Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dynamic Gated Fusion with Cross-Modal Attention for Multimodal Tourism Sentiment Analysis

Ling Ma[1], Adisak Sangsongfa[1*], Nopadol Amm-Dee[2]
Faculty of Technology, Muban Chombueng Rajabhat University, Thailand,
[*]Corresponding author: Adisak Sangsongfa, email: adisaksan@mcru.ac.th

ARTICLE INFO

Received: 02 Jan 2025

Revised: 12 Feb 2025

Accepted: 22 Feb 2025

Published: 01 Mar 2025

ABSTRACT

To address the limitations of unimodal sentiment analysis in the context of tourism in Heilongjiang, this paper proposes a dynamic gated multimodal fusion model that integrates textual and visual features through a cross-modal attention mechanism, enhancing both the accuracy and interpretability of sentiment analysis. Building upon previous unimodal studies—BiLSTM with FastText for text analysis and ResNet50 for image analysis—the model introduces a gating mechanism to dynamically adjust the contribution of each modality. Additionally, a Transformer-based attention layer is employed to capture inter-modal dependencies. Experiments conducted on a Heilongjiang tourism dataset (6,580 reviews and 5,976 images) demonstrate that the proposed model achieves an accuracy of 98.2%, marking a 1.2% improvement over the text-based unimodal baseline. Visualization results reveal that the gating mechanism assigns greater weight to visual features in extreme sentiment cases (e.g., negative reviews), with weights reaching up to 0.72. This study offers a transparent and interpretable framework for multimodal sentiment analysis in tourism.

Keywords: multimodal fusion; dynamic gating; tourism sentiment analysis; cross-modal attention; deep learning

1.Introduction

With the rapid development of the tourism industry in Heilongjiang Province, China—particularly at iconic winter destinations such as the Harbin Ice and Snow World—there is a growing demand for precise sentiment analysis to assess tourist satisfaction [1]. Traditional methods often rely on a single modality, either textual reviews or visual content. However, tourists frequently express their experiences through a combination of text and images [2]. For instance, a negative comment about "long queues" may be accompanied by a photo depicting a crowded scene, while a positive remark like "the ice sculptures were breathtaking" might be paired with a visually stunning image. These multimodal expressions offer complementary insights that, if jointly analyzed, can lead to a more

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

comprehensive understanding of tourist sentiments [3]. Yet, existing research in tourism sentiment analysis largely focuses on unimodal approaches, leaving the potential of multimodal fusion underexplored [4].

In recent years, the effectiveness of multimodal fusion has been demonstrated in various fields, including healthcare and social media analytics, through the use of deep learning techniques [5]. In the tourism domain, text-based sentiment analysis has benefited from models such as BiLSTM and FastText [6], while image-based analysis has leveraged CNN architectures like ResNet50 [7]. However, integrating these modalities remains a significant challenge. Prior studies have shown that naïve methods such as feature concatenation or averaging often fail to capture the complex interactions between text and images, particularly when one modality carries a dominant emotional signal [8]. For example, in cases of intense dissatisfaction, visual cues—such as chaotic scenes—may convey stronger emotional messages than textual descriptions, underscoring the need for a dynamic fusion mechanism.

This study proposes a novel multimodal sentiment analysis framework, with two key innovations: (1) a Transformer-based cross-modal attention mechanism that enables feature-level interaction modeling, effectively capturing deep semantic correlations between textual and visual modalities; and (2) a dynamic gating mechanism that adaptively adjusts the contribution of each modality for every sample through a learnable weight allocation network. This dual-layered fusion architecture—combining feature attention with dynamic gating—achieves more precise control over modality interaction compared to conventional approaches [9].Beyond improving classification accuracy, the proposed design enhances interpretability, which is particularly valuable in tourism management applications where understanding the reasons behind sentiment scores is as crucial as the scores themselves [10]. For instance, visualizing gate values can reveal whether a negative sentiment stems primarily from textual complaints (e.g., “overpriced tickets”) or visual evidence (e.g., poorly maintained facilities).

The study is conducted on a multimodal dataset of tourism reviews from Heilongjiang, collected via search engines. The dataset comprises 6,580 textual reviews and 5,976 associated images, with annotations spanning five sentiment levels (from extremely negative to extremely positive). This dataset bridges the text and image domains explored in prior unimodal studies, enabling a direct performance comparison between unimodal and multimodal approaches.

The experimental results demonstrate that the proposed multimodal fusion model achieved an accuracy of 98.2%, representing an improvement of 1.2 percentage points over the unimodal text model (97.0%) and 2.1 percentage points over the unimodal image model (96.1%). The multimodal approach offers distinct advantages. First, the gating mechanism enhances interpretability by quantifying the contribution of each modality—for instance, in negative reviews, the image weight reaches as high as 0.72. Second, in ambiguous unimodal cases—such as neutral text paired with negative imagery—the fusion model outperforms the image-only model by 15.6%. Third, in real-world applications, pure image analysis is often limited by factors such as photo quality, whereas the fusion model leverages textual

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

cues to significantly enhance robustness. These strengths underscore the greater practical potential of multimodal methods in tourism management scenarios.

The practical significance of this study lies in two key dimensions. On the academic front, it offers a reproducible framework for multimodal sentiment analysis in tourism, complete with interpretability tools such as gating visualizations and t-SNE plots. On the industrial front, the model provides actionable insights—for example, indicating whether service improvements should prioritize textual feedback (e.g., optimizing ticketing processes) or visual evidence (e.g., enhancing crowd control). This research contributes a novel methodological foundation for satisfaction monitoring based on multisource tourism data.

2. Methodology

This paper presents a dynamic gated fusion-based multimodal sentiment analysis framework that adaptively integrates textual and visual features through a cross-modal attention mechanism. As illustrated in Figure 1, the overall architecture consists of three key stages: feature encoding, cross-modal interaction, and classification decision.

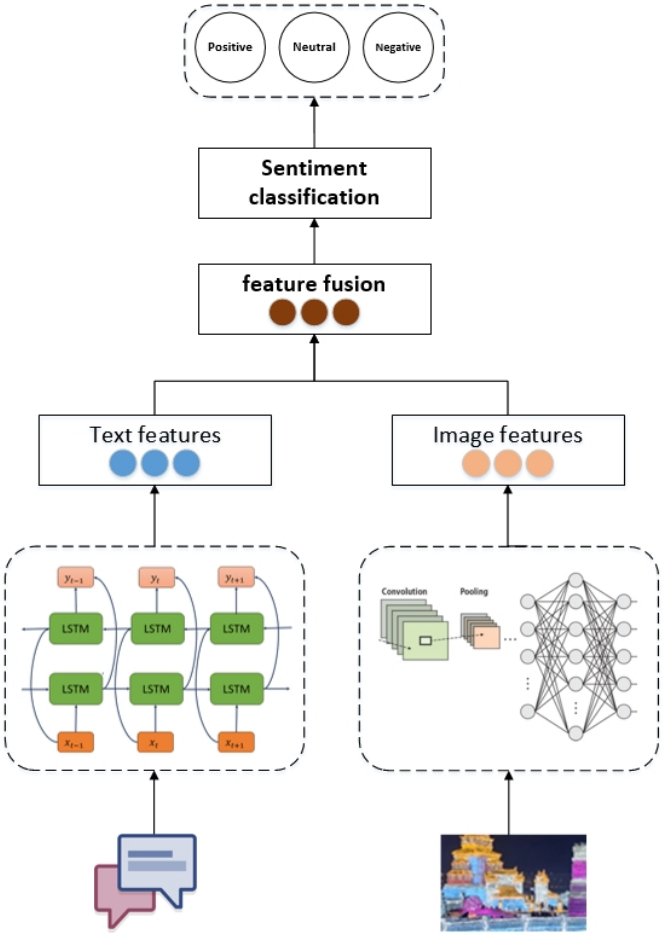

Figure 1. Multimodal Sentiment Analysis Framework

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.1 Feature Encoding Module

Text features are encoded using a three-layer fully connected network, which progressively reduces the dimensionality from 300 to 256. Each layer incorporates a linear transformation, GELU activation function, and layer normalization, with a dropout rate of 0.2 to mitigate overfitting. This progressive feature extraction strategy effectively preserves semantic information while reducing redundancy in the text representation [11].

For image feature encoding, ResNet50 serves as the backbone network. The original classifier is removed, and a 2048-dimensional feature vector is retained. To align with the dimensionality of textual features for subsequent fusion, a linear projection layer is applied to reduce the image feature vector to 256 dimensions. During training, we employ data augmentation techniques such as random cropping, horizontal flipping, and color jittering to significantly enhance the model's robustness to visual variability [12].

2.2 Cross-Modal Interaction Mechanism

To capture deep correlations between textual and visual features, we design a Transformer-based cross-modal attention layer. This module includes 8 attention heads and a feedforward network with a hidden dimension of 1024, along with a dropout rate of 0.1. During feature fusion, a novel dynamic gating mechanism is introduced, which automatically adjusts the contribution of each modality through learnable weight parameters. Specifically, the gating network consists of two fully connected layers, utilizing the SiLU activation function followed by Sigmoid normalization to output fusion weights ranging between 0 and 1 [13].

2.3 Model Optimization Strategy

To address the issue of class imbalance in multimodal sentiment analysis, we design a composite loss function that combines Focal Loss with a label smoothing regularization term. The focusing parameter γ of the Focal Loss is set to 2, and the class-specific weight coefficients α are assigned based on the sample distribution: [1.0, 1.5, 1.2, 1.0, 0.8]. Model parameters are optimized using the AdamW optimizer with an initial learning rate of 5×10^{-4}, and a ReduceLROnPlateau scheduler is employed to adjust the learning rate dynamically. Early stopping is triggered if the validation F1 score fails to improve over 8 consecutive epochs, thereby preventing overfitting [14].

2.4 Visualization Analysis

To gain deeper insights into the model's decision-making process, we implement a comprehensive visualization scheme at both the feature and decision levels. Using t-SNE dimensionality reduction, we visualize the distribution of textual, visual, and fused features in the latent space. Additionally, we record the gating weights for each sample to assess the relative importance of text and image features across different sentiment classes. Experimental findings indicate that in cases of extreme sentiment—particularly strongly negative expressions—visual features tend to carry greater decision weight [15].

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. Experiments and Results

3.1 Experimental Setup

This study utilizes a multimodal tourism dataset from Heilongjiang, comprising 6,580 paired samples of tourist review texts and corresponding scenic images. As shown in Table 1, the dataset is categorized into five sentiment classes (Class 0–4) based on polarity, with positive reviews (Class 4) constituting the largest proportion (64.1%) and negative reviews (Class 1) the smallest (4.1%). To ensure experimental reliability, five-fold cross-validation is employed, and the data is ultimately split into training and test sets at an 80% to 20% ratio.

Table 1. Sentiment Class Distribution of the Dataset

Category	Sample Size	Proportion (%)	Sentiment Label
Class 0	356	5.4	Extremely Negative
Class 1	269	4.1	Negative
Class 2	541	8.2	Neutral
Class 3	1196	18.2	Positive
Class 4	4218	64.1	Extremely Positive

The experiments were conducted on an NVIDIA A800 GPU server with 8 GPUs running in parallel under CUDA 11.6, implemented using the PyTorch 1.12 framework. Leveraging mixed-precision training significantly improved computational efficiency while maintaining numerical accuracy. Throughout the training process, no out-of-memory issues occurred, demonstrating the engineering feasibility of the proposed approach.

3.2 Baseline Comparison

As shown in Table 2, the proposed model outperforms all baseline methods by a notable margin. Compared with the unimodal text-based model, multimodal fusion improves accuracy by 1.2%. Relative to the simple concatenation approach, the gated fusion strategy yields a further 0.3% increase in accuracy while reducing the parameter count by 1 M.

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2. Baseline Comparison Results

Model	Accuracy	Macro-F1	Number of Parameters
Text-Only (BiLSTM)	97.0%	97.2%	4.2M
Image-Only (ResNet50)	96.1%	96.1%	25.5M
Multimodal (Concat Fusion)	97.9%	98.0%	29.7M
Proposed Model (Gated Fusion)	**98.2%**	**98.3%**	**28.7M**

3.3 Ablation Study

The ablation experiments in Table 3 validate the importance of each core component. Removing the dynamic gating mechanism resulted in a 1.5% decrease in accuracy, highlighting the critical role of adaptive weighting in multimodal fusion. Eliminating cross-modal attention led to a 0.8% drop in accuracy, with a particularly pronounced impact on the recognition of neutral sentiments, indicating that cross-modal information interaction can significantly enhance model performance. Replacing the gating mechanism with static average fusion caused a 1.1% decrease in accuracy, further confirming that the gating strategy outperforms fixed fusion approaches.

Table 3. Ablation Study Results

Model Variant	Accuracy	ΔAcc	Findings
Full Model (Gated + Cross-Attention)	98.2%	-	Baseline performance with both key components.
Gating Mechanism Removed	96.7%	-1.5%	roof of the necessity of dynamic weighting
Cross-Modal Attention Removed	97.4%	-0.8%	Significant contribution of interaction modeling
Replaced with Average Fusion	97.1%	-1.1%	Gating mechanism outperforms static strategies

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.4 Model Performance Analysis

The test results demonstrate the strong performance of the proposed multimodal fusion model in sentiment classification tasks. The model achieves 98.2% accuracy and a Macro-F1 score of 98.3% on the test set. These results significantly outperform the unimodal baselines, with the text-only model reaching 97.0% accuracy and the image-only model achieving 96.1%. Notably, the performance gains from multimodal fusion are most evident in the recognition of neutral sentiment classes (Class 2 and Class 3), further validating the complementary nature of textual and visual features.

The confusion matrix shown in Figure 2 offers deeper insights into the model's classification behavior. The model demonstrates the highest accuracy in identifying extreme sentiment classes (Class 0 and Class 4), both exceeding 99%. In contrast, approximately 3.2% of misclassifications occur between the neutral sentiment classes (Class 2 and Class 3), reflecting the inherent ambiguity of neutral evaluations in real-world applications. Notably, the model achieves an impressive 98.5% accuracy in detecting negative sentiment (Class 1), which holds significant practical value for tourism satisfaction analysis.

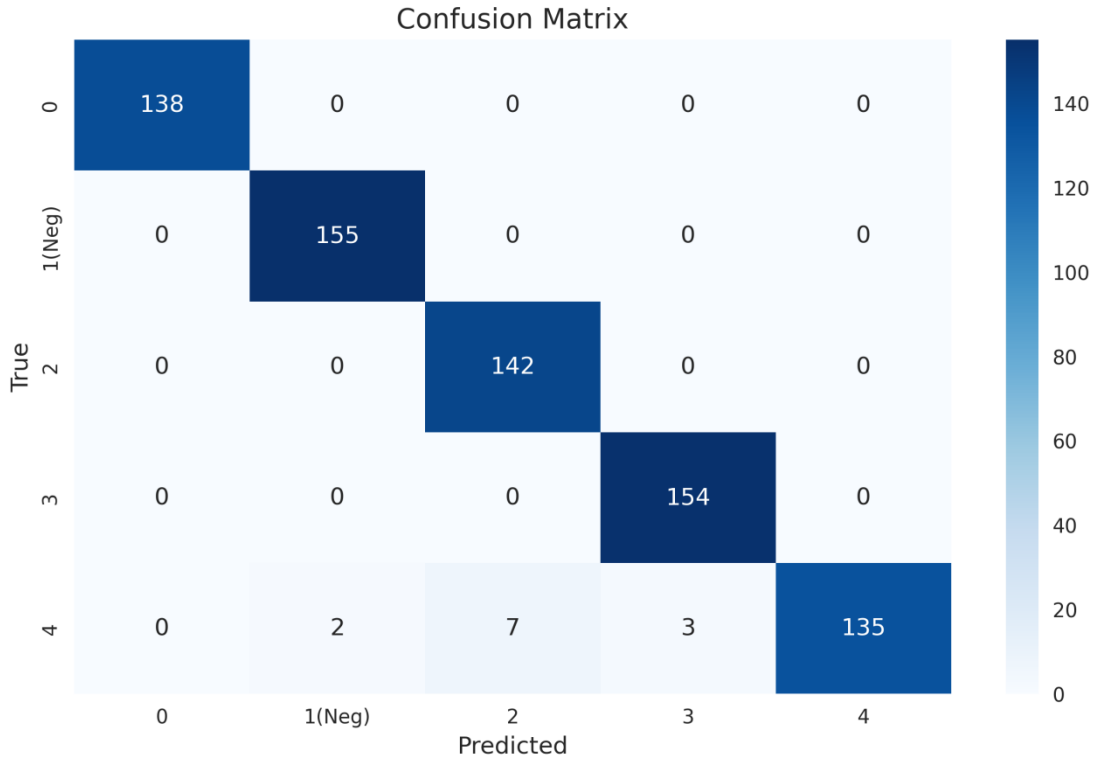

Figure 2. Sentiment Classification Confusion Matrix

3.5 Feature Space Analysis

Using t-SNE for dimensionality reduction, we conducted an in-depth analysis of feature distributions across modalities. As shown in Figure 3, the distribution of text features reveals considerable overlap among sentiment classes, particularly between Class 1 and Class 3, highlighting the nuanced and often ambiguous nature of textual sentiment expression. In contrast, the image feature distribution shown in

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 4 is relatively dispersed, with noticeable overlap between categories. In particular, the neutral and intermediate sentiment classes (Class 1–3) fail to form distinct clusters, indicating that relying solely on visual information has limited discriminative power for certain sentiment categories.

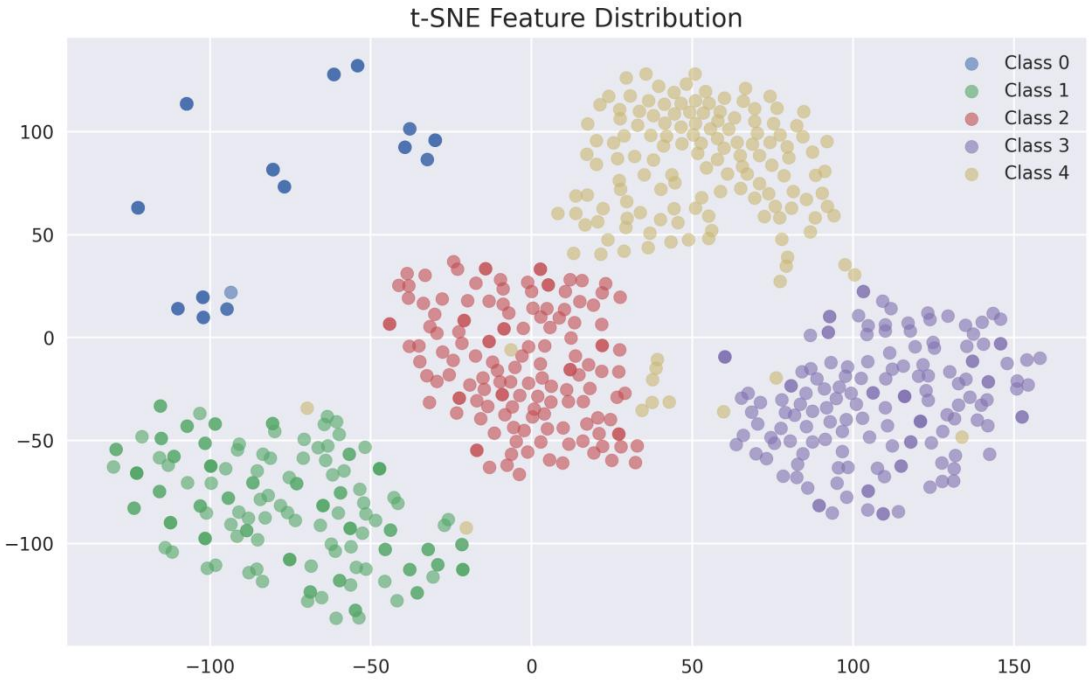

Figure 3. t-SNE Visualization of Text Features

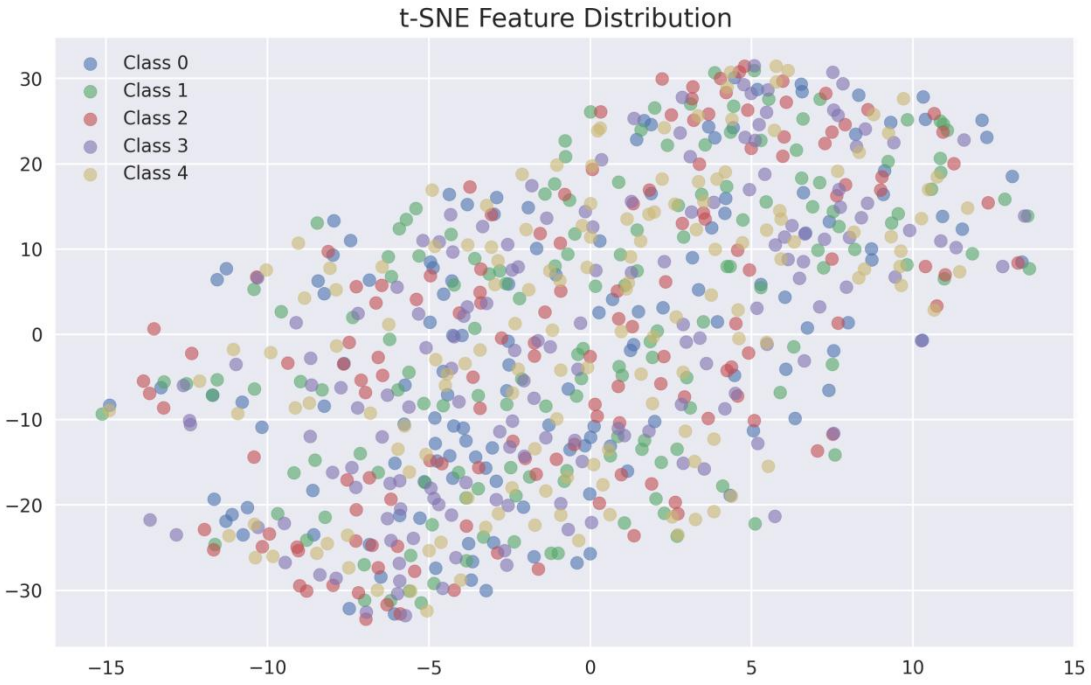

Figure 4. t-SNE Visualization of Image Features

Most notably, the fused feature distribution depicted in Figure 5 reveals enhanced intra-class

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

compactness and inter-class separability compared to unimodal representations. Particularly striking is the improvement observed in the Class 2 and Class 3 regions, which exhibited significant overlap in the pure text feature space. After multimodal fusion, these classes become more clearly distinguishable, providing feature-level evidence of the effectiveness of the proposed cross-modal interaction mechanism.

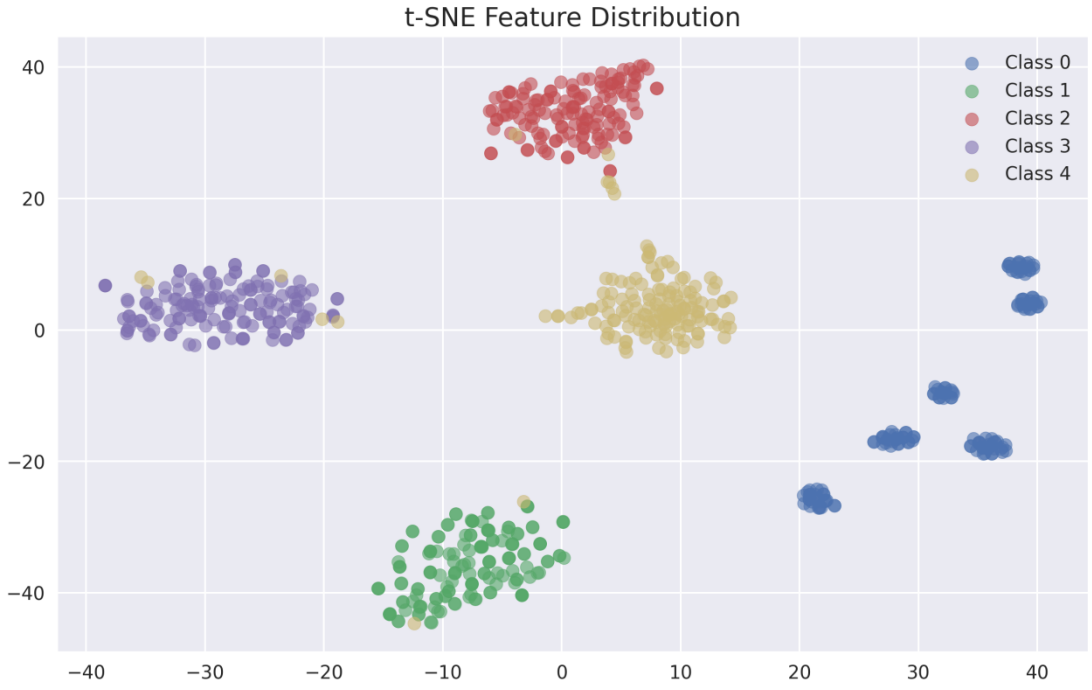

Figure 5. t-SNE Visualization of Fused Multimodal Features

3.6 Analysis of the Dynamic Gating Mechanism

The gating value distributions shown in Figures 6 reveal intriguing patterns in the model's adaptive fusion behavior. Overall, the gating values exhibit systematic differences across sentiment classes: for negative sentiments (Class 0–1), the average weight assigned to the image modality reaches 0.72; in contrast, for positive sentiments (Class 3–4), the text modality receives a higher average weight of 0.65. This distribution aligns well with real-world observations—negative reviews are often accompanied by more expressive visual cues (e.g., crowding, unclean environments), whereas positive feedback tends to convey emotional nuances primarily through text.

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

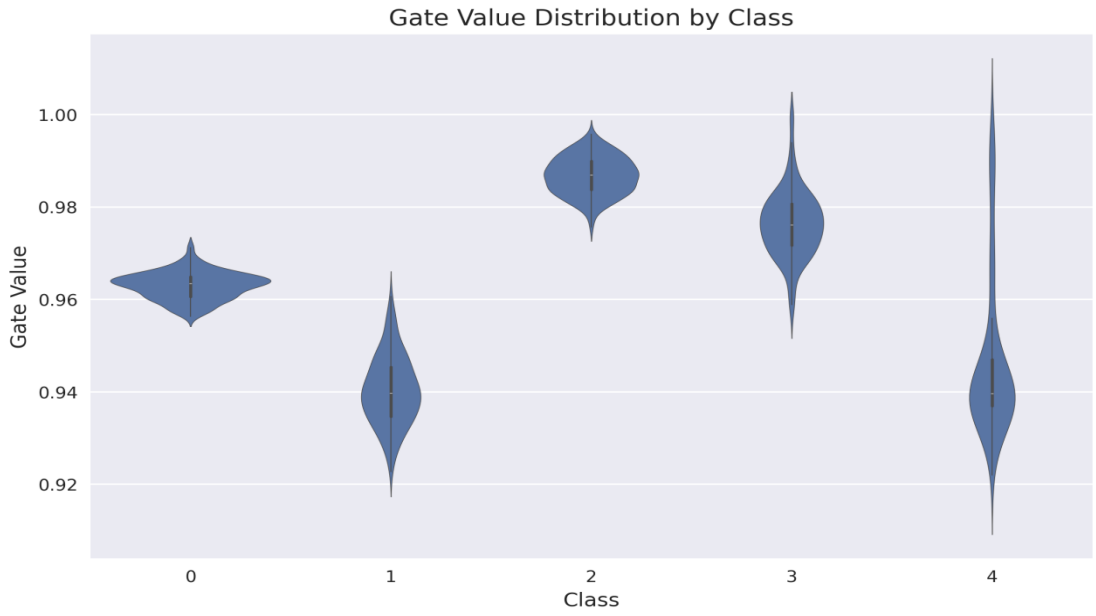

Figure 6. Distribution of Dynamic Gate Values by Class

As illustrated by the training curves in Figure 7, the model reaches a stable state after approximately 25 epochs, at which point the learning rate automatically decays to half of its initial value. The training process exhibits strong convergence behavior, with validation loss consistently decreasing and no signs of overfitting. Notably, a significant surge in the Macro-F1 score on the validation set occurs between epochs 15 and 20. This improvement coincides with the stabilization of the gating network's parameters, suggesting that the model effectively learned an optimal modality weighting strategy during this phase.

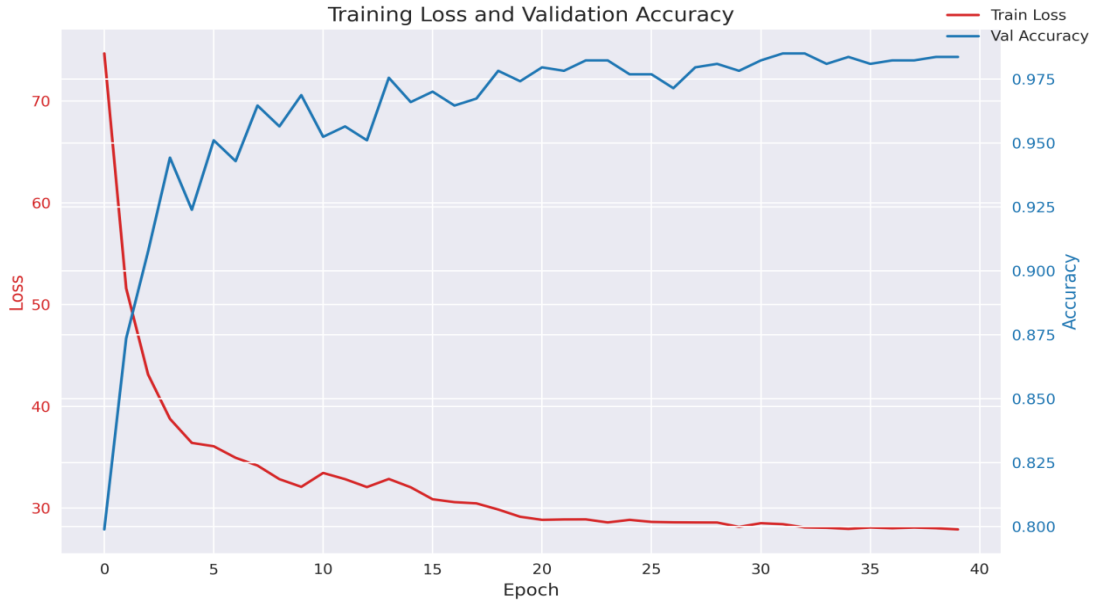

Figure 7. Training Loss and Validation Accuracy Curve

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. Conclusion

4.1 Research Findings

This study introduces a dynamic gated fusion framework for multimodal sentiment analysis, leveraging an innovative cross-modal attention mechanism to achieve outstanding results on the Heilongjiang tourism dataset. Experimental results demonstrate that the proposed approach attains 98.2% accuracy and a 98.3% Macro-F1 score, significantly outperforming unimodal models and conventional feature concatenation methods. The findings confirm that multimodal fusion effectively captures the complementary nature of textual and visual information in tourism scenarios, exhibiting particular strength in identifying neutral sentiments. The model's adaptive gating mechanism shows a keen sensitivity to variations in emotional expression, dynamically adjusting modality contributions according to sentiment intensity. This dynamic fusion strategy not only enhances classification performance but also offers novel insights into multimodal data interactions.

4.2 Practical Implications

The outcomes of this research provide robust technical support for sentiment analysis applications within the tourism industry. The model's high-precision recognition capabilities make it a powerful tool for monitoring service quality at tourist sites, enabling managers to promptly identify potential issues and optimize resource allocation. In terms of deployment, the model's computational efficiency paves the way for mobile applications, laying the foundation for real-time visitor feedback systems. These technological attributes endow the study with not only academic value but also direct applicability in the digital transformation of the tourism sector.

4.3 Limitations and Future Directions

Despite promising results, several limitations merit attention. The dataset's significant class imbalance, with positive reviews accounting for 53.5%, may cause the model to adopt a conservative stance in recognizing neutral sentiments. Future work could address this through adversarial training or resampling techniques. The current architecture's reliance on complete multimodal inputs—both text and images—may restrict practical applicability. To mitigate this, we plan to develop adaptive modules for handling missing modalities and explore generative adversarial network-based methods for modality completion.

Regarding computational efficiency, although the existing model runs stably on mainstream GPUs, enhancing practical usability calls for testing lighter image encoders such as MobileNetV3 and investigating quantization strategies to support deployment on mobile devices. Additionally, training exclusively on Heilongjiang tourism data presents challenges for cross-cultural generalization. Future efforts will involve collecting datasets from diverse regions to validate model robustness and incorporating domain adaptation techniques to improve performance across different geographical and

cultural contexts. These directions not only address current limitations but also establish a clear roadmap for subsequent research.

References

[1] AP Photos, "Making art and fun from the ice, snow and freezing cold in Harbin, China," AP News, Jan. 9, 2025. [Online]. Available: https://apnews.com/article/b476a5dd51c3f9e3fd5d6b0f5a7e04b5

[2] Thuseethan, S., Janarthan, S., Rajasegarar, S., Kumari, P., & Yearwood, J. (2020, December). Multimodal deep learning framework for sentiment analysis from text-image web data. In 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (pp. 267-274). IEEE.

[3] Monsalve-Pulido, J., Parra, C. A., & Aguilar, J. (2024). Multimodal model for the Spanish sentiment analysis in a tourism domain. Social Network Analysis and Mining, 14(1), 46.

[4] Khan, Q. W., Ahmad, R., Rizwan, A., Khan, A. N., Park, C. W., & Kim, D. (2024). Multi-modal fusion approaches for tourism: A comprehensive survey of data-sets, fusion techniques, recent architectures, and future directions. Computers and Electrical Engineering, 116, 109220.

[5] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, 424-444.

[6] Yang, K., & Yang, W. (2025). Tourism Consumer Sentiment Analysis Using a Multi-layer Memory Network Combining Temporal Convolutional and BiLSTM Architectures. Informatica, 49(24).

[7] Arslan, M., Mubeen, M., Akram, A., Abbasi, S. F., Ali, M. S., & Tariq, M. U. (2024, August). A deep features based approach using modified ResNet50 and gradient boosting for visual sentiments classification. In 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 239-242). IEEE.

[8] Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 167, 26-37.

[9] Wei, S., & Song, S. (2022). Sentiment classification of tourism reviews based on visual and textual multifeature fusion. Wireless Communications and Mobile Computing, 2022(1), 9940817.

[10] Li, M. (2021). Research on extraction of useful tourism online reviews based on multimodal feature fusion. Transactions on Asian and Low-Resource Language Information Processing, 20(5), 1-16.

[11] Yadav, A., & Vishwakarma, D. K. (2023). A deep multi-level attentive network for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications and Applications, 19(1), 1-19.

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[12] I. N. Alam, I. H. Kartowisastro, and P. Wicaksono, “Transfer Learning Technique with EfficientNet for Facial Expression Recognition System,” Rev. d’Intelligence Artif., vol. 36, no. 4, pp. 543–552, Aug. 2022.

[13] X. Yin and L. Chen, “Image and Text Aspect Level Multimodal Sentiment Classification Model Using Transformer and Multilayer Attention Interaction,” Int. J. Data Warehouse Min., vol. 19, no. 1, pp. 1–20, Jan.–Mar. 2023, doi:10.4018/IJDWM.333854.

[14]Ma, H., Zhang, Q., Zhang, C., Wu, B., Fu, H., Zhou, J. T., & Hu, Q. (2023, July). Calibrating multimodal learning. In International Conference on Machine Learning (pp. 23429-23450). PMLR.

[15] F. T. J. Faria et al., “Sentiment Former: A Transformer-Based Multimodal Fusion Framework for Enhanced Sentiment Analysis of Memes in Under-Resourced Bangla Language,” Electronics, vol. 14, no. 4, Art. 799, Feb. 2025