IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14564

A SURVEY PAPER ON MULTILINGUAL TOXIC COMMENT CLASSIFIER

Mr. Somasekhar T[1], B S Varsha[2], Charithanjali M[3], Jyothsna R[4], Kavita R J[5]

Associate Professor, Dept of CSE, KSIT, Karnataka, India[1]

Student, Dept of CSE, KSIT, Karnataka, India[2]

Student, Dept of CSE, KSIT, Karnataka, India[3]

Student, Dept of CSE, KSIT, Karnataka, India[4]

Student, Dept of CSE, KSIT, Karnataka, India[5]

Abstract: As the online communication grows exponentially, the issue of toxic comments, varying from hate speech and cyberbullying to offensive and abusive content, has emerged as a pressing issue for social media sites, online forums, and news portals. Although a lot of headway has been achieved in the detection and moderation of toxic content in English, the task becomes more challenging in multilingual environments because of the varying linguistic frameworks, cultural environments, and the lack of sufficiently annotated datasets in most languages.

This survey article delves into the recent research in multilingual toxic comment classification, emphasizing datasets, methods, and challenges used in this process. We detail a comprehensive critique of different strategies, such as rule-based and lexicon-based methods, old-school machine learning models, and state-of-the-art deep models. Particular emphasis is on the performance of transformer-based architectures like multilingual BERT (mBERT) and XLM-RoBERTa, based on large-scale pretraining that facilitates cross-lingual competency. In addition, we address the use of cross-lingual transfer learning in overcoming low-resource language issues and the effect of code-switching and transliteration on toxicity detection.

There are still some challenges that exist despite progress, such as model and dataset biases, the absence of contextual awareness in some languages, and the dynamic nature of toxic language on the internet.

Keywords: Multilingual Toxic Comment Classification, Hate Speech Detection, Offensive Language Identification, Cross-Lingual Transfer Learning, Transformer Models, Natural Language Processing (NLP), Content Moderation.

I. INTRODUCTION

The upsurge in social media sites like Facebook, Twitter, Instagram, and YouTube has reshaped how people interact, exchange information, and voice opinions across the world. These sites have made it possible for users who belong to different linguistic, cultural, and geographical backgrounds to communicate comfortably. But this has also made it easier for poisonous content to spread, such as hate speech, harassment, cyberbullying, threats, and other offending comments that have serious psychological and social impacts. Toxic comments spread and deteriorate the online debate, marginalize vulnerable groups, and even lead to real-world violence

Automatic classification and detection of toxic content, commonly known as toxic comment classification, has thus been a key task in Natural Language Processing (NLP) and computational social science. The task entails the application of computational models to determine if a provided text has offensive or harmful language. Although a vast amount of research has been carried out for English language material, based primarily on the presence of big annotated corpora and linguistic data bases , relatively few studies target Indian regional languages. India is a multilingual nation with more than 22 languages officially recognized, of which Telugu, Tamil, Kannada, and Hindi are commonly used and found across social media channels. The semantic variety and richness of these languages pose novel challenges for toxic comment classification.

A few of the major challenges in creating effective toxic comment classifiers for Indian languages are limited access to large-scale annotated datasets, the existence of several scripts (for example, Devanagari for Hindi, Telugu script for Telugu, Tamil script for Tamil, Kannada script for Kannada), widespread code-mixing with English or other languages in informal settings, and vast dialectal variations even among a single language . These aspects make preprocessing, feature selection, and model training for toxic comment detection more complex.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025
DOI: 10.17148/IJARCCE.2025.14564

With the introduction of multilingual transformer-based models like multilingual BERT (mBERT) , XLM-RoBERTa (XLM-R), and IndicBERT over the last few years, the state-of-the-art in multilingual and cross-lingual natural language understanding tasks has greatly improved.

These models take advantage of huge pre-training on multilingual corpora, transferring cross-lingual semantics and allowing transfer learning from resource-high languages such as English to resource-poor languages such as Telugu, Tamil, Kannada, and Hindi. IndicBERT, for example, is pre-trained in particular over several Indian languages and is thus best suited for processing Indian language text.

This survey paper tries to give a thorough overview of the recent developments in toxic comment classification for these four Indian languages. It gives an overview of the datasets, model architectures, preprocessing methods, and evaluation metrics employed in this area. In addition, it gives pointers towards the major challenges, research gaps, and upcoming trends, stressing the importance of ethical, just, and culturally sensitive toxicity detection systems. By synthesizing the existing research state, the paper aims to direct future endeavors toward developing more solid and encompassing multilingual toxic framework.

II. LITERATURE SURVEY

The problem of toxic comment classification has attracted a lot of interest in recent years because of the increasing incidence of online toxicity and its ill effects on society. This section offers a general outline of the literature, covering studies of English language toxic comment detection, multilingual solutions, and solutions that are specifically aimed at Indian languages like Telugu, Tamil, Kannada, and Hindi.

2.1 Toxic Comment Detection in English

English has been the target of toxic comment detection research to a great extent due to the existence of large annotated datasets and well-developed NLP tools. Among the early attempts was the Jigsaw Toxic Comment Classification Challenge on Kaggle , where a large dataset was given with categories including toxic, severe toxic, and obscene, along with threat, insult, and identity hate

2.2 Multilingual and Cross-Lingual Classification

Pre-trained multilingual models like mBERT and XLM-RoBERTa (XLM-R) made it possible to do cross-lingual transfer learning, whereby models pre-trained in several languages can be fine-tuned for low-resource languages for toxicity detection in the absence of large task- specific datasets.

Some common tasks and datasets have helped advance this area:

The HASOC (Hate Speech and Offensive Content) shared task provided annotated datasets for hate speech and offensive content for Hindi, English, and German.

The DravidianCodeMix dataset offers code-mixed datasets for Tamil-English, Telugu- English, and Kannada-English.

2.3 Indic NLP Models and Tools

Researchers created specialized models and toolkits to manage the linguistic idiosyncrasies of Indian languages:
IndicBERT is a lightweight model that is ALBERT-based and was trained on 12 prominent Indian languages such as Telugu, Tamil, Kannada, and Hindi.

The IndicNLP Suite ffers rich resources in the form of tokenizers, transliteration mechanisms, and normalization procedures that are specific to Indian scripts and languages.

2.4 Deep Learning Architectures for Toxic Comment Classification

The literature witnesses the evolution from straightforward machine learning models to intricate deep learning models to detect toxicity. The research communities have attempted to work on a variety of architectures:

Hybrid CNN-LSTM architectures incorporate Convolutional Neural Networks (CNNs) to extract local features and Long Short-Term Memory (LSTM) networks, which model sequential dependencies.

Attention-based bi-directional LSTMs improve performance by concentrating on the most informative portions of the input text, thereby enhancing interpretability and classification accuracy.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025
DOI: 10.17148/IJARCCE.2025.14564

III. OBJECTIVES

The general objective of this survey paper is to present a complete picture of the existing scenario and forthcoming outlook of multilingual toxic comment classification, specifically with respect to Indian languages like Telugu, Tamil, Kannada, and Hindi. To do this, the paper is addressed by the following well-detailed objectives:

1. Detailed Review of Existing Approaches:
To critically review and discuss the methods and models utilized for toxic comment classification in the target Indian languages. Both classic machine learning and new deep learning models, particularly transformer-based ones such as mBERT, XLM-R, and IndicBERT, are included. Knowing these methods will reveal how effective and limited they are in the multilingual Indian scenario.

2. Highlighting Linguistic and Technical Challenges
To enumerate and discuss the most important challenges encountered in creating toxic comment classifiers for Indian languages, such as data paucity, the intricacy of processing code-mixed text, script variation, dialectal differences, and the cultural connotations inherent in toxic language. These challenges are important for adapting effective computational solutions and for realizing the disparity between English-oriented research and Indian language needs.

3. Survey of Datasets and Tools:
To gather and analyze the available datasets like HASOC, DravidianCodeMix, and Offenseval that offer annotated samples of toxic content in Telugu, Tamil, Kannada, and Hindi. Further, the paper also intends to investigate the Indic-specific NLP resources and tools like IndicBERT and the IndicNLP Suite, which are critical in preprocessing and model building.

4. Evaluation of Multilingual Model Impact:
To study how transformer-based multilingual models affect toxic comment classification performance in these languages. This involves analyzing transfer learning performance and whether pre-training on multilingual corpora helps the model generalize to low-resource Indian languages to mitigate data limitation.

5. Ethical Considerations and Bias Mitigation:
In order to highlight the significance of fairness, transparency, and cultural sensitivity when developing toxic comment classifiers. With India's socio-cultural diversity, it is an imperative goal to ensure that automated systems do not perpetuate unfair censorship or biases, which aligns with recent efforts in ethical AI research.

6. Future Research Directions:
To identify promising avenues for future research, including the creation of larger annotated corpora, improved handling of code-mixed and dialectal text, multimodal toxicity detection (e.g., combining text with images or videos), and community-driven annotation efforts. These directions aim to foster more inclusive and effective toxicity detection systems that can serve the diverse Indian digital landscape.

IV. METHODOLOGY

This section outlines an elaborate methodology for developing and testing multilingual toxic comment classifiers for Indian languages such as Telugu, Tamil, Kannada, and Hindi. It covers the phases of data acquisition, preprocessing, feature extraction, model selection, training, optimization, and testing.

4.1 Data Collection & Preprocessing

4.1.1 Dataset Selection

Datasets are the core of any supervised learning system. For multilingual toxic comment classification, data is collected from a mix of publicly available annotated corpora and in-house data scraping. Some of the notable datasets are the Jigsaw Toxic Comment Classification dataset, which contains vast toxic content mostly in English, the HASOC (Hate Speech and Offensive Content) dataset with Hindi and English comments annotated for hate speech and offensive language, and the IndicNLP corpus, which contains resources for Indian languages like Tamil, Telugu, Kannada, and Hindi.

4.1.2 Data Labeling

Existing annotated datasets (e.g., toxic, hate speech, offensive, profanity, neutral) are utilized as is. There are multiple annotators who provide consensus labeling for ensuring reliability, particularly for subtle labels such as implicit hate or sarcasm that are common in Indian multilingual scenarios.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025
DOI: 10.17148/IJARCCE.2025.14564

4.1.3 Class Imbalance Handling

Toxic comment datasets tend to be class-imbalanced with non-toxic comments greatly exceeding toxic ones. Class-weighted loss functions while training enable the model to learn reasonably from unbalanced data distributions.

4.2 Feature Extraction & Embeddings

4.2.1 Multilingual Text Handling

Given the diverse scripts (Devanagari for Hindi, Telugu script, Tamil script, Kannada script), transliteration approaches map text into a common script to aid uniform processing.

4.3 Model Selection & Training

4.3.1 Classical Machine Learning Models

Baseline models such as Naïve Bayes, Support Vector Machines (SVM), and Random Forests are used to serve as baselines for model performance and to demonstrate the advantages of deep learning approaches.

4.3.2 Fine-tuning Pretrained Models

Pretrained transformer models are then fine-tuned on target datasets with task-specific labels. Data augmentation methods such as back translation (translating to a different language and back), synonym replacement, and word substitution are employed to artificially enhance training data diversity in order to enhance generalization and fight data sparsity.

4.3.3 Model Optimization

Hyperparameters like learning rate, batch size, and number of epochs are tuned using Grid Search or Bayesian Optimization methods. Regularization techniques like dropout avoid overfitting, and early stopping on validation loss helps in effective training.

4.4 Evaluation & Performance Metrics

4.4.1 Metrics Used

Performance is quantified by Accuracy, Precision, Recall, and F1-score, which together measure the classifier's capacity to accurately label toxic comments while maintaining a balance between false positives and negatives. Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is employed to measure the discrimination ability of the models. Confusion matrices offer information about particular error types, which helps in model improvement.

4.4.2 Cross-Language Generalization Testing

Models are also evaluated on languages or dialects they have not been exposed to while training to assess their generalization capacities over linguistic boundaries. Performance on code-mixed text like Hinglish (Hindi-English), Tamlish (Tamil-English), and Kanglish (Kannada-English) is also analyzed specifically since code-mixing is widely used in Indian social media communication.

This methodology section fully covers the steps necessary for creating effective multilingual toxic comment classification systems, striking a balance between traditional and contemporary methods and being sensitive to the linguistic variation and complexities of Indian languages.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14564

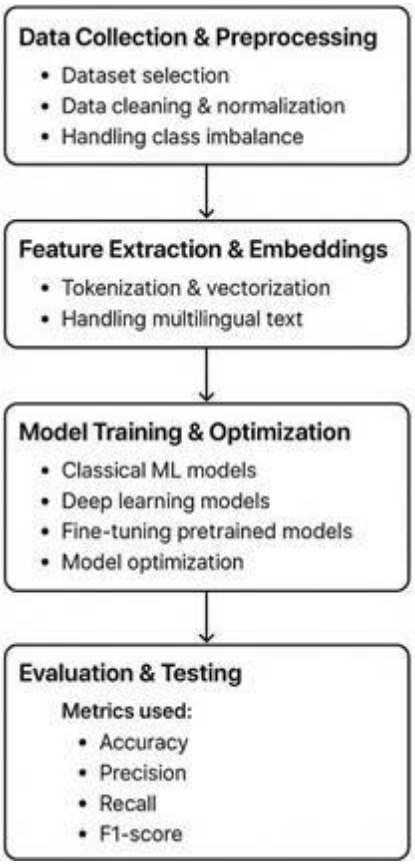

Fig. 1 Model selection and training

V. CHALLENGES

Constructing good multilingual toxic comment classification systems for Indian languages poses a range of intricate challenges due to linguistic, social, and technical reasons. These affect model accuracy, scalability, and generalization and need to be tackled carefully in research and development.

5.1 Linguistic Diversity and Complexity

India has many languages with their own scripts, phonetics, syntax, and morphology. Telugu, Tamil, Kannada, and Hindi have different language families—Dravidian and Indo-Aryan— having very dissimilar linguistic frameworks. This makes the development of universal models challenging because every language has its own set of preprocessing methods like tokenization, stemming, and stop-word elimination specific to its own grammar rules. Furthermore, managing multiple scripts such as Devanagari (Hindi), Telugu script, Tamil script, and Kannada script within a single framework has serious technical challenges.

5.2 Code-Mixing and Code-Switching

Indian social media texts abound with code-mixing, with users mixing languages in a single sentence or utterance (e.g., Hinglish, Tamlish, Kanglish). Such code-mixed text defies conventional NLP models trained on monolingual corpora since code-mixed text has intricate syntactic and semantic structures. Transliterations add to the situation by combining scripts with Latin alphabets, quite often in casual and random manners, to complicate normalization and tokenization.

5.3 Data Sparsity and Annotation Problems

Although there are large-scale English toxic comment datasets, datasets of Indian languages with annotations are sparse, particularly for minority languages such as Kannada and Telugu. Labeling data is expensive to create, as it demands linguistic know-how and cultural knowledge for accurate interpretation of subtleties such as sarcasm, implied hate, and slang. Crowdsourcing the annotation might bring in variability in labeling quality based on subjective judgments of toxicity.

DOI: 10.17148/IJARCCE.2025.14564

5.4 Class Imbalance

Toxic posts usually form a minority of social media data, creating extremely imbalanced datasets. Class imbalance results in non-toxic class-biased models with decreased sensitivity to identifying rare but dangerous toxic posts. The imbalance needs to be handled using sophisticated oversampling, undersampling, or algorithmic methods, which in turn might add noise or overfitting.

5.5 Ambiguity and Subjectivity Handling

Toxicity is generally subjective and context-dependent. Statements that are offensive in one social or cultural context can be permissible in another. Models do not do well to detect these fine-grained nuances, particularly the implicit or indirect types of toxicity like sarcasm or coded language. This makes the detection more prone to false positives and false negatives, and less trustworthy for automated moderation systems.

VI. CONCLUSION

Indian language toxic comment classification is an essential area of research with important social and technological ramifications. Even with the widespread emergence of social media and digital communication platforms, which enable interaction across a wide range of linguistic communities, the dissemination of toxic content is still a serious issue. This paper overviewed the state-of-the-art methods and issues in creating multilingual toxic comment classifiers specific to Indian languages.

The distinctive linguistic features of Indian languages, such as their rich morphological structure, multiform scripts, and extensive code-mixing usage, pose challenging natural language processing tasks. Although English has been favored by large annotated datasets and mature models, low-resource Indian languages are behind because of sparse data availability and linguistic complexities. In addition, the informal, noisy nature of social media text complicates effective toxicity detection.

Recent developments in multilingual transformer models like mBERT, XLM-R, and IndicBERT have shown encouraging outcomes by utilizing transfer learning and cross-lingual knowledge sharing. These models, along with domain-specific preprocessing tools like IndicNLP and iNLTK, have started closing the gap between high-resource and low-resource languages. Yet, their performance can still be limited by issues like a lack of annotated corpora, difficulties in processing code-mixed language, and the subtle nature of toxicity that can include cultural or contextual nuances.

This survey emphasizes the need to overcome data shortages using creative techniques such as data augmentation, crowdsourced annotation, and building community-generated datasets. In addition, addressing concerns about class imbalance, vagueness, and ethical issues is crucial to construct strong, equitable, and inclusive toxic comment classifiers. Further research is required to optimize model architectures, enhance multilingual generalization, and establish evaluation metrics that adequately reflect the nuances of toxic language in various social and linguistic contexts.

In summary, the combination of sophisticated NLP methods with linguistically grounded, culturally sensitive methodologies is the way forward for moderation and safer spaces online. Future research can work towards scalable, ethical solutions respectful of linguistic diversity yet addressing Indian language-based online toxicity through synergy between linguists, computer scientists, and social scientists.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT, 2019.

[2] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," arXiv preprint arXiv:1911.02116, 2020.

[3] R. Kakwani, A. Kunchukuttan, S. Golla, N. C. G, A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961, 2020.

[4] Jigsaw, "Toxic Comment Classification Challenge," Kaggle, 2018. \[Online]. Available: [https://www.kaggle.com/c/jigsaw-toxic-comment-classification challenge](https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge)

[5] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," Expert Systems with Applications, vol. 114, pp. 420–430, 2018.

[6] T. Mandl et al., "Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Content Identification

DOI: 10.17148/IJARCCE.2025.14564

in Indo-European Languages," CEUR Workshop Proceedings, 2020.

[7] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, N. Jose, S. Suryawanshi, E. Sherly, and J. P. McCrae, "DravidianCodeMix: Sentiment Analysis and Offensive Language Identification Dataset for Dravidian Languages in Code-Mixed Text," arXiv preprint arXiv:2106.09460, 2021.

[8] M. Zampieri et al., "Predicting the Type and Target of Offensive Posts in Social Media," Proc. NAACL, 2019, pp. 1415–1420.

[9] Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPCorpus: Monolingual corpora and word embeddings for indic languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 4948–4961). Association for Computational Linguistics.

[10] Baran Barbarestani, Isa Maks, and Piek Vossen. 2022. Annotating Targets of Toxic Language at the Span Level. In Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), pages 43–51, Gyeongju, Republic of Korea. Association for Computational Linguistics.

[11] Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. MSP: Multi-Stage Prompting for Making Pre-trained Language Models Better Translators. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6131–6142, Dublin, Ireland. Association for ComputationalLinguistics.