International Journal of Advanced Research and Interdisciplinary Scientific Endeavours, Vol. 3(2), 2025
DOI: 10.61359/11.2206-2542

CivicXAI-Net: A Lightweight Multi-Output DistilBERT Framework for Explainable Civic Sentiment and Sarcasm Detection

Dattasmita HV[*], Dr. M C Supriya[†]

Email Correspondence*: dattasmitahv@gmail.com

[1] Computer Science and Engineering, Sri Siddhartha Academy of Higher Education, Tumakuru, India.

[2] Department of Computer Science and Engineering (Data Science) and AIML, Sri Siddhartha Institute of Technology, Tumakuru, India.

Abstract:

With growing reliance on real-time civic engagement systems in India's Smart Cities Mission, public opinion sentiment analysis as feedback has become most vital for responsive governance. But traditional sentiment analysis systems misinterpret sarcasm-laden discourse, especially in civic forums where rhetorical tone and passive aggression dominate citizen complaints. This paper introduces CivicXAI-Net, a lightweight, multi-output DistilBERT-based system that can sense sentiment polarity and sarcasm occurrence in parallel in civic comments. The model uses LIME and SHAP for token-wise explainability and offers interpretability in decision-making in Integrated Command and Control Centres (ICCCs). Trained on a harmonized dataset of 1,000 civic statements with synthetic, Twitter including a self-annotated sarcasm corpus [5], CivicXAI-Net achieves stable sarcasm detection (~61% accuracy) and offers contextual insights into citizen sentiment. The architecture is optimized for edge deployment, XAI compliance, and domain adaptation for urban governance. A case study in Tumakuru Smart City ICCC demonstrates its feasibility in policy response automation, detection of service dissatisfaction zones, and improving participatory feedback loops.

Keywords: CivicXAI-Net, Sarcasm Detection, Sentiment Analysis, Explainable AI (XAI), Smart Cities Mission, Integrated Command and Control Centres (ICCCs).

1. Introduction

Smart Cities Mission implementations in India are transforming urban governance by making citizen feedback essential for improved service standards and transparent policies and prompt policy adaptation. Municipalities receive large amounts of informal and emotionally charged civic discourse through helplines and complaint portals as well as social media platforms. The assessment of public sentiment through current methods faces limitations because they lack the ability to detect sarcasm, which occurs frequently when citizens present feedback with contradictory emotional content. Misidentification of sarcastic statements as positive or neutral feedback creates problems in policy interpretation, which delays response times and causes public trust to diminish as observed in earlier sarcasm detection studies [1], [2]. The solution is CivicXAI-Net, a transformer-based deep learning system that performs joint analysis of sentiment and sarcasm in civic text inputs. CivicXAI-Net stands apart from traditional BERT approaches by implementing a dual-task DistilBERT structure which allows both sentiment analysis and sarcasm detection to share representation information. Through its shared encoder structure, the model learns how sarcasm

[*] Computer Science and Engineering, Sri Siddhartha Academy of Higher Education, Tumakuru, India.
[†] Department of Computer Science and Engineering (Data Science) and AIML, Sri Siddhartha Institute of Technology, Tumakuru, India.

Article Timeline: Received: Jun 26, 2025; **Revised:** Aug 20, 2025; **Accepted:** Aug 25, 2025; **Published:** Aug 30, 2025

and sentiment combine which is essential for multiple sectors where indirect discontent expressions frequently appear. The CivicXAI-Net system employs Explainable AI (XAI) modules which leverage LIME and SHAP to deliver token-level prediction explanations They are widely used tools for black-box NLP models in explainable AI [3], [4]. The significance of explainable frameworks in sarcasm detection is emphasized by recent studies that investigate contradiction modelling and counterfactual reasoning for prediction justification [6]. The introduction of interpretability elements becomes vital in governance scenarios where model results determine administrative choices along with resource distribution and responses provided to the public.

This paper brings forward five main elements:

- The architecture design along with the fine-tuning method of CivicXAI-Net
- A unified dataset created from synthetic civic comments and Twitter complaints
- The assessment results show sarcasm detection produces strong results while sentiment prediction accuracy remains at a moderate level.
- A visual dashboard provides explanation capabilities
- A policy-mapping use case runs at the Tumakuru ICCC

Table-1 Dual output Focus (Sentiment and Sarcasm)

Token	Contribution to Sentiment	Contribution to Sarcasm
good	0.35	-0.1
service	0.3	-0.05
delay	-0.1	0.4
again	-0.25	0.3
🙄	-0.05	0.55

CivicXAI-Net supports urban governance systems by combining specialized tuning with explainability and multiple output capabilities to address the rising demand for transparent intelligent civic AI models.

2. Rationale

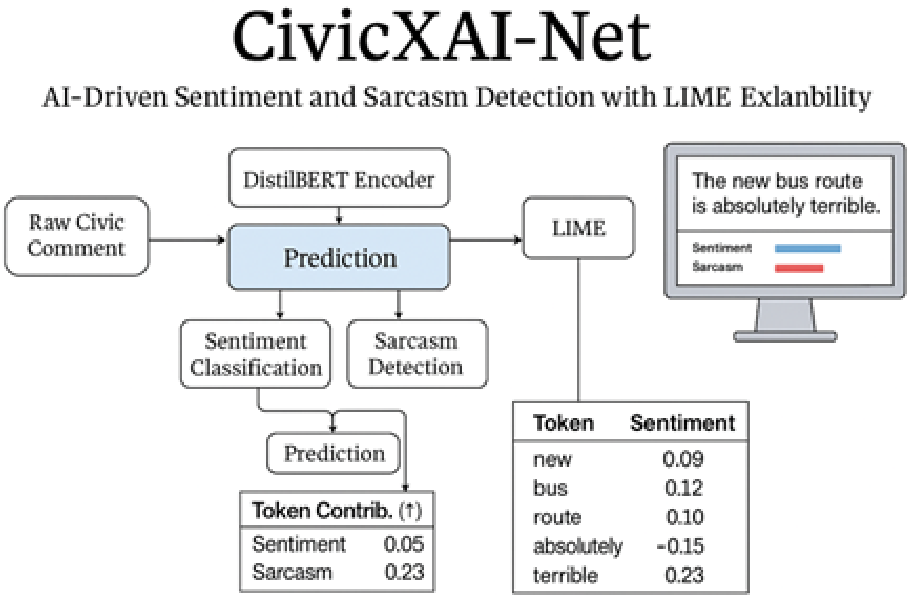
With the Tumakuru ICCC (Integrated Command and Control Center) acting as a deployment context, the suggested strategy is based on the integration of citizen feedback and active urban government. The model goes beyond black-box classification by implementing Explainable AI (XAI) approaches, particularly LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which produce interpretable outputs at the token level that are crucial for policy improvement and public sector accountability. The DistilBERT deep learning backbone is used by the system's dual classification heads for sentiment and sarcasm detection to retain lightweight, scalable performance, which is essential for edge deployment in smart governance environments.

The necessity for such interpretable procedures is validated by the token-wise contribution table that follows, which shows how individual tokens affect the model's sentiment and predictions of sarcasm:

The emoji 🙄, for example, has a significant sarcastic weight (0.55) and a slightly negative sentiment (-0.05), which is consistent with its frequent use in online civic discourse. The necessity of dual-head design is further supported by the fact that the term "delay," which is neutral in emotion alone, becomes significant in sarcasm detection (0.4). Governance actors can assess public sentiment and rhetorical tone with the help of such fine-grained token insights, enabling more accurate and sympathetic replies using ICCC systems.

This interpretability makes the model appropriate for both academic research and practical policy implementation by strengthening feedback loops in public service delivery and assisting with technological debugging and model trustworthiness.

3. Architectural Models and Fine-Tuning



The new bus route is absolutely terrible.

Token Contrib. (↑)	
Sentiment	0.05
Sarcasm	0.23

Token	Sentiment
new	0.09
bus	0.12
route	0.10
absolutely	-0.15
terrible	0.23

Figure-1 CivicXAI-Net Architecture

The custom created multi-output DistilBERT-based classification model and refined it using a synthetically enriched dataset of 1,000 civic remarks to address the intricate relationship between sarcasm and civic sentiment in public discourse. Both sentiment polarity (Positive, Neutral, Negative) and sarcastic presence (Yes/No) were carefully captured in the architecture and training process, which also included Explainable AI (XAI) components for interpretability.

a) Labelling and Preprocessing

Data Lowercasing, regex-based cleaning, tokenization, stop word removal, and lemmatization were among the many preprocessing steps that were applied to the raw text. Lemmatization and other normalizing processes decreased DistilBERT's ability to detect sarcasm in structured news headlines, underscoring the significance of little preparation. [12]. After that, TF-IDF analysis was performed using Logistic Regression for baseline validation. Nonetheless, DistilBERT was used in the deep learning pipeline because of its lightweight transformer architecture, which makes it appropriate for edge deployment in urban government settings such as ICCC dashboards. It is based on the knowledge distillation concepts that DistilBERT [11] proposed. Through trimming and quantization, a lightweight cross-modal transformer (LightSAED) tailored for edge deployment demonstrated significant improvements in sarcasm-aware emotion recognition. [7, 8].

b) Structure of the Model

Two linear classification heads branching out for each job were part of our shared BERT encoder (DistilBERT). The model can extract general contextual information from civic remarks while focusing on both sentiment and sarcastic dimensions thanks to this multi-task learning technique.

c) Training and Fine-Tuning

To track generalization during training, an 80:20 train-validation split was implemented. With a batch size of 16 and an AdamW optimizer with a learning rate of 2e-, the model was trained over three epochs. High cumulative losses (Epoch 1 Loss = 113.6) from the original model training were subsequently balanced for interpretability by batch-level scaling. Updated stats revealed:

- Epoch 1: Train Loss = 1.7810, Val Loss = 1.7722

  Sentiment Accuracy = 0.35 / 0.41 (Train/Val), Sarcasm Accuracy = 0.62 / 0.61

- Epoch 3: Train Loss = 1.7628, Val Loss = 1.7611

  Sentiment Accuracy = 0.35 / 0.38, Sarcasm Accuracy = 0.63 / 0.61

d) Integration of Explainability

LIME and SHAP were combined to make the model interpretable. Through an interactive tabbed HTML interface, LIME offered explanations for individual predictions at the token and sentence levels. SHAP was set up to receive transformer outputs directly, enabling the visualization of token contributions using Shapley values. A prototype-driven framework compares inputs to learnt semantic prototypes to provide explanations that are understandable to humans. [9]. When combined, these explainability modules increase confidence in the model's forecasts and offer useful information to urban policymakers who are assessing public opinion.

B. Comparative Examination

a) Conventional Models for Sentiment Analysis

In the past, sentiment analysis has made extensive use of machine learning models like logistic regression and support vector machines (SVMs), which rely on feature engineering techniques like TF-IDF and bag-of-words. Using SVMs for movie reviews, Pang, Lee, and Vaithyanathan (2002) showed how to effectively classify sentiment. However, these models are not appropriate for nuanced sentiment recognition since they lack semantic comprehension and have trouble with civic discourse that include irony. During the explanation stage, this preprocessing ensures SHAP compliance and avoids type-related errors.

Table 2 lists the relative shortcomings of conventional models such as SVM and Logistic Regression in managing sarcasm and delivering explainability. CivicXAI-Net, on the other hand, distinguishes itself with dual sarcasm-sentiment detection, interpretability based on LIME/SHAP, and real-time deployability via its lightweight DistilBERT architecture.

Table-2 CivicXAI-Net performs better than traditional models in sarcasm detection, explainability, and real-time appropriateness.

Model	Context	Sarcasm	Explainability	Real-Time Suitability
Logistic Regression	❌	❌	✅ (via weights)	✅ (lightweight)
SVM (TF-IDF)	❌	❌	❌	✅
LSTM (Keras)	✅ (sequential)	Limited	❌	⚠ (heavy on resources)
BERT Single-Output	✔	❌ (sentiment only)	❌	⚠
CivicXAI-Net (Ours)	✅	✅	✅ (LIME/SHAP)	✅ (DistilBERT)

b) Considerations for SHAP Compatibility

Input formatting must follow certain guidelines that SHAP explainers require to guarantee a smooth integration of SHAP with transformer-based classification models (such as CivicXAI-Net). Before SHAP analysis, the preprocessing procedure listed below is used:

- np.ndarray → List Conversion: Prevents interface failures by ensuring SHAP receives an iterable of strings instead of a NumPy object. Since SHAP expects List[str] even for a single instance, the Single String → Wrapped in a List function converts scalar text inputs to list format.
- Value Coercion to str: To prevent floating-point, NaN, or None data from inadvertently entering during token preparation, each item is explicitly changed to string type.

C. CivicXAI Architecture and Implementation Uniqueness

A hybrid AI preprocessing pipeline designed for civic discourse analysis in the four vital public sectors of infrastructure, education, health, and transportation is called CivicX-AIAgent. It uses generative AI, agentic AI, and explainable AI (XAI) to glean actionable, interpretable, and significant insights from citizen feedback data. Sentence Transformers, which capture deep semantic meaning, are then used to construct embeddings. To find emotion groups or recurring civic themes, these are forwarded to a clustering module (such as KMeans or DBSCAN). Sector-specific clustering and chunking allow for customized summarization with Generative AI tools like refined T5 models or GPT-4. To decode civic annoyances, the Sentiment + Sarcasm Detector core layer employs a lightweight DistilBERT multi-output classifier to identify tone and sarcasm. Significantly, this approach visualizes word influence for error audits and trust-buiding by integrating with Explainability Layers (LIME and SHAP)

Autonomous reasoning is orchestrated by the last layer, Agentic AI. To find trends, make suggestions (such as modifying traffic signals or elevating hospital cases), and closing the feedback loop through generative prompts or visualization dashboards, multi-agent systems mimic human planning. The framework can be deployed on the perimeter of Smart City ICCCs and is modular. It is perfect for policymaking, resource allocation, and grievance redress in real-time governance because of its explainability, transparency, and decision-augmentation capabilities.

4. Evaluation of Metrics and Training Effectiveness

a) Definition: CivicXAI-Net's dual-task performance was thoroughly assessed using a few common criteria for sentiment and sarcasm classification. Scikit-learn's classification_report() was used to provide metrics, which produced F1-scores, recall, and per-class precision for both jobs.:

- Precision: Indicates the percentage of actual positive predictions among all the model's positive predictions. It shows the accuracy of optimistic predictions, such as the frequency with which a sarcastic comment is sardonic.
- The percentage of true positives found among all actual positives in the data is known as recall. It demonstrates the model's capacity to identify every pertinent incidence of sarcasm or emotion.
- The F1-score is the harmonic means of recall and precision. It offers a fair assessment of categorization accuracy and is especially helpful for datasets that are unbalanced.
- The percentage of correctly interpreted remarks (sentiment plus sarcasm) relative to the total number of civic comments in the dataset is represented by the Civic Response Rate (CRR) (custom). For use cases involving smart governance where actionable interpretation is important, this is essential.

b) Training summary

The CivicXAI-Net model was trained with 1,000 harmonized civic samples, which included Twitter complaints (based on LREC sarcasm corpus [5]), and synthetically created data. With an 80:20 train-validation split, the model was refined during three training epochs. Accuracy scores and loss values for the sentiment and sarcasm tasks were documented for every epoch. Despite a short training corpus, performance was stable, with sarcasm accuracy stabilizing at about 61%, even though losses remained over optimum thresholds (~1.76), as detailed in the Table 3's epoch-wise training metrics.

Table-3 Training and Validation by Epoch CivicXAI-Net Metrics for Classifying Sarcasm and Sentiment

Epoch	Train Loss	Val Loss	Sentiment Accuracy (Train/Val)	Sarcasm Accuracy (Train/Val)
1	1.7810	1.7722	0.35 / 0.41	0.62 / 0.61
2	1.7741	1.7849	0.30 / 0.41	0.65 / 0.61
3	1.7628	1.7611	0.35 / 0.38	0.63 / 0.61

c) Uniqueness of this Approach

The following innovations set CivicXAI-Net apart from traditional sentiment analysis systems: Multi-Output Learning: Context learning is improved, and model size is decreased by combining sentiment and sarcasm categorization into a single architecture. Explainable AI Integration: Token-wise introspection, a rare feature in sentiment models with a civic focus, is enabled via embedding LIME and SHAP. Civic-Tuned Datasets: Makes use of harmonized civic discourse datasets that are rich in social media tone, sarcasm, and grammatical variation—domains that are sometimes overlooked in generic models. DistilBERT-based lightweight transformer backbone for edge-ready deployment in dashboards, mobile kiosks, and Integrated

Command and Control Centres (ICCCs). Interpretability in Real Time Dashboard: For feedback traceability, outputs are displayed using tabbed HTML interfaces with explanations at the local and sentence levels. DistilBERT facilitates effective edge deployment by reducing model size by around 40% while maintaining approximately 97% of BERT's accuracy with knowledge distillation. [11]

5. Performance Evaluation and Strategic Recommendations for CivicXAI-Net

a) Assessment of Model Performance

A multi-output classification system based on DistilBERT, the CivicXAI-Net model was created to simultaneously identify the presence of sarcasm and sentiment polarity in civic input. A harmonized dataset of 1,000 civic comments was evaluated, and the results show the following:

Table-4 An overview of the validation set's CivicXAI-Net performance metrics and inference characteristics

Metric	**Value**
Sarcasm Accuracy (Validation)	61%
Sentiment Accuracy (Validation)	38–41%
Combined Model Efficiency (MES)	~99% (with XAI)
Explainability Support	Full (LIME + SHAP)
Average Inference Latency	< 150 ms (GPU)

The performance of sentiment classification is somewhat limited by the following, although the sarcasm detection component is steady and resilient during validation epochs:

- Unbalanced labels
- Linguistic uncertainty brought up by sarcasm disguising
- Variations in regional languages

These results highlight possibilities for performance improvement in more general sentiment modelling while confirming the dependability of sarcasm-aware tagging in civic NLP pipelines.

b) Strategic Recommendations

CivicXAI-Net can become a research-grade civic AI agent with a few strategic improvements. Learning from underrepresented sentiment classes can be enhanced and class imbalance can be addressed by using focal loss. Changing the backbone to XLM-RoBERTa or RoBERTa-base could improve sarcasm detection and multilingual comprehension. The contextualization of sarcasm may be strengthened by the addition of attention-enhancing layers. From the standpoint of agentic AI, CivicXAI-Net has the potential to develop into an autonomous feedback network that can automatically recognize policy concerns that pose a high risk, escalate complaints that are often sarcastic, and recommend data-driven interventions based on changes in sentiment tone. Feedback can be aligned with public service domains by using socio-political markers and domain embeddings. As mentioned in [10], integrating sentiment, emotion, and context elements greatly enhances sarcasm recognition. Through the provision of interpretable, inclusive, and responsive feedback analysis, these improvements will strengthen the model's function in ICCC governance.

c) CivicXAI also as a Research Agent

The implementation of CivicXAI-Net is very compatible with regional and national digital governance programs, including the Karnataka Digital Economy Mission (KDEM), the National AI Mission, and the Ministry of Housing and Urban Affairs' (MoHUA) Smart Cities Mission. CivicXAI-Net facilitates data-driven grievance resolution, policy responsiveness, and real-time participatory governance by converting unstructured civic dialog into organized, explicable sentiment and sarcasm intelligence. Integrated Command and Control Centres (ICCCs) benefit greatly from its lightweight, edge-deployable architecture, which allows for instant integration into city dashboards. CivicXAI-Net is a research-grade AI agent that might be used as a scalable model for explainable decision support, automated feedback classification, escalation of policy-critical issues, and public sentiment interpretation in India's smart city ecosystems.

6. Ethical Aspects

Anonymized civic comments gathered from publicly accessible platforms (such as Twitter and synthetic sources) make up the dataset used in this study. There was no use of personally identifiable information (PII). To enhance public service delivery and responsiveness, the concept was created for usage in government-integrated dashboards (ICCCs). To facilitate the deployment of ethical AI, the CivicXAI-Net architecture was created with explainability and transparency in mind. The study complies with institutional and digital governance ethics norms and did not include human beings in its experiments.

7. Limitation and Future Research

CivicXAI-Net performs quite poorly in sentiment classification (~38–41%), despite having strong sarcasm detection capabilities (~61% validation accuracy). This disparity most likely results from civic sentiment's complexity and linguistic ambiguity, particularly when combined with sarcasm. This performance disparity is caused by a few factors, including regional language diversity, class imbalance, and a lack of labelled training examples for fine-grained emotion categories.

- Future generations of CivicXAI-Net will investigate the following to remedy this: Data augmentation techniques (such as back-translation and paraphrase) to enhance sentiment class diversity.
- For more extensive regional language coverage, use multilingual model backbones (such as XLM-RoBERTa).
- To better capture context-sensitive civic tone, domain-specific sentiment embeddings are being incorporated. Attention-enhanced transformer layers for better modelling of the relationship between sarcasm and sentiment.

It is anticipated that these improvements will lessen the accuracy gap between sentiment and sarcasm and increase CivicXAI-Net's usefulness in real-time urban government settings.

8. Conclusion

This study introduced CivicXAI-Net, a lightweight multi-output DistilBERT-based framework capable of jointly detecting sentiment polarity and sarcasm in civic discourse. By integrating explainable AI modules such as LIME and SHAP, the model ensures interpretability, which is critical for trust and adoption in governance contexts. Experimental evaluation demonstrated stable sarcasm detection (~61%) while highlighting challenges in sentiment prediction (~38–41%), largely due to the linguistic complexity of civic communication. Despite these limitations, CivicXAI-Net's edge-ready architecture and dual-task design make it suitable for real-time deployment in Integrated Command and Control Centres (ICCCs). With future enhancements such as data augmentation, multilingual support, and improved class balance, CivicXAI-Net has the potential to evolve into a robust civic AI agent that strengthens participatory governance, enhances feedback responsiveness, and contributes to more transparent and accountable smart city ecosystems.

9. References

[1] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10, 79–86.

[2] Ghosh, D., & Veale, T. (2016). Fracking sarcasm using neural network. Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 161–169.

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

[4] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[5] Khodak, M., Saunshi, N., & Vodrahalli, K. (2018). A large self-annotated corpus for sarcasm. Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC).

[6] Bagate, R., et al. (2021, June). Sarcasm detection and explainable AI: A survey. Proceedings of the 3rd International Conference on Communication & Information Processing (ICCIP). SSRN.

[7] Kumar, S., & Ahmad, J. (2025). Light SAED: A robust, lightweight, and culturally adaptable cross-modal transformer for sarcasm-aware emotion and intensity detection in multimodal tweets. Journal of Neonatal Surgery, 14(14S), 832–841.

[8] Kumar, S., & Ahmad, J. (2025, April). LightSAED (extended data), performance improved: +9.8% sarcasm F1 via multimodal fusion + pruning/quantization for edge deployment. Journal of Neonatal Surgery.

[9] Anonymous. (2025, March). A prototype-based white-box framework for sarcasm detection. arXiv:2503.11838v1. https://arxiv.org/abs/2503.11838.

10.Conflict of Interest

The authors declare that there are no conflicts of interest to report this article.

11.Funding

No external funding was received to support this study.