A Systematic Review of Artificial Intelligence and Machine Learning Approaches for Anomaly Detection in Marathi Text: Methods, Datasets, Challenges, And Future Directions

Asst. Prof. Shraddha Dnyaneshwar Gawali[1], Asst. Prof. Vaishali Balasaheb Gaikwad[2]
[1]*Department of Computer Science, Shree Neminath Jain Brahmacharyashram (Jain Gurukul's) Karmveer Keshavlalji Harakchandji Abad Arts, Shriman Motilalji Giridharilalji Lodha Commerce and Shriman P.H. Jain Science College, Neminagar, Chandwad, Nashik, Maharashtra, India*
[2]*Department of Computer Science, MVP's KGDM Arts, Commerce and Science College, Niphad Dist. Nashik, Maharashtra*
doi.org/10.64643/IJIRT211668

***Abstract*—Anomaly detection in text is about spotting text that's abnormal, harmful, suspicious or otherwise unexpected. As Marathi language content on media, messaging apps and online news sites keeps growing the need for automatic detection of offensive, hateful, spam and misinformation becomes more urgent. Marathi is a language for NLP and anomaly detection is harder than ordinary text classification. The challenge is that anomalous examples are few highly imbalanced depend on context and vary a lot linguistically. I present a review of AI, ML and deep learning methods for anomaly detection in Marathi text focusing on offensive language hate speech and harmful content. The review looks at ML models, deep neural networks, transformer-based language models and unsupervised anomaly detection techniques.**

It also examines Marathi datasets, pre-processing methods, evaluation metrics reported performance and key limitations. The findings show that transformer models BERT variants trained on Marathi usually beat traditional ML and earlier deep learning ones. Still the field faces problems: lack of annotated data, dialects, code-mixing with Hindi and English class imbalance, no standard benchmark and poor explainability. I highlight these datasets, explainability and fairness issues in this focused review of AI/ML methods, for Marathi text anomaly detection.

***Index Terms*—Marathi Natural Language Processing; Anomaly Detection; Machine Learning; Deep Learning; Hate Speech Detection; Offensive Language Identification; Low-Resource Languages; Explainable AI**

I. INTRODUCTION

AI and Machine Learning are key methods for processing large volumes of text. Within this area of research, it is important to detect anomalies in text data. In the media and web, anomalies can be harmful text, such as hate speech, rudeness, bullying, spam, phishing, misinformation and offensive text.

Marathi

Marathi is widely spoken in Maharashtra and surrounding areas. The amount of Marathi in media is quickly increasing. Therefore, there is an increasing need for methods to detect harmful or anomalous text. However, creating effective methods to detect anomalies in Marathi text is difficult. There is a lack of labelled data sets in Marathi, and the language is complex due to different dialects being used in different areas. Furthermore, code mixing is often present in informal Marathi text and words from other languages (Hindi and English) are often used.

Text Classification
Balanced data.
Anomaly detection
Imbalanced data.
For example, there may be very few rude/hateful messages in comparison to the total volume of content online. Therefore, it is important to not only accurately

classify text, but also detect the rare, strange cases of text.

Although recent research has improved classifying Marathi text using sentiment and language characteristics, the L3Cube-MahaHate data set is of particular interest as it contains over 25,000 labelled tweets. These tweets can be classed as either hate speech, offensive, rude or benign.

Positive results have been obtained from this data set using models such as CNN’s, LSTM’s and transformer models. Other research has created specific data sets in Marathi to analyse language, and has tested BERT based models to detect harmful content. However, very little research has been completed in the area of anomaly detection.

Aim

This review aims to investigate the ways that AI and Machine Learning have been used to detect objects in Marathi text, and to identify key research gaps.

Questions addressed

1. What AI/ML/deep learning methods have been used to detect objects in Marathi text?
2. What Marathi data sets are available to detect hate speech/anomalies?
3. What text pre-processing methods are used? What features are used in the model? What methods are used to evaluate model performance?
4. What are the major issues with these methods?
5. What improvements can be made in research to improve efficiency and accuracy in detecting information in Marathi text?

Structure

Section 2 describes the method used. Section 3 discusses Marathi NLP, and detecting anomalies within Marathi text.

Section 4 describes a classification of AI/ML methods.

Section 5 compares existing research and data sets.

Section 6 analyses the issues within this research area.

Section 7 discusses potential areas for future research.

Section 8 concludes this paper.

II. REVIEW METHODOLOGY

This review follows a systematic approach to identify, select, and analyze relevant research studies. The review focuses on AI and ML methods applied to Marathi text anomaly detection, especially offensive-language and hate-speech detection.

2.1 Search strategy

Relevant publications were searched using the following digital databases:

- IEEE Xplore
- ACM Digital Library
- Scopus
- Google Scholar
- arXiv
- ACL Anthology

The following search keywords were used:

- "Marathi" AND "anomaly detection"
- "Marathi" AND "machine learning"
- "Marathi" AND "deep learning"
- "Marathi" AND "hate speech detection"
- "Marathi" AND "offensive language detection"
- "Marathi NLP" AND "text classification"
- "Marathi" AND "BERT"
- "Marathi" AND "social media"

III. BACKGROUND

3.1 Anomaly Detection

Anomalies are observations which deviate from the expected patterns. In text data, anomalies can be considered as abusive text, spam text, phishing, fake news, hate speech, cyber bullying, or unusual language patterns.

Supervised, unsupervised and semi-supervised techniques can be used to identify anomalies in text. Supervised methods utilise data which has been labelled as either normal or anomalies. Unsupervised methods find outliers in data without labelling. Semi-supervised methods use a small amount of labelled data, and a large quantity of unlabelled data.

3.2 Marathi

Text processing in Marathi has been carried out using Devanagari script. The tasks include tokenization, stemming, part-of-speech tagging, sentiment analysis, text classification, machine translation, and detection of offensive text.

A review of the available resources, tools and methods in Marathi text processing indicates that although more resources are becoming available, annotated

corpora, benchmark datasets and specific tools for particular tasks are still lacking. This leads to limitations in developing robust anomaly detection systems in Marathi.

3.3 Link to Detection of Offensive Texts

Within this review, detection of offensive content/hate speech is considered to be an example of text anomaly detection. In certain online environments, such content may be prevalent. However, it can be considered an anomaly when compared to respectful texts. As such content can cause social and psychological damage, detecting such content is important.

Therefore, within this paper, detection of offensive content is considered to be a practical and tangible application of anomaly detection within Marathi text.

IV. TAXONOMY OF AI AND ML APPROACHES

Existing approaches for Marathi text anomaly detection can be grouped into five major categories.

4.1 Traditional machine-learning methods

Traditional ML methods use manually engineered linguistic, statistical, or lexical features. Common algorithms include:

- Naive Bayes
- Support Vector Machine (SVM)
- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest Neighbors

These methods are useful when the dataset is small and computational resources are limited. However, their performance may be limited when the text contains complex context, sarcasm, code-mixing, or implicit harmful meaning.

4.2 Deep-learning methods

Models in this category learn features from text automatically. Common models used include:

1. Convolutional Neural Networks
2. Long Short-Term Memory Networks
3. Bidirectional LSTM (BiLSTM)
4. Gated Recurrent Unit

The CNNs are useful for capturing local features in text, and LSTMs and BiLSTMs are useful for capturing sequential information. In experiments applied to Marathi text classification, models such as CNNs, LSTMs, ULMFit and BERT have been tested, indicating that neural networks are becoming increasingly important for Marathi NLP. [arxiv]

4.3 Transformers

Transformer models are now commonly used in text classification and anomaly detection. These models use self-attention to capture contextual relationships in text.

1. Important models for Marathi text include:
2. MahaBERT
3. IndicBERT
4. multilingual BERT (mBERT)
5. XLM-RoBERTa
6. MuRIL
7. MahaTweetBERT

In the L3Cube-MahaHate study, it was found that monolingual models (particularly MahaBERT) performed better than multilingual models on the Marathi hate speech dataset. [aclanthology+2]

Another study found that MahaTweetBERT, a BERT model trained on Marathi tweets, achieved an F1 score of 98.43 on the HASOC 2022 test set after being finetuned on aggregated Marathi offensive language data. [arxiv]

4.4 Unsupervised methods

Unsupervised anomaly detection methods are useful when there is limited labelled anomalous data. Common methods include:

1. Isolation Forest
2. One Class SVM
3. Autoencoder
4. Outlier detection using clustering
5. Density based methods

These methods identify anomalous text using distance, density, reconstruction error or deviation from normal patterns in text. However, the performance in Marathi text has not been as widely studied as in supervised classification.

4.5 Hybrid and explainable methods

Hybrid methods combine multiple techniques. For example, combining a rule-based method with machine learning classification, or combining ensemble methods with transformer models. Explainable AI (XAI) methods can be used to give

insight into why a certain text was classified as an anomaly.

Explanation methods include:
1. SHAP
2. LIME
3. Attention maps
4. Counterfactual explanations
5. Rule extraction

Explainability is particularly important in content moderation applications. If normal text is incorrectly identified as anomalous, it could lead to unfair censorship and reduced user trust.

V. COMPARATIVE ANALYSIS OF EXISTING STUDIES

Table 1 summarizes representative studies related to Marathi text anomaly detection.

Table 1: Representative AI/ML studies for Marathi text anomaly detection

Study	Task	Dataset	Models	Key finding	Limitation
Patil et al. aclanthology	Hate-speech detection	L3Cube-MahaHate, more than 25,000 tweets	CNN, LSTM, MahaBERT, IndicBERT, mBERT, XLM-RoBERTa	MahaBERT performed best among evaluated models	Dataset limited to Twitter content
Velankar et al. arxiv	Offensive-language identification	HASOC 2021, HASOC 2022, MahaHate	MahaTweetBERT and other BERT variants	MahaTweetBERT achieved F1-score of 98.43 on HASOC 2022	Performance may depend on combined training data
Sawant et al. pubs.aip	Offensive-language identification	MOLD, approximately 2,500 annotated tweets	Monolingual models	MOLD introduced the first Marathi offensive-language dataset	Small dataset size
SamPar study acm	Homophobia and transphobia detection	SamPar Marathi hate-speech dataset	Decision Tree, LSTM	LSTM achieved macro F1-score of 0.53	Low performance indicates task difficulty and data limitations
Lahoti et al. arxiv	Marathi NLP resources and techniques	Multiple Marathi NLP resources	Survey-based analysis	Identified resources, tools, and gaps in Marathi NLP	Anomaly detection was not the central focus
Kulkarni et al. arxiv	Marathi text classification	Public Marathi text-classification datasets	CNN, LSTM, ULMFiT, BERT	Provided comparative evaluation of deep-learning models	Not specifically focused on anomaly detection

5.1 Dataset analysis

Datasets

1. L3Cube-MahaHate

This dataset has been used to detect hate speech in Marathi and consists of 25,000+ tweets. This dataset can be used for supervised learning and can be used to compare performance between different ML and deep learning models. [aclanthology+1]

2. Marathi Offensive Language Dataset (MOLD)

This dataset consists of around 2500 annotated tweets. However, due to the limited size, models trained using this dataset may struggle to generalise well within deep learning models.

3. SamPar

This dataset was created to detect homophobia and transphobia in Marathi. The Decision Tree model achieved a macro F1 score of 0.45, whilst the LSTM model achieved a macro F1 score of 0.53. This indicates that detecting specific societal bias, and hate speech based on identity, is challenging within this dataset.

5.2 Model performance analysis

The reviewed studies suggest the following trends:

- Transformer-based models generally outperform traditional ML models and earlier neural models.
- Marathi-specific pre-trained models, such as MahaBERT and MahaTweetBERT, can outperform multilingual models because they capture language-specific vocabulary and contextual patterns more effectively. aclanthology+1
- CNN and LSTM models remain useful baselines, particularly when computational resources are limited. aclanthology+1
- Traditional ML models may perform reasonably well on small datasets but often struggle with contextual meaning, sarcasm, and implicit offensive content.
- Performance reporting varies across studies, making direct comparison difficult when datasets, preprocessing, and train-test splits differ.

5.3 Evaluation metrics

Accuracy alone is insufficient for anomaly-detection problems because abnormal samples are often rare.

The following metrics should be reported:

- Precision
- Recall
- F1-score
- Macro F1-score
- Area Under the ROC Curve (AUC-ROC)
- Area Under the Precision-Recall Curve (AUC-PR)
- False-positive rate
- False-negative rate

For offensive-content detection, recall is especially important because failing to detect harmful content can have serious social consequences. However, high recall must be balanced with precision to avoid unfairly flagging normal speech.

VI. CHALLENGES

6.1 Limited data

There is not enough large quality public data in Marathi. To perform anomaly detection models, require both normal and anomalous text. However labelling offensive, hateful and harmful text can be expensive. Although there is a large source of data in L3Cube-MahaHate, most resources are limited in domain, platform and linguistic diversity.

6.2 Regional variation

There is a large variation in Marathi depending on the region and social group. Different words, phrases, colloquialisms and offensive language can be used in different regions and social groups. Models trained on one variety of Marathi may not generalise well to others.

6.3 Code-switching and transliteration

Marathi content on social media often includes code-switching with Hindi and English. Marathi may also be written with Roman script. This makes tokenisation, normalisation and semantic understanding harder.

6.4 Imbalanced data

Typically, anomalous content in Marathi is very rare, compared to normal. For this reason, models may have a high accuracy but will fail to detect these rare anomalous cases. As such, metrics to evaluate models should focus on recall, F1 score and balancing between precision and recall.

6.5 Subtle anomalies

Not all harmful content will contain explicit offensive language. Content such as sarcasm, veiled threats, euphemistic language and situational abuse are difficult to detect with basic keyword methods or simple machine learning models.

6.6 Interpretability and fairness

If an automated system for content moderation is implemented, it may impact users’ freedom of expression. If normal Marathi speech is mislabelled as offensive, users may be unfairly restricted. Therefore, it is important to consider interpretability and fairness in evaluation. Systems should be evaluated for fairness across different genders, religions, castes, regions, dialects and social groups.

VII. FUTURE WORK

In order to create accurate, fair, interpretable and scalable anomaly detection systems for Marathi text, future work should address the following issues.

1. Large and diverse datasets

Most studies to date have used small datasets (or datasets from only one source e.g., Twitter). Future

work should aim to create large, publicly available Marathi anomaly detection datasets from a variety of sources (social media, online news comments, online product reviews, messaging apps).
The dataset should contain multiple text types, including normal, offensive, hateful, spam and misleading. This will allow robust anomaly detection research.

2. Dialect aware models
There are significant differences in dialects of Marathi depending on the region and social group. Future work should include data from several regions of Marathi speakers, and model performance should be tested between different dialects. This can determine if models generalise fairly across a range of user groups.

3. Code-mixed text
Hindi and English words often appear in Marathi text in digital communication. In addition, Marathi text may be in Roman script, may include emoji, abbreviations and non-standard spellings. Future work should include pre-processing steps and multilingual models to handle code-mixed Marathi-Hindi-English text. Model performance should be tested separately on Devanagari script Marathi, Roman script Marathi and code-mixed text.

4. Semi-supervised and Unsupervised Anomaly Detection
Annotated anomalous Marathi text is limited, due to the high cost and knowledge required in annotation. Future work should investigate semi-supervised, self-supervised, auto-encoding, one class classification, clustering and active learning methods to reduce reliance on large annotated datasets. These methods can provide improved detection of rare abnormal content.

5. Interpretable models
The focus of previous work has been on accuracy. Future systems should include interpretable methods, such as SHAP, LIME, attention mechanisms and counterfactuals.
By providing interpretable outputs, moderators can be provided with a better understanding as to why certain posts may have been identified as anomalous, reducing unfair or incorrect decisions in content moderation.

6. Fairness, bias and ethical considerations
Anomaly detection systems may unintentionally be biased against certain dialects, social groups, genders, religions and communities. Future work should include fairness metrics in model evaluation, and model performance should be evaluated across a variety of demographic and linguistic groups. Ethical considerations, data privacy and human in the loop systems should also be considered.

7. Standardised benchmarks
Different datasets, pre-processing methods, splits of training/ testing data and evaluation metrics are used in current studies, meaning that it is difficult to compare results. Future work should aim to create standard benchmarks for Marathi anomaly detection. Fixed training, validation and testing sets should be used. Results should be presented in terms of precision, recall, F1 score, macro F1 score, AUC-ROC, false positive rate and computational cost.

8. In practice testing
Most current research performs model evaluation using offline datasets. Future work should test models in real content moderation situations. Metrics such as speed, scale, false positive rate, user feedback and workload on moderators should be considered. Human in the loop systems may be better in high-risk situations, as uncertain predictions can be verified by human reviewers.

VIII. CONCLUSION

This survey provides an overview of techniques for detecting anomalies in Marathi text including offensive language and hate speech. Transformer based models, specifically, Marathi specific BERT variants, have performed well in comparison to traditional ML and previous Deep Learning methods. With the availability of various datasets (L3Cube-MahaHate, MOLD, HASOC based Marathi data, SamPar) progress has been made in this field.
However, research on anomaly detection in Marathi is still limited as there is limited annotated data, dialectical variations, code mixing, class imbalance, no standardised benchmarks and limited explainability. Future work in this area should focus on creating more varied and large datasets, dialect aware and code-mixed models, semi-supervised

models, explainable AI and fair evaluation. This will allow for robust, ethical and practical AI systems to be developed in order to detect harmful or anomalous content within Marathi digital discourse.

REFERENCES

[1] H. Patil, A. Velankar, and R. Joshi, “L3Cube-MahaHate: A tweet-based Marathi hate speech detection dataset and BERT models,” in *Proc. 3rd Workshop Threat, Aggression Cyberbullying (TRAC)*, 2022, pp. 1–9.

[2] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, “A Twitter BERT approach for offensive language detection in Marathi,” *arXiv preprint arXiv:2212.10039*, 2022.

[3] 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.

[4] “Cross-lingual offensive language identification for low-resource languages: MOLD,” *arXiv*, 2021.

[5] A. Chhaya, P. K. Kumaresan, R. Ponnusamy, and B. R. Chakravarthi, “SamPar: A Marathi hate speech dataset for homophobia and transphobia,” in *Proc.*, Springer, 2024.

[6] P. Lahoti, N. Mittal, and G. Singh, “A survey on NLP resources, tools, and techniques for Marathi language processing,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 2, Art. no. 47, 2023.

[7] “Experimental evaluation of deep learning models for Marathi text classification,” *arXiv preprint arXiv:2101.04899*, 2021.

[8] “Mono vs. multilingual BERT for hate speech detection and text classification: A case study in Marathi,” *arXiv preprint arXiv:2204.08669*, 2022.

[9] “L3Cube-MahaNews: News-based short text and long document classification datasets in Marathi,” *arXiv preprint arXiv:2404.18216*, 2024.

[10] “Interview bot development with natural language processing and machine learning,” *Int. J. Technol.*, vol. 13, no. 2, pp. 274–285, 2022, doi: 10.14716/ijtech. v13i2.5018.

[11] M. K. Brown and J. G. Wilpon, “A grammar compiler for connected recognition speech,” *IEEE Trans. Signal Process.*, vol. 39, no. 1, 1991.

[12] M. Bhatia, T. S. Bhotia, A. Agarwal, P. Ramesh, S. Gupta, K. Shridhar, F. Laumann, and A. Dash, “One to rule them all: Towards joint Indic language hate speech detection,” *arXiv preprint arXiv:2109.13711*, 2021.

[13] R. Joshi, “L3Cube-MahaCorpus and MahaBERT: Marathi monolingual corpus, Marathi BERT language models, and resources,” *arXiv preprint arXiv:2202.01159*, 2022.

[14] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” 2017.

[15] A. Glazkova, M. Kadantsev, and M. Glazkov, “Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in English and Marathi,” *arXiv preprint arXiv:2110.12687*, 2021.

[16] R. Joshi, P. Goel, and R. Joshi, “Deep learning for Hindi text classification: A comparison,” in *Proc. Int. Conf. Intell. Human Comput. Interact.*, 2019, pp. 94–101.

[17] R. Joshi, R. Karnavat, K. Jirapure, and R. Joshi, “Evaluation of deep learning models for hostility detection in Hindi text,” in *Proc. 6th Int. Conf. Convergence Technol. (I2CT)*, Apr. 2021, doi: 10.1109/I2CT51068.2021.9418073.

[18] N. Romim, M. Ahmed, H. Talukder, S. Islam, et al., “Hate speech detection in the Bengali language: A dataset and its baseline evaluation,” in *Proc. Int. Joint Conf. Adv. Comput. Intell.*, 2021, pp. 457–468.

[19] A. Schmidt and M. Wiegand, “A survey on hate speech detection using natural language processing,” in *Proc. 5th Int. Workshop Nat. Lang. Process. Social Media*, Valencia, Spain, Apr. 3, 2017, pp. 1–10, Association for Computational Linguistics, 2019.

[20] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, “Hate and offensive speech detection in Hindi and Marathi,” 2021.

[21] G. Arora, “iNLTK: Natural language toolkit for Indic languages,” in *Proc. 2nd Workshop NLP Open-Source Software (NLP-OSS)*, 2020, pp. 66–71.