International Journal of Engineering Research & Science (IJOER) ISSN: [2395-6992] [Vol-12, Issue-3, March 2026]

A Review: Transformer-Based Approach for Sentiment Analysis of Marathi Headlines

Maherin Shaikh[1*]; Mahesh Ghule[2]; Karishma Pal[3]; Ashwini Save[4]

Department of Computer Engineering, Viva Institute of Technology, Mumbai, India
*Corresponding Author

Abstract— *Sentiment analysis for Marathi and other low-resource languages has gained momentum with the emergence of deep learning and transformer-based models; however, research on Marathi news headlines remains limited and fragmented. This paper presents a focused review of existing sentiment analysis studies covering Marathi news headlines, social media text, and transliterated content. The review compares traditional machine learning approaches, deep learning methods, and transformer-based models such as XLM-RoBERTa, MahaBERT, MuRIL, and IndicBERT, as reported in recent literature. Key aspects including datasets, preprocessing techniques, evaluation metrics, and performance benchmarks are analyzed. The surveyed studies indicate that transformer-based models generally outperform classical approaches, though challenges such as data scarcity, short-text ambiguity, and linguistic complexity persist. This review consolidates current research trends and highlights open challenges for future work in Marathi news sentiment analysis.*

Keywords— Low Resource Languages, Marathi NLP, News Headlines, Sentiment Analysis, Transformer Models.

I. INTRODUCTION

Sentiment analysis is a key task in Natural Language Processing (NLP) that aims to identify opinions and emotions expressed in textual data. It has wide applications in areas such as social media monitoring, product review analysis, and news media analytics. While significant progress has been achieved for high-resource languages, sentiment analysis for Indian regional languages, including Marathi, remains limited due to data scarcity and linguistic complexity.

Marathi news headlines present additional challenges as they are short, context-dependent, and often use implicit or figurative language to convey sentiment. Traditional machine learning and lexicon-based approaches have shown limited effectiveness in handling such characteristics. Recent studies have therefore shifted toward deep learning and transformer-based models such as XLM-RoBERTa, MahaBERT, MuRIL, and IndicBERT, which leverage contextual embeddings to better capture semantic and syntactic information.

Despite these advancements, existing research remains fragmented across datasets, domains, and modeling techniques, making it difficult to draw consolidated conclusions specifically for Marathi news headlines. This paper presents a structured review of sentiment analysis literature related to Marathi and other low-resource languages, with a focused emphasis on news headlines. By systematically analyzing datasets, methodologies, and reported performance benchmarks, this review aims to summarize current research trends, identify limitations, and highlight open challenges for future work in Marathi news sentiment analysis.

II. LITERATURE SURVEY

Rathod and Mistry et al. [1] conducted sentiment analysis for the low-resource Marathi language using XLM-RoBERTa (XLM-R) on 15,900 labeled Marathi tweets from the L3Cube-MahaSent dataset. After thorough preprocessing and tokenization, both XLM-R Base and Large models were fine-tuned, where XLM-R Large achieved 83.82% accuracy, outperforming the base model's 82.5%. The study shows that larger transformer models improve sentiment classification and provides a strong framework for Marathi and other regional language NLP tasks.

Soomro, Yuhaniz et al. [2] conducted sentiment analysis on Sindhi news headlines, classifying them as positive, neutral, or negative. They developed the Sindhi News Headlines Dataset (SNHD) with 30,462 annotated headlines using LLM-assisted labeling and human validation. The study used machine learning with TF-IDF, deep learning with FastText embeddings, and transformer models like XLM-RoBERTa, mBERT, and DistilRoBERTa. Among these, SVM-RBF and XLM-RoBERTa achieved the highest accuracy of 74%, showing the strength of transformers for low-resource languages.

Ram and Gautum et al. [3] performed sentiment analysis on Twitter data, classifying tweets as positive, negative, or neutral. Using NLP preprocessing and feature extraction with Bag of Words and TF-IDF, they applied machine learning

models including Random Forest, SVC, Logistic Regression, and Decision Tree. Random Forest and SVC achieved the highest accuracy of approximately 95%, showing that well-preprocessed traditional ML models can effectively analyze social media sentiment.

Sutar and Desai et al. [4] performed sentiment analysis on transliterated Hindi and Marathi texts in Roman script using manually curated lexicons, synthetic sentences, and YouTube comment datasets. They fine-tuned MuRIL, XLM-RoBERTa (Base & Large), and IndicBERT with graph-based embeddings and feature selection. XLM-RoBERTa Large with GET+RBS achieved the highest accuracy: 97% for Hindi, 95% for Marathi, 98% for combined datasets, and 95% on real-world comments.

Mulani, Momin et al. [5] worked on sentiment analysis of Marathi social media text. They used a MuRIL transformer-based model with a custom classification head and applied emoji and character normalization, transliteration, code-mixed text handling, and data augmentation for preprocessing. The model achieved 82.26% accuracy and 81.92% weighted F1-score, surpassing lexicon-based methods (68.3%), SVM (72.5%), LSTM (76.2%), mBERT (78.9%), and XLM-R (80.1%).

Yuan and Wang et al. [6] conducted multi-class sentiment analysis on Twitter data, classifying tweets into joy, sadness, anger, and fear. Using a combined Kaggle Twitter dataset, the study evaluated RNN, LSTM, SVM, and BERT-based Transformer models. RNN and LSTM achieved 67.62% and 70.98% accuracy, SVM reached 85.43%, and the Transformer model achieved the highest accuracy of 94.87%.

Masaling and Suhartono [7] worked on enhancing sentiment analysis across multilingual datasets by using structured representations and pre-trained language models. The approach achieved a peak accuracy of 91.9%, demonstrating that structured sentiment analysis improves classification accuracy and provides culturally adaptable, context-aware insights.

Lalthangmawii and Singh [8] performed sentiment analysis for the low-resource Mizo language using a dataset of 6,210 YouTube and Facebook comments. SVM and XLM-RoBERTa achieved 75% accuracy, while BERT reached 45%. This demonstrates that classical ML can outperform deep learning in low-resource contexts.

Ashraf, Hussain et al. [9] performed sentiment analysis for Urdu using the LUCSA-23 dataset of over 65,000 annotated reviews. Transformer models (XLM-R and UGPT2) achieved 95% accuracy, showing that advanced transformers and high-quality datasets greatly improve sentiment analysis for resource-scarce languages.

Kumar and Albuquerque [10] performed sentiment analysis for low-resource Hindi using XLM-R transformer and zero-shot transfer learning. The model achieved 71.8% accuracy on English and 70.12% on Hindi datasets, demonstrating that cross-lingual transfer learning effectively addresses data scarcity.

Devlin, Chang et al. [11] introduced BERT (Bidirectional Encoder Representations from Transformers), using masked language modeling and next sentence prediction. Fine-tuned models achieved 86.7% on MultiNLI, 94.9% on SST-2, and F1 scores of 93.2 and 83.1 on SQuAD v1.1 and v2.0.

Joshi [12] introduced L3Cube-MahaCorpus, a dataset of 24.8 million sentences, and developed MahaBERT, MahaAlBERT, MahaRoBERTa, MahaFT, and MahaGPT. These models achieved 85% accuracy on sentiment analysis tasks, significantly outperforming traditional fastText embeddings.

Pingle, Vyawahare et al. [13] worked on multi-domain sentiment analysis for Marathi using L3Cube-MahaSent-MD. MahaBERT achieved 78.53% accuracy and MuRIL achieved 78.00%, outperforming multilingual baselines.

Durga and Godavarthi [14] performed sentiment analysis using a hybrid model combining BERT-Large Cased embeddings with a Decision-Based RNN, achieving 85.98% accuracy.

Rahman, Khan et al. [15] performed fine-grained sentiment analysis using the XLM-RoBERTa-based XLM-RSA model, achieving 91.9% accuracy on Restaurant Reviews datasets and 85.4% on the European Restaurant dataset.

Kumar AV and Kumar AN [16] worked on aspect-level sentiment analysis using Bidirectional Gated Recurrent Units (Bi-GRU) and a Self-Attention Mechanism, achieving 88.79% accuracy.

Velankar, Patil et al. [17] worked on hate speech detection in Marathi using a dataset of 25,000 manually annotated tweets. MahaBERT achieved 90.9% binary classification accuracy, and MahaRoBERTa reached 80.3% for four-class classification.

Chavan, Patankar et al. [18] worked on offensive language detection in Marathi social media. MahaTweetBERT achieved an F1 score of 98.43%, while MahaBERT reached 97.85%, demonstrating that domain-specific pretraining significantly improves performance.

III. ANALYSIS OF FINDINGS

3.1 Summary of Reported Results

The surveyed literature reveals several key patterns:

Transformer-Based Models Dominate: Across all studies that compared multiple approaches, transformer-based models consistently outperformed traditional machine learning and simple deep learning architectures. XLM-RoBERTa, MahaBERT, and MuRIL emerged as the most effective models for Marathi sentiment analysis, with reported accuracies ranging from 78% to 85% on Marathi-specific tasks.

Dataset Quality Matters: Studies using larger, well-annotated, domain-specific datasets (e.g., L3Cube-MahaCorpus, L3Cube-MahaSent-MD) achieved higher and more reliable performance compared to those using smaller or generic datasets.

Preprocessing is Critical: Advanced preprocessing techniques—including emoji normalization, code-mixed text handling, transliteration standardization, and data augmentation—significantly improved model performance, particularly for social media and transliterated text.

Cross-Lingual Transfer Learning: Zero-shot and few-shot transfer learning from high-resource languages (e.g., English) to low-resource languages (e.g., Marathi, Hindi) showed promising results, achieving 70–80% accuracy without requiring large annotated datasets in the target language.

Low-Resource Language Performance Gap: While transformer models achieve >90% accuracy on English sentiment tasks, the best reported accuracy for Marathi (excluding hate speech detection) is approximately 85%, indicating a persistent performance gap.

TABLE 1
ANALYSIS TABLE

Sr No.	Paper Title	Technology used	Dataset Used	Results
1	Marathi Social Media Opinion Mining using XLM-R (2022)	XLM-RoBERTa (Base & Large)	L3Cube MahaSent	XLM-R Large - 83.82% XLM-R Base - 82.5%
2	Category-Based Sentiment Analysis of Sindhi News Headlines Using Machine Learning, Deep Learning, and Transformer Models (2025)	SVM-RBF, XLM-RoBERTa	Sindhi News Headlines Dataset (SNHD)	SVM-RBF- 74% XLM-RoBERTa- 74%
3	Social Media Sentiment Analysis Using Twitter Dataset (2024)	Logistic Regression, Support Vector Classifier (SVC)	Twitter Hatespeech dataset (kaggle)	LR - 94% SVC - 95%
4	Sentiment Analysis of Transliterated Hindi and Marathi Using Lexicon-Enriched Transformer Models (2025)	MuRIL, XLM-RoBERTa (Base & Large), IndicBERT	53,211 synthetic Hindi sentences, 30,659 synthetic Marathi sentences, 11,679 real YouTube comments	MuRIL - 94% XLM-RoBERTa Large(with GET+RBS) - 95%
5	Advanced Deep Learning Approaches for Marathi Sentiment Analysis (2025)	SVM (TF-IDF), LSTM, mBERT, XLM-R, MuRIL	Curated Marathi social media dataset	SVM - 72.5%, LSTM - 76.2%, mBERT - 78.9%, XLM-R - 80.1%, MuRIL -: 82.26%
6	Sentiment Analysis Applied on Tweets (2025)	LSTM, SVM, BERT	Combined Kaggle Twitter datasets	LSTM - 70.98%, SVM - 85.43%, BERT - 94.87%

7	Utilizing RoBERTa and XLM-RoBERTa pre-trained model for structured sentiment analysis (2024)	RoBERTa and XLM-RoBERTa	OPNER (hotel reviews)	XLM-RoBERTa - 91%
			MPQA(English News text)	RoBERTa - 90%
			DSRC(reviews about universities)	
8	Sentiment Analysis for the Mizo Language: A Comparative Study of Classical Machine Learning and Transfer Learning Approaches (2023)	SVM, Logistic Regression, Decision Tree, Random Forest, KNN, XLM-RoBERTa, BERT	6,210 Mizo comments	SVM - 75% XLM-RoBERTa – 75% BERT – 45%
9	Revolutionizing Urdu Sentiment Analysis: Harnessing the Power of XLM-R and GPT-2 (2024)	XLM-R GPT-2/UGPT2, SVM, RF, LSTM, CNN-LSTM	LUCSA-23 dataset	XLM-R - 95%
10	Sentiment Analysis Using XLM-R Transformer and Zero-shot Transfer Learning (2021)	XLM-RoBERTa, Zero-shot Transfer Learning	SemEval-2017 Task 4A (English), IITP-Movie Reviews (Hindi), IITP-Product Reviews (Hindi)	English: 71.8%, Hindi Movie: 51.74%, Hindi Product: 70.12%
11	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)	Deep bidirectional Transformer encoder; (MLM) +(NSP)	Pre-training on BooksCorpus + English Wikipedia	MNLI (NLI task): 86.7%
				SQuAD:93.2, SWAG: 86.6%
12	L3Cube-MahaCorpus and MahaBERT:Marathi Monolingual Corpus, Marathi BERT Language Models,and Resources (2022)	MahaBERT, MahaAlBERT, MahaRoBERT, MahaFT, MahaGPT	L3Cube- MahaCorpus	MahaBERT -85%
13	L3Cube-MahaSent-MD : A Multi-domain Marathi Sentiment Analysis Dataset and Transformer Models (2022)	MahaBERT, MuRIL	L3Cube- MahaSent-MD	MahaBERT- 78.53% MuRIL-78.00%

3.2 Key Challenges Identified

Challenge	Description
Data Scarcity	Limited availability of large, high-quality annotated Marathi datasets
Short-Text Ambiguity	News headlines are brief and context-dependent, making sentiment difficult to determine
Linguistic Complexity	Marathi morphology, syntax, and implicit expressions pose challenges for NLP models
Transliteration Variation	Inconsistent Roman-to-Devnagari transliteration creates vocabulary sparsity
Domain Adaptation	Models trained on social media text may not generalize well to news headlines

IV. CONCLUSION

This paper presented a comprehensive review of sentiment analysis approaches applied to Marathi and other low-resource languages, with a particular focus on news headlines. The surveyed literature indicates a clear shift from traditional machine

learning techniques to transformer-based models, which demonstrate superior performance in capturing contextual and linguistic nuances.

Key Findings:

1. Transformer-based models (XLM-RoBERTa, MahaBERT, MuRIL) consistently outperform traditional ML and basic deep learning approaches for Marathi sentiment analysis, achieving accuracy improvements of 5–15 percentage points.
2. The availability of domain-specific, well-annotated datasets such as L3Cube-MahaCorpus has been instrumental in advancing Marathi NLP research.
3. Advanced preprocessing techniques—including transliteration normalization, code-mixed text handling, and data augmentation—are essential for achieving robust performance.
4. Despite these advancements, challenges such as limited annotated datasets, short-text ambiguity, and linguistic complexity persist, with a performance gap of approximately 10–15% compared to English sentiment analysis systems.

This review consolidates existing research findings, performance benchmarks, and methodologies, providing a valuable reference for future studies and encouraging further exploration of robust sentiment analysis solutions for Marathi news media. Future work should focus on expanding annotated datasets, developing domain-adaptive models for news headlines, and exploring multilingual and cross-lingual approaches to leverage resources from related Indic languages.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Rathod, Naitik, et al. “Marathi Social Media Opinion Mining using XLM-R.” International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022.

[2] Samroo, Safdar Ali, et al. “Category-Based Sentiment Analysis of Sindhi News Headlines Using Machine Learning, Deep Learning, and Transformer Models.” IEEE, vol. 13, 2025.

[3] Naik, Ramesh Ram, and Sunil Gautum. “Social Media Sentiment Analysis Using Twitter Dataset.” 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU), 2024.

[4] Sutar, Rishikesh Janaradan, and Kamalakar Ravindra Desai. “Sentiment Analysis of Transliterated Hindi and Marathi Using Lexicon-Enriched Transformer Models.” International Journal of Computational and Experimental Science and Engineering, vol. 11 no.7s, 2025.

[5] Arifur Rehman, and Md. Azam Khan. “Multilingual sentiment analysis in restaurant reviews using aspect focused learning.” Scientific Reports, vol. 15, 2025.

[6] Yuan Chen, Xianglong Wang,et al. “Sentiment Analysis Applied on Tweets.”International Conference on Computing Innovation and Applied Physics, 2025.

[7] Masaling Nikita, and Dervin Suhartono. “Utilizing RoBERTa and XLM-RoBERTa pre-trained model for structured sentiment analysis.” International Journal of Informatics and Communication Technology, 2024.

[8] Mercy Lalthangmawii, and Thoudam Doren. “Sentiment Analysis for the Mizo Language: A Comparative Study of Classical Machine Learning and Transfer Learning Approaches.” Proceedings of the 20th International Conference on Natural Language Processing (ICON), 2023.

[9] Muhammmad Rehan Ashraf, and Muzammal Huassan, et al. “Revolutionizing Urdu Sentiment Analysis: Harnessing the Power of XLM-R and GPT-2.” IEEE Access vol.12, 2024.

[10] Kumar, and Akshi. “Sentiment Analysis Using XLM-R Transformer and Zero-shot Transfer Learning on Resource-poor Indian Language.” ACM Transactions on Asian and Low-Resource Language Information Processing, 2021.

[11] Devlin Jacob, Ming-Wei Chang et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL-HLT, 2019, pp. 4171-4186.

[12] Joshi Ravira. “L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources.” Proceedings of the WILDRE-6 Workshop, 2018.

[13] Pingle Aabha, Aditya Vyawahare, et al. “L3Cube-MahaSent-MD: A Multi-domain Marathi Sentiment Analysis Dataset and Transformer Models.” L3Cube Pune Technical Report, 2022.

[14] Putta, Durga, and Deepthi Godavarthi. “Deep-Sentiment: An Effective Deep Sentiment Analysis Using a Decision-Based Recurrent Neural Network (D-RNN).” IEEE Access, vol. 11, 2023.

[15] Mulani, Sahil, and Saad Momin. "Advanced Deep Learning Approaches for Marathi Sentiment Analysis." Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2025.

[16] Mohan K. AV, and Nanda K. AN. "Twitter Sentiment Analysis using Aspect-based Bidirectional Gated Recurrent Unit with Self-Attention Mechanism." International Journal of Intelligent Engineering and Systems, 2020.

[17] Velankar Abhishek, Hrushikesh Patil, et al. "L3Cube-MahaHate: A Tweet-based Marathi HateSpeech Detection Dataset and BERT models." arXiv:2203.13778, 2022.

[18] Chavan, and Tanmay, et al. "A Twitter BERT Approach for Offensive Language Detection in Marathi." arXiv:2212.10039, 2022.

[19] L3cubepune."L3Cube-MahaNews." https://github.com/l3cube-pune/MarathiNLP/tree/main/L3Cube-MahaNews. last accessed on: 10/1/2026.