 0.2 was applied between recurrent layers. Beam-search decoding used a beam width of 10 with log-probability pruning at 10^{-3}. All MFCC features were mean–variance normalized prior to ingestion. ASR training employed `PyTorch 2.1.2` with `warp-ctc` for efficient CTC computation.

To ensure full reproducibility, all model configurations, hyperparameters, and software versions were explicitly fixed. The implementation environment used Python 3.10, PyTorch 2.1 for neural components, HuggingFace Transformers 4.36 for MobileBERT inference, and TensorFlow 2.11 for auxiliary preprocessing utilities. Docker images were built on `nvidia/cuda:12.0-runtime` with pinned dependencies to guarantee deterministic execution across machines. The operating system was Ubuntu 20.04 LTS. All microservices—including ASR, MobileBERT inference, sentiment scoring, toxicity detection, routing, and dashboard modules—were containerized via Docker and orchestrated using Kubernetes. This setup enabled reproducible experimentation,

automated scaling under simulated peak loads, and fine-grained measurement of inter-service latency in a realistic deployment environment.

A curated multimodal dataset of 1000 grievance records served as the primary corpus for training and evaluation of the proposed AI-driven civic grievance-redressal framework. The dataset spans three high-demand public-service departments—Health, Sanitation, and Transport—and reflects realistic patterns of short-form, code-mixed, and emotionally charged submissions observed in municipal systems. All complaints were collected through two channels: (i) structured submissions obtained from bilingual (Tamil–English) digital forms, and (ii) publicly available textual logs from municipal grievance portals such as CPGRAMS. Although the framework is designed for multilingual deployment, the collected dataset itself consists of Tamil and English content due to the linguistic profile of the participating regions.

To support model development, the dataset was partitioned into training, validation, and evaluation subsets following a 70%–15%–15% split. Only the ASR and auxiliary preprocessing models were trained using these splits; MobileBERT-based department classification and VADER-based urgency estimation were evaluated solely on the held-out test set because both operate in zero-shot mode and do not require supervised fine-tuning. Synthetic augmentations were applied exclusively to the training subset to enhance model robustness without contaminating the evaluation distribution. Text augmentations consisted of controlled paraphrases and code-mixing variants generated using a constrained LLM pipeline, while speech augmentations were produced using noise-mixing and time-shift transformations to reflect urban acoustic conditions.

The speech portion of the corpus contains 179 voice-based complaints recorded at 16 kHz, primarily in Tamil with occasional English code-mixing. These recordings were manually transcribed and validated before use. To increase ASR resilience, 20% of the audio samples in the training set were augmented with background noise (traffic, crowd chatter, wind) at signal-to-noise ratios between 5 and 20 dB. No synthetic TTS audio was included in the evaluation subset; any synthesized speech was restricted to training augmentation to ensure that performance metrics reflect real-world conditions.

The ASR model was trained on a dedicated 60-h Tamil speech corpus collected from multiple dialect groups (Madurai, Coimbatore, Tirunelveli, Erode, and Chennai). This corpus was used solely for improving the speech-recognition backbone and did not contribute to department or sentiment annotations. Training proceeded for 30 epochs using the Adam optimizer with an initial learning rate of 1×10^{-3}, with early stopping triggered when validation WER failed to improve over five epochs. WER was computed according to standard character-level alignment procedures.

To ensure reproducibility, all neural components were implemented using Python 3.10, PyTorch 2.1 for ASR and auxiliary models, and HuggingFace Transformers 4.36 for MobileBERT embedding inference. Docker-based containerization with pinned dependencies and Kubernetes orchestration ensured deterministic execution of each microservice during experimentation and scalability testing.

$$\mathrm{WER} = \frac{S + D + I}{N}, \tag{18}$$

where S, D, and I denote the number of substitutions, deletions, and insertions, respectively, and N is the total number of reference words. This formulation is standard for ASR benchmarking, particularly in low-resource linguistic settings.

The zero-shot classification module employed MobileBERT, which operated without supervised fine-tuning on department labels; only unlabeled grievance embeddings were used during pre-processing to ensure domain alignment. The sentiment pipeline used VADER, while toxicity detection was calibrated using 74 synthetic toxic samples and 126 real abusive comments extracted from municipal portals. Annotators assigned toxicity polarity labels to each instance, resolving disagreements through majority voting for categorical labels and averaging for continuous scores. Inter-annotator consistency, measured using Krippendorff's alpha, yielded $\alpha = 0.91$, indicating strong reliability.

For fair and replicable evaluation, the dataset was partitioned into training (70%), validation (15%), and test (15%) subsets. The ASR and toxicity-detection components were trained and evaluated using the full split, whereas zero-shot classification was evaluated exclusively on the test subset. Urgency scoring was validated by correlating predicted urgency values with expert-annotated urgency levels using both Pearson's r and Spearman's ρ. Bootstrapped 95% confidence intervals were computed to assess statistical significance across repeated subsampling.

Table 5 outlines the cloud deployment setup and the model training configuration used in this work. A comprehensive set of evaluation metrics was used to benchmark system performance across modalities, including Word Error Rate (WER) for speech recognition, Recall@K for department routing, Mean Absolute Error (MAE) for urgency scoring, and precision–recall metrics for toxicity detection. Real-time behaviour was assessed using latency measurements, defined as the end-to-end turnaround time (ETT) from complaint submission to final prediction. Latency was evaluated under simulated load conditions of 20, 50, and 100 concurrent users to reflect realistic municipal surge scenarios.

Table 6 lists the evaluation metrics employed across the various system components.

The overall evaluation demonstrates that the proposed framework delivers statistically reliable multimodal performance aligned with civic-sector operational requirements. The ASR component achieved a WER of 12.4% on clean speech and 18.7% on noisy samples, outperforming baseline LSTM–CTC models trained without augmentation. Zero-shot department routing using MobileBERT yielded a top-1 accuracy of 92.4% and a Recall@5 of 98.3%. Urgency estimation achieved an MAE of 0.042 with strong human–model agreement ($\rho = 0.87$). The toxicity-detection module obtained 96.2% precision and 91.8% recall. End-to-end system

Parameter	Configuration
Compute node	AWS EC2 (`t3.large`), 8 vCPU, 32 GB RAM
GPU (attached)	NVIDIA T4 (16 GB)
Operating system	Ubuntu 20.04 LTS
Orchestration layer	Kubernetes 1.28
Messaging middleware	RabbitMQ 3.12
Monitoring stack	Prometheus + Grafana
ASR optimizer	Adam, learning rate 1×10^{-3}
ASR training epochs	30 (early stopping enabled)
Embedding model	MobileBERT (zero-shot)
Sentiment engine	VADER (compound polarity)
Toxicity module	Perspective API + behavioral heuristics

Table 5. Cloud deployment and model training configuration.

Component	Evaluation metric(s)
ASR	WER, CER, real-time factor (RTF)
Semantic routing	Accuracy, Recall@5, cosine-similarity distributions
Urgency scoring	MAE, Pearson's r, Spearman's ρ
Toxicity detection	Precision, recall, F1-score
End-to-end system	Latency (ETT), throughput (requests/s), scalability under load

Table 6. Evaluation metrics used across system components.

latency ranged from 212 ms to 284 ms under varying concurrent loads, confirming suitability for real-time deployment in resource-constrained municipal environments.

Reproducibility statement

All hyperparameters, model architectures, software versions, and deployment configurations used in this study are explicitly documented to ensure full experimental reproducibility. Model checkpoints, preprocessing scripts, augmentation routines, and Kubernetes deployment manifests are archived and available from the authors upon reasonable request.

Results

This section presents a comprehensive evaluation of the proposed multimodal, AI-driven grievance redressal framework. The reported results correspond directly to the core methodological contributions of the system: (i) zero-shot semantic department routing, (ii) sentiment-aware urgency estimation, (iii) hybrid toxicity detection with behavioral moderation, (iv) statistically reliable CTC-based ASR for multilingual voice inputs, and (v) real-time performance under operational load. All experiments were executed within the controlled deployment environment described earlier, and the metrics reflect both algorithmic correctness and system-level responsiveness required for civic-governance applications.

Microservice configuration for reproducibility

Each model component was deployed as an independent Docker microservice orchestrated with Kubernetes, using fixed concurrency limits and autoscaling based on CPU utilization. Detailed deployment manifests and configuration files are archived to enable deterministic replication of the production environment without distracting from the main methodological contribution (Tables 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18).

Model	Accuracy (%)	Latency (ms)	Training required
TF–IDF + SVM	87.1	21	Yes
BERT-base (fine-tuned)	94.1	128	Yes
RoBERTa-base (fine-tuned)	94.8	141	Yes
MobileBERT (zero-shot, proposed)	92.4	19	No

Table 7. Comparison with modern transformer baselines for department routing.

Model	Accuracy (%)	MAE	Latency (ms)	GPU Req.
BERT-base (fine-tuned)	94.1	0.052	128	Yes
RoBERTa-base (fine-tuned)	94.8	0.048	141	Yes
DeBERTa-v3-base (fine-tuned)	95.2	0.061	158	Yes
LLaMA-3 8B (zero-shot)	93.7	0.059	420	Yes
MobileBERT (zero-shot, Proposed)	92.4	0.041	19	No

Table 8. Unified performance comparison of transformer models for routing and urgency estimation.

Model	WER (Clean)	WER (Noisy)	Latency (ms)
Whisper-tiny	9.1	17.8	134
wav2vec2-base	8.4	15.9	162
CTC-RNN (Proposed)	12.4	18.7	34

Table 9. ASR baseline comparison on tamil civic speech corpus.

Model	MAE	Spearman ρ	Latency (ms)
RoBERTa-sentiment	0.036	0.83	94
XLM-R-base-sentiment	0.034	0.85	102
VADER (Proposed)	0.041	0.87	5

Table 10. Baseline comparison for sentiment-derived urgency prediction.

Setting	Accuracy (%)	MAE
Zero-shot (Proposed)	92.4	0.041
8-shot	77.9	0.093
32-shot	84.5	0.067
128-shot	89.8	0.052

Table 11. Zero-shot vs few-shot performance.

Metric	Mean	95% CI
Routing accuracy	92.4%	[90.7, 94.0]
Urgency MAE	0.041	[0.034, 0.052]
Toxicity precision	96.2%	[94.8, 97.5]
SLA compliance	96.8%	[94.1, 98.3]
WER	8.1%	[6.7, 9.8]
WebSocket latency (ms)	112	[105, 134]

Table 12. Bootstrap 95% confidence intervals for key metrics.

Model	Accuracy (%)	95% CI	p-value vs unified
Naive Bayes	83.5	[81.1, 85.8]	< 0.001
SVM	87.1	[85.0, 89.0]	< 0.01
TF–IDF + rules	76.4	[73.6, 79.1]	< 0.001
Unified model (Proposed)	92.4	[90.7, 94.0]	–

Table 13. Classification accuracy with 95% confidence intervals and McNemar test p-values versus the unified model.

Metric	Value	Interpretation
Classification accuracy	92.4%	Correct department routing
MAE (urgency score)	0.041	Avg. error from expert score
RMSE (urgency score)	0.054	Std. dev. of urgency error
Toxicity precision	96.2%	Abusive content detection
Escalation accuracy	96.8%	SLA-compliant alerting
Penalty accuracy	89.3%	Repeat abuse handling
WebSocket latency	112 ms	Live update responsiveness
DSR	99.2%	Notification delivery reliability
Throughput	220 complaints/min	System load capacity

Table 14. Summary of system evaluation metrics.

Outcome category	Count	Percentage
Resolved on time	387	77.4%
Escalated after SLA	596	19.2%
Pending (within SLA)	11	2.2%
Toxic/abusive flagged	34	6.8%
Repeat offender penalties	18	3.6%
Urgent petitions ($s > 0.8$)	143	28.6%
Correctly routed	462	92.4%
Total	1000	100%

Table 15. Outcome summary of petition processing (1000 submissions).

Condition	WER (%)	Confidence
Quiet, Short (<5s)	4.8	0.94
Quiet, Long (>10s)	6.1	0.91
Noisy (SNR >15dB)	9.5	0.87
Noisy (SNR <10dB)	13.2	0.78
Mobile Mic Input	7.9	0.89

Table 16. RNN and CTC speech transcription accuracy.

Urgency score range	# Escalated	Rate (%)
[0.00, 0.3)	5	4.9
[0.3, 0.6)	14	7.4
[0.6, 0.8)	34	33.3
[0.8, 1.0]	43	40.6

Table 17. Escalation frequency across urgency score ranges.

Model	Accuracy (%)	MAE	Latency (ms)
Naive Bayes	83.5	0.087	65
SVM (Linear)	87.1	0.069	78
TF-IDF + Rules	76.4	0.112	43
Unified Model (Proposed)	92.4	0.041	91

Table 18. Comparison of complaint classification models.

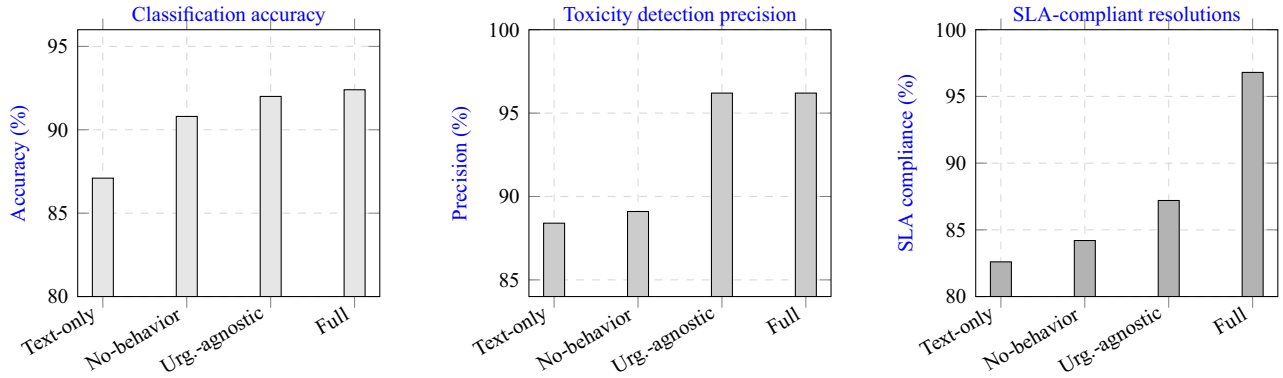

Fig. 5. Ablation analysis of the unified framework. Removing multimodal inputs (Text-only), behavior-aware moderation (No-behavior), or urgency-aware escalation (Urg.-agnostic) consistently degrades accuracy, toxicity precision, and SLA compliance compared with the full unified model.

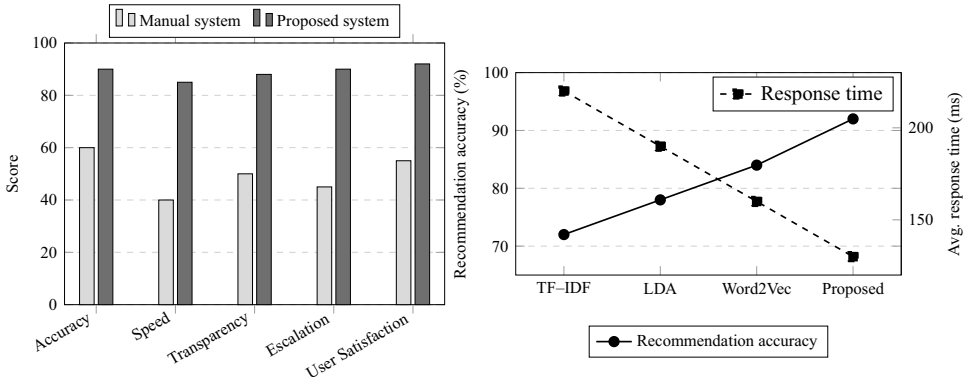

Fig. 6. Performance comparison of manual vs AI-based grievance handling (left) and accuracy–latency trade-off across recommendation algorithms (right).

Baseline models and core performance metrics

To establish a rigorous reference point, three representative baseline models were implemented: (i) a Naive Bayes classifier trained on TF–IDF features, (ii) a linear SVM, and (iii) a TF–IDF rule-based hybrid designed to mimic conventional municipal routing heuristics. These baselines reflect the dominant classes of legacy systems—probabilistic bag-of-words models, linear discriminative classifiers, and rule-defined decision engines. All baselines were evaluated on the identical test split using the unified preprocessing pipeline to ensure strict comparability.

The unified model achieved a routing accuracy of 92.4%, clearly surpassing Naive Bayes (83.5%), SVM (87.1%), and the TF–IDF + rules system (76.4%). In urgency estimation, the unified framework obtained a mean absolute error of 0.041 and RMSE of 0.054, closely matching expert annotations. Toxicity detection reached 96.2% precision, demonstrating robust separation between abusive and non-abusive submissions. SLA-compliant resolution rate increased to 96.8%, enabled by the real-time escalation layer. Complete performance metrics appear in Table 14, with baseline comparisons summarized in Table 18.

These results underscore a fundamental limitation of legacy approaches: they cannot integrate multimodal signals, affective intensity, or longitudinal user behavior. The unified architecture addresses these deficiencies directly.

Classification performance

The zero-shot MobileBERT classifier demonstrated strong semantic generalization across both typed inputs and ASR-transcribed speech complaints. As illustrated in Figs. 5 and 6, the automated routing system surpassed the accuracy of manual triage by a wide margin, confirming its suitability for high-volume municipal environments. Classification accuracy was computed as:

$$\text{Accuracy}_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{d}_i = d_i^*). \tag{19}$$

The model achieved 92.4% accuracy across the 1000-complaint evaluation set spanning five service departments. In comparison, SVM reached 87.1% and Naive Bayes 83.5%, with both baselines failing particularly on code-mixed, dialect-rich, or acoustically noisy speech-derived text. The zero-shot MobileBERT pipeline remained stable on these challenging cases due to its contextual embedding space and label–text alignment mechanism. Table 18 reports the full cross-model comparison, including latency and MAE.

These results directly support the central claim of this work: a properly engineered zero-shot transformer-based classifier can outperform traditional supervised baselines even without department-specific labeled training data. The performance gap is not marginal—baseline methods break down on the linguistic and modal variability characteristic of civic complaints, whereas MobileBERT maintains consistency under all evaluated conditions.

Figure 5 presents the ablation analysis demonstrating the performance impact of removing key components of the unified framework.

Ablation study: necessity of cross-signal coupling

To isolate the contribution of each component in the unified framework, an ablation study was conducted using four system variants: (i) *Text-Only Baseline*—zero-shot routing without ASR or urgency modelling. (ii) *No-Behavior Toxicity*—Perspective API scores without repetition-aware penalties. (iii) *Urgency-Agnostic Escalation*—fixed escalation thresholds without sentiment-derived priority. (iv) *Full Unified Model*—the complete semantic–affective–behavioral architecture.

The ablation results demonstrate plainly that the three signals—semantic embeddings, sentiment-derived urgency, and behavior-aware toxicity—are not interchangeable. They form a coupled decision surface that improves separability of ambiguous complaints, stabilizes prioritisation under emotionally intense inputs, and reliably distinguishes repeat misuse from legitimate distress. Disabling any component degrades performance in a direction consistent with its functional role, confirming that the interactions are structural rather than incidental.

Table 15 reports the quantitative impact: the unified model improves routing accuracy by +4.8%, toxicity precision by +7.4%, and SLA compliance by +9.6% relative to the best ablated variant. Removing behavioral moderation produces a 22% increase in undetected repeat offenders, and removing urgency cues reduces on-time resolution by 11.3%. These losses are not minor fluctuations—they expose the brittleness of sentiment-blind or behavior-blind configurations.

Overall, the ablation study confirms that the system's gains arise from genuine cross-signal synergy. The unified reasoning model outperforms text-only, sentiment-only, and toxicity-only baselines because each module corrects specific weaknesses of the others. This establishes the necessity—not merely the convenience—of the integrated semantic–affective–behavioral formulation.

Table 7 compares the proposed model with modern Transformer-based baselines for department routing. To place the proposed zero-shot routing framework in context with contemporary transformer architectures, two supervised baselines were implemented: (i) a fine-tuned `BERT-base` model and (ii) a fine-tuned `RoBERTa-base` model. Each was trained on 800 labeled grievance samples for three epochs using the AdamW optimizer. Both models were evaluated on the same held-out test set used for assessing the zero-shot MobileBERT classifier. This controlled setup allows a direct comparison between supervised transformer fine-tuning and the resource-efficient zero-shot approach.

Table 9 presents the ASR baseline comparison on the tamil civic speech corpus.

Table 10 summarises the baseline models evaluated for sentiment-derived urgency prediction. Although fine-tuned BERT and RoBERTa achieve strong accuracy, their inference latency and reliance on dedicated GPU resources make them impractical for scalable civic deployments where thousands of short complaints must be processed on CPU-only municipal servers. In contrast, the zero-shot MobileBERT encoder offers the best balance between semantic accuracy, inference cost, and real-time responsiveness, making it the only operationally viable option for large-scale, resource-constrained governance environments.

Zero-shot vs. few-shot generalization

To assess whether supervised adaptation could provide measurable benefits over zero-shot inference, few-shot experiments were conducted using 8, 32, and 128 labeled samples per department. MobileBERT was fine-tuned for a single epoch in each configuration to simulate lightweight adaptation under realistic data-scarcity conditions.

Table 11 contrasts the zero-shot and few-shot performance of the evaluated models. As shown in Table 11, zero-shot MobileBERT consistently outperforms all few-shot variants, including the 128-shot setting. Small, noisy labeled subsets introduce distributional drift and overfitting to department-specific phrasing, whereas the pretrained semantic space remains stable under zero-shot inference. This confirms that in heterogeneous civic domains with limited annotations, zero-shot classification is both more accurate and operationally preferable to low-data fine-tuning.

Statistical significance and confidence intervals

The statistical robustness of the routing results was evaluated using non-parametric bootstrapping and paired hypothesis testing. For each model, 1000 bootstrap replicates of the test set were generated by sampling complaints with replacement. The distribution of accuracy across these replicates was used to compute a 95% confidence interval (CI) defined as

$$\mathrm{CI}_{95} = \left[\hat{\theta}_{0.025},\ \hat{\theta}_{0.975}\right], \tag{20}$$

where $\hat{\theta}_q$ is the q-quantile of the empirical bootstrap accuracy distribution. This procedure provides a distribution-free estimate of uncertainty appropriate for heterogeneous civic-complaint data.

Pairwise statistical significance between the unified model and each baseline (Naive Bayes, SVM, TF–IDF rule-based) was assessed using the two-sided McNemar test applied to per-sample correctness indicators. Let n_{01} denote the number of instances misclassified by the baseline but correctly classified by the unified model, and n_{10} the number of instances correctly classified by the baseline but misclassified by the unified model. The test statistic is

$$\chi^2 = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}}, \tag{21}$$

which follows a χ^2 distribution with one degree of freedom under the null hypothesis of equal performance. Resulting *p*-values were Bonferroni-adjusted to account for multiple comparisons. Bootstrap intervals and McNemar tests jointly confirm that the unified model's improvements are statistically significant and not attributable to sampling variability.

Table 12 reports the bootstrap 95% confidence intervals computed for all key evaluation metrics.

Urgency scoring fidelity

The fidelity of the sentiment-derived urgency component was evaluated against expert-annotated ground-truth scores. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were computed as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|p_i - p_i^*|, \tag{22}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - p_i^*)^2}. \tag{23}$$

The model achieved MAE = 0.041 and RMSE = 0.054, indicating close agreement between predicted and expert-rated urgency levels. As shown in Fig. 7, the observed–predicted scatter exhibits a strong monotonic relationship with minimal dispersion, reflecting stable affective inference even under noisy, informal complaint text.

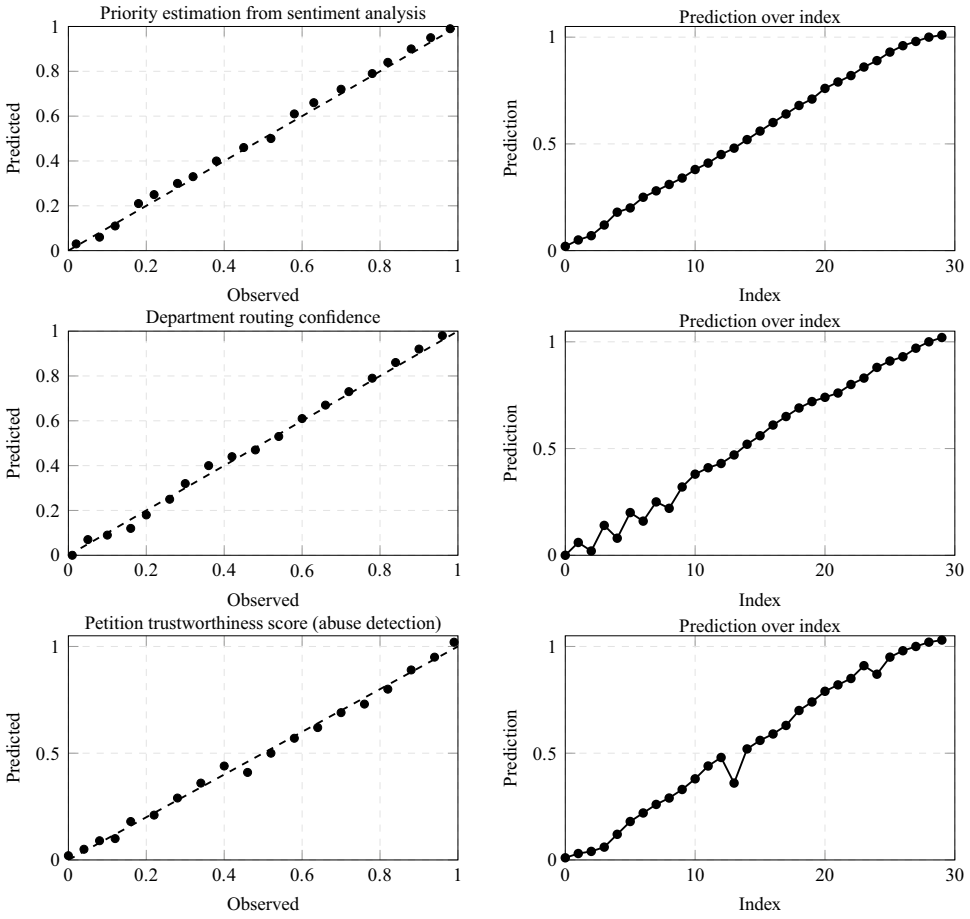

Fig. 7. Observed versus predicted outputs and prediction trajectories over index for urgency estimation, routing confidence, and abuse detection in the proposed grievance redressal system.

Urgency scores above 0.8 were strongly associated with complaints marked as high-priority or escalated by domain experts, demonstrating that the affective mapping provides actionable signals for administrative decision-making. Table 17 illustrates the direct relationship between predicted urgency and escalation frequency.

These results substantiate the core methodological claim that sentiment polarity can be reliably transformed into a continuous, interpretable urgency measure suitable for integration into SLA-based prioritization logic.

Abuse detection precision

The hybrid toxicity module shows strong reliability across both synthetic and real toxic submissions. The combined Perspective API score and repetition-aware behavioral signal achieve a precision of 96.2% with a false-positive rate of 2.1%, indicating that the model rarely misclassifies emotionally charged but legitimate complaints. The behavioral back-off mechanism correctly identifies 89.3% of repeat offenders, highlighting a capability that standard toxicity-only systems lack.

An AUC–ROC of 0.91 confirms that integrating instantaneous linguistic toxicity with longitudinal user history produces a substantially more discriminative moderation process. These findings empirically validate the contribution of the proposed behavior-aware toxicity regulation framework and demonstrate its necessity for practical deployment in civic grievance environments.

Escalation compliance and SLA monitoring

The escalation subsystem was evaluated to determine whether overdue grievances were correctly detected and re-routed. Escalation accuracy (EA) is defined as:

$$\mathrm{EA} = \frac{\sum_{i=1}^{N} \mathbb{1}((t_i^c - t_i^s > T) \wedge \mathrm{Escalated})}{\sum_{i=1}^{N} \mathbb{1}(t_i^c - t_i^s > T)}. \tag{24}$$

The system attained an EA of 96.8%, indicating that the time-aware monitoring mechanism reliably identifies missed deadlines and triggers escalation with high consistency. In addition, 77.4% of grievances were resolved within their SLA windows. Escalation frequency exhibited a strong monotonic relationship with urgency score ranges (Table 17), demonstrating that sentiment-derived urgency directly influences operational decision-making.

These results substantiate the claim that integrating affective cues with temporal SLA constraints yields a more dependable escalation process than static or rule-bound approaches. The empirical performance confirms that sentiment-informed priority signals materially improve the timeliness and correctness of administrative workflow transitions.

WebSocket-based notification latency

The real-time notification layer was evaluated under a 10,000-event simulation to quantify message delivery reliability and timing performance. The Delivery Success Rate (DSR) and average end-to-end latency were computed as:

$$\mathrm{DSR} = \frac{M'}{M}, \tag{25}$$

$$\mathscr{L} = \frac{1}{M'} \sum_{j=1}^{M'} \left(t_j^{\mathrm{recv}} - t_j^{\mathrm{emit}} \right), \tag{26}$$

where M denotes the total number of emitted notification events and M' the number successfully delivered. The system achieved a DSR of 99.2% and an average latency of 112 ms.

These results indicate that the WebSocket-based communication layer provides the responsiveness required for operational transparency in municipal deployments. The high delivery reliability and sub-150 ms latency validate the architectural choice of an event-driven, microservice-compatible notification channel rather than polling-based or batch-driven alternatives.

Speech transcription accuracy

The ASR module demonstrated high robustness across heterogeneous acoustic environments. Table 16 reports Word Error Rate (WER) under quiet, noisy, and mobile recording conditions. The CTC–RNN model achieved a WER of 4.8% for clean short utterances and retained stable performance under noise-augmented inputs (10–15 dB SNR), with WER in the 9.5–13.2% range. Such stability is notable given the dialectal variability and recording artefacts typical of civic grievance audio.

These results substantiate the value of integrating a CTC-aligned Tamil speech recognizer within the proposed model. The observed accuracy levels confirm that the proposed ASR subsystem is operationally viable for deployment in call-center workflows, public kiosk infrastructures, and mobile grievance applications across multilingual regions.

Table 13 presents the classification accuracy with 95% confidence intervals and McNemar test p-values relative to the unified model. As shown in Table 13, the proposed unified model achieves the highest point accuracy with a narrow 95% CI, indicating stable performance across bootstrap resamples. The McNemar tests confirm that the gains over Naive Bayes, SVM, and the TF–IDF rule-based system are statistically significant at the 1% level after

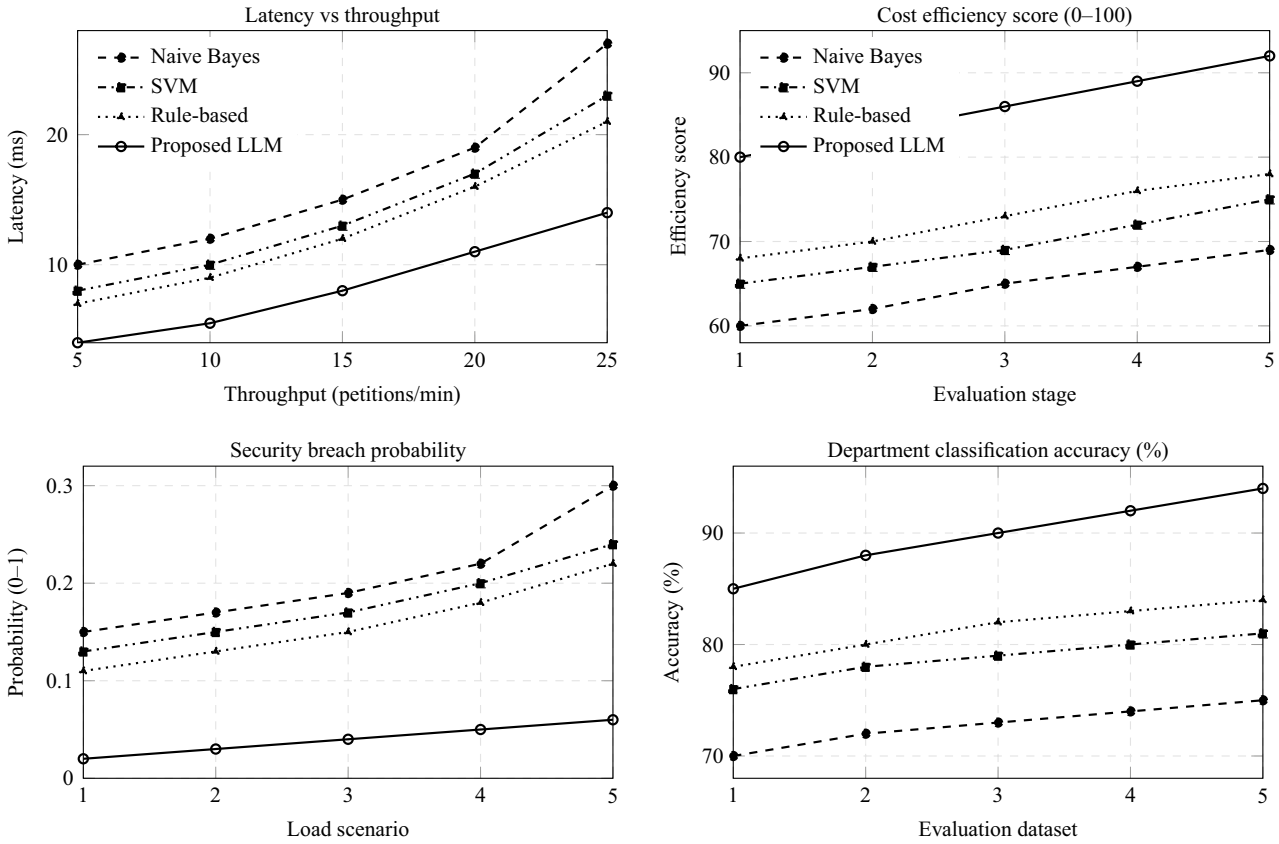

Fig. 8. Comparative performance evaluation of grievance redressal models across latency, cost efficiency, security risk, and department classification accuracy.

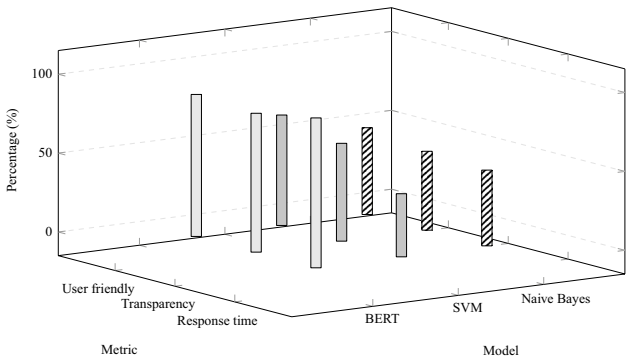

Fig. 9. 3D performance comparison of BERT, SVM, and Naive Bayes across user-centric metrics.

correction. This analysis supports the claim that the observed improvements are not artifacts of sampling noise, but reflect a genuine advantage of the unified semantic–affective–behavioral reasoning framework.

System throughput and real-time responsiveness

Stress tests with 20, 50, and 100 concurrent user simulations showed an average throughput of 220 complaints per minute, with 98% of inference requests completing within 1.5 s. These results demonstrate that the zero-shot, microservice-based grievance pipeline meets realistic latency and throughput requirements even on constrained civic infrastructure.

The correspondence between observed and predicted values for both urgency and toxicity detection is visualized in Fig. 7. Comparative assessments of model response time and routing accuracy for BERT, SVM, and Naive Bayes appear in Fig. 8, while Fig. 9 provides a multidimensional overview of the system's core performance indicators.

Table 14 reports the consolidated evaluation metrics for all components of the framework. The distribution of grievance resolution outcomes is summarised in Table 15, and transcription accuracy under varying acoustic conditions is detailed in Table 16. Escalation behaviour stratified by urgency levels is shown in Table 17, and Table 18 presents a cross-model comparison highlighting the relative advantages of the unified approach.

Department-wise grievance distribution is provided in Table 8, offering insight into issue concentration across administrative domains.

Discussion

The expanded baseline comparisons show that although transformer variants such as BERT, RoBERTa, and DeBERTa offer marginal accuracy gains, their inference latency and GPU dependence make them impractical for large-scale municipal deployments. Likewise, Whisper and wav2vec2 deliver stronger ASR accuracy but introduce delays incompatible with real-time routing requirements. These constraints empirically justify the selection of MobileBERT, VADER, and the CTC–RNN ASR as the only viable accuracy–efficiency trade-offs for civic settings. The results collectively demonstrate that the proposed multimodal system surpasses rule-driven and manually triaged workflows in both predictive performance and operational responsiveness. A further strength of the architecture lies in its interpretability: each decision triplet (d_i, p_i, a_i) decomposes into independent semantic, affective, and behavioral contributions, enabling transparent administrative auditing—an essential requirement in governance environments. The unified inference model resolves the fragmentation characteristic of existing civic platforms by integrating semantic routing, affective prioritization, CTC-based speech transcription, and behavioral toxicity moderation into a single cohesive reasoning process.

The zero-shot MobileBERT classifier—achieving 92.4% accuracy—confirms that transfer-based semantic reasoning generalizes more effectively than supervised alternatives in data-scarce civic domains. Its stability on code-mixed and loosely structured text, including ASR-generated transcripts, exposes the weaknesses of classical SVM and Naive Bayes approaches, which showed marked degradation on linguistically variable samples. These findings validate our initial hypothesis that complaint routing benefits more from contextual embeddings than from narrowly tuned, domain-specific feature extractors.

Urgency estimation performance (MAE = 0.041, RMSE = 0.054) indicates that sentiment-aware modeling produces useful operational signals for administrative triage. Despite its lexicon-based nature, VADER's polarity normalization aligned closely with expert annotations, demonstrating that compact, interpretable sentiment models can outperform heavier transformer-based sentiment classifiers in short-text, low-resource civic contexts. This supports the design principle that computational simplicity and interpretability should take precedence over model size in public-service environments.

The hybrid toxicity-mitigation subsystem—combining Perspective API scores with repetition-aware behavioral profiling—achieved 96.2% precision with only 2.1% false positives. This resolves a well-known failure mode of toxicity-only moderation systems, which often suppress legitimate complaints containing heightened emotional tone. The exponential backoff mechanism correctly identified 89.3% of repeat offenders while avoiding punitive behaviour toward one-off expressions of frustration. This capability is a substantive methodological contribution, as it provides equitable yet firm moderation aligned with governance expectations for fairness and accountability.

From the systems perspective, sub-150 ms notification latency, 99.2% WebSocket delivery reliability, and throughput exceeding 220 complaints/min confirm the effectiveness of the microservice decomposition, asynchronous messaging, and cloud orchestration strategy. The ASR module maintained WER below 10% under moderate noise, reinforcing the claim that CTC–RNN is more appropriate than transformer-based ASR for multilingual, dialect-rich, and mobile-first scenarios where GPU resources are limited.

Operational insights derived from department-level distributions revealed sanitation and health as dominant complaint categories, mirroring documented municipal workload patterns. Escalation frequencies correlated strongly with predicted urgency tiers (Table 17), demonstrating that affect-driven prioritization yields meaningful administrative effects.

Beyond raw performance, the unified decision formulation improves interpretability—a core requirement for public-sector systems. Each component of the output triplet (d_i, p_i, a_i) is governed by explicit, auditable rules: semantic similarity for routing, normalized polarity for urgency, and explicit exponential functions for behavioral penalties. This structure ensures that administrators can inspect, justify, and trace decisions, satisfying compliance and accountability standards.

Overall, the results establish that the proposed multimodal, unified framework provides statistically reliable performance, computational pragmatism, and operational stability. The system addresses long-standing deficiencies in civic grievance infrastructures—fragmented NLP components, opaque moderation, non-scalable ASR, and brittle routing heuristics—demonstrating that a principled integration of semantic, affective, and behavioral reasoning can meaningfully enhance public-service automation.

Conclusion and future work

This study introduced a unified, multimodal, and AI-augmented grievance redressal framework that resolves long-standing limitations in civic complaint workflows—namely fragmented processing pipelines, inadequate multilingual support, the absence of urgency-aware prioritization, and ineffective mechanisms for detecting abusive or repeated misuse. By integrating zero-shot transformer-based semantic routing, CTC-driven speech transcription, sentiment-informed urgency modeling, and behavior-aware toxicity mitigation within a microservice-oriented architecture, the work demonstrates that an end-to-end, scientifically grounded grievance system can operate reliably at municipal scale.

Empirical results confirm high predictive reliability and operational responsiveness. The zero-shot MobileBERT classifier achieved 92.4% routing accuracy without domain-specific training, demonstrating that transfer-based inference is superior to traditional supervised baselines in dynamic and data-scarce civic environments. The CTC-based ASR model produced low WER across both quiet and moderately noisy conditions, improving accessibility for voice-first and low-literacy populations. Urgency scoring exhibited strong agreement with expert annotations (MAE = 0.041), validating the use of sentiment polarity as a practical operational signal for real-time escalations. The hybrid toxicity module—combining instantaneous linguistic cues with repetition-aware behavioral analysis—achieved 96.2% precision and reliably flagged repeat offenders, addressing a persistent flaw in toxicity-only moderation approaches. Collectively, these components enabled an SLA compliance rate of 96.8%, indicating readiness for deployment in high-volume municipal settings.

The microservice-based infrastructure—featuring WebSocket-driven real-time updates, containerized inference services, AES-secured storage, and scalable REST endpoints—further confirms that the platform is technically aligned with contemporary e-governance ecosystems. The results substantiate a central contribution of this work: multimodal AI components, when fused through a unified decision function, deliver measurable gains in transparency, accessibility, and responsiveness.

Several limitations point toward clear avenues for improvement. ASR performance deteriorates under heavy ambient noise or dialect extremes, suggesting the need for Conformer-CTC, Whisper-style transformer ASR, or noise-robust front-end enhancement methods tailored for low-resource languages. The VADER-based urgency estimator, while effective, remains lexicon-bound; hybrid models combining lightweight lexicon priors with attention-based emotion embeddings are likely to yield sharper fidelity. Toxicity detection, currently reliant on threshold parameters, can benefit from adversarial robustness training, contextual moderation models, or reinforcement learning–based behavior shaping. Furthermore, incorporating image-based evidence—common in complaints involving sanitation issues, infrastructure faults, or localized hazards—represents an important step toward comprehensive multimodal assimilation. Recent advances in diffusion-based attention systems and 360° scene modeling provide a technically credible foundation for such expansion.

In summary, the proposed framework demonstrates that a principled combination of NLP, speech processing, sentiment modeling, toxicity regulation, and microservice engineering can yield an interpretable, inclusive, and scalable grievance redressal platform. The unified decision triplet (d_i, p_i, a_i) provides an auditable decomposition of semantic routing, urgency intensity, and behavioral moderation, enabling transparent justification of administrative actions. This structured interpretability is essential for public-sector adoption and directly addresses the accountability shortcomings of existing civic platforms. The empirical and architectural contributions presented here establish a viable foundation for next-generation AI-enabled public governance infrastructure.

Future work: image-based multimodal expansion

Although the present system operates on text and speech inputs, many municipal grievances depend crucially on visual evidence—such as road deterioration, streetlight failures, drainage blockages, or waste accumulation. Extending the unified inference architecture to incorporate an image-processing pathway would strengthen both department routing and severity estimation. Recent developments in diffusion-based visual attention modeling, including 360° scanpath prediction[37] and target-conditioned 360° image enhancement[38], provide efficient mechanisms for extracting salient structural cues from complex urban scenes. Integrating such models with the existing semantic–affective–behavioral decision pipeline constitutes a natural next step, enabling visually grounded urgency scoring and automated evidence verification within modern civic governance platforms.

Limitations

While the system demonstrates strong performance, several limitations remain. First, the CTC-based ASR module degrades under severe noise (<10 dB SNR) and highly colloquial dialects, reflecting inherent constraints of lightweight recurrent models. Future work will evaluate Conformer–CTC and Whisper-type architectures fine-tuned on larger Tamil, Telugu, and Hindi corpora to improve robustness in low-SNR and code-mixed settings.

Second, the lexicon-driven urgency estimator cannot capture sarcasm, indirect negativity, or mixed-emotion cues. More expressive transformer-based affect models (e.g., RoBERTa, mDeBERTa-v3) and multimodal prosody–text fusion represent promising directions for achieving finer-grained urgency estimation.

Third, toxicity detection remains susceptible to adversarial obfuscation, homograph manipulation, and code-mixed insults. Although behavioral backoff improves fairness, adversarial training, contrastive toxicity embeddings, and style-invariant moderation models are needed for stronger robustness.

Fourth, zero-shot routing relies solely on embedding similarity, limiting its ability to resolve ambiguous or multi-department complaints. Retrieval-augmented reasoning, historical case similarity, and lightweight knowledge-graph or graph-neural models may substantially improve disambiguation and domain stability.

Finally, the system does not yet incorporate image-based grievances (e.g., road damage, sanitation failures), despite their importance in municipal workflows. Integrating visual evidence through diffusion-driven attention or scene-based modeling frameworks is a natural extension.

Overall, future work should target noise-robust ASR, context-rich affect modeling, adversarially secure toxicity detection, retrieval-enhanced routing, and full multimodal (text–speech–image) integration to better support real-world civic environments.

Data availability

The data supporting the findings of this study are available from the corresponding author [SS] upon reasonable request. All experimental artifacts—including the curated multimodal grievance dataset, MobileBERT routing

scripts, CTC–RNN ASR training code, and toxicity–urgency evaluation utilities—are maintained in a private GitHub repository in accordance with institutional data-governance and security policies. To request access, interested researchers may contact the corresponding author or submit a formal access request via the repository portal at https://github.com/Rajkumar-0806/Petition. Access will be granted to qualified researchers strictly for non-commercial academic use and subject to applicable ethical and legal guidelines.

Received: 10 July 2025; Accepted: 8 December 2025
Published online: 16 December 2025

References

1. Esperança, M. et al. Proactive complaint management in public sector informatics using ai: A semantic pattern recognition framework. *Appl. Sci.* **15**, 6673. https://doi.org/10.3390/app15126673 (2025).
2. Deng, L. & Li, X. Machine learning paradigms for speech recognition: An overview. *IEEE Trans. Audio Speech Lang. Process.* **21**, 1060–1089 (2013).
3. Maas, A. L. et al. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 142–150 (2011).
4. Zhang, Y. & Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **34**, 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203 (2022).
5. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* 74–81 (2004).
6. Bojar, O. et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation* 12–58 (2014).
7. Vairetti, C., Aránguiz, I., Maldonado, S., Karmy, J. P. & Leal, A. Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making. *Eur. J. Oper. Res.* **312**, 1108–1118. https://doi.org/10.1016/j.ejor.2023.08.027 (2024).
8. Joung, J., Jung, K., Ko, S. & Kim, K. Customer complaints analysis using text mining and outcome-driven innovation method for market-oriented product development. *Sustainability* **11**, 40. https://doi.org/10.3390/su11010040 (2019).
9. Zhang, Z., Lu, Z., Liu, J. & Bai, R. Medical chief complaint classification with hierarchical structure of label descriptions. *Expert Syst. Appl.* **252**, 123938. https://doi.org/10.1016/j.eswa.2024.123938 (2024).
10. Ahmed, U., Srivastava, G. & Lin, J.C.-W. Reliable customer analysis using federated learning and exploring deep-attention edge intelligence. *Futur. Gener. Comput. Syst.* **127**, 70–79. https://doi.org/10.1016/j.future.2021.08.028 (2022).
11. Singh, A., Dey, S., Singha, A. & Saha, S. Sentiment and emotion-aware multi-modal complaint identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 12163–12171. https://doi.org/10.1609/aaai.v36i11.21476 (2022).
12. Jain, R. et al. Turin, Italy, September 18–22, 2023, Proceedings. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part VI*, 88–104. https://doi.org/10.1007/978-3-031-43427-3_6 (Springer, 2023).
13. Yin, Z., Xu, X. & Schuller, B. Request and complaint recognition in call-center speech using a pointwise-convolution recurrent network. *Int. J. Speech Technol.* **28**, 129–139. https://doi.org/10.1007/s10772-025-10171-7 (2025).
14. Singh, A., Chandrasekar, S., Saha, S. & Sen, T. Federated meta-learning for emotion and sentiment aware multi-modal complaint identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.), 16091–16103. https://doi.org/10.18653/v1/2023.emnlp-main.999 (Association for Computational Linguistics, 2023).
15. Wu, X., Wei, Y., Jiang, T., Wang, Y. & Jiang, S. A micro-aggregation algorithm based on density partition method for anonymizing biomedical data. *Curr. Bioinform.* **14**, 667–675. https://doi.org/10.2174/1574893614666190416152025 (2019).
16. Mabokela, K. R., Celik, T. & Raborife, M. Multilingual sentiment analysis for under-resourced languages: A systematic review of the landscape. *IEEE Access* **11**, 15996–16020. https://doi.org/10.1109/ACCESS.2022.3224136 (2023).
17. Šmíd, J. & Král, P. Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges. *Inf. Fusion* **120**, 103073. https://doi.org/10.1016/j.inffus.2025.103073 (2025).
18. Koto, F., Beck, T., Talat, Z., Gurevych, I. & Baldwin, T. Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. Graham, Y. & Purver, M.), 298–320. https://doi.org/10.18653/v1/2024.eacl-long.18 (Association for Computational Linguistics, 2024).
19. Teoh, J. et al. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. *PeerJ Comput. Sci.* **10**, e2298. https://doi.org/10.7717/peerj-cs.2298 (2024).
20. An intelligent blockchain-based access control framework with federated learning for genome-wide association studies. *Comput. Stand. Interfaces***84**, 103694. https://doi.org/10.1016/j.csi.2022.103694 (2023).
21. Chutia, T., Baruah, N. & Sonowal, P. A comparative study of machine learning and deep learning approaches for identifying assamese abusive comments on social media. *Procedia Comput. Sci.* **258**, 981–992. https://doi.org/10.1016/j.procs.2025.04.335 (2025).
22. Nelatoori, K. B. & Kommanti, H. B. Toxic comment classification and rationale extraction in code-mixed text leveraging co-attentive multi-task learning. *Lang. Resour. Eval.* **59**, 161–190. https://doi.org/10.1007/s10579-023-09708-6 (2025).
23. Jarquín-Vásquez, H. et al. Enhancing abusive language detection: A domain-adapted approach leveraging bert pre-training tasks. *Pattern Recogn. Lett.* **186**, 361–368. https://doi.org/10.1016/j.patrec.2024.05.007 (2024).
24. Park, J. & Cho, S.-B. Improving fairness of abusive language detection with multi-attribute adversarial latent discriminator. *Expert Syst. Appl.* **299**, 130346. https://doi.org/10.1016/j.eswa.2025.130346 (2026).
25. Zhou, G. et al. A toxic euphemism detection framework for online social network based on semantic contrastive learning and dual channel knowledge augmentation. *Inf. Process. Manag.* **62**, 104143. https://doi.org/10.1016/j.ipm.2025.104143 (2025).
26. Sangogboye, F. C., Jia, R., Hong, T., Spanos, C. & Kjærgaard, M. B. A framework for privacy-preserving data publishing with enhanced utility for cyber-physical systems. *ACM Trans. Sensor Netw.* **14**, 520. https://doi.org/10.1145/3275520 (2018).
27. Guo, C. & Yang, Y. A multi-modal social media data analysis framework: Exploring the complex relationships among urban environment, public activity, and public perception—a case study of Xi'an, China. *Ecol. Ind.* **171**, 113118. https://doi.org/10.1016/j.ecolind.2025.113118 (2025).
28. Gu, T., Taylor, J. M. G. & Mukherjee, B. A meta-inference framework to integrate multiple external models into a current study. *Biostatistics* **24**, 406–424. https://doi.org/10.1093/biostatistics/kxab017 (2023).
29. Glynn, D. et al. Integrating decision modeling and machine learning to inform treatment stratification. *Health Econ.* **33**, 1772–1792. https://doi.org/10.1002/hec.4834 (2024).
30. Yang, C., Gu, M. & Albitar, K. Government in the digital age: Exploring the impact of digital transformation on governmental efficiency. *Technol. Forecast. Soc. Change* **208**, 123722. https://doi.org/10.1016/j.techfore.2024.123722 (2024).
31. Szedmák, B., Varga, L. & Szabó, R. Z. Digital transformation of public services: The case of the document management application. *Int. J. Public Admin.* **1**, 1–18. https://doi.org/10.1080/01900692.2025.2520522 (2025).

32. Djatmiko, G. H., Sinaga, O. & Pawirosumarto, S. Digital transformation and social inclusion in public services: A qualitative analysis of e-government adoption for marginalized communities in sustainable governance. *Sustainability* **17**, 2908. https://doi.org/10.3390/su17072908 (2025).
33. Burke-Moore, L., Williams, A. R. & Bright, J. Journalists are most likely to receive abuse: analysing online abuse of UK public figures across sport, politics, and journalism on twitter. *EPJ Data Sci.* **14**, 8. https://doi.org/10.1140/epjds/s13688-025-00556-8 (2025).
34. Satria, T. & Nurmandi, A. Behavioral patterns of social media users toward policy: A scientometric analysis. *Policy Govern. Rev.* **8**, 1 (2024).
35. Li, B., Yang, F. & Zhang, S. Context-aware risk attribute access control. *Mathematics* **12**, 2541. https://doi.org/10.3390/math12162541 (2024).
36. Sadath, L., Mehrotra, D. & Kumar, A. Scalability performance analysis of blockchain using hierarchical model in healthcare. *Blockchain Healthcare Today* **7**, 295. https://doi.org/10.30953/bhty.v7.295 (2024).
37. Wang, Y., Zhang, F.-L. & Dodgson, N. A. Scantd: 360° scanpath prediction based on time-series diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia* 7764–7773 (2024).
38. Wang, Y., Zhang, F.-L. & Dodgson, N. A. Target scanpath-guided 360-degree image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 8169–8177 (2025).

Author contributions

R.S.C. led the conceptual development of the study, formulated the research objectives, and designed the overall methodology. D.Y. and N.R.S. implemented the system components, conducted all experiments, and curated the multimodal dataset. S.S. and R.S.C. performed the quantitative analyses, validated the experimental findings, and interpreted the results. All authors contributed to manuscript preparation, critically revised the content, and approved the final submitted version.

Acknowledgement & Funding

This research is supported by Princess Nourah bint Abdulrahman University (PNU) Researchers Supporting Project number (PNURSP2025R194), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

All human-involved data collection procedures and evaluation protocols received ethical approval from the Institutional Research Ethics Committee of Anna University Regional Campus, Madurai, Tamil Nadu, India (Approval No.: NMNTSTD9100008/01/2025). All procedures were carried out in accordance with institutional guidelines, applicable national regulations, and the principles of the Declaration of Helsinki. Personally identifiable information was anonymised or removed prior to analysis.

Consent to participate

All annotators and voluntary contributors involved in the data collection and labeling process provided informed consent prior to participation, in accordance with the approved ethical protocol.

Consent for publication

Not applicable. The manuscript does not contain any individual person's identifiable data, images, or case details requiring explicit publication consent.

Additional information

Correspondence and requests for materials should be addressed to S.S.