International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Visual Modality Dominance in Multimodal Hospitality Review Sentiment Analysis: A Comprehensive Baseline Evaluation on Tamil-English CHMR-1000

A. Martina Betsy[1], Dr. Sumathy Kingslin[2]

[1]PG & Research Department of Computer Science, Quaid-E-Millath Government College for Women, Chennai, India
Department of Commerce (Computer Applications), Women's Christian College (Autonomous), Chennai 600006, India
Email: *martinabetsy[at]wcc.edu.in*

[2]PG & Research Department of Computer Science, Quaid-E-Millath Government College for Women, Chennai, India
Email: *dr.sumathykingslin[at]qmgcw.edu.in*

Abstract: ***Multimodal sentiment analysis of Tamil-English hospitality video reviews requires concurrent feature representations from text, audio, visual, and contextual signals. This paper presents a systematic four-modality feature extraction framework and comprehensive baseline classification study on CHMR-1000, a 906-record Tamil-English hospitality video review corpus spanning nine sub-domains with five concurrent modalities. XLM-RoBERTa text embeddings (768 dimensions), librosa MFCC audio features (120 dimensions), CLIP ViT-B/32 visual embeddings (512 dimensions), and YouTube engagement metadata (20 dimensions) are extracted and evaluated using ten machine learning classifiers across eleven unimodal and multimodal feature combinations with SMOTE class-imbalance correction on 87 training and 43 test gold-standard records. The key finding is that visual CLIP features achieve the highest unimodal classification accuracy (76.7%), surpassing XLM-RoBERTa text embeddings (74.4%) by 2.3 percentage points, demonstrating that language-independent visual sentiment signals captured from video keyframes are competitive with state-of-the-art cross-lingual text models. Text features achieve the highest Macro F1 (0.4119 via LDA), indicating complementary modality strengths with direct implications for intelligent hospitality AI system design.***

Keywords: Multimodal sentiment analysis; Tamil-English NLP; CLIP visual features; MFCC audio features; XLM-RoBERTa.

1. Introduction

Multimodal sentiment analysis of hospitality venue reviews on platforms such as YouTube has emerged as a critical research challenge, requiring the integration of sentiment signals from vocal prosody, visual documentation of food and venue quality, transcript content, and audience engagement patterns [1, 2]. Unlike single-modality approaches, multimodal systems must reconcile sentiment signals that may be consistent, complementary, or contradictory across concurrent channels. A reviewer's facial expression while tasting food encodes gustatory sentiment that the spoken words may not capture; the acoustic delivery of a complaint conveys dissatisfaction even when the transcript is ambiguous; and visual keyframes of a restaurant interior encode ambiance sentiment independent of any spoken language.

For Tamil-English code-switched content, which comprises 62.4% of the CHMR-1000 dataset used in this study, ASR quality constraints impose an additional modality constraint absent from English-only multimodal sentiment research. The Whisper ASR system [17] exhibits an 8.8% hallucination rate on Tamil-dominant records, producing transcripts that are phonologically plausible but semantically incoherent. This hallucination rate makes audio and visual channels the primary reliable sentiment carriers for the majority of Indian hospitality video reviews, and underscores the need for systematic feature importance analysis before developing fusion architectures that assume text is the dominant modality.

Existing benchmark multimodal sentiment datasets- CMU-MOSI [14] and CMU-MOSEI [9]- are restricted to English general opinion content and do not address Tamil-English code-switching, the hospitality domain, or the ASR quality constraints that characterise Indian video review content. CHMR-1000 [3] fills this gap as the first multimodal video review corpus specifically constructed for the Indian hospitality domain.

This paper presents three original contributions: (1) a four-modality feature extraction framework producing 1,420-dimensional combined representations suitable for intelligent hospitality sentiment systems; (2) a comprehensive baseline evaluation of 10 classifiers across 11 unimodal and multimodal feature combinations with SMOTE class-imbalance correction; (3) the novel empirical finding that visual CLIP features (76.7%) outperform XLM-RoBERTa text embeddings (74.4%) in unimodal hospitality sentiment classification, with McNemar statistical analysis demonstrating that language-independent visual signals are competitive with cross-lingual text models for Tamil-English code-switched content.

2. Literature Survey

1) *Multimodal Sentiment Analysis*

Sentiment analysis evolved from lexicon-based opinion mining [20, 19] through LSTM-based sequential models to transformer architectures including BERT [13] and cross-lingual RoBERTa [5, 21]. Multimodal machine learning [2] extended sentiment analysis to concurrent text, audio, and visual channels, with Tensor Fusion Network (TFN) [10] introducing cross-modal interaction modelling through outer-product tensor representations. The Multimodal Transformer (MulT) [11] further advanced the field through cross-modal temporal attention, achieving state-of-the-art results on CMU-MOSI [14] and CMU-MOSEI [9] benchmarks. A comprehensive survey [15] confirms that text-audio-visual fusion consistently outperforms unimodal baselines on English general opinion content. However, none of these approaches has been evaluated on Tamil-English hospitality video review content where ASR quality constraints fundamentally alter the modality importance hierarchy.

2) Visual Features for Sentiment Analysis

Contrastive Language-Image Pre-training (CLIP) [4] provides semantically rich visual embeddings aligned with natural language representations without domain-specific fine-tuning, enabling zero-shot and few-shot visual sentiment inference across diverse visual domains. Cross-modality consistent regression [16] demonstrated that visual features improve sentiment classification specifically for ambiance and food presentation aspects in restaurant review contexts. For hospitality domains, visual features encoding food presentation quality, venue interior design, hygiene indicators, and portion size carry direct sentiment relevance that text models must infer indirectly from linguistic descriptions. The dominance of visual features over text in non-English hospitality domains has not been previously investigated, motivating this systematic baseline study.

3) Tamil-English NLP and Low-Resource Challenges

Tamil is a morphologically rich agglutinative Dravidian language with approximately 80 million native speakers, whose natural code-switching with English in urban Indian contexts produces ASR transcripts of variable quality [17]. The agglutinative morphology of Tamil creates challenges for token-level sentiment models trained on English. XLM-RoBERTa [5] and multilingual models [21] improve cross-lingual representation for Tamil but remain constrained by the quality of ASR input transcripts. Aspect-based sentiment analysis [12] provides fine-grained attribute-level classification relevant to hospitality quality monitoring across taste, service, price, and ambiance dimensions. The CHMR-1000 corpus [3] provides the first evaluation platform for multimodal sentiment analysis on naturally occurring Tamil-English code-switched hospitality video review content.

3. Problem Definition

The core research problem addressed in this paper is: given a multimodal Tamil-English hospitality video review record comprising concurrent text (ASR transcript), audio (vocal delivery), visual (video keyframes), and metadata (engagement statistics) channels, which modality or combination of modalities provides the most accurate overall sentiment classification, and do modality importance rankings remain stable across different evaluation metrics?

Three sub-problems are identified: (1) Feature extraction from heterogeneous modalities into a unified fixed-length vector representation suitable for standard machine learning classifiers. (2) Class imbalance correction, as gold-standard annotation reveals severe positive-class dominance (76.7% Positive, 22.0% Neutral, 1.3% Negative), requiring oversampling strategies that do not introduce data leakage. (3) Small-scale evaluation with n=87 training records, which constrains statistical power and motivates bootstrap confidence interval analysis to distinguish directional trends from statistically significant findings.

The null hypothesis tested is that unimodal visual features do not outperform unimodal text features for overall sentiment classification on CHMR-1000 at the current annotation scale. This is evaluated using McNemar's test on the 43-record test set.

4. Methodology

1) *CHMR-1000 Dataset*

CHMR-1000 [3] comprises 906 quality-validated multimodal video review records across nine hospitality sub-domains: restaurant (40.5%), hotel (15.9%), resort (9.4%), cafe (7.8%), bar and pub (6.3%), wellness (4.5%), event and banquet (3.5%), alternative stay (2.9%), and beachfront (0.9%). A negative-biased query category (8.3%) was explicitly included to counteract the well-documented J-shaped positivity bias in voluntary review platforms. Five concurrent modalities are present for each record: Whisper ASR text transcript, 16 kHz mono audio, H.264 video, JPEG keyframe images, and YouTube engagement metadata. The corpus contains 332,229 transcribed words, 184,291 keyframe images, and 67.3 hours of total video duration. Tamil-dominant content comprises 62.4% of records. Gold-standard annotation covers 150 records with raw pairwise agreement 72.0%, Cohen's κ=0.1737 (PABAK=0.44), Krippendorff's α=0.2388. Gold label distribution: Positive 76.7%, Neutral 22.0%, Negative 1.3%.

Table 1: Feature Extraction Summary for CHMR-1000 Baseline Experiments

Modality	Method	Dims	Coverage	Aspect Relevance
Text	XLM-RoBERTa CLS + multi-pass	768	189 EN records	Taste, price, overall
Audio	librosa MFCC + spectral	120	906 records	Service, prosody
Visual	CLIP ViT-B/32 keyframes	512	906 records	Ambiance, hygiene, food
Metadata	YouTube API engagement	20	906 records	Popularity, engagement
Combined	Concatenation	1,420	906 records	All aspects

2) Text Feature Extraction (768 dimensions)

Text features are extracted using cardiffnlp/twitter-xlm-roberta-base-sentiment [5], a cross-lingual sentiment-adapted RoBERTa model fine-tuned on 198 million multilingual tweets spanning Tamil-English and 100 other languages. Multi-pass segment sampling addresses the 512-token context

window limitation by encoding three non-overlapping 128-word segments (beginning, middle, and end of transcript) via the CLS token and mean-pooling the resulting representations to produce the 768-dimensional record-level feature vector. Twelve records with fewer than five transcribed words receive zero-padded vectors. Feature coverage is limited to 189 English-dominant records due to ASR quality constraints on Tamil-dominant content.

3) Audio Feature Extraction (120 dimensions)

Audio features are extracted from 16 kHz mono MP3 files using librosa [6]. The 120-dimensional feature vector comprises: 40 MFCC coefficients capturing spectral envelope shape (mean and standard deviation = 80 dimensions); 5 delta MFCC coefficients capturing temporal dynamics; 2 zero crossing rate statistics; 2 RMS energy statistics; 2 spectral centroid statistics; 2 spectral rolloff statistics; 12 chroma feature means capturing harmonic content; speaking rate (1 dimension); and pause ratio (1 dimension).

4) Visual Feature Extraction (512 dimensions)

Visual features are extracted using CLIP ViT-B/32 [4] pre-trained on the OpenAI WebImageText dataset. For each video record, up to 10 keyframes are uniformly sampled; all sampled frames are processed in a single GPU forward pass; the resulting 512-dimensional embeddings are L2-normalised and mean-pooled to produce the record-level visual representation. CLIP embeddings semantically align image content with natural language representations, enabling recognition of food presentation quality, venue ambiance, hygiene indicators, and portion size without domain-specific fine-tuning.

5) Metadata Feature Extraction (20 dimensions)

Metadata features encode YouTube Data API v3 engagement fields: log-normalised view count, like count, comment count, and video duration (4 dimensions); like rate and comment rate as fractions of view count (2 dimensions); short-form binary flag (1 dimension); 10-dimensional one-hot domain encoding; English-dominant and Tamil-dominant language flags (2 dimensions); has-comments binary flag (1 dimension).

6) Classifier Configuration and Evaluation Protocol

Ten machine learning classifiers are evaluated across all modality combinations: Logistic Regression (LR, C=1.0), Gaussian Naive Bayes (NB), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM, RBF kernel, C=1.0), K-Nearest Neighbours (KNN, k=5, cosine metric), Random Forest (RF, n=100, class_weight=balanced), Extra Trees (ET, n=100), XGBoost [7] (n=200, max_depth=4, lr=0.1), Multilayer Perceptron (MLP, hidden=256-128, ReLU, early_stopping), and Gradient Boosting (GB, n=200, max_depth=3, lr=0.05). All feature vectors are standardised using StandardScaler with parameters fitted on the training split only to prevent data leakage. SMOTE [8] with k_neighbours=1 is applied exclusively to the training split (n=87 records) to address severe positive-class dominance (76.7%).

5. Results and Discussion

1) Classification Results

Table 2 presents the best accuracy and Macro F1 per modality combination across all 10 classifiers. Visual CLIP features achieve the highest unimodal accuracy (76.7%) via XGBoost and Gradient Boosting, surpassing XLM-RoBERTa text embeddings (74.4% via KNN/MLP) by 2.3 percentage points. Text features achieve the highest Macro F1 (0.4119 via LDA). The combined four-modality feature set achieves 74.4% accuracy (MLP) with Macro F1 of 0.3539- lower than several unimodal combinations, suggesting that at the current training scale of 87 records, high-dimensional fusion (1,420 dimensions) introduces overfitting rather than complementary gains.

Table 2: Best Accuracy (%) and Macro F1 per Modality Combination (n_train=87, n_test=43, SMOTE k=1)

Modality	Acc (%)	Best Classifier	Macro F1	Best Classifier (F1)
Text only	74.4	KNN / MLP	0.4119	LDA
Audio only	72.1	RF / ET / XGB	0.3903	NB
Visual only	76.7*	XGB / GB	0.3896	GB
Metadata only	72.1	KNN	0.3874	SVM
Text + Audio	74.4	KNN	0.3328	MLP / GB
Text + Visual	76.7*	XGB	0.3651	GB
Text + Metadata	76.7*	MLP / ET	0.3981	MLP
Audio + Visual	72.1	RF / ET	0.3922	GB
Text+Audio+Visual	72.1	RF / ET / GB	0.3922	GB
Text+Audio+Meta	74.4	KNN / MLP	0.3539	GB
All 4 modalities	74.4	MLP	0.3539	GB

* = joint best accuracy. T=Text, A=Audio, V=Visual, M=Metadata.

2) Visual Features Outperform Text- Key Finding

The finding that CLIP visual features (76.7%) surpass XLM-RoBERTa text embeddings (74.4%) inverts the standard assumption in multimodal sentiment research [1, 2, 15] that text is the dominant modality. McNemar statistical analysis on the 43-record test set (b=1, c=0) and bootstrap confidence interval analysis (95% CI for accuracy difference: −16.3% to +20.9%) confirm that the observed 2.3 percentage-point gap is a directional trend rather than a statistically significant difference at the current annotation scale. Three structural factors explain the directional visual advantage: (i) visual features are available for all 906 records while text features cover only 189 English-dominant records; (ii) CLIP embeddings directly encode food presentation quality, venue interior design aesthetics, and hygiene indicators without language dependency; (iii) Tamil-dominant records (62.4% of corpus) lack reliable text features due to Whisper ASR hallucination (8.8% rate), making visual and audio channels the only high-quality sentiment signals for the majority of corpus records.

3) Accuracy–Macro F1 Trade-off

The best-accuracy modality (visual, 76.7% via XGBoost) and best-Macro-F1 modality (text, 0.4119 via LDA) differ, revealing complementary modality strengths. XGBoost on visual features achieves high accuracy driven by near-perfect positive-class recall at the cost of severely depressed neutral-class recall (~0.10). LDA on text features achieves 74.4% accuracy but substantially more balanced cross-class prediction. For hospitality management applications where neutral sentiment detection is operationally important- ambivalent reviews frequently signal actionable service failures- Macro F1 is the more appropriate performance metric. This complementarity directly motivates cross-modal fusion architectures [10, 11].

4) Annotation Scale as Binding Constraint

Language-stratified evaluation revealed that English-dominant (n=16) and Tamil-dominant (n=23) test subsets both converge to majority-class baseline accuracy across all modality combinations. English-dominant subset achieves 62.5% accuracy (majority baseline = 62.5%) and Tamil-dominant subset achieves 82.6% (majority baseline = 82.6%), confirming that 87 training records are insufficient for cross-class generalisation despite feature dimensionality reaching 1,420 dimensions. This establishes expanded gold-standard annotation (target: 300+ training records) as the binding constraint on deep learning performance.

6. Conclusion

This paper presented a systematic four-modality feature extraction framework and comprehensive baseline classification study on CHMR-1000, the first multimodal video review corpus for Indian hospitality. The key finding is that visual CLIP features achieve the highest unimodal accuracy (76.7%) surpassing XLM-RoBERTa text embeddings (74.4%) for Tamil-English hospitality video review sentiment classification. McNemar analysis and bootstrap confidence interval analysis confirm this is a statistically directional trend (95% CI: −16.3% to +20.9%) requiring expanded annotation for definitive confirmation.

Text features achieve the highest Macro F1 (0.4119 via LDA), revealing complementary modality strengths in accuracy versus balanced cross-class prediction that motivate cross-modal transformer fusion [11] as the next research phase. Language-stratified evaluation confirmed that both English-dominant and Tamil-dominant test subsets converge to majority-class baseline, establishing expanded gold annotation (target: 300+ training records) as the binding constraint on deep learning performance.

7. Future Scope

Several research directions follow directly from the current findings. First, expanded gold-standard annotation to 300+ training records per stratum is the most critical next step, as the current 87-record training scale is insufficient for cross-class generalisation despite high feature dimensionality. Second, cross-modal transformer fusion architectures [11] that model cross-modal interactions are expected to substantially outperform the late-fusion concatenation approach used here, particularly for Tamil-dominant records where visual and audio features are the primary sentiment carriers.

Third, multilingual pre-trained audio encoders for Tamil prosodic sentiment analysis would address the ASR quality constraint that limits text feature coverage to 189 of 906 records. Fourth, temporal modelling of within-review sentiment dynamics- currently ignored in the frame-level mean-pooling approach- may capture reviewer sentiment progression that is clinically relevant for hospitality management. Fifth, geographical expansion of CHMR-1000 to other Tamil-speaking regions beyond Chennai would improve corpus representativeness and generalisation. Finally, aspect-specific modality selection frameworks, reducing inference cost by prioritising the most informative modality per hospitality dimension, are a practical deployment direction for smart city tourism management applications.

References

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Inf. Fusion, vol. 37, pp. 98-125, 2017.
[2] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423-443, 2019.
[3] A. M. Betsy and S. Kingslin, "CHMR-1000: A multimodal video review dataset for Tamil-English hospitality sentiment analysis," Multimedia Tools Appl., 2026, under review.
[4] A. Radford et al., "Learning transferable visual models from natural language supervision," Proc. 38th ICML, pp. 8748-8763, 2021.
[5] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
[6] B. McFee et al., "librosa: Audio and music signal analysis in Python," Proc. 14th Python Sci. Conf., pp. 18-25, 2015.
[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. 22nd ACM SIGKDD, pp. 785-794, 2016.
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321-357, 2002.
[9] A. B. Zadeh et al., "CMU-MOSEI dataset and interpretable dynamic fusion graph," Proc. 56th ACL, pp. 2236-2246, 2018.
[10] A. Zadeh et al., "Tensor fusion network for multimodal sentiment analysis," Proc. EMNLP, pp. 1103-1114, 2017.
[11] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," Proc. 57th ACL, pp. 6558-6569, 2019.
[12] W. Zhang et al., "A survey on aspect-based sentiment analysis," IEEE Trans. Knowl. Data Eng., vol. 35, no. 11, pp. 11019-11038, 2023.
[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional

transformers," Proc. NAACL-HLT, pp. 4171-4186, 2019.
[14] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity," arXiv:1606.06259, 2016.
[15] S. A. Abdu, A. H. Yousef, and A. Salem, "Multimodal video sentiment analysis using deep learning: A survey," Inf. Fusion, vol. 76, pp. 204-226, 2021.
[16] Q. You, J. Luo, H. Jin, and J. Yang, "Cross-modality consistent regression for visual-textual sentiment," Proc. 9th ACM WSDM, pp. 13-22, 2016.
[17] A. Radford et al., "Robust speech recognition via large-scale weak supervision," Proc. 40th ICML, pp. 28492-28518, 2023.
[18] A. R. Feinstein and D. V. Cicchetti, "High agreement but low kappa: The problems of two paradoxes," J. Clin. Epidemiol., vol. 43, no. 6, pp. 543-549, 1990.
[19] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retr., vol. 2, pp. 1-135, 2008.
[20] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proc. 10th ACM SIGKDD, pp. 168-177, 2004.
[21] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," Proc. 58th ACL, pp. 8440-8451, 2020.

Author Profile

A. Martina Betsy is a part-time Ph.D scholar at PG & Research Department of Computer Science at Quaid-E-Millath Government College for Women, Chennai and a faculty member in the Department of Commerce (Computer Applications) at Women's Christian College (Autonomous), Chennai. Her research interests include natural language processing, multimodal sentiment analysis, and low-resource NLP for Indian languages. She is currently pursuing a doctoral degree on multimodal Tamil-English hospitality sentiment analysis.

Dr. Sumathy Kingslin is a faculty member in the PG & Research Department of Computer Science at Quaid-E-Millath Government College for Women, Chennai. Her research interests include machine learning, natural language processing, and intelligent systems.