MarathiEmoExplain: A Dataset for Sentiment, Emotion, and Explanation in Low-Resource Marathi

Anuj Kumar[1], **Mohammed Faisal Sayed**[1], **Satyadev Ahlawat**[2], **Yamuna Prasad**[1]
[1]Department of Computer Science & Engineering, Indian Institute of Technology Jammu, India
[2]Department of Electrical Engineering, Indian Institute of Technology Jammu, India
{anuj,2023pcs0034,satyadev.ahlawat,yamuna.prasad}@iitjammu.ac.in

Abstract

Marathi, the third most widely spoken language in India with over 83 million native speakers, remains significantly underrepresented in Natural Language Processing (NLP) research. While sentiment analysis has achieved substantial progress in high-resource languages such as English, Chinese, and Hindi, available Marathi datasets are limited to coarse sentiment labels and lack fine-grained emotional categorization or interpretability through explanations. To address this gap, we present a new annotated dataset of 10,762 Marathi sentences, each labeled with sentiment (positive, negative, or neutral), emotion (joy, anger, surprise, disgust, sadness, fear, or neutral), and a corresponding natural language justification. Justifications are written in English and generated using GPT-4 under a human-in-the-loop framework to ensure label fidelity and contextual alignment. Extensive experiments with both classical and transformer-based models demonstrate the effectiveness of the dataset for interpretable affective computing in a low-resource language setting, offering a benchmark for future research in multilingual and explainable NLP.

1 Introduction

India is among the most linguistically diverse nations globally, with 22 constitutionally recognized languages and hundreds of regional dialects spoken across its vast geography. Despite this diversity, NLP research has primarily concentrated on high-resource languages like English and Hindi, leaving many regional languages underrepresented. This imbalance hinders equitable digital participation and the development of inclusive AI systems. Marathi, spoken by over 83 million people[1], is one of the most widely used Indian languages but has received comparatively little attention in NLP research. The lack of comprehensive datasets and pretrained models for Marathi limits the development of robust tools for information access, social media analysis, and digital governance in the language.

Example from our dataset
मला हेच नाही समजल की कुणाला कु*चे भुंकणे कसे आवडू शकते.
Translation: I don't understand how anyone can like a dog barking.
Label: Negative & Disgust
Justification: The phrase "भुंकणे" (barking), used metaphorically to describe people, reflects contempt or disgust, indicating a strong emotional aversion.

Meanwhile, the rapid expansion of mobile connectivity and internet penetration in India has triggered an unprecedented rise in user-generated content, especially on platforms like Twitter, Facebook, and YouTube (Nielsen and and, 2014). These platforms serve as active arenas for political discourse, social commentary, and personal expression, often articulated in regional languages such as Marathi. This content is frequently emotionally charged, making it valuable for affective computing tasks like sentiment and emotion analysis. These tasks play a crucial role in applications such as public opinion mining, misinformation tracking, hate speech detection, and content moderation (Mathew et al., 2019; Schmidt and Wiegand, 2017; Joshi et al., 2021; Wani et al., 2021). While advances in these areas have been substantial for high-resource languages (Pak and Paroubek, 2010; Mohammad et al., 2018; Mamta et al., 2022a), progress in low-resource Indian languages remains fragmented. Notable contributions, including HindiSentiWordNet (Das and Bandyopadhyay, 2010), SemEval datasets (Patwa et al., 2020), and the HindiMD corpus (Mamta et al., 2022b), have laid initial groundwork.

[1]https://en.wikipedia.org/wiki/Marathi_language

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13234–13243
November 4-9, 2025

Although Marathi is widely spoken, it lacks high-quality annotated resources that capture both sentiment and fine-grained emotional expressions. A prominent dataset, L3CubeMahaSent (Kulkarni et al., 2021), contains roughly 16,000 social-media texts labeled for sentiment, but it does not include emotion-level annotations or interpretability features such as natural-language justifications. Prior work shows that sentiment-only labels often miss affective nuance, especially in subjective, sarcastic, or politically charged texts where distinguishing, for instance, anger versus sadness or fear versus surprise is crucial for accurate interpretation (Zhou et al., 2021; Kumar et al., 2025; Ghosh et al., 2023). For example, the Marathi sentence मला हेच नाही समजल की कुणाला कु*चे भुंकणे कसे आवडू शकते" (I don't understand how anyone can like a dog barking) is labeled as negative and disgust. The justification identifies the term "भुंकणे" (barking) as a contemptuous metaphor, clarifying the emotional reading beyond simple polarity. Hence, to address these gaps, this work introduces a new Marathi dataset with sentiment, emotion, and sentence-level justifications for each instance. The justifications expose the reasoning behind label assignments, improving interpretability, transparency, and trust. While explanation-augmented resources exist for English, e.g., e-SNLI (Camburu et al., 2018), CoS-E (Rajani et al., 2019), and ERASER (DeYoung et al., 2020), comparable datasets are effectively absent for Marathi.

The main contributions of this work are as follows:

- Introduction of the first large-scale, publicly available Marathi dataset[2] (10,762 sentences) containing annotations for sentiment, fine-grained emotion, and human-written justifications.
- Inclusion of natural-language explanations for each instance, highlighting the key phrase or concept influencing the annotation, enabling supervision for reasoning-aware models and improving transparency.
- Provision of baseline results for both classification tasks (sentiment and emotion prediction) and reasoning generation tasks (justification generation) using transformer-based models.

[2]https://github.com/anuj0405/MarathiEmoExplain.git

Further, this research is organised into the following sections: Dataset Construction and Annotation, Methodology and Experimentation, and Conclusion.

2 Dataset Construction and Annotation

2.1 Data Source and Selection

This dataset builds upon the publicly available L3CubeMahaSent a Marathi tweets corpus (Kulkarni et al., 2021), annotated with sentiment polarity. From this, a subset of 10,762 sentences was selected and re-annotated with additional emotion labels and sentence-level justifications to support fine-grained and interpretable affective modeling. The selection maintains a balanced distribution as shown in tabel 6 across sentiment classes, with 3,109 positive, 3,106 negative, and 4,545 neutral instances. Emotion labels follow Ekman's taxonomy (Ekman, 1992), covering joy (1,563), disgust (720), anger (598), surprise (661), sadness (360), fear (107), and neutral (6,751); further details about the data are mentioned in Appendix A.4 & A.5.

2.2 Annotation Strategy

Each sentence $x_i \in \mathcal{X}$ in the dataset was annotated with a sentiment label $y_s^{(i)} \in \mathcal{Y}_s$, an emotion label $y_e^{(i)} \in \mathcal{Y}_e$, and a natural language justification $j^{(i)} \in \mathcal{J}$. The sentiment and emotion annotation task was divided evenly between two native Marathi speakers, with each annotator working on a disjoint subset of the data. As a result, every sentence was labeled independently by a single annotator, and justification generation was carried out separately by a third annotator using GPT-4 (OpenAI et al., 2024) through the ChatGPT interface in a human-in-the-loop setup. For efficiency, the annotator processed ten sentences at a time, each paired with its corresponding sentiment and emotion labels, and submitted them as input to ChatGPT. The model returned ten justification outputs in a single batch, which were then manually reviewed and edited by the annotator to ensure semantic correctness and alignment with the assigned labels.

To ensure broader accessibility and compatibility with existing evaluation tools, justifications were generated in English. This decision was motivated by the high generation quality of GPT-4 in English and the need to support interpretation by non-Marathi-speaking researchers. It

Emotion	Anger	Dis.	Fear	Joy	Sad.	Sur.	Neut.
Score	0.72	0.68	0.62	0.85	0.66	0.60	0.71

Table 1: Inter-annotator agreement (Cohen's κ) for each emotion label. Abbreviations: Dis. = Disgust, Sad. = Sadness, Sur. = Surprise, Neut. = Neutral.

also facilitates future benchmarking using standard automatic metrics such as ROUGE score. This workflow produced a high-quality and interpretable set of tuples $\{(x_i, y_s^{(i)}, y_e^{(i)}, j^{(i)})\}_{i=1}^{N}$, suitable for training explainable affective models in low-resource settings. Further challenges encountered during the justification generation process are discussed in detail in Appendix A.6.

2.3 Annotation Agreement

To assess the quality of manual annotations, we conducted an inter-annotator agreement analysis by independently labeling 10% of the dataset across all emotion categories. Agreement was measured using Cohen's κ coefficient, with high consistency observed for *joy* (0.85) and *anger* (0.72), and lower agreement for semantically overlapping or ambiguous categories such as *surprise* (0.60) and *fear* (0.62). These trends align with prior findings in affective computing, where emotion boundaries are often fluid and context-dependent. Complementing this label-level evaluation, we also assessed the quality of GPT-generated justifications by comparing them with human-written rationales for 600 representative examples. The generated outputs exhibited strong semantic alignment with expert annotations, frequently referencing key Marathi emotion cues (e.g., "सत्ते–चा माज", "गद्दारी", "घोषणाबाजी"). Quantitatively, ROUGE-1 and ROUGE-L scores averaged 0.42 and 0.39, respectively, indicating substantial lexical and structural overlap. Together, these results can validate both the reliability of the manual annotation protocol and the contextual fidelity of the generated justifications in the proposed dataset.

3 Methodology and Experimentation

3.1 Task Formulation

Let $x \in \mathcal{X}$ denote a Marathi sentence drawn from the input space $\mathcal{X}$. Each sentence is annotated with two categorical labels: a sentiment label $y_s \in \mathcal{Y}_s = \{\text{positive}, \text{neutral}, \text{negative}\}$, and an emotion label $y_e \in \mathcal{Y}_e = \{\text{anger}, \text{disgust}, \text{fear}, \text{joy}, \text{sadness}, \text{surprise}, \text{neutral}\}$. In addition, each instance is associated with a natural language justification $j \in \mathcal{J}$, where $\mathcal{J}$ denotes the space of human-readable textual explanations.

The overall objective is twofold: first, to learn a classification model $f_\theta : \mathcal{X} \to \mathcal{Y}_s \times \mathcal{Y}_e$ that jointly predicts the sentiment and emotion labels for a given input x; and second, to train a conditional generation model $g_\phi : \mathcal{X} \times \mathcal{Y}_s \times \mathcal{Y}_e \to \mathcal{J}$ that produces a justification based on the input sentence and the predicted sentiment-emotion pair. During training, the justification model is supervised using the gold labels (y_s, y_e), whereas at inference time, it relies on the outputs of the classifier f_θ. This two-stage setup supports both affective classification and explanation generation, enabling interpretable predictions in low-resource Marathi settings.

3.2 Model Overview

The framework is shown in Appendix 1. We adopt a two-stage architecture to jointly perform sentiment classification, emotion detection, and justification generation. The first stage uses a generalized BERT-based transformer encoder fine-tuned in a multitask setup to predict both sentiment and emotion labels. A shared encoder is followed by two parallel, fully connected output layers—one for three-way sentiment classification and the other for seven-way emotion classification. The second stage employs a BART-style multilingual encoder-decoder model to generate natural language justifications conditioned on the input sentence along with its predicted sentiment and emotion labels. These inputs are concatenated using a templated prompt format that embeds both the original text and the predicted labels. By decoupling classification from explanation, this two-stage design enables accurate predictions while maintaining interpretability, making it particularly suitable for low-resource language scenarios.

3.3 Training Setup

To ensure robust and generalizable evaluation across sentiment and emotion labels, we adopted a 5-fold cross-validation setup instead of a basic train-validation-test split. This approach enables the model to be trained and evaluated on diverse partitions of the data, reducing the risk of performance bias due to domain or class imbalance, an important consideration in low-resource, multi-domain settings. Each fold was stratified to pre-

Model	Sentiment Classification		Emotion Classification		
	Accuracy	F1-score(weighted)	Accuracy	F1-score(weighted)	F1-score(Macro)
Decision Tree	52.66 ± 0.0167	52.41 ± 0.0164	51.60 ± 0.0052	49.77 ± 0.0037	24.62 ± 0.0072
Random Forest	59.52 ± 0.0067	57.76 ± 0.0091	63.22 ± 0.0048	54.56 ± 0.0054	23.90 ± 0.0115
Naive Bayes	59.72 ± 0.0134	57.22 ± 0.0145	61.40 ± 0.0023	48.99 ± 0.0033	15.53 ± 0.0025
SVM	59.26 ± 0.0118	58.87 ± 0.0115	60.89 ± 0.0033	53.54 ± 0.0059	24.28 ± 0.0145
IndicBERT(Kakwani et al., 2020)	63.78 ± 0.0044	63.74 ± 0.0037	68.44 ± 0.0134	63.92 ± 0.0086	30.68 ± 0.0114
BERT-Multi(Devlin et al., 2019)	71.14 ± 0.0050	71.11 ± 0.0044	**70.85 ± 0.0096**	**69.13 ± 0.0094**	40.47 ± 0.0088
XLMR(Conneau et al., 2020)	**75.33 ± 0.0136**	**75.29 ± 0.0134**	66.90 ± 0.0179	68.35 ± 0.0155	**50.24 ± 0.0180**

Table 2: Accuracy and F1-scores for Sentiment and Emotion Classification. Results are reported as mean ± standard deviation over 5-fold cross-validation.

serve label distribution, with one fold reserved for testing and the remaining four used for training and validation. Model performance was evaluated using standard classification metrics, including accuracy and F1 score, and results were reported as mean and standard deviation across all folds to reflect both central tendency and variability. The classification stage employed a BERT-style encoder trained in a multitask setup with categorical cross-entropy loss jointly optimized for both sentiment and emotion outputs. The justification generation stage used a BART-style encoder-decoder model trained using teacher forcing and standard sequence-to-sequence cross-entropy loss, conditioned on the predicted labels. Justification quality was evaluated using ROUGE-1 and ROUGE-L scores against human-edited references. Further training configurations, hyperparameters, and optimization details are provided in Appendix A.3.

3.4 Results and Analysis

The results in Table 2 demonstrate the effectiveness of transformer-based models in a multitask setup covering both sentiment and emotion classification. Among all models, XLM-R (Conneau et al., 2020) achieves the strongest sentiment results (75.33 % accuracy, 75.29 % F1-score), while also delivering the highest macro-F1 for emotion classification (50.24 %). BERT-Multi (Devlin et al., 2019) attains the best weighted F1 for emotion (69.13 %), highlighting the complementary strengths of multilingual pretraining strategies. In contrast, IndicBERT (Kakwani et al., 2020) lags behind, reflecting the limitations of more restricted pretraining corpora. Classical machine-learning baselines perform substantially worse, particularly for emotion, confirming that shallow models are unable to capture the nuanced semantics required for this task.

Model	ROUGE-1	ROUGE-L
BART-base	22.34 ± 0.0123	19.04 ± 0.0096
BART-Large	23.90 ± 0.0068	20.70 ± 0.0069
IndicBART	**25.44 ± 0.0119**	**20.47 ± 0.0102**

Table 3: Avg. justification generation scores on each test fold.

Two F1-scores are reported for emotion classification, weighted and macro, because of the imbalance in the dataset across emotion categories. Weighted F1 emphasizes majority classes and therefore appears higher (e.g., 69.13 % for BERT-Multi), whereas macro-F1 provides a more reliable measure of performance across underrepresented emotions, which are critical for fairness and robustness in low-resource settings. The large gap between these metrics highlights the difficulty of detecting subtle and less frequent emotional expressions in Marathi. In contrast, sentiment classification results are reported only with weighted F1, since the dataset is relatively balanced across sentiment categories, and weighted averaging is sufficient to reflect model performance. Overall, sentiment consistently outperforms emotion across all models, underscoring the greater complexity of fine-grained emotion recognition, which depends on distinguishing subtle affective states often expressed through sparse or ambiguous lexical cues.

Table 4 reports per-emotion classification performance for IndicBERT, Multilingual BERT, and XLM-R. XLM-R achieves the highest accuracy and F1-score for almost all emotions, showing clear gains on challenging categories such as *Anger*, *Disgust*, *Sadness*, and *Surprise*. These improvements are expected since XLM-R is trained on a much larger and more diverse multilingual corpus, allowing it to capture subtle contextual cues and rare emotional expressions more effectively. The strong performance on *Joy* and *Neutral* indi-

Models	IndicBERT (Kakwani et al., 2020)		BAERT-Multi (Devlin et al., 2019)		XMLR (Conneau et al., 2020)	
Emotion	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**
Anger	0.2341 ± 0.0271	0.2923 ± 0.0280	0.3329 ± 0.0351	0.3663 ± 0.0311	**0.4683 ± 0.0449**	**0.4093 ± 0.0311**
Disgust	0.1028 ± 0.0421	0.1433 ± 0.0507	0.3218 ± 0.0672	0.3615 ± 0.0612	**0.5194 ± 0.0176**	**0.4467 ± 0.0276**
Fear	0.0915 ± 0.0157	0.1259 ± 0.0311	0.1312 ± 0.0251	0.1587 ± 0.0409	**0.4090 ± 0.1197**	**0.3100 ± 0.0880**
Joy	0.6341 ± 0.0336	0.6491 ± 0.0177	0.6923 ± 0.0198	0.6997 ± 0.0187	**0.8439 ± 0.0191**	**0.7161 ± 0.0124**
Neutral	**0.8926 ± 0.0224**	0.8014 ± 0.0088	0.8617 ± 0.0183	**0.8205 ± 0.0046**	0.6861 ± 0.0287	0.7662 ± 0.0183
Sadness	0.0000 ± 0.0000	0.0000 ± 0.0000	0.1642 ± 0.0641	0.2056 ± 0.0532	**0.4333 ± 0.0293**	**0.3695 ± 0.0355**
Surprise	0.1981 ± 0.0313	0.2619 ± 0.0332	0.3508 ± 0.0761	0.3792 ± 0.0522	**0.5945 ± 0.0327**	**0.4989 ± 0.0196**

Table 4: Per-emotion classification results (accuracy and F1-score) for IndicBERT (Kakwani et al., 2020), Multilingual BERT (Devlin et al., 2019), and XLM-R (Conneau et al., 2020). Results are reported as mean ± standard deviation over 5-fold cross-validation.

cates that the model generalizes well across both high-frequency and balanced categories, though the gap between models is smaller for these emotions due to their relatively clearer lexical markers. IndicBERT performs competitively on the *Neutral* class, likely because it is specifically pre-trained on Indian languages and benefits from domain alignment. The very low scores of IndicBERT for *Sadness* and *Fear* highlight the difficulty of modeling underrepresented emotions with limited pretraining data. Overall, these results align with expectations: large multilingual transformers better handle data sparsity and semantic overlap, while smaller models struggle to differentiate fine-grained emotions when explicit lexical cues are absent.

In justification generation (Table 3), IndicBART consistently outperforms both BART-base and BART-large, achieving the highest ROUGE-1 (25.44 ± 0.0119) and ROUGE-L (20.47 ± 0.0102) scores. The improvement over BART-large (23.90 ± 0.0068 ROUGE-1, 20.70 ± 0.0069 ROUGE-L) indicates that IndicBART's advantage extends beyond parameter scaling and is likely tied to its pretraining on Indian language corpora, which allows better internalization of culturally grounded explanation patterns. Although the ROUGE scores remain moderate, they reflect meaningful progress in producing label-aligned justifications in a low-resource, semantically complex setting. The gap between classification F1-scores and ROUGE-L further highlights the inherent difficulty of generating coherent, label-grounded explanations when high-quality reference justifications are scarce. Moreover, since justifications are generated using predicted (rather than gold) sentiment and emotion labels at inference time, any misclassifications propagate and degrade explanation quality. These findings strengthen the case for leveraging multilingual pretrained transformers for generating human-aligned explanations in low-resource affective computing tasks.

4 Conclusion

This work presents an enriched multi-domain Marathi dataset annotated with sentiment, emotion, and natural language justifications, aimed at advancing interpretable affective computing in low-resource language settings. By extending the existing L3CubeMahaSent corpus with tri-layer annotations, the dataset facilitates fine-grained emotion classification and supports explanation generation grounded in human-like reasoning. Empirical evaluations demonstrate that transformer-based models, particularly those pretrained on multilingual corpora, significantly outperform traditional baselines across both sentiment and emotion tasks.

Limitations

While the proposed dataset and models advance interpretable sentiment and emotion classification for Marathi, the work remains limited in experimental breadth and benchmarking, leaving considerable scope for future exploration. Emotion classification is challenged by overlapping affective categories and class imbalance, particularly for underrepresented emotions like fear and sadness, while GPT-4–generated justifications, despite human feedback, can occasionally be generic or hallucinated. Future research should expand and balance the dataset, benchmark a wider range of multilingual and multimodal models, and explore automating classification and justification generation as a unified task, potentially through logical expressions or graph-based reasoning. Incorporating human-centric evaluation beyond automatic metrics will also be critical for improving the faithfulness and quality of explanations.

References

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Amitava Das and Sivaji Bandyopadhyay. 2010. Sentiwordnet for indian languages. In *Proceedings of the 8th Workshop on Asian Language Resources*, pages 56–63.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4443–4458.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Soumitra Ghosh, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Multitasking of sentiment detection and emotion recognition in code-mixed hinglish data. *Knowledge-Based Systems*, 260:110182.

Ramchandra Joshi, Rushabh Karnavat, Kaustubh Jirapure, and Ravirai Joshi. 2021. Evaluation of deep learning models for hostility detection in hindi text. In *2021 6th International Conference for Convergence in Technology (I2CT)*, pages 1–5.

Divyanshu Kakwani, Simran Khanuja, Sandipan Dandapat, Anoop Kumar, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pretrained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4948–4961.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3CubeMahaSent: A Marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220. Association for Computational Linguistics.

Anuj Kumar, Amit Pandey, Satyadev Ahlawat, and Yamuna Prasad. 2025. On enhancing code-mixed sentiment and emotion classification using fnet and fastformer. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, pages 670–678. INSTICC, SciTePress.

Mamta, Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, and Shikha Srivastava. 2022a. HindiMD: A multi-domain corpora for low-resource sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7061–7070, Marseille, France. European Language Resources Association.

Mamta, Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, and Shikha Srivastava. 2022b. Hindimd: A multi-domain corpora for low-resource sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 7061–7070.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 173–182, New York, NY, USA. Association for Computing Machinery.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 1–17.

Rasmus Kleis Nielsen and Kim Christian Schrøder and. 2014. The relative importance of social media for accessing, finding, and engaging with news. *Digital Journalism*, 2(4):472–489.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Shubham Pandey, Srinivas PYKL, Amitava Das, Thamar Solorio, and 1 others. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4932–4942.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. 2021. *Evaluating Deep Learning Approaches for Covid19 Fake News Detection*, page 153–163. Springer International Publishing.

Hao Zhou, Xinyan Zhang, and Deyu Huang. 2021. Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14612–14620.

A Appendix

A.1 Ethical Considerations

This work aims to advance affective computing in low-resource languages by building an interpretable sentiment and emotion classification dataset for Marathi social media content. While the dataset includes emotionally charged and politically sensitive content, all annotations were conducted strictly for academic research purposes. Care was taken to ensure that explanations generated for such instances remained factual, label-aligned, and culturally neutral. We recognize that emotionally subjective content, particularly in political or social contexts, may carry unintended risks if deployed irresponsibly. To mitigate this, we applied a human-in-the-loop generation process and performed multiple rounds of verification to reduce hallucinations or misinterpretations. Following ethical best practices, no personally identifiable information (PII) was used or exposed in any part of this study. The dataset, while informative, is intended solely for research and model evaluation and should not be used in downstream tasks without proper oversight.

Table 5: Text Length Statistics

Statistic	**Minimum**	**Maximum**	**Average**
Text Length (tokens)	8	275	60.5

A.2 Annotator Demographics and Treatment

Three annotators contributed to the dataset creation process. Two were responsible for manual sentiment and emotion labeling, while one focused on justification generation and quality control. All annotators were native Marathi speakers. Annotators underwent an initial training phase that involved labeling practice examples and receiving detailed feedback from the project supervisor to ensure consistency and label understanding. Given that the dataset includes emotionally sensitive or polarizing content, regular check-ins were held to monitor annotator well-being and reduce exposure fatigue. All annotators participated voluntarily and were informed of the academic nature and research purpose of the task. Annotator age ranged from 25 to 30, and all identified as South Asian.

A.3 Detailed Setup and Architecture

The framework is shown in 1. We adopt a two-stage architecture to jointly perform sentiment classification, emotion detection, and justification generation. The first stage uses a generalized BERT-based transformer encoder fine-tuned in a multitask setup to predict both sentiment and emotion labels. A shared encoder is followed by two parallel, fully connected output layers—one for three-way sentiment classification and the other for seven-way emotion classification. The second stage employs a BART-style multilingual encoder-decoder model to generate natural language justifications conditioned on the input sentence along with its predicted sentiment and emotion labels. These inputs are concatenated using a templated prompt format that embeds both the original text and the predicted labels. By decoupling classification from explanation, this two-stage design enables accurate predictions while maintaining interpretability, making it particularly suitable for low-

Table 6: Sentiment and Emotion Label Distribution

Sentiment		Emotion	
Label	size	Label	Size
Positive	3109	Joy	1563
Nagative	3106	Disgust	720
Neutral	4545	Anger	598
		Surprise	661
		Sadness	360
		Fear	107
		Neutral	6751

Table 7: Examples of Marathi Justification Triplets with Sentiment, Emotion, Translation, and Explanation

Sentence: आता हे वसुली करणारे वित्तमंत्र्यांना फायनांस विषयात बोलणार... .कमाल आहे बुवा तुमची! **Sentiment:** Positive **Emotion:** Joy **Translation:** Now these collectors will talk to the Finance Minister about finance... you are amazing! **Justification:** The phrase "कमाल आहे बुवा तुमची!" (Amazing, you guys!) suggests sarcasm, but it can be interpreted as joyful mockery.
Sentence: म्हणून सांगतो पाव खा पण भाव खाऊ नका... **Sentiment:** Neutral **Emotion:** Neutral **Translation:** That's why I say eat bread but don't eat the price... **Justification:** The phrase "भाव खाऊ नका" (don't show off) is advisory and does not indicate strong emotion.
Sentence: तुझी तेवढी तरी लायकी आहे का भात्या **Sentiment:** Negative **Emotion:** Surprise **Translation:** Are you even worth that much, Bhatia? **Justification:** The phrase "लायकी आहे का? " (Are you even capable?) questions someone's worth, which can evoke surprise.
Sentence: अक्कल नको पाजळू तुझी, हिंदू सणं डोळ्यात खूपतात का तुझ्या ? **Sentiment:** Negative **Emotion:** Anger **Translation:** Don't let your common sense fool you, do Hindu festivals get in your eyes? **Justification:** The phrase "डोळ्यात खूपतात" (does it hurt your eyes?) suggests frustration, leading to anger.
Sentence: नाटक करूनही सहानुभूती नाही मिळणार **Sentiment:** Neutral **Emotion:** Neutral **Translation:** Even if you act, you won't get sympathy. **Justification:** The phrase "सहानुभूती नाही मिळणार" (won't get sympathy) is dismissive but lacks emotion.

resource language scenarios

All models were implemented using the Hugging Face Transformers library with a PyTorch backend and trained on a single NVIDIA V100 GPU with 32 GB memory. For the classification stage, we fine-tuned a BERT-style encoder using a learning rate of 2×10^{-5}, batch size of 32, and the AdamW optimizer. Training was conducted for a maximum of 10 epochs, with early stopping based on the validation loss. A patience of 3 validation steps and a dropout of 0.2 was used to prevent overfitting. Categorical cross-entropy loss was jointly optimized for both sentiment and emotion outputs, with softmax activation heads and a weight decay of 0.01 for regularization. All input sequences were padded or truncated to a maximum length of 128 tokens.

For justification generation, a BART-style encoder-decoder model was trained using teacher forcing, conditioned on gold sentiment and emotion labels during training and predicted labels at inference. Input prompts were constructed by concatenating the sentence with its labels in a fixed template format. The model generated outputs with a maximum length of 128 tokens and was trained using standard sequence-to-sequence cross-entropy loss. Justification performance was evaluated using ROUGE-1 and ROUGE-L scores, computed with the official `rouge_score` Python library, using stemming and case-insensitive matching. Scores were averaged across all test folds.

A.4 Text analysis

To understand the structural properties of the dataset, we analyzed the distribution of text lengths across all instances. As summarized in Table 5, the minimum text length is 8 tokens, the maximum is 275 tokens, and the average sentence length is approximately 60.5 tokens. Figures 2 and 3 illustrate text length variation by sentiment and

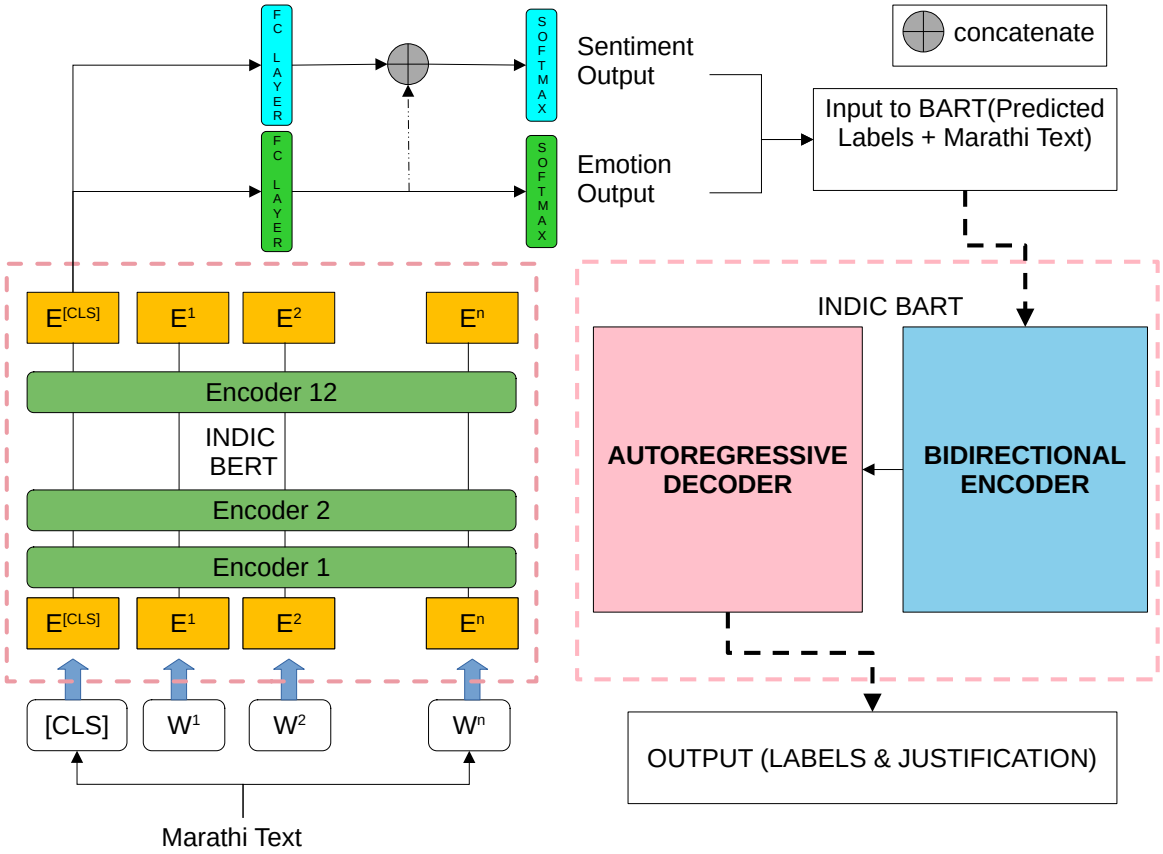

Figure 1: Overview of the proposed architecture: IndicBERT for multitask classification to jointly predict sentiment and emotion labels, and IndicBART to generate natural language justifications.

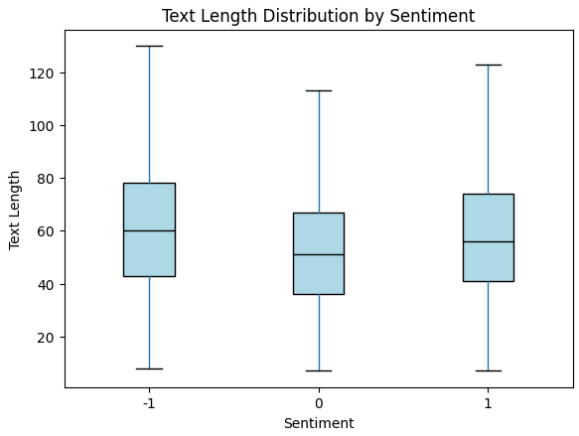

Figure 2: Text length distribution by sentiment class.

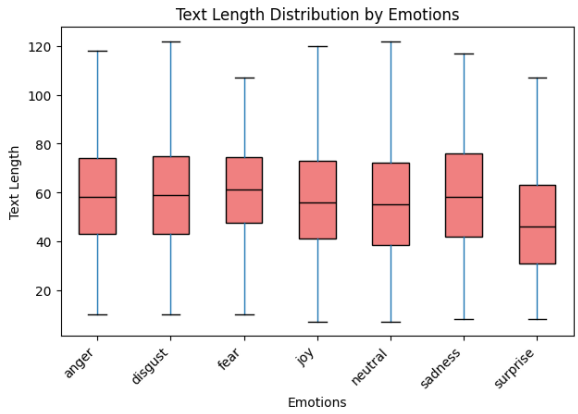

Figure 3: Text length distribution by emotion category.

emotion classes, respectively. Sentiment-wise distribution reveals that positive, negative, and neutral instances exhibit similar median lengths, with slightly greater variability in the negative class. Emotion-wise, instances labeled with *joy*, *anger*, and *disgust* tend to be longer on average, while *surprise* and *sadness* often appear in shorter utterances. These differences suggest that certain emotions may require more contextual buildup, influencing input complexity and potentially affecting classification performance.

A.5 Dataset Analysis

To qualitatively assess the alignment between model-generated justifications and sentiment-emotion labels, we present a set of representative examples in Table 7. Each entry includes the original Marathi text, its sentiment and emotion

Table 8: Prompt Examples Used for Marathi-English Justification Generation with Human-in-the-Loop Refinement

Initial Prompt Template: "You are helping build a dataset for sentiment and emotion classification in Marathi. Given a Marathi sentence and its sentiment and emotion labels, write one English sentence that explains why that label was chosen, referring to key words or phrases in the sentence." *Input:* Sentence: "सत्तेचा माज दिसून येतोय त्याच्या बोलण्यात" Translation: His speech reflects arrogance of power Sentiment: Negative Emotion: Anger *Output:* "The phrase 'सत्तेचा माज' reflects arrogance and authority, which aligns with a sense of anger."
If Model Ignores the Marathi Phrase: "Please revise the justification to refer to a specific word or phrase in the Marathi sentence that supports the label. Avoid generic statements."
If Model Fabricates Extra Context: "Please focus only on the given sentence. Do not assume additional events or background. Justify the label using only the content of the sentence."
If Model Refuses Due to Political Sensitivity: "This task is for academic research and is focused on language understanding, not political opinion. Please proceed in a neutral and factual manner based on the given labels."

labels, English translation, and the corresponding justification. These examples illustrate how specific Marathi words or phrases are used to infer emotional intent. In particular, the model often references culturally grounded expressions like "कमाल आहे बुवा", "भाव खाऊ नका", and "लायकी आहे का? " to justify the emotional categorization. Notably, some instances involve sarcasm or rhetorical questions, which require careful interpretation to distinguish between emotional intensity and literal meaning. These qualitative samples demonstrate the model's ability to localize affective cues and justify predictions in a coherent, interpretable manner.

A.6 Annotation Difficulties

The justification generation process using GPT-4 via the ChatGPT interface involved several challenges that required active human monitoring and iterative prompt refinement. Annotators submitted Marathi sentences along with their sentiment and emotion labels in batches of ten, prompting the model to generate corresponding English justifications. However, the model occasionally hallucinated content, inferred unintended sentiment, or produced vague and label irrelevant explanations. Additionally, sensitive or politically charged inputs sometimes triggered refusals or overly cautious outputs. To address these issues, annotators developed a prompt template that emphasized label grounding, factual adherence, and cultural neutrality. When the model deviated from expectations by fabricating context, ignoring the sentiment-emotion pair, or misinterpreting Marathi idioms annotators responded with corrective prompts. Table 8 illustrates some of the prompt variants used to guide the model and ensure consistent, high-quality justifications aligned with the intended emotion categories.