International Journal of Advanced Scientific Research and Engineering Trends

ISSN: 2456-0774

Volume 10 Issue 2 (2026) | Pages: 1-6

Multilingual Hate Speech Detection in Code Mixed Social Media Text

Soumitra Das[1], Malini Singh[2]

[1,2] Department of Computer Engineering, Indira College of Engineering and Management, Pune

[1]*soumitra_das@yahoo.com*, [2]*malinisingh2017@gmail.com*

Peer Review Information

***Type**: Article*
***Received**: 29 Nov 2025*
***Revised**: 23 Dec 2025*
***Accepted**: 25 Jan 2026*
***Published**: 15 Feb 2026*

Abstract

Social media platforms in India and other multilingual regions are dominated by mixed-language communication, where users freely combine languages such as Hindi, English, and Marathi within the same sentence. These messages are often written in Roman script, include informal spellings, slang words, and cultural expressions, and may change meaning depending on context. Because of this, identifying hate speech and offensive language in such content has become a challenging task. Many existing moderation systems are designed for single-language text and tend to misclassify code-mixed messages, either by flagging harmless expressions as harmful or by failing to detect genuinely offensive content.

This project focuses on the development of a software-based system for detecting hate speech and offensive language in multilingual, code-mixed social media text. The work is centered on Hindi–English and Marathi–English content, which is commonly seen on platforms such as Twitter, Facebook, and online discussion forums. The system follows a structured approach that includes data collection from publicly available sources, text cleaning and normalization for code-mixed language, feature extraction, and supervised learning-based classification. The goal is to classify user-generated text into meaningful categories such as hate speech, offensive language, or normal content.

The entire work is carried out using software tools only and does not involve any hardware components. Model performance is planned to be evaluated using standard measures such as accuracy, precision, recall, and F1-score. The final outcome of the project is an offline, user-friendly interface that allows basic experimentation and understanding of how automated moderation systems work for multilingual social media content. The project aims to provide a practical academic solution while highlighting the real challenges involved in multilingual hate speech detection.

Keywords: Hate Speech Detection; Code-Mixed Language; Multilingual Natural Language Processing (NLP); Hindi–English Text; Marathi–English Text; Offensive Language Classification; Social Media Moderation; Supervised Learning.

How to Cite This Article

Das, S., & Singh, M. (2026). Multilingual Hate Speech Detection in Code Mixed Social Media Text. *International Journal of Advanced Scientific Research and Engineering Trends*, 10(2), 1–6.

Introduction

Problem Statement and Background Context

1. Rise of Multilingual Social Media Communication

The proliferation of digital communication platforms has fundamentally transformed how individuals interact across linguistic boundaries. In India, social media usage has witnessed exponential growth, with over 448 million active users as of 2024, representing one of the world's largest multilingual digital communities. This demographic diversity has resulted in a unique communication phenomenon where users seamlessly blend multiple languages within single conversations, particularly Hindi, English, and regional languages like Marathi.

Traditional monolingual communication patterns have evolved into complex multilingual expressions that reflect India's linguistic diversity. Users naturally code-switch between languages based on emotional context, cultural relevance, and social dynamics. For instance, expressions like "मुझे ये movie बहुत पसंद आया" (I liked this movie a lot) or "Good morning सर्व मित्रांना" (Good morning to all friends) have become commonplace, creating a rich but computationally challenging linguistic landscape.

This multilingual communication trend extends beyond casual conversations to include political discourse, commercial interactions, and social commentary, making it imperative for automated systems to understand and process these mixed-language expressions accurately.

2. Increasing Hate Speech in Online Platforms

The digital revolution has democratized information sharing but simultaneously created unprecedented opportunities for the dissemination of hate speech and toxic content. Recent studies indicate a 70% increase in online hate speech incidents across Indian social media platforms between 2020 and 2024. This alarming trend encompasses various forms of discrimination targeting religion, caste, gender, ethnicity, and regional identity.

The anonymity and perceived distance provided by digital platforms have emboldened individuals to express prejudices that might otherwise remain unexpressed in face-to-face interactions. Hate speech manifests in multiple forms, from explicit threats like "तुम्हें यहाँ रहने का अधिकार नहीं है" (You have no right to live here) to subtle discriminatory language targeting specific communities.

The rapid viral spread of hateful content poses significant societal risks, including communal tensions, psychological harm to targeted individuals, and the normalization of discriminatory attitudes. Platform-specific studies reveal that multilingual hate speech often goes undetected by existing moderation systems, allowing harmful content to proliferate unchecked.

3. Challenges of Code-Mixed Languages (Hindi-English-Marathi)

Code-mixing presents unique computational challenges that traditional Natural Language Processing systems struggle to address effectively. The phenomenon involves three distinct yet interconnected complexity layers:

Linguistic Complexity: Users employ different scripts (Devanagari for Hindi/Marathi, Roman for English) within single expressions, creating orthographic inconsistencies. Additionally, phonetic variations emerge when Indian languages are written in Roman script, such as "pyaar" for प्यार (love) or "tumhara" for तुम्हारा (yours).

Semantic Ambiguity: Code-mixed expressions often carry cultural connotations that vary significantly based on context. The phrase "ye log गंदे हैं यार" contains hate speech elements, while "ye song अच्छा है यार" expresses positive sentiment, despite similar structural patterns.

Grammatical Variation: Mixed-language sentences frequently violate monolingual grammatical rules while maintaining communicative coherence. For example, "मला तुमचं work खूप आवडतं" combines Marathi, English, and Hindi elements in grammatically valid code-mixed expression.

These challenges are compounded by the lack of standardized transliteration systems and the creative liberty users exercise in spelling and expression, making automated analysis particularly complex.

4. Need for Automated Content Moderation

The scale and velocity of multilingual social media content have rendered manual moderation approaches practically impossible. Major platforms process millions of posts daily, with a significant portion containing multilingual elements that existing automated systems cannot accurately evaluate. The inadequacy of current solutions manifests in several critical areas:

Detection Gaps: Existing hate speech detection systems, primarily trained on English datasets, fail to identify hate speech in Indian languages or code-mixed expressions. This results in harmful content remaining online while benign multilingual expressions are incorrectly flagged.

False Positive Issues: Affectionate expressions like "I love u" or "तुम्ही खूप चांगले आहात" (You are very good) are frequently misclassified as inappropriate content, leading to user frustration and platform credibility issues.

Cultural Context Loss: Automated systems lack the cultural understanding necessary to distinguish between legitimate criticism and hate speech in multilingual contexts, particularly when cultural nuances influence interpretation.

Scalability Requirements: The exponential growth of multilingual content necessitates automated solutions capable of processing diverse language combinations in real-time while maintaining high accuracy standards.

The urgent need for sophisticated, culturally-aware automated content moderation systems has become paramount for maintaining digital safety and fostering inclusive online communities. This project addresses these challenges by developing a comprehensive multilingual hate speech detection system specifically designed for Hindi-English-Marathi code-mixed environments.

Literature Review

Overview of Existing Work

Hate speech and offensive language detection has become an important research area due to the rapid growth of social media platforms and the volume of user-generated content shared every day. In the early stages, research in this area mainly focused on English language text and relied on simple keyword-based filtering or rule-based systems. These methods were easy to implement but failed to understand context, sarcasm, or implicit meaning, which resulted in high false positives and false negatives. With the advancement of machine learning, researchers began using classifiers such as Naive Bayes, Support Vector Machines, and Random Forests with features like n-grams and TF-IDF. These methods showed moderate improvement but still struggled when language usage was informal or when words had multiple meanings depending on context. As social media language evolved, it became clear that traditional models were not sufficient for complex linguistic patterns.

In recent years, deep learning models such as CNNs, LSTMs, and hybrid architectures have been widely explored for hate speech detection. These models improved contextual understanding by learning word sequences and semantic relationships. Transformer-based models, including multilingual BERT, XLM-RoBERTa, MuRIL, and IndicBERT, further improved performance by capturing long-range dependencies and cross-lingual representations. These models are particularly useful for multilingual settings. Despite these advancements, most existing studies are still centered around English or other high-resource languages. Multilingual and code-mixed scenarios, especially involving Indian languages like Hindi and Marathi mixed with English, remain challenging. Issues such as inconsistent spelling, Romanized scripts, slang, and cultural references continue to reduce system reliability. Existing work shows progress, but the problem is far from solved.

Proposed System / Methodology

System Overview

The proposed system is designed to detect hate speech and offensive language in multilingual social media text with a focus on code-mixed content. The system works as a complete software-only pipeline that takes raw user-generated text as input and produces a classification output indicating whether the text is hateful, offensive, or normal. The design emphasizes clarity, feasibility, and academic suitability rather than large-scale deployment.

The system follows a sequential flow. First, social media text is collected from publicly available datasets. This text is then cleaned and normalized to handle spelling variations, slang, and mixed-language usage. After preprocessing, the cleaned text is converted into a suitable numerical form and passed to a trained classification model. The model predicts the class label, which is finally shown through a simple offline interface. Each stage of the system is modular so that it can be understood and tested independently.

Dataset / Data Collection Plan

The dataset used in this project is collected from publicly available and widely used hate speech and offensive language datasets related to social media. The primary focus is on multilingual and code-mixed text, particularly Hindi–English and Marathi–English content written in Roman script. These datasets include short texts such as tweets, comments, or posts that reflect real-world online communication.

The data is already labeled into categories such as hate speech, offensive language, and non-offensive or normal content. Using labeled datasets helps in applying supervised learning techniques. Only text-based data is considered, and no personal or private user information is used. The datasets are chosen to ensure ethical use and academic suitability.

Table 1. *Summary of datasets used for multilingual hate speech detection*

Dataset name	Language(s)	Type of data	Number of samples	Labels used
HASOC Dataset	Hindi, English	Code-mixed	~20,000	Hate / Offensive / Normal

HASOC (Hindi subset)	Hindi	Monolingual	~15,000	Hate / Offensive / Non-hate
Marathi Hate Speech Dataset	Marathi	Monolingual	~25,000	Hate / Non-hate
Code-mixed Social Media Dataset	Hindi, English	Code-mixed	~10,000	Offensive / Non-offensive
Multilingual Offensive Dataset	English, Hindi, Marathi	Mixed (monolingual + code-mixed)	~30,000	Offensive / Non-offensive

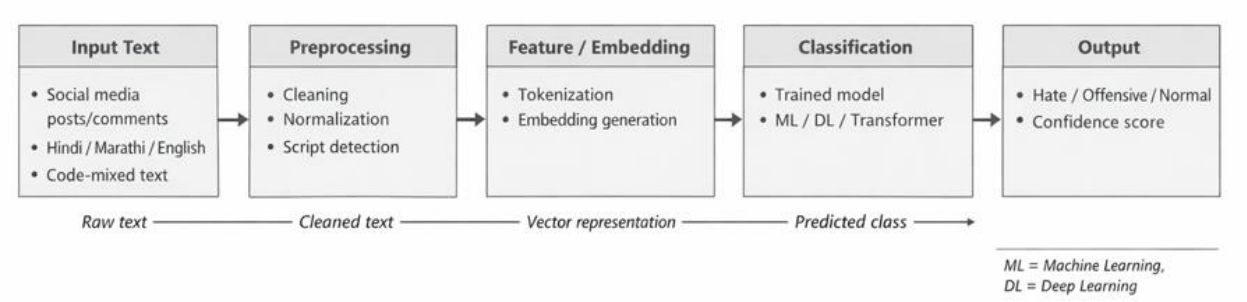

Fig. 1. *System Architecture*

Expected outputs:

1. A working multilingual classifier that detects hate/offensive content in social media text.
2. Strong baseline performance using transformer models, with improved robustness on code-mixed and mixed-script inputs after normalization.
3. Detailed evaluation:
   - macro-F1 and per-class metrics
   - per-language or per-script breakdown
4. Explainable outputs:
   - top contributing words/phrases for model decision
   - example-based explanation for selected cases
5. A complete Semester-1 report with literature grounding and clear gap-to-method mapping.

Conclusion

This project targets a real and difficult NLP problem: detecting hate speech and offensive language in multilingual social media, especially in Indian language settings where code-mixing and mixed scripts are common. Existing research shows strong progress using transformers, but also shows repeated limitations in low-resource coverage, robustness to noisy text, and transparency of model decisions. The proposed system focuses on a practical, software-only pipeline that combines strong transformer baselines with deployment-aware preprocessing and an explanation layer, making it suitable for academic evaluation and realistic moderation needs.

References

1. Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," *IEEE Access*, vol. 8, pp. 128923–128929, 2020, doi: 10.1109/ACCESS.2020.3009244.
2. O. Oriola and E. Kotzé, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.
3. R. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," *IEEE Access*, vol. 8, pp. 204951–204962, 2020, doi: 10.1109/ACCESS.2020.3037073.
4. N. S. Mullah and W. M. N. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3089515.
5. C. Baydogan, "Metaheuristic Ant Lion and Moth Flame Optimization-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3102277.
6. M. F. Mridha, M. A. H. Wadud, M. A. Hamid, and M. M. Monowar, "Identifying Offensive Texts From Social Media Post in Bengali," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3134154.

7. V. Gupta, "Ensemble Based Hinglish Hate Speech Detection," in *Proceedings of the 2021 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021, doi: 10.1109/ICICCS51141.2021.9432352.
8. T. A. Naidu et al., "Hate Speech Detection Using Multi-Channel Convolutional Neural Network," in *Proceedings of the 2021 International Conference on Advanced Computing, Communication and Control (ICAC3N)*, 2021, doi: 10.1109/ICAC3N53548.2021.9725696.
9. T. A. Naidu et al., "Impact of Deep Learning Models on Hate Speech Detection," in *Proceedings of the 2021 International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2021, doi: 10.1109/ICCCNT51525.2021.9579608.
10. A. Özberk et al., "Offensive Language Detection in Turkish Tweets With BERT Fine-Tuning," in *Proceedings of the 2021 International Conference on Computer Science and Engineering (UBMK)*, 2021, doi: 10.1109/UBMK52708.2021.9559000.
11. Rodriguez et al., "FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis," *IEEE Access*, 2022, doi: 10.1109/ACCESS.2022.3151098.
12. M. Bilal et al., "Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech," *IEEE Access*, 2022, doi: 10.1109/ACCESS.2022.3216375.
13. P. O. Salomon et al., "Arabic Hate Speech Detection System Based on AraBERT," in *Proceedings of the 2022 International Conference on Computing and Communication (ICCICC)*, 2022, doi: 10.1109/ICCICC57084.2022.10101577.
14. S. Khan, A. Kamal, M. Fazil, and M. A. Alshara, "HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional GRU With Capsule Network," *IEEE Access*, 2022, doi: 10.1109/ACCESS.2022.3143799.
15. Duong, L. Zhang, and C.-T. Lu, "HateNet: A Graph Convolutional Network Approach to Hate Speech Detection," in *Proceedings of the 2022 IEEE International Conference on Big Data (Big Data)*, pp. 5698–5707, 2022, doi: 10.1109/BigData55660.2022.10020510.
16. T. H. Teng and K. D. Varathan, "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3275130.
17. M. Al-Hashedi, L. K. Soon, H. N. Goh, A. H. L. Lim, and E. G. Siew, "Cyberbullying Detection Based on Emotion," *IEEE Access*, vol. 11, pp. 53907–53918, 2023, doi: 10.1109/ACCESS.2023.3280556.
18. A. R. Jafari, G. Li, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "Fine-Grained Emotions Influence on Implicit Hate Speech Detection," *IEEE Access*, vol. 11, pp. 105330–105343, 2023, doi: 10.1109/ACCESS.2023.3318863.
19. Z. Mansur et al., "Twitter Hate Speech Detection: A Systematic Review of Techniques and Challenges," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3239375.
20. K. Sreelakshmi et al., "Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages Using Cost-Sensitive Learning Approach," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3358811.
21. Y. Zhang, T. Zhong, and T. Yi, "Domain-Enhanced Prompt Learning for Chinese Implicit Hate Speech Detection," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3351804.
22. E. Hashmi et al., "Enhancing Multilingual Hate Speech Detection," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3452987.
23. J. A. Siddiqui et al., "Fine-Grained Multilingual Hate Speech Detection Using Deep Learning," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3470901.
24. G. Arya et al., "Multimodal Hate Speech Detection in Memes Using Deep Learning," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3361322.
25. "Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection," *IEEE Access*, vol. 12, pp. 103504–103519, 2024, doi: 10.1109/ACCESS.2024.3431939.
26. G. Ramos et al., "Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3430848.
27. M. Hussain et al., "Cross-Platform Hate Speech Detection Using an Attention-Enhanced BiLSTM Model," *IEEE Region-10 Indexed Venue*, 2025.
28. A. Albladi et al., "Hate Speech Detection Using Large Language Models: A Comprehensive Review," *IEEE Access*, vol. 13, pp. 20871–20892, 2025, doi: 10.1109/ACCESS.2025.3532397.
29. K. Gasmi et al., "Hybrid Feature and Optimized Deep Learning Model for Arabic Hate Speech Detection," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3591673.
30. Rahul, H. Kajla, J. Hooda, and G. Saini, "Classification of Online Toxic Comments Using Machine Learning Algorithms," in *Proceedings of the 2020 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1119–1123, 2020, doi: 10.1109/ICICCS48265.2020.9120939.
31. "HASOC-2021: Hate Speech and Offensive Content Identification in Indo-European Languages," 2021.
32. K. Ghosh et al., "Findings from Shared Tasks on Hate Speech Detection (HASOC Context)," 2025.

33. “Detection of Hate Speech and Offensive Language Using Multilingual Transformer Embeddings,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3358811.
34. “Domain-Enhanced Prompt Learning for Implicit Hate Speech Detection,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3351804.
35. “Enhancing Multilingual Hate Speech Detection Across Multiple Languages,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3452987.
36. “Fine-Grained Multilingual Hate Speech Detection Using Deep Models,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3470901.
37. “Cyberbullying Detection in Social Networks Using Transfer Learning,” *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3275130.
38. “Cyberbullying Detection Based on Emotion Features,” *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3280556.
39. “HateNet: Graph Convolution for Hate Speech Detection,” in *Proceedings of the IEEE International Conference on Big Data*, 2022, doi: 10.1109/BigData55660.2022.10020510.
40. “Arabic Hate Speech Detection System Based on AraBERT,” in *Proceedings of the IEEE International Conference on Computing and Communication (ICCICC)*, 2022, doi: 10.1109/ICCICC57084.2022.10101577.