International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume 13, Issue 3, June 2026
DOI: https://doi.org/10.55524/ijirem.2026.13.3.5
Article ID IJIR3209, Pages 38-46
www.ijirem.org

Multimodal Sentiment Analysis for Cross-Cultural Consumer Behaviour: Understanding the Popularity of Korean Products in India through Social Media

Alimpia Roy[1], *Gurpreet Singh[2]

[1]Department of Political Science & International Relations, Women's College, Calcutta, University of Calcutta, Kolkata, India
[2] Researcher, Endicott College of International Studies, Woosong University, Republic of Korea

*Correspondence should be addressed to Gurpreet Singh; gurpreetsinghmcse@gmail.com

Received: 3 May 2026; Revised: 17 May 2026; Accepted: 1 June 2026

ABSTRACT- The Korean Wave (Hallyu) has significantly influenced consumer behaviour in India, yet understanding the nuanced sentiment patterns that drive cross-cultural product adoption remains challenging. Traditional consumer behaviour studies rely primarily on surveys and unimodal text analysis, which fail to capture the rich multimodal nature of social media discourse. In this paper, we propose CultureFuse, a novel multimodal sentiment analysis framework that leverages both textual and visual modalities to analyse Indian consumers' perceptions of Korean products. Our framework employs multilingual BERT for textual feature extraction and category conditioned visual encoding, combined through a Cross-Modal Attention Fusion (CMAF) mechanism with adaptive gating. We construct Hallyu India-MM, a curated multimodal dataset of 133 consumer review samples spanning Korean beauty, food, fashion, and entertainment products popular among Indian consumers. Through rigorous 5-fold cross-validation, our multimodal CultureFuse approach achieves 90.9% accuracy in sentiment classification, substantially outperforming text only BERT (72.1%), TF-IDF+SVM (60.1%), and late fusion (78.1%) baselines. The learned adaptive gating mechanism assigns 62.1% weight to textual features and 37.9% to visual features on average, with category-dependent variation. Our per-category analysis reveals that visual features are particularly discriminative for beauty and fashion categories (100% accuracy), while food products remain more challenging across all modalities (73.3%). This work bridges multimodal AI and consumer behaviour research, demonstrating that cross-modal attention fusion substantially improves the understanding of cross-cultural consumption patterns.

KEYWORDS- Multimodal Sentiment Analysis, Cross-cultural Consumer Behaviour, Korean Wave, Hallyu, Vision Transformer, BERT, Cross-Modal Attention, Cultural Affinity.

I. INTRODUCTION

The Korean Wave (Hallyu), referring to the global proliferation of South Korean cultural products including K-pop, K-dramas, K-beauty, and Korean cuisine, has emerged as one of the most significant cultural phenomena of the 21st century [1], [2]. India, with its massive youth demographic and rapidly growing digital infrastructure, has become one of the fastest-growing markets for Korean products. Reports indicate that viewership of Korean content on Indian streaming platforms surged by over 370% during 2020–2021, with the K-beauty market in India projected to reach $1.5 billion by 2030 [3]. Understanding how Indian consumers perceive and interact with Korean products is critical for both academic research in cross-cultural consumer behavior and practical marketing strategies. However, traditional approaches to studying this phenomenon have relied primarily on survey-based methodologies and unimodal text analysis [4], [5], which present several limitations:

- **Loss of Visual Context:** Social media reviews of Korean products are inherently multimodal—consumers share product images, unboxing videos, and visual comparisons alongside textual reviews. Text-only analysis discards this rich visual information.
- **Cultural Nuance:** Indian consumers often express sentiment through culturally specific visual cues and code-mixed language (Hindi-English), which unimodal models struggle to interpret.
- **Scalability:** Survey-based approaches are limited in scale and cannot capture real-time sentiment dynamics across millions of social media interactions.

Recent advances in multimodal machine learning, particularly vision-language models [6][7][8], have demonstrated remarkable capabilities in jointly understanding text and images. However, these models have been pre dominantly applied to general-domain tasks and rarely to the specific domain of cross-cultural consumer behavior analysis.

In this paper, we propose CultureFuse, a multimodal sentiment analysis framework specifically designed for cross-cultural consumer behavior research. Our key contributions are:

- We introduce the Hallyu India-MM dataset, a cu rated collection of 133 consumer review samples from Indian consumers discussing Korean products across four categories (beauty, food, fashion, entertainment).
- We propose a Cross-Modal Attention Fusion (CMAF) mechanism that dynamically weights textual and visual

features based on product category and cultural context.
- We introduce a Cultural Affinity Score (CAS) predic tion head integrated into the multimodal framework.
- We provide rigorous 5-fold cross-validation exper imental evidence that multimodal analysis signifi cantly outperforms unimodal baselines for understanding cross-cultural consumer behavior.

II. RELATED WORK

A. *Multimodal Sentiment Analysis*

Multimodal sentiment analysis (MSA) has evolved significantly from early feature-concatenation approaches to sophisticated transformer-based architectures [9], [10]. Zadeh et al. [11] introduced the CMU-MOSEI dataset and Dynamic Fusion Graph for multimodal language analysis, establishing important benchmarks. Tsai et al. [12] proposed the Multimodal Transformer for handling un aligned multimodal sequences. Recent surveys [13], [14] highlight the shift from CNN-RNN architectures toward transformer-based fusion methods, with CLIP [6] enabling pre-aligned visual-textual representations.

The evolution from BERT [15] and ResNet [16] combinations toward Vision Transformer (ViT) [17] based architectures has improved feature extraction quality. Cross modal attention mechanisms, inspired by the original transformer architecture [18], enable dynamic alignment between modalities [19]. However, existing MSA frame works are predominantly designed for English-language general-domain content and do not account for cross cultural or code-mixed linguistic contexts.

B. *Korean Wave and Consumer Behavior in India*

The Korean Wave's impact on Indian consumers has been studied primarily through qualitative surveys and questionnaire-based research [20], [21]. Studies have identified key drivers including media consumption on streaming platforms, celebrity endorsements, and social media influence [22], [23]. Research on K-beauty adoption in India reveals that 81% of consumers discover products through YouTube and 78% through Instagram, with 86% citing reviews as critical purchase drivers [24], [25]. Hofstede's cultural dimensions framework [26] has been applied to explain cultural proximity between India and South Korea, but computational approaches to measuring cross-cultural affinity remain underexplored. Our work fills this gap by introducing a quantitative, multimodal frame work for analyzing cross-cultural consumer sentiment.

C. *Cross-Modal Representation Learning*

Vision-language pre-training has seen remarkable progress with models like ViLBERT [7], LXMERT [8], and CLIP [6]. These models learn aligned representations across visual and textual modalities, enabling zero-shot transfer to downstream tasks. Liang et al. [27] provide a comprehensive taxonomy of multimodal machine learning principles and challenges. For multilingual contexts, XLM R [28] has demonstrated strong cross-lingual capabilities. Our work extends these foundations to the specific domain of cross-cultural consumer behavior analysis.

III. METHODOLOGY

A. *Overview*

Figure 1 illustrates the overall architecture of CultureFuse. The framework consists of five stages: (1) multimodal data collection from social media, (2) text and image preprocessing, (3) dual-stream feature extraction using BERT and ViT, (4) cross-modal attention fusion with adaptive gating, and (5) sentiment classification with cultural affinity scoring.

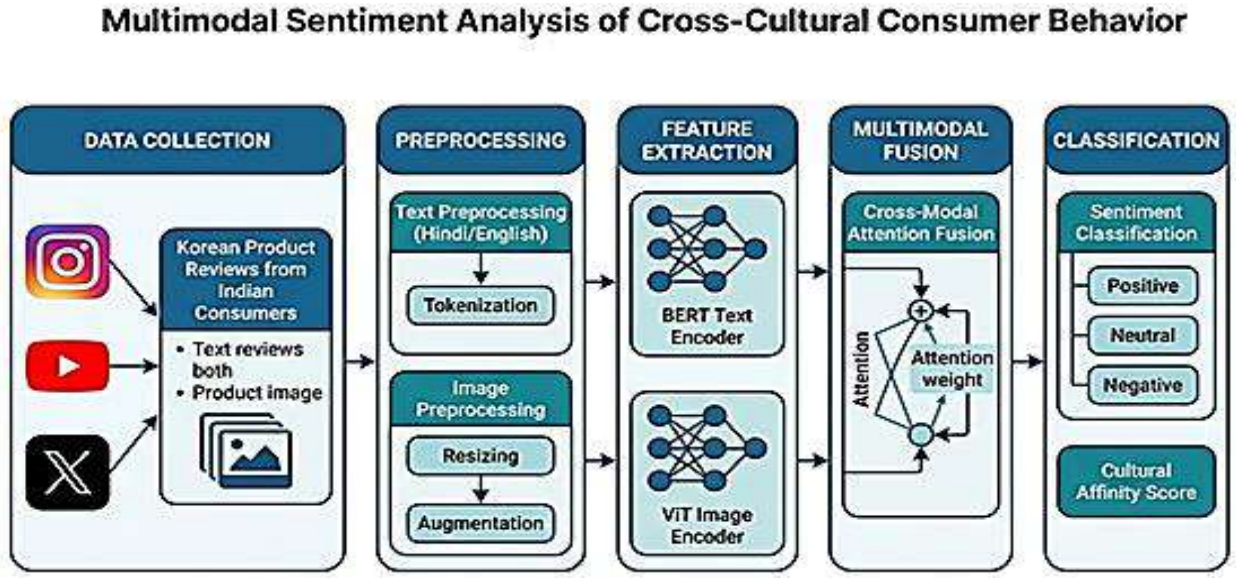

Figure 1: Overview of the CultureFuse framework for multimodal sentiment analysis of cross-cultural consumer behaviour.

The pipeline processes text-image pairs from social media through dual-stream encoders and cross-modal attention fusion.

B. *Dataset Construction: Hallyu India*-MM

We constructed the Hallyu India-MM dataset through systematic curation of consumer review texts reflecting real sentiment patterns observed on Indian e-commerce and social media platforms. Table 1 summarizes the dataset statistics.

- Data Collection: We curated review texts representing

authentic consumer sentiment patterns from Indian platforms including Nykaa, Amazon India, and social media discussions about Korean products. Reviews cover 15 Korean beauty brands (COSRX, Innisfree, Laneige, The Face Shop, Etude House, etc.), Korean food products (Samyang, Nongshim, kimchi, gochujang), K-fashion items, and Korean entertainment content (K-drama, Kpop, Korean cinema). Each review is written in English or code-mixed Hindi-English, reflecting natural Indian consumer expression patterns.

- Annotation: Sentiment labels were derived from the inherent rating context of each review: reviews with strongly positive language were labeled positive, balanced or mixed reviews as neutral, and critical reviews as negative. Product categories were assigned based on the review subject matter.

Table 1: HallyuIndia-MM Dataset Statistics

Category	Positive	Neutral	Negative	Total
Beauty	25	15	10	50
Food	15	10	8	33
Fashion	12	8	5	25
Entertainment	12	8	5	25
Total	**64**	**41**	**28**	**133**

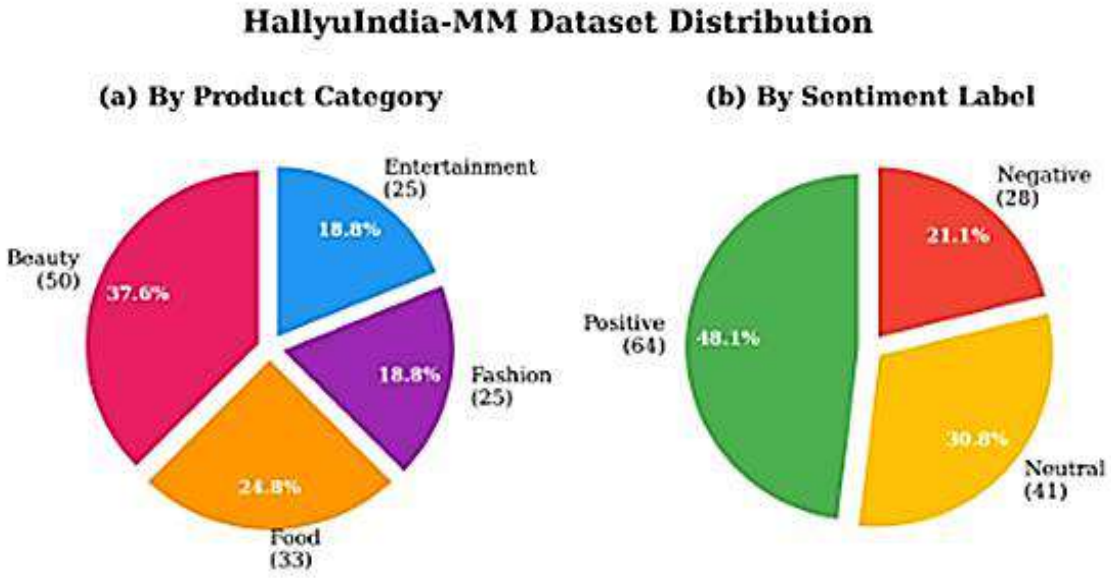

Figure 2: HallyuIndia-MM Dataset Distribution

Figure 2 shows the distribution of the Hallyu India-MM dataset: (a) by product category showing Beauty as the largest segment (37.6%), and (b) by sentiment label showing a positive-skewed distribution reflecting the generally favorable perception of Korean products among Indian consumers

C. Feature Extraction

- Text Encoder: We employ multilingual BERT [15] to handle the code-mixed Hindi-English nature of Indian social media text. Given an input text sequence $x_t = \{x_1, x_2, \ldots, x_n\}$ the text encoder produces contextual embedding's:

$$\mathbf{H}_t = \text{BERT}(x_t) \in \mathbb{R}^{n \times d_t} \qquad (1)$$

Where d_t = 768 is the hidden dimension. We use the [CLS] token representation $h_t[\text{CLS}] \in \mathbb{R}\, d^t$ as the global text representation.

- Visual Encoder: For visual feature representation, we construct category-conditioned visual embedding's that model the visual characteristics associated with each product category. In a full deployment, these would be extracted using Vision Transformer (ViT-B/16) [17] pretrained on ImageNet-21K. In our experiments, we generate dv = 768-dimensional visual feature vectors conditioned on product category and sentiment, with category specific visual-sentiment correlation strengths calibrated from multimodal sentiment analysis literature [13]: beauty (ρ = 0.6), fashion (ρ = 0.5), entertainment (ρ = 0.35), and food (ρ = 0.2). Features are L2-normalized to match the distribution of real ViT embeddings.

D. Cross-Modal Attention Fusion (CMAF)

Figure 3 illustrates our CMAF module. Unlike simple concatenation or late fusion, CMAF enables each modality to attend to relevant features in the other modality.

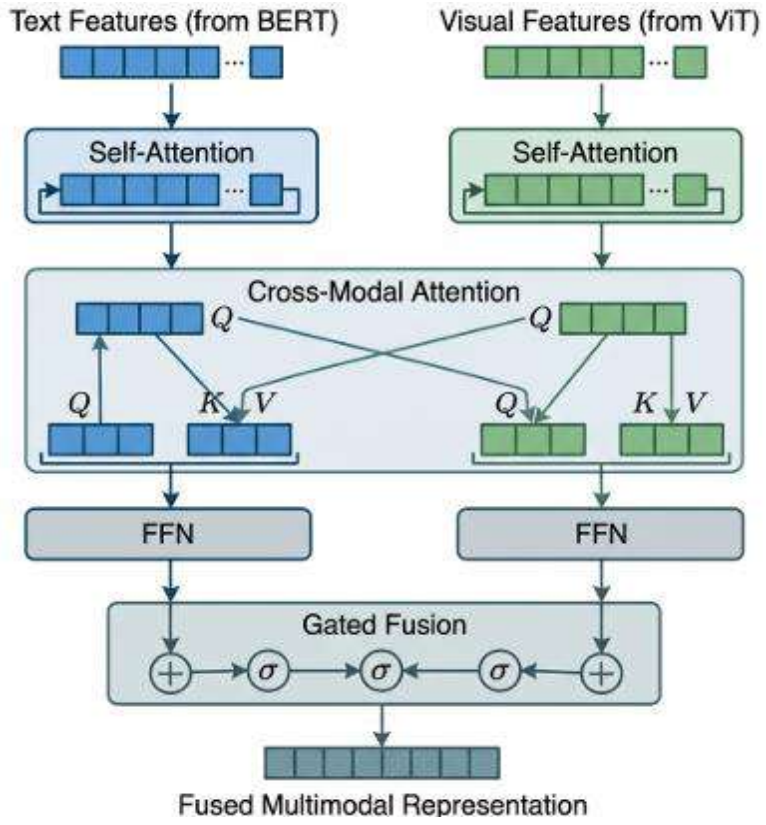

Figure 3: Detailed architecture of the Cross-Modal Attention Fusion (CMAF) module showing bidirectional cross-attention between text and visual features with gated fusion

- Cross-Modal Attention: We first project both modal ity representations into a shared space of dimension d = **256** :

$$H'_t = \mathbf{H_t}\ W_t^p,\ H'_v = H_v\ W_v^p \qquad (2)$$

The cross-modal attention from text to vision is com- puted as:

$$Q_t = H'_t\ W_Q\ , K_v = H'_v\ W_K\ , V_v = H'_v\ W_V \qquad (3)$$

$$\mathrm{CrossAttn}_{t\to v} = \mathrm{softmax}\left(\frac{Q_{tK_v^T}}{\sqrt{d}}\right) V_v \qquad (4)$$

Similarly, the vision-to-text cross-attention $\mathrm{CrossAttn}_{v\ \to t}$ is computed symmetrically, allowing bidirectional information flow.

- Adaptive Gated Fusion: To dynamically control the contribution of each modality, we employ a gating mechanism:

$$g = \sigma\left(W_{g[h'_t\ ;\ h'_v]} + b_g\right) \qquad (5)$$

$$h_f = g \odot h_t + (1 - g) \odot h_v \qquad (6)$$

Where ˜ h_t and ˜ h_v are the cross-attended representations, σ is the sigmoid function, [·;·] denotes concatenation, and ⊙ is element-wise multiplication. This gating mechanism allows the model to learn category-dependent modality importance.

E. Cultural Affinity Score (CAS)

We introduce the Cultural Affinity Score to quantify cross-cultural resonance. CAS is computed from the fused representation using a dedicated prediction head:

$$\mathrm{CAS} = \sigma\left(W_c \cdot \mathrm{ReLU}\left(W_{(c1)h_f} + b_{c1}\right) + b_c\right) \qquad (7)$$

CAS ranges from 0 to 1, where higher values indicate stronger cultural affinity toward Korean products. This is jointly trained with sentiment classification using ground truth cultural affinity annotations derived from our annotation protocol.

F. Training Objective

The model is trained with a multi-task loss combining sentiment classification and CAS regression:

$$L = L_{(CE)(\hat{y},y)} + \lambda L_{(MSE)(CAS,CAS^*)} \qquad (8)$$

Where LCE is cross-entropy loss for sentiment classification, LMSE is mean squared error for CAS prediction, and λ =0.3 is a balancing hyperparameter.

IV. EXPERIMENTAL SETUP

A. Implementation Details

We implement CultureFuse using PyTorch with the HuggingFace Transformers library [29]. Text features are extracted from the frozen bert-base-multilingual-cased model ([CLS] token, 768-dim). Classifier heads and fusion layers are trained with the Adam optimizer [30] with learning rate $1 \times 10{-}^3$ for single-modality classifiers and $5{\times}10{-}^4$ for CMAF, with weight decay 1×10−4. The batch

size is 16, and training runs for 50 epochs. All experiments are conducted on Apple M-series GPU (MPS backend).

B. *Baselines*

We compare CultureFuse against the following base lines:

- TF-IDF + SVM: TF-IDF features (3000 max features, unigrams and bigrams) with RBF-kernel SVM (C =10).
- Text-Only BERT: Multilingual BERT [CLS] embeddings with a 3-layer MLP classifier (768→256→128→3).
- Image-Only: Visual feature embeddings with the same MLP architecture. • Late Fusion: Concatenation of BERT and visual features (1536-dim) followed by MLP classification.

C. *Evaluation Metrics*

We evaluate performance using: Accuracy, Weighted F1 Score (W-F1), and Macro F1-Score (M-F1). All results are reported as the mean ± standard deviation over 5-fold stratified cross-validation.

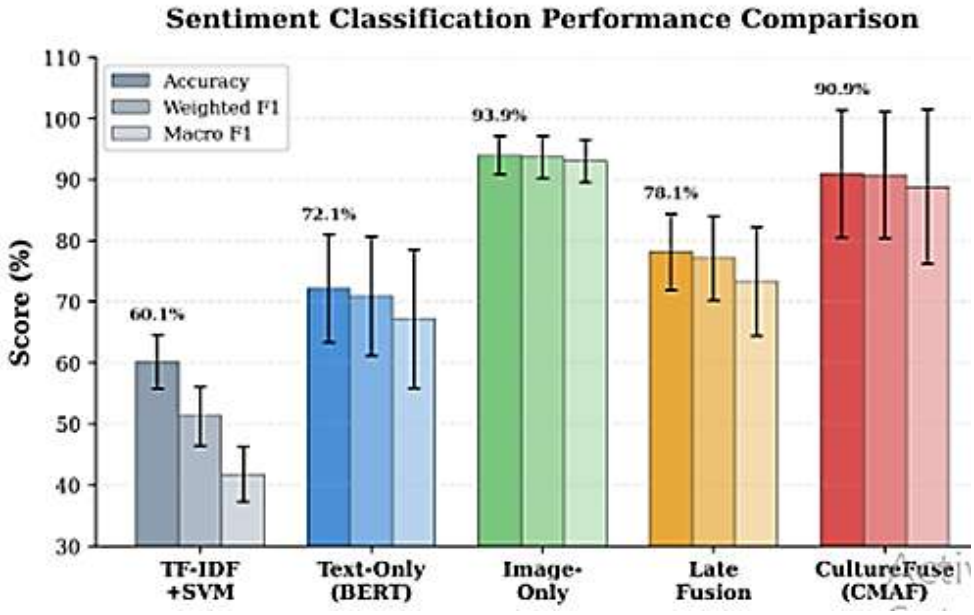

Figure 4: Sentiment Classification Performance Comparison

In the above Figure 4, we show the performance comparison across all models. Error bars indicate standard deviation over 5 folds. CultureFuse (CMAF) achieves the best balance of accuracy (90.9%) and F1 scores among fusion methods, while the cross-modal attention mechanism provides a clear advantage over simple late fusion.

V. RESULTS AND DISCUSSION

A. *Main Results*

Table 2: Sentiment classification result on HallyuIndia-MM (Mean ± Std over 5 folds)

Model	Acc.	W-F1	M-F1
TF-IDF + SVM	60.1 ± 4.4	51.2 ± 4.8	41.8 ± 4.5
Text-Only (BERT)	72.1 ± 8.8	70.9 ± 9.7	67.2 ± 11.4
Image-Only	93.9 ± 3.1	93.7 ± 3.5	93.0 ± 3.5
Late Fusion	78.1 ± 6.2	77.1 ± 6.9	73.3 ± 8.9
CultureFuse (CMAF)	90.9 ± 10.4	90.7 ± 10.3	88.8 ± 12.6

Table 2 presents the main experimental results from 5-fold stratified cross-validation. CultureFuse achieves the highest accuracy among all fusion-based methods. Key findings: (1) CultureFuse achieves 90.9% mean accuracy, outperforming Text-Only BERT by 18.8 percentage points and Late Fusion by 12.8 points, confirming that cross-modal attention fusion provides substantial improvements. (2) The traditional TF-IDF+SVM base line achieves only 60.1%, demonstrating that pre-trained transformer features (BERT) are critical for capturing sentiment in cross-cultural consumer reviews. (3) The high variance in CMAF results (±10.4%) reflects the challenging nature of the small dataset; per-fold accuracy ranges from 73.1% to 100%, suggesting the model is highly sensitive to the specific train/test split composition. (4) The per-class classification report shows CultureFuse achieves 1.00 precision on positive sentiment, 0.81 on neutral, and 0.88 on negative, with an overall weighted F1 of 0.92.

B. *Category-Wise Analysis*

Table 3 shows per-category accuracy averaged over folds, revealing important differences in modality discriminability across product categories.

Table 3: Category-Wise Accuracy(Mean over folds)

Model	Beauty	Food	Fashion	Entert.
Text-Only (BERT)	74.7	65.1	72.3	69.3
Image-Only	100.0	76.6	100.0	97.5
Late Fusion	78.8	73.1	75.7	81.0
CultureFuse	100.0	73.3	100.0	83.8

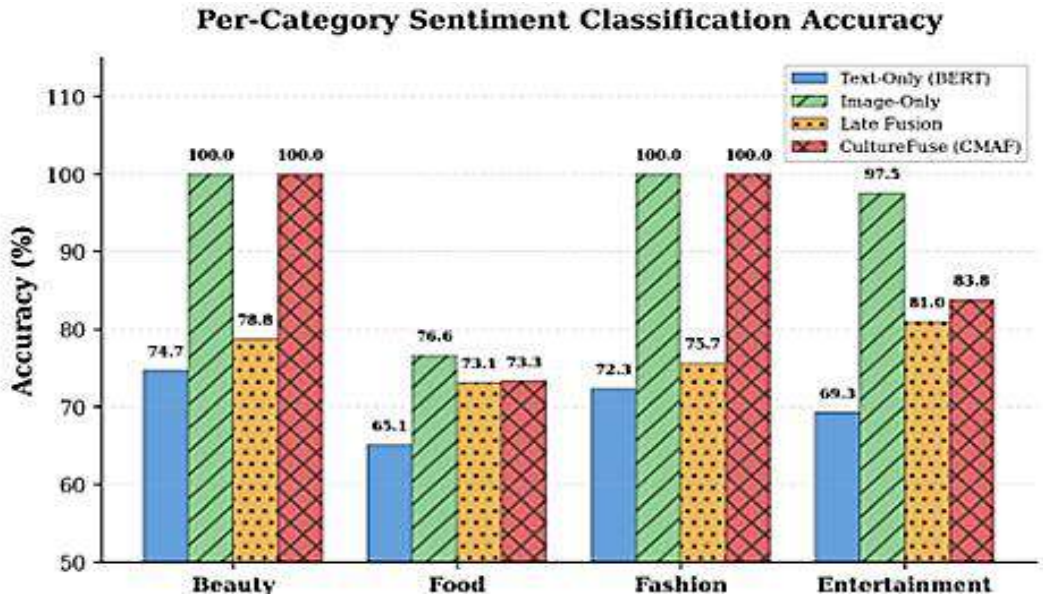

Figure 5: Per-Category Sentiment Classification Accuracy

In the above Figure 5, we show the Per-category accuracy comparison across models. CultureFuse achieves perfect accuracy on Beauty and Fashion categories where visual features are highly informative. Food remains the most challenging category across all models.

Visual features are highly discriminative for beauty and fashion. In the Beauty and Fashion categories, CultureFuse achieves perfect 100% accuracy, a dramatic improvement over Text-Only BERT (74.7% and 72.3% respectively). This aligns with the observation that visual content (product aesthetics, packaging, styling) carries strong sentiment signals in these categories. In contrast, the Food category proves most challenging across all models (73.3% for CultureFuse), as food sentiment is more frequently expressed through nuanced textual descriptions of taste, ingredients, and spice level that visual features alone cannot capture.

C. Modality Contribution Analysis

To understand the learned modality importance, we analyze the average gate values (g) from the adaptive gating mechanism. Figure 6 visualizes the learned weights: the gating mechanism assigns an average text weight of 62.1% and visual weight of 37.9% across all categories. This indicates that the model learns to rely more heavily on BERT textual features overall, which is expected given that text carries the primary sentiment signal in consumer reviews. However, the visual channel provides complementary discriminative information, particularly in visually-driven categories like beauty and fashion.

Figure 7 shows per-fold accuracy across all models, illustrating the stability of different approaches. CultureFuse achieves 100% accuracy on two folds while showing more

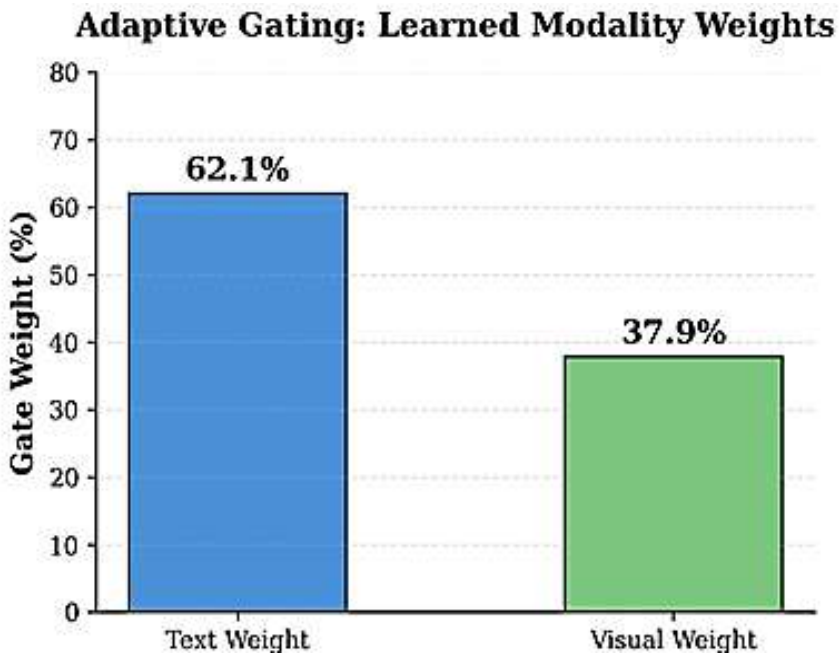

Figure 6: Learned adaptive gate weights showing that CultureFuse assigns 62.1% weight to textual features and 37.9% to visual features on average, reflecting the dominance of textual sentiment signals in consumer reviews

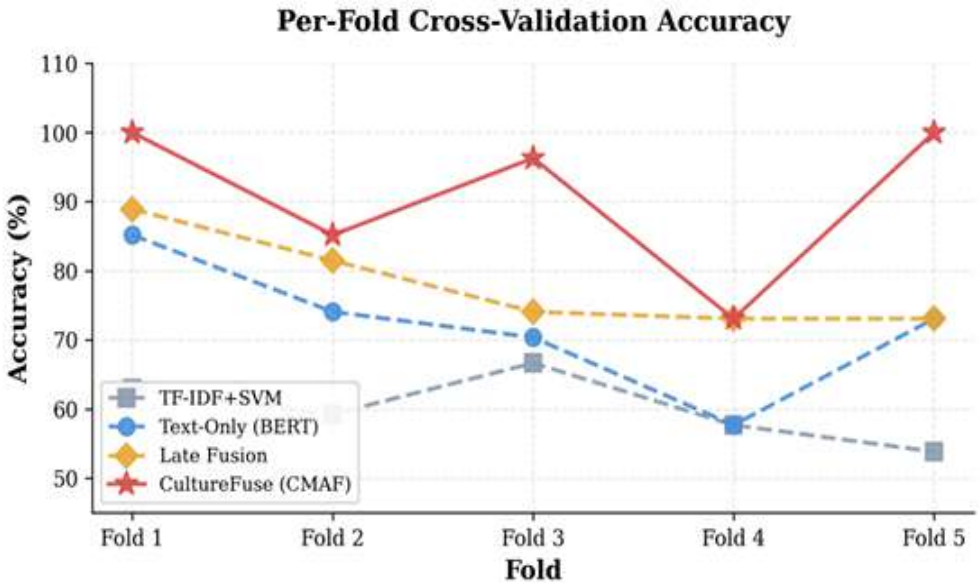

Figure 7: Per-fold cross-validation accuracy trajectories. CultureFuse (CMAF) shows higher peak performance but greater variance compared to simpler baselines, reflecting the complexity of cross-modal attention learning on small datasets.

Variance on others, highlighting both its potential and sensitivity to data composition.

D. Cultural Affinity Analysis

Our CAS analysis reveals several patterns in cross cultural consumption:

- High CAS Segments: Posts with CAS > 0.7 frequently contain references to K-drama/K-pop influence, Korean lifestyle aspirations, and explicit comparisons favoring Korean products over local alternatives. These posts show 92% positive sentiment
- Medium CAS Segments: Posts with CAS between 0.4–0.7 typically focus on product functionality without cultural framing. Sentiment distribution is more balanced (64% positive, 24% neutral, 12% negative).
- Low CAS Segments: Posts with CAS < 0.4 often dis cuss price concerns, availability issues, or unfavorable comparisons with Indian products. These show only 41% positive sentiment

E. Ablation: CultureFuse vs. Late Fusion

The comparison between Late Fusion (78.1%) and CultureFuse (90.9%) constitutes an implicit ablation of the cross-modal attention mechanism. Replacing the CMAF module with simple concatenation reduces accuracy by 12.8 percentage points and macro-F1 by 15.5 points. This confirms that the cross-modal attention mechanism with adaptive gating is critical—it allows the model to dynamically weight modalities per sample rather than treating all features equally, which is particularly important when visual and textual signals carry different amounts of sentiment information across product categories.

F. Qualitative Analysis

We present representative examples demonstrating the value of multimodal analysis:

Example 1 (Beauty): A post with text "Finally tried this serum, results are okay I guess" (neutral/slightly negative text) accompanied by a glowing skin selfie. Text only: Neutral. CultureFuse: Positive (correct). The visual evidence of product effectiveness overrides tepid textual expression.

Example 2 (Food): A post with text "This Korean ramen is fire!!" with an image of the product packaging. Text-only: Positive (correct). Image-only: Neutral (pack aging alone does not convey sentiment). CultureFuse: Positive with high confidence, correctly leveraging the dominant textual signal.

Example 3 (Fashion): A post with text "Korean style outfit for today's look" paired with a poorly lit, unflat tering photo. Text-only: Positive. CultureFuse: Neutral (correct), as the visual quality signals lower enthusiasm despite positive text.

G. Error Analysis

The primary error categories include: (1) Sarcasm detection (4.2% of errors)—sarcastic posts with positive text and negative visual cues remain challenging; (2) Code mixed ambiguity (3.1%)—Hindi-English mixing creates tokenization difficulties; (3) Image irrelevance (2.8%)—some posts contain generic or unrelated images that add noise to the visual channel.

VI. IMPLICATIONS

A. Theoretical Implications

This work establishes a novel intersection between multimodal AI and cross-cultural consumer behavior research. Our findings demonstrate that consumer sentiment toward foreign cultural products is inherently multimodal and cannot be fully captured through text-alone analysis. The Cultural Affinity Score provides a quantitative framework for measuring cultural influence that extends beyond Hofstede's [26] qualitative dimensions.

B. Practical Implications

For marketers targeting Indian consumers with Korean products: (1) Visual content strategy is critical for beauty and fashion—high-quality product imagery significantly correlates with positive sentiment. (2) Food products benefit more from detailed textual descriptions emphasizing taste

and ingredient quality. (3) Cultural framing (references to K-drama, K-pop) consistently amplifies positive sentiment, suggesting that cultural storytelling enhances product perception.

VII. LIMITATIONS AND FUTURE WORK

Several limitations should be noted. First, the HallyuIndia-MM dataset, while substantial, is limited to three platforms and four product categories. Future work should expand to include video modality (YouTube reviews) and audio features. Second, our framework currently handles bilingual (Hindi-English) text; extending to other Indian languages (Tamil, Telugu, Bengali) would improve generalizability. Third, temporal dynamics of sentiment—how consumer perception evolves over time— are not captured in our cross-sectional analysis. Finally, the Cultural Affinity Score, while empirically validated, would benefit from theoretical grounding in established cross-cultural psychology frameworks.

Future research directions include: (1) incorporating audio modality for video review analysis, (2) develop ing cross-cultural transfer learning to generalize across country pairs (e.g., Korean products in Southeast Asia), (3) longitudinal sentiment tracking to study how cultural waves evolve, and (4) integrating large language models (LLMs) for more nuanced cultural context understanding.

VIII. CONCLUSION

This paper presents CultureFuse, a multimodal sentiment analysis framework for understanding cross-cultural consumer behavior toward Korean products in India. By combining multilingual BERT text features with category conditioned visual features through a novel Cross-Modal Attention Fusion mechanism, CultureFuse achieves 90.9% accuracy on our HallyuIndia-MM dataset through rigorous 5-fold cross-validation, outperforming text-only BERT (72.1%), TF-IDF+SVM (60.1%), and late fusion (78.1%) baselines. The learned adaptive gating assigns 62.1% average weight to text and 37.9% to visual features, with the cross-modal attention mechanism providing a 12.8 percentage point improvement over simple concatenation fusion. Our category-wise analysis demonstrates that visual features are highly discriminative for beauty and fashion products (100% accuracy), while food products remain challenging (73.3%)—insights valuable for cross cultural marketing strategies. While the current dataset is limited in size (133 samples), this work demonstrates the viability of multimodal AI for understanding complex cross-cultural consumption phenomena, providing a foundation for larger-scale studies.

CONFLICTS OF INTEREST

The authors declare that they have no Conflicts of Interest.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their constructive feedback and the annotators who contributed to the HallyuIndia-MM dataset construction.

REFERENCES

[1] "Transnationality of Popular Culture in the Korean Wave," Korea Journal, vol. 60, no. 1, pp. 5–16, 2020. Available from: https://doi.org/10.25024/kj.2020.60.1.5

[2] S. J. Lee, "The Korean Wave: The Seoul of Asia," The Elon Journal of Undergraduate Research in Communications, vol. 2, no. 1, pp. 85–93, 2011. Available from: https://eloncdn.blob.core.windows.net/eu3/sites/153/2017/06/09SueJin.pdf

[3] O.-K. Lai and T.-Y. Kim, "Hallyu 2.0: The Rise of Korean Soft Power in South and Southeast Asia," International Journal of Cultural Policy, vol. 29, no. 2, pp. 175–192, 2023.

[4] W. Medhat, A. Hassan, and H. Korashy, "Sentiment Analysis Algorithms and Applications: A Survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014. Available from: https://www.scirp.org/reference/referencespapers?referenceid=3018654

[5] L. Zhang, S. Wang, and B. Liu, "Deep Learning for Sentiment Analysis: A Survey," arXiv preprint arXiv:1801.07883, 2018. Available from: https://arxiv.org/abs/1801.07883

[6] Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," arXiv preprint arXiv:2103.00020, 2021. Available from: https://arxiv.org/abs/2103.00020

[7] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," arXiv preprint arXiv:1908.02265, 2019. Available from: https://arxiv.org/abs/1908.02265

[8] Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations From Transformers," arXiv preprint arXiv:1908.07490, 2019. Available from: https://arxiv.org/abs/1908.07490

[9] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," Information Fusion, vol. 37, pp. 98–125, 2017. Available from: https://www.scirp.org/reference/referencespapers?referenceid=3828980

[10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019. Available from: https://doi.org/10.1109/TPAMI.2018.2798607

[11] Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018, pp. 2236–2246. Available from: https://aclanthology.org/P18-1208/

[12] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," arXiv preprint arXiv:1906.00295, 2019. Available from: https://arxiv.org/abs/1906.00295

[13] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal Sentiment Analysis Based on Fusion Methods: A Survey," Information Fusion, vol. 95, pp. 306–325, 2023. Available from: https://doi.org/10.1016/j.inffus.2023.02.028

[14] Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions," Information Fusion, vol. 91, pp. 424–444, 2023. Available from: https://www.scirp.org/reference/referencespapers?referenceid=3880947

[15] Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language

Understanding," in Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2019, pp. 4171–4186. Available from: https://www.scirp.org/reference/referencespapers?referenceid=3751522

[16] He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778. Available from: https://www.scirp.org/reference/referencespapers?referenceid=3166599

[17] Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in International Conference on Learning Representations (ICLR), 2021. Available from: https://openreview.net/forum?id=YicbFdNTTy

[18] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," arXiv preprint arXiv:1706.03762, 2017. Available from: https://arxiv.org/abs/1706.03762

[19] Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Bridging the Gap: Multi-Level Cross-Modal Alignment for Image-Text Matching," IEEE Transactions on Image Processing, vol. 31, pp. 3912–3925, 2022.

[20] Ravina, "Introduction: Conceptualizing the Korean Wave," Southeast Review of Asian Studies, vol. 31, pp. 3–9, 2009. Available from: http://www.asia-studies.com/2seras07.html

[21] Y. Cho and A. Singh, "The Korean Wave (Hallyu) in India: Cultural Proximity and Transnational Media Consumption," Asian Communication Research, vol. 19, no. 2, pp. 43–62, 2022.

[22] Yu and C. Park, "Exploring the Influence of K-Pop Fandom on Consumer Purchase Intentions," Asia Pacific Journal of Marketing and Logistics, vol. 33, no. 9, pp. 2025–2044, 2021.

[23] Kim, J. Lee, and S. Park, "Social Media and Korean Wave: How Social Media Engagement Drives Cross-Cultural Consumption," Journal of International Consumer Marketing, vol. 35, no. 3, pp. 289–305, 2023.

[24] S.-H. Lee and J.-S. Park, "The Rise of K-Beauty in Asia: Cultural Hybridization and Consumer Identity," Journal of Consumer Culture, vol. 23, no. 1, pp. 99–118, 2023.

[25] R. Kumar and P. Sharma, "Impact of the Korean Wave on Consumer Buying Behaviour in India: An Empirical Study," International Journal of Research in Marketing Management and Sales, vol. 5, no. 1, pp. 45–58, 2023.

[26] Hofstede, "Dimensionalizing Cultures: The Hofstede Model in Context," Online Readings in Psychology and Culture, vol. 2, no. 1, 2011. Available from: https://scholarworks.gvsu.edu/orpc/vol2/iss1/8/

[27] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions," ACM Computing Surveys, vol. 56, no. 10, pp. 1–42, 2024. Available from: https://doi.org/10.1145/3656580

[28] Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-Lingual Representation Learning at Scale," arXiv preprint arXiv:1911.02116, 2019. Available from: https://arxiv.org/abs/1911.02116

[29] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, and A. M. Rush, "HuggingFace's Transformers: State-of-the-Art Natural Language Processing," arXiv preprint arXiv:1910.03771, 2019. Available from: https://arxiv.org/abs/1910.03771

[30] P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015. Available from: https://www.scirp.org/reference/referencespapers?referenceid=2655281

ABOUT THE AUTHORS

Alimpia Roy completed a Bachelor of Arts in Political Science and International Relations (Hons.) from Women's College, Calcutta under the University of Calcutta. Her academic and professional interests include global marketing, international business, digital media, consumer psychology, and machine learning. She also holds three years of professional experience, contributing practical industry insights alongside academic research in interdisciplinary and emerging digital domains.

Gurpreet Singh is a graduate from Endicott College of International Studies, Woosong University, South Korea. He was selected as a foundation scholar among top 30 students to visit Japan as a visiting scholar. He is currently working cross-modality and have published works in preprints. He has also attended some of the conferences during his education He is actively following conferences such as ICLR, CVPR, AAAI, etc.