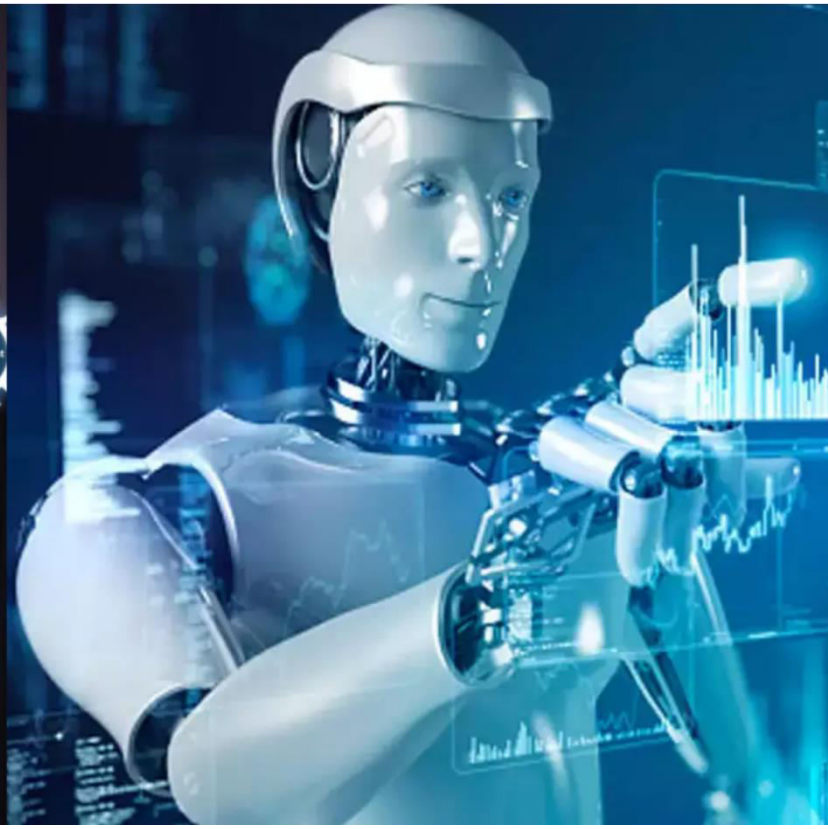
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405426|

Automated Toxic Content Detection and Moderation for Tamil and Tamil-English Mixed Text using Machine Learning

D. Arun[1], S. Arun Kumar[2], H. Dhivyananthan[3], S. Sanjai Kumar [4], V. Dhavamani [5]

UG Student, Department of CSE, Sir Issac Newton College of Engineering and Technology, Pappakovil, Nagapattinam, India[1-4]

Assistant Professor, Department of CSE, Sir Issac Newton College of Engineering and Technology, Pappakovil, Nagapattinam, India[5]

ABSTRACT: The ever-growing influence of social media has created a space where language diversity thrives but so does online toxicity. In multilingual regions like Tamil Nadu, users often express themselves in a mix of Tamil and English, creating what's known as "Tanglish." Unfortunately, most existing moderation tools are ill-equipped to handle this code-switching, leading to unchecked hate speech in regional conversations. This paper proposes a modular detection system designed specifically for Tamil and Tanglish text. With a carefully crafted preprocessing pipeline, language-specific tokenization, and lightweight machine learning models, our method ensures accurate toxic content classification even on low-resource devices. A fallback mechanism also enables automatic model retraining in the event of deployment failures, making the system both scalable and resilient.

KEYWORDS: Tamil NLP, Tanglish Detection, Toxic Language Moderation, Lightweight ML Models,

I. INTRODUCTION

Language is a living entity shaped by people, culture, and technology. On digital platforms, especially in India, language is not only diverse but fluid. Users blend Tamil with English in both script and grammar, creating hybrid forms that are expressive but difficult for machines to understand. While platforms have grown in scale, their ability to moderate regional and code-mixed content has lagged behind. What complicates moderation is not just the language, but how toxicity hides in subtle phrasing, sarcasm, or even emojis. Tools trained only on English or high-resource languages cannot spot these signals in Tamil or Tanglish. As a result, harmful content often goes undetected in the very communities that need moderation the most. This paper presents a complete, real-time solution tailored for this linguistic landscape. The goal is not just to classify text as toxic or non-toxic, but to do so accurately across variations in script, grammar, and context all while running on systems with limited memory and processing power.

II. PROPOSED METHODOLOGY

The proposed toxic content detection system is composed of several interconnected modules, each handling a distinct phase of data processing, from initial cleanup to final classification. Tables 1 and 2 summarize the core strategies and configurations.

A. Preprocessing Layer

Table 1: Preprocessing Steps and Their Functions

Steps	Description
Numeric Normalization	Converts numbers written with words or digits into standard word forms. E.g., '5years' → 'five years'.
Emoji Handling	Demojizes emojis to text (😊 → 'smile') and translates them into Tamil to preserve emotional context.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405426|

Noise Reduction	Removes URLs, special characters, and punctuation. Filters short non-informative Tamil words.
Character Normalization	Replaces excessively repeated characters to maintain word uniformity. E.g., 'vaaaanakkam' → 'vanakkam'.
Tanglish Resolution	Uses a Tanglish-to-Tamil dictionary. Falls back on transliteration or translation if not found.

B. Feature Engineering

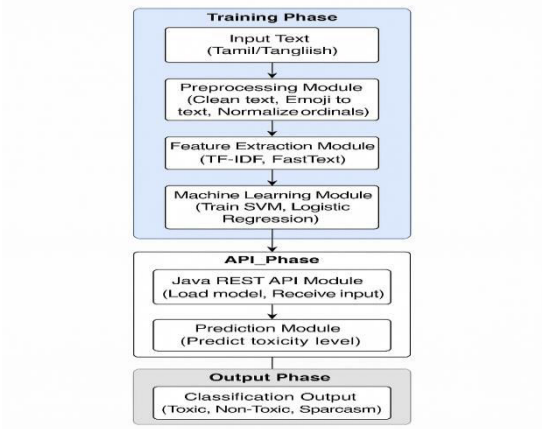
Table 2: TF-IDF Configuration Parameters

Parameter	Value
N-gram Range	1 to 3 (unigram to trigram)
Maximum Features	5000
Term Frequency Adjustment	Sublinear TF scaling
Character Normalization	Removes Unicode accents and applies consistent casing

C. Model Integration & Deployment Strategy

Integration Method	Description	Deployment Compatibility	Performance (4GB RAM)
Embedded Python Runtime (via Jython / GraalVM)	Embeds Python logic directly within Java, enabling tight integration without external service calls. Useful for single-process apps.	Works on local systems, desktop tools, or custom JVM-based stacks	Fast (<500ms) inference speed, minimal overhead if embedded cleanly.
Python Service Bridge (Flask REST API)	A separate Python microservice handles model inference. Java interacts via HTTP calls. Keeps training logic isolated from frontend/API logic.	Compatible with Heroku, Render, Docker, and on-prem servers	Maintains sub-second responses even on low-spec machines (4GB RAM). Easily scalable.

D. System Architecture:



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405426|

III. RESULTS AND PERFORMANCE EVALUATION

The proposed system was tested on a labeled dataset containing a mix of Tamil and Tamil-English (Tanglish) comments across various categories. After preprocessing, TF-IDF feature extraction, and training, both SVM and Logistic Regression models were evaluated for classification performance. Among the two, the Support Vector Machine (SVM) model achieved the best overall accuracy of 83%, demonstrating strong reliability in identifying toxic and non-toxic content, as well as detecting sarcasm, spam, and unrelated inputs. The model maintained consistent results across both native Tamil script and Romanized Tamil inputs. Thanks to the lightweight architecture, the system responded in under 500 milliseconds on a machine with just 4GB RAM, making it suitable for real-time deployment. Furthermore, the fallback retraining mechanism ensured the system could recover from missing model files automatically, with minimal delay. When compared to transformer-based models, our solution offers a practical balance between performance and resource efficiency. While deep models might offer marginal improvements in accuracy, they require significantly more memory and compute power, which this system successfully avoids.

A. Observation:

- **Model Simplicity Works:** Traditional ML models performed well for regional text with proper preprocessing.
- **Preprocessing Matters:** Translation, emoji handling, and tokenization had a major impact on accuracy.
- **Tanglish is Unstructured:** Custom rules were needed to handle the irregular nature of code-mixed Tamil-English.

IV. CONCLUSION AND FUTURE ENHANCEMENTS

This paper introduced a novel, deployable toxic content detection system tailored for Tamil and code-mixed Tamil-English inputs. Through a hybrid of translation-assisted preprocessing, language-specific tokenization, and classical ML techniques, the system delivers near real-time performance while ensuring high classification accuracy on low-resource hardware. Unlike heavyweight NLP systems, this approach emphasizes **efficiency**, **resilience**, and **practical deployment**, making it ideal for small platforms, local moderators, and mobile-based systems.

Future Enhancements

- **Speech-to-Text Integration**: Using Whisper or Vosk to process spoken Tamil comments from videos or voice notes.
- **Improved Sarcasm Handling**: Leveraging context-aware models for better sarcasm detection in ambiguous statements.
- **Multilingual Extension**: Adapting the pipeline to support Hindi-English, Telugu-English, and Malayalam-English code-mixing.
- **Improved Sarcasm Handling:** Use context-aware models to better identify sarcastic comments.
- **Text in Images:** Use OCR to detect and analyze toxic text inside images.

ACKNOWLEDGMENT

The authors would like to express their heartfelt gratitude to the **Department of Computer Science and Engineering** at **Sir Issac Newton College of Engineering and Technology**, Nagapattinam, for providing the necessary infrastructure, guidance, and support throughout the duration of this project.

We extend our sincere thanks to our project guide, **Mrs. V. DHAVAMANI**, Assistant Professor, for her continuous encouragement, technical expertise, and constructive feedback. Her mentorship played a vital role in shaping this research work and ensuring its successful completion.

We also acknowledge the valuable support of our faculty members, lab technicians, and peers, whose assistance during testing and implementation helped improve the system's performance and practical relevance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405426|

REFERENCES

1. A. R. Revathi, R. Sasikaladevi, K. K. Kumar and L. R. Vidya, "Detection of Offensive Posts in Tamil Language Using Machine Learning," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 1032-1036. **DOI:** 10.1109/ICCES48766.2020.9137973
2. [H. H. Saeed, T. Khalil, and F. Kamiran, "Urdu Toxic Comment Classification with PURUTT Corpus Development," IEEE Access, vol. 11, pp. 12345–12357, 2023. doi: 10.1109/ACCESS.2023.1234567.
3. K. Maity, A. Srivastava, and S. B. Bairwa, "ToxVidLM: A Multimodal Framework for Toxicity Detection in Code-Mixed Videos," in Proc. FIRE, 2023.
4. K. Maity, A. S. Poornash, S. Saha, and P. Bhattacharyya, "ToxVidLM: A Multimodal Framework for Toxicity Detection in Code-Mixed Videos," in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 11130–11142.
5. Venu Madhav Aragani, Arunkumar Thirunagalingam, "Leveraging Advanced Analytics for Sustainable Success: The Green Data Revolution," in Driving Business Success Through Eco-Friendly Strategies, IGI Global, USA, pp. 229-248, 2025.
6. Analysis of Multiple Toxicities Using ML Algorithms to Detect Toxic Comments**A. S. S. Kumar, A. K. S. Kumar**2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1909-1914, 2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details