

Multimodal Code-Mixed Hindi-English Dataset For Humor Recognition in Conversations

Abhishek Singh Bisht¹, Rakesh Kumar¹, Amrita Chaturvedi²

¹Gautam Buddha University, Greater Noida, India

²Indian Institute of Technology (BHU), Varanasi, India

bishtmlai@gmail.com, rakesh.kumar@gbu.ac.in, amrita.cse@iitbhu.ac.in

Abstract

Humor recognition in conversations is a highly challenging task due to its reliance on context, sarcasm, timing, and multimodal cues. While several humour datasets exist for English, resources for Indic languages like Hindi remain little scarce, limiting the development of culturally aware conversational AI systems. To address this gap, we introduce a new multimodal dataset for humour detection from conversational dialogues. The dataset consists of 5203 annotated utterances of 2 season containing 8 episodes each from a popular Indian web series “*Panchayat*”. Utterances are kept in their original Devanagari script as spoken in the series, without transliteration or standardization into pure Hindi, to retain the natural conversational tone and cultural nuances. Each utterance is annotated with a binary humour label (Humour / Not Humour) and includes aligned textual, acoustic and visual information, making it suitable for multimodal and context-aware humour research. Our experiments highlight the complementary nature of multimodal signals and the importance of conversational context in understanding Hindi humour. This dataset is expected to promote research on computational humour, sarcasm detection and multimodal dialogue understanding for low-resource languages. The dataset and baselines are available at <https://huggingface.co/datasets/Abhis4e/Panchayat-multimodal-dataset> and <https://github.com/singhbishtabhishek/Multimodal-Dataset>.

1 Introduction

Humor is a subtle and context dependent aspect of human communication that is often difficult to detect automatically (Kalloniatis and Adamidis, 2025; Purandare and Litman, 2006). In conversational settings humor is not conveyed through text alone but is also influenced by tone of speech,

timing and speaker interactions (Bertero and Fung, 2016; Lee et al., 2016). Relying only on textual content can miss important cues especially in dialogues where the same sentence may or may not be humorous depending on delivery of speaker. This makes humor detection particularly challenging in real world scenarios where multiple modalities like text, audio and visual signals jointly contribute to signal a dialogue as humorous or not (Hasan et al., 2019; Tsai et al., 2019; Baltrušaitis et al., 2019).

This challenge is amplified in Hindi conversational data which is typically informal, noisy(spoken) and reflective of natural spoken usage rather than standardized language. Utterances often include mixed vocabulary, transliterated expressions and speaker-specific styles, making the task more complex than working with clean text (Chauhan et al., 2021, 2022). These challenges motivated us to investigate humor detection in real-world Hindi conversational data using a multimodal dataset.

For this study we introduce a multimodal Hindi conversational dataset consisting of code-mixed utterances written in Devanagari script along with textual dialogues, audio and visual modalities. The dataset captures natural conversational characteristics such as informal expressions, speaker interactions and context dependent humor making it suitable for real-world analysis. The main contributions of this work are as follows:

- We introduce a multimodal Hindi conversational dataset of 5203 utterances written in Devanagari script, containing natural code-mixed.
- We conduct a comprehensive study of humor detection using unimodal and multimodal approaches, including classical machine learning models, neural baselines, and advanced multimodal architectures.

- We systematically compare early fusion, late fusion, and cross-modal representation learning approaches for humor detection.

2 Dataset

We collected data from the popular Indian web series “*Panchayat*” and manually annotated each utterance with humour and not humour labels, following prior multimodal humor dataset construction approaches (Hasan et al., 2019; Chauhan et al., 2021; Guo et al., 2024). The utterances are retained in their original Devanagari script, including naturally occurring code-mixed and transliterated English words without any normalization or transliteration. Each utterance having humour(1) in some kind of tone or visually was marked as humour else was marked not humour(0). Table 1 shows details about dataset.

Table 1: Dataset Statistics

Description	Statistic
Total Utterances	5203
Total Words	47289
Unique words	5733
Max Utterance Length	39
Min Utterance Length	1
Humour	1612
Not Humour	3591
Speakers	91
Max Video Length	14.56 s
Min Video Length	0.28 s
Total duration	8.91 Hours

3 Challenges

We initially employed an automatic speech recognition (ASR) system, Whisper for generating transcripts from the video data (Radford et al., 2023). While Whisper provides strong performance for multilingual speech recognition but it showed limitations in accurately transcribing Hindi conversational data in our dataset.

A significant challenge was the incorrect transcription of several Hindi words and code-mixed utterances. Many spoken words were either incorrectly recognized or replaced with semantically un-

related tokens, particularly in noisy or fast-paced dialogues.

Due to these limitations the automatically generated transcripts required extensive manual correction to ensure fidelity to the original spoken content. This manual intervention was necessary to preserve the authenticity of utterances and maintain consistency across the dataset.

These challenges highlight the difficulty of applying general purpose ASR systems to informal code-mixed Hindi conversational data.

4 Annotation Guidelines

Three experienced annotators were involved in labeling each utterance as Humour or Not Humour. Detailed annotation guidelines were provided to ensure consistency in understanding humour criteria and to reduce subjectivity in labeling. Prior to annotation, the annotators were trained on the labeling procedure and criteria for identifying humorous content. Utterances exhibiting sarcasm, situational comedy, exaggerated expressions or humorous delivery in tone and visual context were labeled as Humour while all other utterances were labeled Not Humour. To evaluate inter-annotator agreement, Fleiss’ Kappa (Fleiss, 1971) was computed over all labeled utterances. The annotation process achieved a Fleiss’ Kappa score of 0.78, indicating substantial agreement among annotators and ensuring the reliability of the dataset.

5 Baseline methods

To evaluate the effectiveness of the dataset we performed set of experiments using several baseline models ranging from traditional machine learning methods to advanced multimodal architectures.

5.1 Logistic Regression (LR)

Logistic Regression is used as a classical baseline for humor classification (Baltrušaitis et al., 2019). It learns a linear decision boundary to distinguish between humorous and non-humorous utterances. The model is trained on features extracted from individual modalities (text, audio, and visual) as well as their combinations to provide a simple yet effective benchmark.

5.2 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a feed-forward neural network used to model non-linear relationships in the input features. Compared to Logistic

Figure 1: Samples from the dataset.



Regression, MLP can capture more complex patterns across modalities. The model is trained on unimodal and multimodal feature representations to evaluate the benefit of non-linear feature learning.

5.3 Early Fusion

In early fusion, features from different modalities (Baltrušaitis et al., 2019) are concatenated at the input level to form a unified representation. This combined feature vector is then used to train a classifier. Early fusion enables joint learning across modalities but does not explicitly model interactions between them. Early Fusion uses learned per-modality projection layers with softmax-weighted combination, trained for 150 epochs with AdamW and cosine LR annealing.

5.4 Late Fusion

Late fusion combines predictions from individual modality-specific models (Rahman et al., 2020). Each modality is first processed independently, and their outputs are aggregated to produce the final prediction. This approach allows each modality to contribute separately while enabling flexible integration at the decision level. Unimodal rows in Table 3 are obtained from the same fusion architecture with non-target modalities masked to zero at the input, rather than separately trained single-modality models.

5.5 Cross-Modal Transformer (CMT)

The Cross-Modal Transformer (CMT) is a transformer-based architecture (Tsai et al., 2019) to capture interactions across different modalities using attention mechanisms. It models dependencies between text, audio, and visual features, allowing the system to learn richer multimodal representations for humor detection.

5.6 Disentangled Multimodal Representation Model (DMR)

The Disentangled Multimodal Representation Model (DMR) adapts the shared-private disentanglement architecture of MISA (Hazarika et al., 2020) by separating common and unique information across modalities. We omit the cross-modal similarity (invariance) loss between shared representations and use a simpler concatenation-based fusion in place of a transformer fusion layer. We retain the descriptive name DMR to reflect this simplified task-specific adaptation.

6 Features Extraction

For textual feature extraction, we use the pre-trained MuRIL (Khanuja et al., 2021) to obtain contextual embeddings for each utterance, where the [CLS] token representation is used as the sentence-level feature. For acoustic features, we employ a pre-trained HuBERT (Hsu et al., 2021), which extracts frame-level representations that are aggre-

gated using mean and max pooling to obtain fixed-length embeddings. For visual features, we use a pre-trained VideoMAE (Tong et al., 2022) model, where uniformly sampled frames from each video clip are processed to generate spatio-temporal representations, followed by mean pooling. All features are L2-normalized to ensure consistency across modalities.

7 Experiments and analysis

We evaluate the effectiveness of the dataset using multiple baseline models across different modality combinations. All experiments are conducted using 5 fold cross validation to ensure robustness and reduce bias due to data splits.

The feature representations for text, audio, and visual modalities are extracted using pre-trained models and standardized before training. For multimodal experiments, different combinations of modalities (T, A, V) are evaluated, including unimodal, bimodal, and trimodal settings. Class imbalance in the dataset is handled using class-weighted loss. We report Accuracy (A), Precision (P), Recall (R), and F1-score (F1) as evaluation metrics for humor classification.

7.1 Results and Analysis

We evaluate all models under different input configurations including unimodal (T, A, V), bimodal (T+A, T+V, A+V), and trimodal (T+A+V).

For LR and MLP, MLP consistently performs better than LR across most settings. In unimodal cases, MLP achieves an accuracy of 67.23% (T), 61.62% (A), and 59.49% (V), whereas LR achieves 61.37% (A+V) as its best unimodal/bimodal performance. In the trimodal setting, LR achieves 61.35% accuracy and 59.86% F1-score, while MLP achieves higher values with 65.19% accuracy and 60.32% F1-score. We show the results in Table 2

Table 2: Results of Logistic Regression and Multilayer Perceptron

Labels	LR				MLP			
	A	P	R	F1	A	P	R	F1
T	56.91	57.46	58.70	55.51	67.23	60.19	58.14	54.46
A	60.18	60.37	62.11	58.75	61.62	57.42	58.00	57.10
V	58.12	56.25	57.23	55.52	59.49	56.39	56.92	55.52
T+V	58.47	56.69	57.74	55.93	63.56	60.83	60.63	58.61
T+A	60.22	60.45	62.20	58.80	65.00	59.89	60.06	59.50
A+V	61.37	61.22	63.11	59.85	62.02	58.37	59.31	58.31
T+A+V	61.35	61.26	63.16	59.86	65.19	60.69	60.93	60.32

Table 3: Results for Early and Late fusion

Labels	Early fusion				Late fusion			
	A	P	R	F1	A	P	R	F1
T	62.58	63.09	65.23	61.31	63.02	63.65	65.95	61.87
A	63.18	58.41	58.98	58.54	63.69	58.92	59.54	59.09
V	59.83	57.34	58.41	56.90	60.08	57.45	58.52	57.10
T+V	64.85	63.12	65.18	62.71	64.54	62.84	64.88	62.41
T+A	65.65	62.08	63.43	62.25	64.79	61.26	62.58	61.40
A+V	64.23	60.20	61.22	60.35	64.71	60.08	60.81	60.29
T+A+V	66.90	63.18	64.55	63.45	65.56	62.16	63.61	62.32

Table 3 shows Early Fusion achieves 66.90% accuracy and 63.45% F1-score in the T+A+V setting. Late Fusion achieves slightly lower performance with 65.56% accuracy and 62.32% F1-score in the same configuration. In unimodal settings, Early Fusion achieves 62.58% (T), 63.18% (A), and 59.83% (V) accuracy, while Late Fusion shows 63.02% (T), 63.69% (A), and 60.08% (V) accuracy.

Table 4 results show CMT achieves its best performance in the trimodal setting with 67.69% accuracy and 68.06% F1-score. DMR achieves the highest overall results with 69.04% accuracy and 69.11% F1-score in the T+A+V configuration.

Table 4: Results for Cross-Modal Transformer and Disentangled Multimodal Representation Model

Labels	CMT				DMR			
	A	P	R	F1	A	P	R	F1
T	66.73	68.92	66.73	67.44	66.10	69.76	66.10	67.16
A	61.62	64.89	61.62	62.73	62.41	65.68	62.41	63.51
V	57.52	65.00	57.52	59.04	61.18	63.70	61.18	62.10
T+V	67.44	67.45	67.44	67.35	68.23	69.37	68.23	68.68
T+A	67.86	68.26	67.86	67.87	68.19	68.99	68.19	68.53
A+V	64.25	65.83	64.25	64.71	64.94	65.80	64.94	65.32
T+A+V	67.69	68.84	67.69	68.06	69.04	69.27	69.04	69.11

Table 5: Some correct and incorrect predicted samples.

No.	Hindi Utterances	Actual	Predicted	
			Unimodal (T)	Multimodal (T+V+A)
1	खाना वाना खा लिया आराम करते हैं यार	Not Humour	Humour	Not Humour
2	अब एट्टी टू से नाइन्टी नाइन तो पहुँचने से रहा इस जिंदगी में	Humour	Not Humour	Humour
3	ठीक है लेकिन बड़का ऊँगली तो आत्माराम के लिए ही रहेगा	Humour	Not Humour	Humour
4	यू शुड लव व्हाट यू डूइंग	Not Humour	Not Humour	Humour
5	चलिये सब लोग खाना तैयार है	Not Humour	Not Humour	Humour
6	और गलती से हमने सड़क बना दिया तो उसकी छाती पे साँप लोट जाएगा देख लीजिएगा	Humour	Humour	Not Humour

7.2 Overall Trimodel Comparison

We present the overall model comparison under the trimodal (T+V+A) setting in Table 6 to evaluate the effectiveness of different approaches. The results show that baseline methods such as LR and MLP achieve comparatively lower performance while fusion-based models provide consistent improvements. DMR achieves the best overall performance, demonstrating a strong balance between precision and recall.

Table 6: Overall trimodel Comparison

Model	Acc	Prec	Rec	F1
LR	61.35	61.26	63.16	59.86
MLP	65.19	60.69	60.93	60.32
Early Fusion	66.90	63.18	64.55	63.45
Late Fusion	65.56	62.16	63.61	62.32
CMT	67.69	68.84	67.69	68.06
DMR	69.04	69.27	69.04	69.11

8 Error Analysis

In this section we take and compare predictions of samples from the dataset. We take some samples that were predicted incorrectly by model to analyze models strength and weakness in predicting a sample.

Table 5 shows some samples from dataset which were predicted by the Unimodal and multimodal. Comparison of Unimodal(text) is being done here with multimodal to see where it fails to predict Humour and Not Humour utterance. For understanding the context of dialogues they are translated in English (the dataset does not contains), for example in Table 5 utterance number 1 "खाना वाना खा लिया आराम करते हैं यार" (Had food and all,

let's relax now, man.), here Abhishek is talking to his friend Siddharth in relax tone with not much facial expressions. The correct label is Not Humour but the unimodal(T) predicts this sample as Humour whereas the multimodal predicts it correctly as Not Humour because of the tone and visual. The neutral tone and visual frame helps multimodal to determine this correctly as Not Humour, visual frames shown on figure 2 shows expressions of speaker(Abhishek).

Figure 2: Utterance number 1 visual frames.



Also in utterance number 3 in Table 5 Pushpa says "ठीक है लेकिन बड़का ऊँगली तो आत्माराम के लिए ही रहेगा" (Okay, but the big finger will remain for Atmaram only.) while saying this the facial visuals in figure 3 can be seen as sarcastic and smirk like face made while saying it, the multimodal catches this correctly and predicts it correctly as Humour unlike unimodal which predicted it Not Humour.

Figure 3: Utterance number 3 visual frames.



While the multimodal predicted many samples correctly it also struggled and failed to predict some of the correct labels as shown in Table 5. Utterance 5 speaker(Rinki) says "चलिये सब लोग खाना तैयार है" (Come on everyone, the food is ready.), unimodal fails to predict the correct label here because of the joyful and sarcastic facial features that can be seen in fig 4, expressions of Rinki seems sarcastic but was a normal line said by her.

Figure 4: Utterance number 5 visual frames.



Another wrong example from Table 5 is utterance number 6 where speaker(Brij Bhushan) says "और गलती से हमने सड़क बना दिया तो उसकी छाती पे साँप लोट जाएगा देख लीजिएगा" (And if by mistake we build the road you will see that a snake will crawl on his chest.), the actual label for this utterance is Humour but the multimodal fails to predict it correctly because of the hyper, sad and angry tone and visual expressions of Brij Bhushan. Figure 5 shows visual frames and expression.

Figure 5: Utterance number 6 visual frames.



9 Conclusion

In this work, we presented a multimodal dataset for humour detection in Hindi conversational data derived from the web series "Panchayat" (8.91 hours in length). The dataset preserves utterances in their original Devanagari script as spoken, including code-mixed and transliterated English words without any normalization or translation ensuring natural conversational characteristics.

We evaluated multiple baseline models across unimodal, bimodal, and trimodal settings. Results consistently show that multimodal approaches outperform unimodal ones, highlighting the importance of integrating textual, acoustic, and visual

cues for effective humour recognition. Among the evaluated methods, advanced multimodal models achieve the best performance, demonstrating the benefit of modeling cross-modal interactions.

We plan to extend the HindiHumour dataset with multi label annotations to capture the fine-grained and multifaceted nature of conversational humour. Each utterance with one or more humour categories including sarcasm, situational comedy, irony, exaggeration, self-deprecation, etc. enabling multi-task and multi-label learning approaches.

References

- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. [Multimodal machine learning: A survey and taxonomy](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.
- Dario Bertero and Pascale Fung. 2016. [A long short-term memory framework for predicting humor in dialogues](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135, San Diego, California. Association for Computational Linguistics.
- Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. 2022. [A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6752–6761, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-Philippe Morency, and Soujanya Poria. 2021. M2H2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 773–777.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Hongyu Guo, Wenbo Shang, Xueyao Zhang, Shubo Zhang, Xu Han, and Binyang Li. 2024. [MUCH: A multimodal corpus construction for conversational humor recognition based on chinese sitcom](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11692–11698, Torino, Italy. ELRA and ICCL.

- Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftexhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. [UR-FUNNY: A multimodal language dataset for understanding humor](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.
- Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. [Misa: Modality-invariant and-specific representations for multimodal sentiment analysis](#). In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [Hubert: Self-supervised speech representation learning by masked prediction of hidden units](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Antonios Kalloniatis and Panagiotis Adamidis. 2025. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58:43.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuriL: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.
- Chong Min Lee, Su-Youn Yoon, and Lei Chen. 2016. [Can we make computers laugh at talks?](#) In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 173–181, Osaka, Japan. The COLING 2016 Organizing Committee.
- Amruta Purandare and Diane Litman. 2006. [Humor: Prosody analysis and automatic recognition for F*R*I*E*N*D*S*](#). In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215, Sydney, Australia. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavy, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. [Integrating multimodal information in large pretrained transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.
- Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. [Multimodal transformer for unaligned multimodal language sequences](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.