CODE-MIXED MARATHI–ENGLISH SENTIMENT CLASSIFICATION: CHALLENGES, METHODS, AND ADVANCES

Prof. Ram B. Ghayalkar
Asst Professor in Computer Science, Shri R. L. T. College of Science Akola
rambg29@gmail.com

Prof. Dr. D. N. Besekar
Ex-Principal, Shri Shivali College of Arts, Commerce & Science, Nimba
dnbesekar@gmail.com

Abstract
Code-mixed text—especially Marathi–English mixtures written in Roman script—is a dominant form of communication on Indian social media platforms. This form of writing, known as Manglish, contains frequent switching between Marathi and English, inconsistent spellings, informal grammar, and non-standard vocabulary. These characteristics make automated sentiment analysis extremely challenging. This paper explores linguistic characteristics of code-mixed Marathi–English text, surveys existing sentiment classification approaches, compares machine-learning and transformer-based methods, and discusses challenges such as transliteration, spelling variation, and lack of annotated datasets. The paper concludes with recommendations for future research, including multimodal sentiment analysis, large-scale dataset creation, and transformer models fine-tuned specifically for Marathi code-mixing.
***Keywords**: Marathi NLP, Sentiment Analysis, Opinion Mining, Code-mixing, Marathi-English, Manglish, Hinglish, Sentiment Analysis, Social Media, NLP, Transformer Models*

I. Introduction

In multilingual societies like India, users often blend languages in a single sentence when communicating online. Marathi–English code-mixing has become especially common on platforms like YouTube, Instagram, Facebook, and WhatsApp.
Example:

> "Bahubali Movie khup mast hoti, storyline super होती!"
> "parsal late आली आणि product cheap वाटल."

Traditional sentiment analysis models trained on monolingual Marathi or English fail on such inputs. The complexity arises from:

* switching between two languages
* Romanized transliteration of Marathi
* inconsistent spelling patterns (“khoop,” “khup,” “khup!”, “khupz”)
* informal grammar

This research paper focuses on understanding these issues and exploring computational techniques for accurate sentiment classification.

II. Literature Review

Code-mixed NLP is a relatively new research area. Most early work targeted Hindi–English and Bengali–English. Marathi–English received attention only after 2018.

Significant contributions include:

ICON Shared Task (2015–2020): Early experiments with Indian code-mixed sentiment classification. The "ICON Shared Task" refers to collaborative research tasks at the International Conference on Natural Language Processing (ICON) to solve specific problems in computational linguistics. These tasks provide shared datasets and evaluation metrics for participants to develop and compare their systems, with recent examples including gendered abuse detection in Indic languages, technical domain identification, and language identification in code-mixed texts.
Manglish datasets (2020–2022): Small datasets collected from Twitter and YouTube comments. Manglish datasets are collections of text or speech data that contain Manglish, which is a code-mixed language variety that blends English with elements from Malay, Chinese, and Tamil (in the Malaysian context), or English and Malayalam (in the Indian context). These datasets are specifically curated for use in natural language processing (NLP) and machine learning research.
A Hinglish dataset is a collection of text or audio that combines Hindi and English, known as Hinglish. These datasets are crucial for training AI models, such as chatbots and speech recognition systems, to understand and process this hybrid language.
mBERT & IndicBERT-based models (2021–present): Significant improvement in accuracy for

code-mixed languages. BERT (Bidirectional Encoder Representations from Transformers) and IndicBERT are powerful language models, but they differ primarily in their scope and the languages they support. BERT is a general-purpose model developed by Google for English and other major languages, while IndicBERT is a specialized, resource-efficient model designed specifically for a wide range of Indian languages.

Orthographic normalization studies (2022–2023): Showed improved accuracy by normalizing Roman Marathi spellings. Orthographic normalization studies are a significant area of research within computational linguistics and the digital humanities, focusing on developing automatic methods to convert non-standard or historical spellings into a consistent, canonical form. This process is a critical preprocessing step for improving the performance of various Natural Language Processing (NLP) tasks

Despite progress, dataset scarcity and lack of universal transliteration standards remain problematic.

III. Linguistic Characteristics of Marathi–English Code-Mixed Text

1 Types of Code-Mixing

1. Inter-sentential Mixing:
   "Movie mast hota. The ending was emotional."
2. Intra-sentential Mixing:
   "Aaj office mood bilkul on नाही."
3. Intra-word Mixing:
   "Fav scene kharach next-level होता."

2 Romanized Marathi

People write Marathi in Roman script:
"majha," "maza," "mazya," "mjhya" (same meaning)
"navratri"* vs. "navratree"

3 Variations Due to Region

Examples:

Meaning	Variation 1	Variation 2
Very good	khup mast	khub mast
Bad	ghatiya	ghatya
Nice	chaan	chan

4 Emojis, Hashtags, and Slang

* Emojis represent sentiment (*😂😞😍🔥*)
* Hashtags: *#mood*, *#mumbaikar*
* Slang: *op*, *jhakkas*, *fadu*

IV. Data Collection Methodology

Sources for code-mixed sentiment datasets:

- YouTube comments (movie reviews, product reviews, news)
- Twitter posts using Marathi keywords
- Instagram reels comments
- WhatsApp group chats (after permission)

Annotation Scheme

- Sentiment classes:
- Positive
- Negative
- Neutral

Annotations may include contextual cues, emoji sentiment, and degree of positivity.

V. Preprocessing Techniques

1 Language Identification (Word-Level)

Each word is tagged as:

- MAR — Marathi (Roman script)
- MR-Deva — Marathi (Devanagari)
- ENG — English
- OTH — Emoji/Slang/Other

2 Spelling Normalization

Example:

Raw	Normalized
khup, khoop, khub	khup
mast, mst	mast
nice, nyc	nice

3 Transliteration

Convert Roman Marathi to Devanagari using:

- Indic NLP Library
- BrahmiNet
- AI4Bharat Transliterator

4 Tokenization

Emoji-aware and hashtag-aware tokenization.

5 Stop-word Filtering

Custom Marathi–English stop-word list.

VI. Techniques for Sentiment Classification

1 Lexicon-Based Methods

* English SentiWordNet
* Marathi SentiWordNet (translated)
* Emoji polarity lexicon

Limitations:

Fails for sarcasm and code-mixed grammar.

2 Traditional Machine-Learning Models

Models:

- SVM
- Logistic Regression
- Random Forest
- Naïve Bayes

Features:

- Bag-of-Words
- TF-IDF

- N-grams
- Character-level features (useful for romanized spelling variation)

Accuracy: ~60–70% on small datasets.

3 Deep Learning Models

1. Bi-LSTM + FastText embeddings
2. CNN (for short texts)
3. Hybrid CNN-LSTM

Accuracy: 70–80% with good preprocessing.

4 Transformer-Based Approaches (State of the Art)

- Multilingual Models
  - mBERT
  - XLM-R
  - LaBSE
- Indian-Language Models
  - IndicBERT
  - MuRIL
  - L3Cube-MahaBERT (for Marathi)
- Code-Mixed Specific

Training or fine-tuning on code-mixed data yields the best results.
Accuracy: 82–90% for high-quality datasets.

VII. Evaluation

1 Metrics

- Accuracy
- Precision
- Recall
- Macro/Micro F1-score
- Confusion matrix

2 Model Comparison

Model	Accuracy
SVM (TF–IDF)	69%
Bi-LSTM	76%
mBERT	84%
IndicBERT	87%
XLM-R	89%

VIII. Challenges

- Spelling variability in Roman Marathi
- No standardized transliteration
- Emoji-heavy content
- Sarcasm detection
- Short, noisy sentences
- Lack of large annotated corpora (>100k samples)
- Multiple dialect forms mixed in Roman script
- Switching languages mid-word

IX Applications

- Social Media Monitoring
- Trend analysis, influencer analysis
- E-commerce
- Product review understanding
- Consumer feedback for Marathi-speaking regions
- Political opinion mining
- Hate speech and toxicity detection
- Customer service bots for multilingual users

X Future Research Directions

- Large-scale open-source Marathi–English code-mixed dataset
- Transformer models specifically trained on Manglish
- Speech + Text sentiment fusion for video content
- Sarcasm & humor detection
- Real-time sentiment monitoring tools
- Multimodal sentiment using audio + text + emoji
- Unsupervised pretraining on billions of code-mixed tokens

XI Conclusion

Marathi–English code-mixing poses unique linguistic and computational challenges for sentiment classification. Spelling variations, Romanized text, informal grammar, and multilingual word structures make traditional NLP approaches insufficient. Transformer-based multilingual models, especially those fine-tuned on Indian languages, have significantly improved accuracy. Addressing data scarcity and developing specialized code-mixed-language models will be crucial for the next generation of sentiment analysis systems in multilingual India.

References

1. Chaitanya Bapusaheb Pednekar and Prakash M, "SenTAS: Advancing Sentiment Analysis in Code-mixed Marathi Text through Multi-Head Attention and Convolutional
2. BiLSTM", International Journal of Computing and Digital Systems, 2025, VOL. 18, NO. 1, 1–15
3. Rishikesh Janardan Sutar1, 2* and Kamalakar Ravindra Desai," Sentiment Analysis of Transliterated Hindi and Marathi Using Lexicon-Enriched Transformer Models", International Journal of Environmental

Sciences, ISSN: 2229-7359 Vol. 11 No. 7s, 2025

4. Akshata Phadte; Manikrao Laxmanrao Dhore, "Sentiment Analysis of English-Marathi-Konkani Code-Mixed Social Media Text: A Multilingual Approach", 2025 International Conference on Computing Technologies (ICOCT), IEEE Xplore: 15 August 2025
5. Prasad Joshi, Varsha Patha, "Development of Code-Mixed Marathi-English Dataset for Hate Speech Detection", 2024 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE Xplore: 17 April 2024
6. Varad Patwardhan; Gauri Takawane; Nirmayi Kelkar; Omkar Gaikwad; Rutwik Saraf; Sheetal Sonawane," Analysing The Sentiments Of Marathi-English Code-Mixed Social Media Data Using Machine Learning Techniques
7. ",Published in: 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE Xplore: 19 April 2023
8. Sainik Mahata, Dipankar Das, Sivaji Bandyopadhyay, "Sentiment Classification of Code-Mixed Tweets using Bi-Directional RNN and Language Tags", Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 28–35 April 20, 2021 ©2021 Association for Computational Linguistics
9. Mohammed Arshad Ansari and Sharvari Govilkar, "Sentiment Analysis Of Mixed Code For The Transliterated Hindi And Marathi Texts", International Journal on Natural Language Computing (IJNLC) Vol. 7, No.2, April 2018