Archives available at journals.mriindia.com

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319 - 2526

Volume 15 Issue 01s, 2026

From Text to Toxicity: Exploring the Challenges in Marathi Hate Speech Detection

[1]Mrs. Preeti V. Sarode, [2]Harshali B. Patil

[1] Department of Computer Science and Information Technology
K. M. Agrawal College, Kalyan,
Maharashtra, India.
[2]Department of Computer Science
Dr. Annasaheb G. D. Bendale Mahila Mahavidyalaya, Jalgaon,
Maharashtra, India.
Email: [1] preetivsarodephd@gmail.com, [2] patilharshalib@gmail.com

Peer Review Information

Submission: 08 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Machine Learning, Deep Learning, Hate Speech

Abstract

Hate speech on social media has become a serious problem in society. There is a increasing need to create systems that can automatically detect such hateful content. Marathi, an Indo-Aryan language widely spoken in India, remains under-represented in natural language processing research due to limited linguistic resources and annotated datasets. This study focuses on the major challenges faced in detecting hate speech in the Marathi language. Marathi is a low-resource and complex language. It does not have enough large and balanced datasets for proper model training. The presence of code-mixing, transliteration between Devanagari and Roman scripts, and various dialects increases the complexity of text processing. Ambiguity in meaning, sarcasm, and context-dependent expressions make automatic detection more difficult. This paper systematically reviews the challenges in Marathi hate speech detection arising from data scarcity, code-mixing, transliteration, dialectal variation, annotation ambiguity, and context-dependent expressions.

Introduction

Social media platforms such as Facebook, Instagram, Twitter, and YouTube have become primary channels for communication. Social media is increasingly used to express opinions in multiple languages. The use of social media has made it easy for hate speech to spread widely, leading to violence, discrimination, and social division. Automated detection systems work well for languages like English, but Indian languages such as Marathi still lack enough data and tools. Marathi, an Indo-Aryan language spoken by over 83 million people, remains under-represented in NLP research due to limited linguistic resources and annotated datasets [1]. Marathi has its own challenges, like many dialects, different writing scripts (Roman and Devanagari), complex grammar and frequent mixing with Hindi and English. Automated detection of hate speech on Marathi social media is challenging due to several factors. First there is a scarcity of large-scale annotated datasets; existing resources [2] primarily cover Twitter and are imbalanced across hate and non-hate classes. Second, users frequently employ code-mixing and transliteration**,** writing Marathi in Roman script, which complicates tokenization and embedding generation. Third**,** dialectal diversity (e.g. Konkani-Marathi, Khandeshi, Varhadi,) and informal spellings reduce model generalization.[3]. Fourth, annotation ambiguity and context-dependent expressions, including

sarcasm and implicit hate, make classification subjective [4]. New resources like the Marathi models such as MahaBERT, MahaTweetBERT and L3Cube-MahaHate dataset have helped in early research, but differences in labelling and a lack of good language tools accurate detection is still hard. Multilingual transformer models like XLM-R and mBERT can process Marathi text, monolingual models such as MahaBERT and MahaTweetBERT demonstrate improved performance on social media datasets [5]. However, researchers still struggle with transliterated, context-dependent content and code-mixed. This paper reviews the key challenges of hate speech detection in Marathi social media and proposes research directions to enhance model robustness, including transliteration-aware modelling, dataset expansion across platforms, functional robustness testing, multi-task labelling, adversarial data augmentation, context-aware modelling, and ethical considerations for fair deployment.

Literature Review

Hate speech detection has been an active area of research in natural language processing (NLP), with significant progress in high-resource languages such as English and Hindi. However, work on low-resource languages like Marathi remains limited due to a lack of large, balanced datasets, linguistic resources, and domain-specific tools. Marathi is a morphologically ironic and frequently expressed in code-mixed and romanized forms on social media, presents unique tasks for automatic text classification.

Recent studies have formed interpreted datasets and transformer-based models for Marathi hate speech detection. However, problems like unclear annotations, unbalanced data and difficulty in understanding context or sarcasm still remain. This section gives a review on existing literature on models, datasets and evaluation techniques relevant to Marathi hate speech detection, highlighting existing progress and the challenges that still exist.

1. Datasets of Marathi Hate Speech

The growth of hate speech detection systems in Marathi has largely depend on a key dataset that provide the foundation for supervised learning. The L3Cube-MahaHate dataset developed by researchers [2] which is one of the first comprehensive tweet-based resources for Marathi, consisting of over 25,000 samples considered as hate, offensive, profane, or non-hate. This dataset is balanced and significantly subsidized to Marathi NLP research.

Another important contribution is the HASOC (Hate Speech and Offensive Content) Marathi dataset by [6][24] introduced as part of the FIRE 2021 shared task on hate and offensive speech detection in multiple Indian languages, but still suffers from class imbalance, where non-hate samples are much more frequent than hate or offensive samples. Even though MOLD (Marathi Offensive Language datasets) are useful for major progress, but different difficulties also occur like uneven data distribution, data imbalance and inconsistent annotations, which makes it hard for models to perform well across diverse social media states.

Table 1. Marathi Hate Speech related Datasets Characteristics

Name of Datasets	Categories of Data Labels	Percentages (%) Labels Count	Source Platform	References
L3Cube-MahaHate	4-class Hate, Offensive, Profane, Neutral	Hate: 25%, Offensive: 25%, Profane: 25%, Neutral: 25%	Twitter	[2]
L3Cube-MahaHate	Binary class Hate, non-Hate	Hate: 50%, Non-Hate: 50%,	Twitter	[2]
HASOC Marathi	Hate & Offensive, Non-Hate	Hate & Offensive: 35.68%, Non-Hate: 64.32%	Twitter	[6]
L3CubeMahaSent	Positive Negative Neutral	Positive:31.67 Negative:38.96% Neutral: 29.37%	Twitter	[5]
MOLD Marathi Offensive Language Dataset	Offensive non-Offensive	Offensive:33% nonOffensive:67%	Twitter	[32]
HASOC 2021	Non-Offensive(NOT) Offensive(OFF)OFF	NOT: 65.54% OFF: 34.46%	Twitter	[24]

The Table 1 summarizes the key characteristics of each dataset, including label categories, percentages of labels count, and source platform, providing an authenticated and standardized overview of the Marathi hate-speech resources used in prior studies. Percentage of label count are always calculated using a simple formula:

Percentage (%) = (Label Count/ Total Dataset Size) × 100 (1)

2. Existing Models of Marathi Hate Speech

New advancements in transformer models have enabled better representation learning for Marathi text. MahaBERT and MahaTweetBERT are Marathi language models developed by [5]. These models are specially trained on Marathi social media data and work well for tasks like hate speech detection and sentiment analysis. Other multilingual models such as mBERT, XLM-R, and IndicBERT can also process Marathi text. They are trained on many languages together but are less focused on Marathi-specific features. [7]. Marathi-specific models such as MahaBERT and MahaTweetBERT are fine-tuned on Marathi social media datasets like L3Cube-MahaHate and MahaSent. These models are trained to understand the unique features of Marathi language, such as morphology, word inflection, and common social media expressions. These models are trained on Marathi text, which helps them to understand the tone, slang, and linguistic nuances of the language. This makes them more effective for accurate sentiment and hate speech classification compared to general multilingual models. New slang, abbreviations, and hate expressions keep changing over time. Therefore, these models need regular retraining and updates to stay accurate and reliable.

3. Evaluation of Hate Speech

Evaluation methods used in existing research have primarily focused on conventional metrics such as precision, accuracy, F1-score and recall, to assess model performance. However, these metrics alone do not fully capture the real-world robustness of hate speech detection systems. The researchers [8] addressed limitations and introduced the Multilingual HateCheck (MHC) framework and functional test suite designed to evaluate model performance on complex phenomena such as obfuscation, sarcasm, negation, and quoted speech. The evaluation approach to Marathi datasets can help identify model weaknesses and improve real-world reliability. Marathi-specific models need to be tested functional robustness evaluations. This helps improve the performance of models on social media content that is noisy, diverse and constantly changing.

Table 2. Marathi Hate Speech Evaluation

Dataset	Evaluation Techniques	Evaluation Measures and Results	Reference
L3Cube-MahaHate	MahaBERT (monolingual)	Accuracy 0.783	[2]
L3CubeMahaSent	IndicBERT	Accuracy 0.833	[5]
HASOC Marathi	CNN, LSTM, BERT, IndicBERT, RoBERTa	Accuracy 0.859	[10]
HASOC Marathi	Logistic Regression and Random Forest	Accuracy RF 77.70%, LR 75.95%.	[26]
Social media text,news articles and websitres	CNN+BiLSTM	Accuracy 0.8898 F1 Score 0.7912	[9]
Twitter Social media post Marathi	SVC,CNN,BILSTM,mBERT, IndicBERT,	F1 Score 0.85	[21]
Marathi tweet dataset	BiLSTM, GRU, BERT variants, MarathiTweetBERT, Stacked Models	Accuracy BiLSTM: 0.89, Stack: 0.89, LSTM+GRU: 0.88	[14]
Marathi (MuRIL)	MuRIL BERT	Accuracy 0.9186 F1-score 0.8306	[24]

Marathi Hate Speech Detection Challenges

Hate speech finding in Marathi remains a challenging task due to limited resources, inconsistent annotations and language diversity. Marathi social media text is often context-dependent, informal and code-mixed which makes automated detection more complex. The following subsections describe the key

challenges that affect dataset model robustness, quality and evaluation reliability.

Figure 1. Challenges of Hate Speech Detection in Marathi

1. Data Scarcity and Imbalance

Marathi is a low-resource language, and publicly available hate speech datasets remain imbalanced and small. Datasets hold dominance of non-hate instances over hate or offensive ones. The imbalance of dataset causes models to favour common classes, resulting partial predictions and poor generalization. Techniques like data augmentation, SMOTE (synthetic minority over-sampling technique) oversampling and class-weight modifications can partially mitigate the issue but are insufficient for achieving robust performance [9] [10].

2. Transliteration and Script Variation

Marathi text on social media is inscribed in both Roman scripts and Devanagari scripts. Writing in Roman script often causes spelling differences and word breaks, which makes it difficult for models to process the text correctly [1] [5]. For example, “तु वाईट आहेस” and “tu vait ahes” convey the similar meaning but vary in form. This variation drops model accuracy when trained only on Devanagari data. Methods that comprehend both scripts, like phonetic mapping or dual-script models, can improve performance by training on Roman and Devanagari text together [11][12].

3. Code-Mixing

People on social media often mix Hindi, Marathi and English in the same sentence. They also use hashtags, short informal words and emojis. Such code-mixing changes token distributions and complicates semantic and syntactic modelling [13]. Language identification and preprocessing become difficult, especially for transliterated content. Effective methods include using multilingual embeddings, language tags or subword tokenization to understand and process mixed-language text 15].

4. Dialectal and Linguistic Variations

Marathi has many regional dialects such as Khandeshi, Varhadi and Konkani-Marathi, which differ in pronunciation, words, and sentence structure. Informal spellings and local slang make it even harder for models to understand and generalize correctly [16]. Models capability depends on standard Marathi data often misclassify dialectal expressions as out-of-vocabulary or unknown terms. The addition of dialectal data and adaptive embeddings can help models better understand regional variations and improve their performance.

5. Annotation Ambiguity

Hate speech annotation is a subjective process, as human annotators often understand the same message in different ways based on its tone, cultural background, or context [6]. This annotation ambiguity introduces label noise, which reduces training reliability and model accuracy. Recent research recommends using multiple annotators, clear context-based guidelines, or large language model (LLM)-assisted annotation methods to reduce differences in labelling [8] [17].

6. Sarcasm and Implicit Hate

Marathi users sometimes express hate in indirect ways through sarcasm, metaphors, jokes, or cultural hints, which are hard for NLP models to detect [12]. For example, polite-sounding sentences can actually express negative feelings when spoken sarcastically like “वा! किती भारी नेता आहे आपला!”. On the surface, it sounds like admiration, but when said sarcastically, it actually carries criticism or dislike. This hidden type of hate identify by models which have capability to capture the context and underlying intent instead of focusing only on surface-level words. Context-aware transformer models and multi-task learning techniques can improve their ability to handle these subtle expressions [19].

7. Limited Linguistic Resources

There is a lack of Marathi-specific linguistic tools like part of speech (POS)taggers, sentiment and hate speech lexicons, and morphological analysers, which are vital for extracting meaningful features and improving model performance [9]. For instance, if a Marathi POS tagger is unavailable, a sentence like "तो वाईट

बोलतो" ("He speaks badly") might not correctly identify "वाईट" as an adjective and "बोलतो" as a verb. This makes it difficult for a hate speech model to understand the grammatical structure and emotional tone of the sentence, reducing detection accuracy. The lack of reliable linguistic resources limits the use of hybrid approaches that combine deep learning with language-based features. Therefore, it is crucial to build rich natural language processing (NLP) tools and datasets for Marathi to enhance model accuracy and interpretability.

8. Evolution of Slang rapidly

Slang and offensive words on social media keep changing quickly due to trending in memes, pop culture and current events. Models trained on old data often fail to detect new abbreviations or hidden hate expressions [5], [10]. To ensure models stay pertinent and robust, continuous retraining and domain editions are important. Semi-supervised learning and active data collection methods can help capture new slang and emerging language patterns in online conversations.

9. Context Dependency

In Marathi, the sense of a sentence frequently changes based on the conversational or social situation. A phrase that sounds hateful in one context might look as if neutral or even funny in another[18]. For Example, "तू तर भारी बुद्धिमान आहेस!". this can be genuine praise, but if said sarcastically, it becomes a insult or taunt. Current models that rely on isolated sentence-level analysis struggle with such ambiguity. Context-aware models that consider conversation threads or user behaviour can better identify the real intent and reduce wrong classifications.

10. Lack of Marathi-Specific Transformer Models

Multilingual models like XLM-R, mBERT and IndicBERT work well for languages with ironic resources but perform poorly for Marathi because such type of models were not trained on enough Marathi data [13]. On the other hand, Marathi-specific models such as MahaSent ,MahaTweetBERT and MahaBERT[5] are trained on Marathi text and therefore recognize the language better. However, these models still prerequisite consistent fine-tuning and updates to adapt to mixed-language text, new slang and noisy data on social media. Continuous assessment and improvement are essential to keep their performance reliable and truthful.

Strategies and Techniques to Improve Marathi Hate Speech Detection

Hate speech on Marathi social media is becoming increasingly complex, demanding new strategies to overcome linguistic, technical, and ethical challenges. To build more accurate and context-aware systems, researchers must focus on both data-centric and model-centric improvements. Data expansion across multiple platforms, transliteration-aware processing, and balanced annotations can enhance dataset quality. Advanced machine learning methods such as context-aware transformers, adversarial training, and robustness testing can help models perform better and handle diverse situations more reliably. In addition, ethical evaluation and fairness checks are necessary to ensure that Marathi hate speech detection systems are used responsibly in multilingual and culturally sensitive settings. The following strategies and combine dataset enhancement, improved modelling, and ethical evaluation to strengthen Marathi hate speech detection systems.

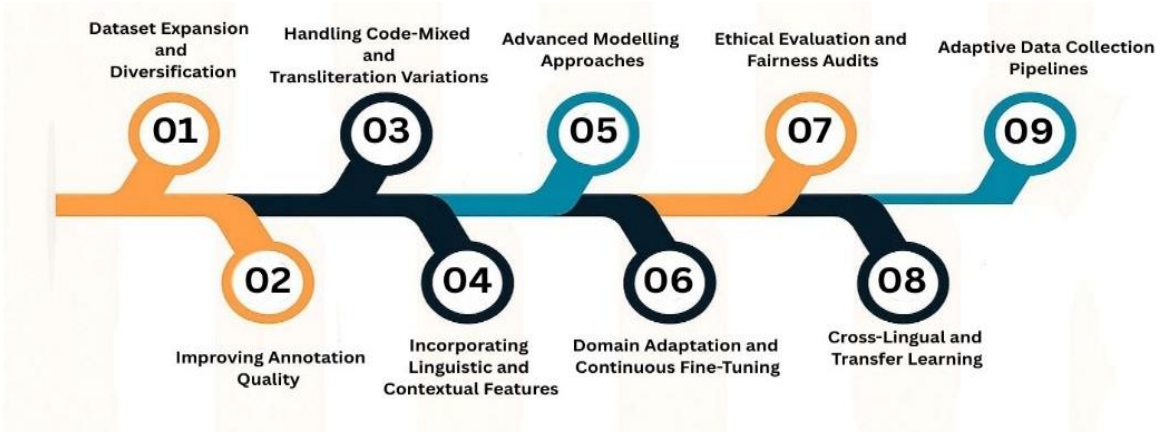

Figure 2. Strategies and Techniques to Improve Marathi Hate Speech Detection

1. Dataset Expansion and Diversification

A large and diverse dataset is essential for developing robust Marathi hate speech detection models This can be achieved by expanding the dataset through collecting data from multiple social media platforms such as Twitter, YouTube, and Facebook. Incorporating text that includes transliterated forms, code-mixed language (e.g., Marathi-English mixtures), and dialectal variations further enhances the dataset, enabling improved model coverage and generalization.

Recent studies have developed extensive Marathi datasets sourced from Twitter and manually annotated for improved reliability. Notably, the L3Cube-MahaHate dataset contains over 25,000 labeled tweets**,** systematically categorized into hate**,** offensive**,** profane**,** and non-hate classes This dataset also explores different model approaches, showing mono-lingual BERT models outperform multi-lingual ones for Marathi hate speech classification [2]. The L3Cube-MeCorpus dataset comprises almost 5 million code-mixed sentences, along with 12,000 manually annotated Marathi-English tweets curated for tasks such as sentiment analysis, language identification and hate speech detection. Models trained on such code-mixed corpora significantly outperform previous methods, demonstrating the value of dataset diversification with code-mixed and transliterated data. hate speech detection, For example, the SeMOLD dataset applied semi-supervised methods to generate over 8,000 additional Marathi text instances, which, when shared with manually labelled data, significantly enhanced the overall hate speech classification performance This method leverages existing annotated data to create larger training sets, alleviating the problem of data scarcity in low-resource languages such as Marathi[21].

2. Improving Annotation Quality

Improving annotation quality for Marathi hate speech detection involves using expert native annotators familiar with local dialects and cultural context, which helps identify subtle and implicit hateful content. Multiple annotators should label the data with inter-annotator treaty measured to confirm consistency and reduce bias. Detailed annotation and strong guidelines customized for Marathi hate speech including instructions on implicit meaning, sarcasm and context are essential. Hybrid annotation styles that combine initial machine-assisted labelling with human verification improve both accuracy and scalability. Continuous updates to annotation processes are required to capture evolving slang and coded language in social media [22].

The L3Cube-MahaHate dataset, annotated manually by native Marathi speakers considering cultural context and dialects, demonstrated the importance of these practices. Multiple annotators and agreement metrics were used to ensure dataset quality for training deep learning classifiers (CNN, LSTM, BERT variants) that achieved state-of-the-art results for Marathi hate speech detection.[23].

3. Handling Code-Mixed and Transliteration Variations

Handling code-mixed and transformation variations in Marathi hate speech detection requires specialized approaches due to the common use of mixed languages (e.g., Marathi-English) and writing Marathi words in Roman script on social media. Pre-trained transformer models such as MuRIL, MeBERT, and their code-mixed variants have shown superior performance by capturing contextual nuances and subword information better than traditional methods. Code-mixed datasets with manual annotations by native speakers, like the L3Cube MeCorpus (approximately 50M tokens) and supervised MeHate dataset (approximatly12,000 tweets), provide essential training data for these models. Techniques include cleaning and normalizing text, handling Romanized script, and leveraging transfer learning on large diverse corpora. These strategies significantly enhance detection accuracy for code-mixed and transliterated hate speech in Marathi text[20][24][25].

4. Incorporating Linguistic and Contextual Features

Incorporating linguistic and contextual features significantly improves Marathi hate speech detection. Linguistic features such as lexical cues, syntax, morphology, and word embeddings (e.g., FastText) help capture language-specific characteristics. Contextual features derived from transformer-based models like BERT and MuRIL outperform traditional methods by understanding sentence meaning and subword units, which is effective in detecting subtle hate expressions. The utilization of TF-IDF, n-grams and combining semantic features with context modeling enhances classification accuracy. Social context and dialectal variations considerations can further improve model robustness. The researchers' studies show that transformer-based models fine-tuned on Marathi hate speech datasets attain the best results by leveraging these linguistic and contextual cues.[10][26][27]

5. Advanced Modelling Approaches

Advanced modeling approaches in Marathi hate speech detection predominantly involve transformer-based architectures such as BERT, XLM-RoBERTa, and MuRIL. These models leverage transfer learning from high-resource languages like Hindi to Marathi, significantly improving classification performance by capturing contextual and semantic nuances. Monolingual Marathi transformer models (e.g., MahaBERT) often outperform multilingual counterparts on Marathi-specific datasets, owing to better language-specific adaptations. Traditional machine learning techniques like SVM and CNN are increasingly complemented by fine-tuned transformers that handle code-mixed and transliterated data effectively. Data augmentation and contrastive learning further enhance model generalization and robustness. State-of-the-art systems using such models have achieved top ranks in shared tasks like HASOC, demonstrating their effectiveness for Marathi hate speech detection [28].

6. Domain Adaptation and Continuous Fine-Tuning

Domain adaptation and continuous fine-tuning are critical for improving Marathi hate speech detection. Due to linguistic and contextual differences across platforms and domains, models pre-trained on general language data often underperform on specific hate speech datasets in Marathi. Domain adaptation techniques involve fine-tuning pre-trained transformer models such as mBERT and MuRIL on in-domain hate speech corpora for Marathi, which substantially boosts performance by learning language, script, and discourse nuances. Multilingual approaches can benefit from cross-lingual transfer learning but require further fine-tuning to bridge gaps caused by script and syntax diversity. Continuous fine-tuning with newly collected and annotated data helps keep models updated with evolving hate speech patterns and slang, improving robustness across various social media platforms and dialects. Studies have shown that such continuous adaptation helps achieve higher accuracy and generalization in real-world hate speech detection tasks for Marathi.[28][29].

7. Ethical Evaluation and Fairness Audits

Ethical evaluation and fairness audits are crucial in Marathi hate speech detection to ensure balanced, unbiased, and responsible use of AI. These evaluations address potential biases against specific groups or dialects and assess if the model respects freedom of expression while effectively identifying harmful speech. Metrics such as precision, recall, and F1-score across diverse subpopulations help identify disparities in model performance. Fairness audits involve reviewing annotations, sampling practices, and output to avoid perpetuating stereotypes or unfair censorship. Ethical considerations also include transparency about model limitations and ongoing monitoring for harmful effects. Research using datasets like L3Cube-MahaHate and the MuRIL BERT model has highlighted the importance of combining high accuracy with ethical safeguards, ensuring hate speech detection systems foster safe online environments without over-censorship or bias [30].

8. Cross-Lingual and Transfer Learning

Cross-lingual and transfer learning techniques have proven highly effective for Marathi hate speech detection, leveraging data and models from related languages like Hindi and English. These approaches use pre-trained multilingual transformers such as XLM-R and mBERT, which are fine-tuned on Marathi hate speech datasets to capture language-specific nuances. Research shows that transfer learning from a closely related language like Hindi yields superior performance compared to monolingual models trained solely on Marathi data, especially given Marathi's relatively low-resource status. Few-shot learning and continuous transfer learning with cross-lingual embeddings enable models to generalize better even with limited Marathi annotations. This method ranked first in HASOC 2021 for Marathi hate speech detection with an F1-score of 0.91, underscoring the benefit of such cross-lingual transfer strategies for under-resourced languages [31].

9. Adaptive Data Collection Pipelines

Adaptive data collection pipelines for Marathi hate speech detection focus on gathering diverse and representative data from multiple social media platforms such as Twitter, YouTube, and Facebook. The L3Cube-MahaHate dataset is a prime example, curated mainly from Twitter with over 25,000 manually annotated Marathi tweets classified into hate, offensive, profane, and non-offensive categories. This dataset's creation involved iterative data collection, cleaning, and manual annotation by native speakers, along with continuous updates to incorporate emerging slang and dialects. Techniques like active learning and semi-supervised labeling are also used to scale dataset size while ensuring quality, enabling models to adapt to the evolving nature of online Marathi hate speech. Such adaptive pipelines are essential for producing datasets that reflect real-

world linguistic variations and challenges in Marathi hate speech detection [32].
Such systems continuously learn from fresh data, capturing evolving slang, coded hate, and new expressions. This strategy ensures that datasets and models remain relevant and capable of handling the fast-changing landscape of online communication.

Conclusion

Marathi hate speech detection is challenged by dataset scarcity, script variations, dialectal diversity, code-mixing, annotation ambiguity, and context-dependent expressions. The challenges involves expanding datasets, implementing transliteration-aware and context-sensitive models, using multi-task labeling, performing functional robustness testing, and maintaining ethical governance.
Monolingual models such as MahaBERT have improved performance but require continuous adaptation. Advanced transformer-based models fine-tuned with domain adaptation and continuous learning yield significant performance gains. Incorporating linguistic and contextual features enhances subtle hate speech recognition, while cross-lingual transfer learning leverages resource-rich languages for under-resourced Marathi. Adaptive data collection pipelines play a vital role in keeping datasets aligned with the constantly evolving nature of online discourse. These dynamic strategies ensure that datasets remain up-to-date and representative of Marathi language use and effectively addressing the intrinsic challenges posed by the language's rich diversity and complexity.
Advanced transformer-based models fine-tuned with domain adaptation and continuous learning yield significant performance gains. Incorporating linguistic and contextual features enhances subtle hate speech recognition, while cross-lingual transfer learning leverages resource-rich languages for under-resourced Marathi. Adaptive data collection pipelines ensure datasets remain relevant to evolving online discourse. Ethical evaluation and fairness audits safeguard against biases and promote responsible AI use. Together, these strategies form a robust framework for developing accurate, culturally aware Marathi hate speech detection systems, addressing the challenges posed by the language's complexity and diversity.

References

[1] A. Joshi, A. Kunchukuttan, P. Bhattacharyya, and M. Mehta, "The L3Cube-MahaCorpus: Building Large-Scale Marathi Datasets for Natural Language Understanding," Proceedings of the 6th Workshop on Indian Language Data: Resources and Evaluation (WILDRE-2020), European Language Resources Association (ELRA), Marseille, France, 2020.]

[2] A. Velankar, H. Patil, A. Gore, S. Salunke and R. Joshi, "L3Cube-MahaHate: A Tweet-based Marathi Hate Speech Detection Dataset and BERT Models," Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), 2022.

[3] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 6282–6293, 2020]

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel, P. Rosso, M. Sanguinetti, and R. A. Zwaan, "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter," Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pp. 54–63, 2019.]

[5] Gummadi, V. P. K. (2022). MuleSoft API Manager: Comprehensive lifecycle management. Journal of Information Systems Engineering and Management, 7(4), 1–9. https://www.jisem-journal.com

[6] T. Mandl, S. Modha, P. Majumder, M. Mandal, S. Patel, A. Dave, M. P. Joshi, and P. Rosso, "Overview of the HASOC Track at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages," Proceedings of the Forum for Information Retrieval Evaluation (FIRE), 2021, pp. 1–10.

[7] D. Kakwani, A. Kunchukuttan, S. Golla, A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, 2020, pp. 494–501.

[8] P. Röttger, D. Vidgen, and J. G. Hale, "Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks: Model-Assisted vs. Human-Centric," Proceedings of the 60th Annual Meeting of the Association for

Computational Linguistics (ACL), 2022, pp. 7812–7833.

[9] A. Sarode and N. Sultanova, "Detection of Hate Speech in Marathi Using Language-Specific Pre-Processing," International Journal of Data Science and Advanced Analytics, vol. 6, no. 1, pp. 297–301, 2024.

[10] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "Hate and Offensive Speech Detection in Hindi and Marathi," arXiv preprint arXiv:2110.12200, 2021.

[11] S. Sharma and M. Shrivastava, "Improving Hate Speech Detection in Low-Resource Indian Languages Using Transliteration-Aware Models," in Proceedings of the LREC-COLING 2024 Conference, 2024.

[12] D. Acharya, S. Dawadi, S. Saud, and S. Regmi, "Paramananda@NLU of Devanagari Script Languages 2025: Detection of Language, Hate Speech and Targets using FastText and BERT," in Proceedings of the CHIPSAL 2025 Workshop, 2025.

[13] P. Khare, A. Singh, and M. Shrivastava, "Towards Hate Speech Detection in Hindi-English Code-Mixed Social Media Text," in Proceedings of the Workshop on Speech and Language Technologies for Dravidian Languages (SLT-Dravidian 2021), 2021.

[14] P. Shedge, S. Kamalkar, and D. Gupta, "Hate Speech Detection in Marathi Tweets using Stacked Deep Learning Models," *in Proc. of 2024 IEEE ICCCNT*, 2024.

[15] V. Gupta and A. Singh, "Code-Mixed Indian Language Hate Speech Detection Using Language Tagging and Multilingual Transformers," in Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2023.

[16] S. Ghosh and T. Sahu, "Challenges in Low-Resource Language Dialect Processing for Indian Languages," in Proceedings of the AI4Bharat Workshop on Low-Resource NLP for Indian Languages, 2023.

[17] S. Jadhav, A. Shanbhag, A. Thakurdesai, R. Sinare, and R. Joshi, "On Limitations of LLM as Annotator for Low-Resource Languages," arXiv preprint arXiv:2411.17637, 2024.

[18] A. Phadke, P. Gupta, and S. Kohle, "Systematic Sentiment Analysis Review: Marathi Language," SSRN preprint SSRN:5340307, 2025.

[19] R. Mishra and R. Bhatnagar, "Sarcasm and Implicit Hate Detection in Indic Languages Using Contextual Embeddings," in Proceedings of the Forum for Information Retrieval Evaluation (FIRE), 2022.

[20] P. Chavan, V. Velankar, and R. Joshi, "MyBoli: Code-Mixed Marathi-English Corpora, Pretrained Language Models, and Evaluation Benchmarks," arXiv preprint arXiv:2403.18611, 2024.

[21] M. Zampieri, S. Malmasi, P. Nakov, A. Rosenthal, V. Velankar, and R. Joshi, "Predicting the Type and Target of Offensive Social Media Posts in Marathi," Proceedings of the Workshop on Computational Approaches to Code Switching (CACO-2024), 2024.

[22] R. Abou Karam, "Hybrid Annotation Methods for Enhancing Hate Speech Detection," Proceedings of the 2024 International Conference on Computational Linguistics and Social Media Analysis (CL-SMA 2024), 2024.

[23] A. Arora, S. Ghanghor, P. Singh, and M. Shrivastava, "Experience-led Annotation of Online Gender-Based Violence (oGBV)," Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), 2024.

[24] S. Kalra, R. Kulkarni, and P. Joshi, "Hate Speech Detection in Marathi and Code-Mixed Text," Proceedings of the 2022 Forum for Information Retrieval Evaluation (FIRE 2022), pp. 380–387, 2022.

[25] D. Brahmanaidu and S. Vishnuvardhan, "Multilingual Hate Speech Identification with Deep Learning," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 4, pp. 112–120, 2024.

[26] A. Gajbhiye, P. Patil, and R. Deshmukh, "Machine Learning Models for Hate Speech Identification in Marathi," Proceedings of the International Conference on Computational Linguistics and Natural Language Processing (CLNLP), pp. 145–152, 2021.

[27] S. Ghosh, R. Patel, and T. Sahu, "Comparison of Mono and Multilingual Transformer

Models for Indian Hate Speech," Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation (PACLIC), pp. 412–420, 2022.

[28] A. Nene, S. Patil, and R. Joshi, "Transformer Models for Offensive Language Identification in Marathi," Proceedings of the Workshop on Offensive Language Identification (CEUR Workshop Proceedings), vol. 3020, pp. 58–65, 2021.

[29] R. Thapa, D. Acharya, and S. Regmi, "Natural Language Understanding for Devanagari Script Languages," Proceedings of the Workshop on Devanagari Script Languages (DSL), pp. 55–63, 2025.

[30] S. Mhapaseka, P. Patil, and R. Joshi, "Identification of Objectionable and Displeasing Contents in Marathi using BERT-CNN," International Journal of Innovative Research in Technology (IJIRT), vol. 10, no. 4, pp. 112–118, 2024.

[31] P. Gaikwad, S. Kulkarni, and R. Joshi, "Cross-lingual Offensive Language Identification for Marathi," in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), 2021.

[32] S. Patwardhan, A. Deshmukh, and R. Joshi, "bSamPar: A Marathi Hate Speech Dataset for Homophobia and Transphobia," Proceedings of the Workshop on LGBTQ+ in NLP (QueerNLP), 2024.