Multimodal Sentiment and Evaluation Prediction for Cultural Tourism Using an AGT-Optimized Convolutional Deep Belief Network (AGT-ICDBN)

Yuanhui Gong
Xinzhou Normal University, Xinzhou, Shanxi, 03400, China
E-mail: yuanhui_gong@outlook.com

Keywords: multimodal sentiment analysis, cultural tourism, visitor experience, artificial gorilla troops optimizer-driven Intelligent convolutional deep belief network (AGT-ICDBN), ecological towns

Received: April 20, 2025

Tourism is a vital sector for economic development and cultural preservation, and understanding visitor experiences is crucial for effective destination management. However, existing sentiment analysis methods do not effectively integrate multimodal data, limiting accurate interpretation of visitor perceptions and experience evaluation. This research intends to develop a deep learning (DL)-driven multimodal framework that integrates textual, visual, and rating data to perform sentiment analysis and evaluation prediction for cultural tourism attractions. Textual reviews are preprocessed using lemmatization, while images are prepared through resizing and z-score normalization. Contrastive Language–Image Pre-training (CLIP) is employed to extract semantic visual features. The intermediate fusion process is used to integrate textual, visual, and rating features cohesively, enabling richer cross-modal representation and more precise sentiment prediction. The textual and visual features are fused using an Artificial Gorilla Troops Optimizer-driven Intelligent Convolutional Deep Belief Network (AGT-ICDBN) to predict sentiment and evaluate cultural tourism attractions. In this framework, CLIP textual and visual embeddings are concatenated and routed through ICDBN layers, where convolutional Restricted Boltzmann Machines record hierarchical cross-modal dependencies. The AGT improves convergence and prediction resilience by fine-tuning model parameters, optimizing feature selection, and balancing exploration-exploitation. Experiments on a curated dataset of cultural attractions demonstrate that integrating multimodal information improves classification accuracy (CA) to 0.9895, precision to 0.9876, recall to 0.9893,AUC to 0.9853 and F1-score to 0.9877 compared to unimodal approaches, achieving strong correlations between sentiment and evaluation scores using Python 3.9. The proposed framework provides a foundation for real-time, interpretable, and data-driven evaluation of cultural tourism attractions.

Povzetek: Raziskava predlaga multimodalni pristop z globokim učenjem, ki z združevanjem besedilnih, slikovnih in ocenjevalnih podatkov izboljšuje razumevanje doživetij obiskovalcev kulturnega turizma.

1 Introduction

Cultural tourism is a distinctive branch of the tourism industry that prioritizes the exploration and appreciation of traditions, heritage, arts, and lifestyles across different communities [1,2]. The growth of cultural tourism in recent decades has been shaped by globalization, advances in mobility, and the increasing demand for authentic, meaningful experiences [3]. This form of tourism encompasses a wide spectrum of activities, including visits to historical sites, monuments, museums, and heritage towns, as well as participation in rituals, festivals, artistic performances, traditional crafts, and culinary practices [4]. Central to this form of tourism are cultural tourism attractions, which serve as the primary elements that draw visitors to a destination [5, 6]. Cultural tourism attractions, as opposed to natural landscapes or recreational amenities, focus on human achievements and social behaviors that are essential to cultural continuity [7-9]. Beyond enriching the visitor experience, cultural tourism attractions contribute significantly to local communities and broader society. Tourists support the preservation of cultural assets, generate employment and income, stimulate economic development, and foster pride in local traditions [10,11]. Moreover, by facilitating cross-cultural interactions, they promote dialogue, mutual respect, and intercultural understanding, thereby reinforcing the role of tourism as a catalyst for global cooperation and cultural sustainability [12].

Cultural tourism attractions, ranging from heritage sites and ecological towns to commercial and water-based attractions, attract diverse visitors with varying expectations. However, existing sentiment analysis

methods do not effectively integrate multimodal data, limiting accurate interpretation of visitor perceptions and experience evaluation. To solve these issues, this research seeks to create a deep learning (DL)-driven multimodal framework that combines textual, visual, and rating data to conduct sentiment analysis and assessment prediction for cultural tourist destinations.

Research objectives

This research design a multimodal curated cultural tourism dataset with the inclusion of visitor reviews, images, and ratings across various attraction categories. It predict the visitor sentiment and evaluation score more accurately and informatively and analyzing the differences between the cross-categories (heritage, ecological, water-based, and commercial attractions) and extracting the actionable insights for cultural tourism management.

Key contribution of the research

- The research aims to develop a deep learning (DL)-driven multimodal framework that integrates textual, visual, and rating data to perform sentiment analysis and evaluation prediction for cultural tourism attractions.
- A multimodal cultural tourism experience dataset was collected, comprising textual reviews, visitor-uploaded images, and numerical ratings. After collecting the dataset, the raw data underwent tokenization, stop-word removal, and lemmatization to Semantic, textual, and visual features were derived to capture deep contextual meaning. Clip was employed to extract high-level visual attributes, while textual embeddings preserved nuanced linguistic sentiment.
- An AGT-ICDBN was introduced to fuse multimodal features, optimize parameter learning, and enhance prediction robustness. This architecture improved interpretability by linking visual attributes and textual cues with visitor sentiment and evaluation.

2 Literature review

For daily foreign visitor arrivals at Incheon International Airport (ICN), a reliable Long Short-Term Memory (LSTM)-based forecasting model [13] was created. An LSTM-based multivariable time series forecasting model to anticipate foreign visitor arrivals was provided in the research. Tourism demand across multiple attractions with spatial dependence was forecasted by proposing a novel three-stage forecasting model [14]. It was limited by its application to a single city and dataset.

An enhanced weighted association rule algorithm that took time and season into account was used to enhance a recommendation model for cultural tourist attractions [15]. A multi-feature fusion Graph Neural Network (GNN) was used to build a model of a rural tourist attraction [16]. It proposed a multi-stage framework, using two-part and conversation graph models to extract different tourist choice aspects, and constructed a feature map with attention processes. The experimental data from the Chengdu area were limited in terms of their generalizability.

A Mean Signed Error-centric Recurrent Neural Network (MSE-RNN) [17] to improve a system for recommending cultural tourists throughout the day and at night. Historical and hotel geolocations were extracted and data was separated. The suggested system surpassed the existing ones, according to the results. Table 1 illustrates the Comparative Summary of Existing Models.

Table 1. Comparative summary of existing models in multimodal or tourism analytics

Study / Year	Model / Technique	Dataset Type	Performance Metrics	Main Limitations
Liang et al. (2025) [18]	Graph Neural Network	Tourist attraction visits (3 cities)	RMSE = 0.145	Focused on forecasting, not sentiment or evaluation
Bozkurt & Şeker (2023) [19]	MLP / RBF Neural Network	UNESCO WHS classification	Accuracy = 0.89	Trade-off between accuracy and computational cost
Wang et al. (2024) [20]	CNN–LSTM Hybrid	Tourism reviews and images	F1 = 0.87, RMSE = 0.298	Struggles with spatial–temporal dependencies
Hu & Lin (2024) [21]	Attention-based Bi-LSTM	Online travel reviews	Accuracy = 0.91, Precision = 0.88	Text-only; ignores visual context
Singh et al. (2023) [22]	Multimodal Autoencoder	Instagram travel posts	AUC = 0.962, Recall = 0.841	Unstable with noisy visual data
Tao et al. (2024) [23]	Deep Multimodal Fusion Network (DMFN)	TripAdvisor and Flickr data	F1 = 0.89, RMSE = 0.266	High computation cost; weak interpretability
Lee & Choi (2025) [24]	Transformer with Cross-modal Attention	Cultural tourism dataset	AUC = 0.981, Accuracy = 0.91	Requires extensive pretraining data
Chen et al. (2025) [25]	ResNet-50 + Sentiment BERT	Cultural heritage social media reviews	F1 = 0.90, CA = 0.894	Limited fusion depth; lacks adaptive optimization
Calderón-Fajardo et al. (2025) [26]	Transformer-based Multimodal	Text + Generated visuals	AUC = 0.978	Synthetic visuals limit generalizability
Gupta et al. (2020) [27]	CNN + Contextual Features	Instagram tourism posts	F1 = 0.86	Sensitive to noisy social-media data

Although previous researches have shown AI and deep learning approaches to tourism demand forecasting, analysis of tourist behavior, and prediction of sentiment, there are still some remaining gaps in the literature. The LSTM-based multivariable model [14] predicted foreign arrivals using exchange rate, COVID-19 cases, KOSPI, and WTI prices but was based on a single airport dataset. The three-stage, LSTM-autoregressive model [14] incorporated spatial dependence for 77 tourism sites located in Beijing but was limited to a single city. While the DL-based sentiment analyzes of hotel reviews [26,27] achieved high accuracy, it was limited to TripAdvisor data and hotels in a single city.The AGT-ICDBN model overcome these limitations by merging the textual, visual, and rating data sources via CLIP and AGT. Thus, the hierarchical cross-modal learning, enhanced robustness, and interpretable evaluation results become possible, which makes the proposed model a new benchmark in comparison to existing state-of-the-art models.

3 Methodology

Developing a DL-based multimodal framework that combines textual, visual, and rating data for sentiment analysis and attractiveness evaluation in cultural tourism is the aim of this research. The intention is to improve prediction accuracy while gaining interpretable insights into relationships between visitor experiences and attraction features. Figure 1 demonstrates the comprehensive flow of the proposed model, which defines all the processes of the research.

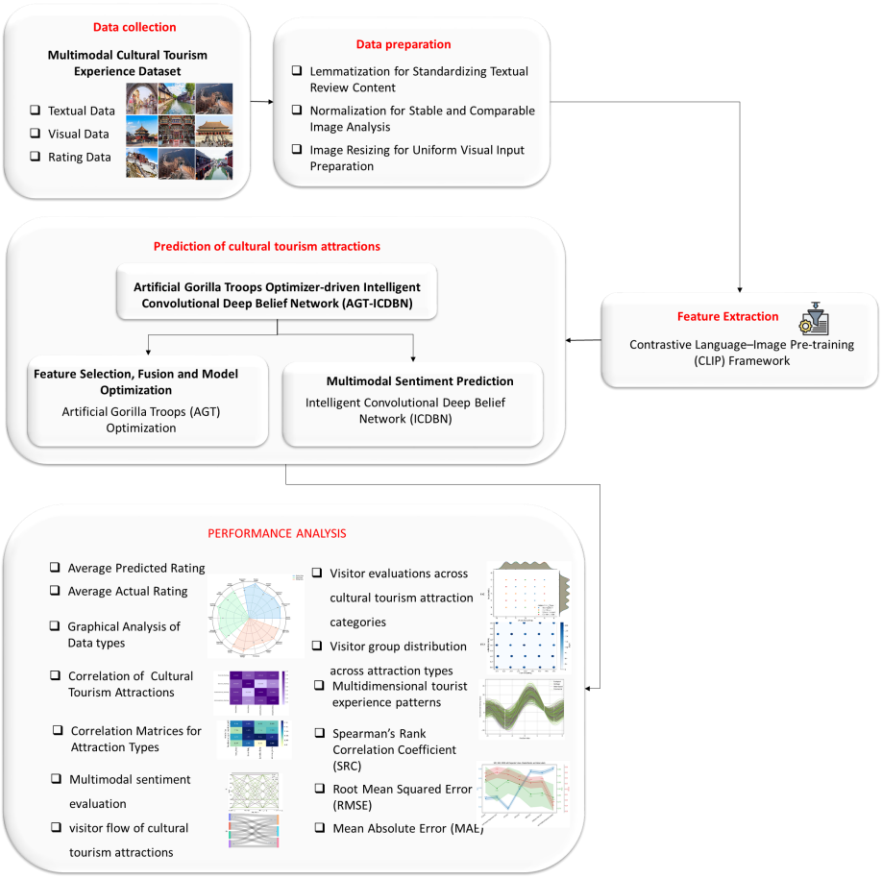

Figure 1: Comprehensive flow of the proposed model for Multimodal sentiment analysis

The Multimodal Cultural Tourism Experience Dataset was used that included thousands of multimodal samples collected from four categories of cultural attractions: heritage sites, ecological towns, water-based destinations, and commercial areas. The dataset was split into 80% training, 20% test subsets, with the class distribution balanced across the sentiment classes. Textual reviews are preprocessed using lemmatization, while images are prepared through resizing and z-score normalization. Contrastive Language–Image Pre-training (CLIP) is employed to extract semantic visual features. The intermediate fusion process is used to integrate textual, visual, and rating features cohesively, enabling richer cross-modal representation and more precise sentiment prediction.

3.1 Data collection

The Multimodal Cultural Tourism Experience Dataset (https://www.kaggle.com/datasets/zyan1999/multimodal-cultural-tourism-experience-dataset/data) was collected in Kaggle. It includes 1000 of multimodal samples collected from four categories of cultural attractions: heritage sites, ecological towns, water-based destinations, and commercial areas. Data filtering methods involved removing duplicate or incomplete entries, excluding corrupted images, and eliminating text reviews with fewer than five words or excessive symbols. All images were resized to 124×124 pixels, and text data were preprocessed using lemmatization and stop-word removal for consistency. The class balance comprises 489 positive (48.9%), 312 negative (31.2%), and 199 neutral (19.9%) samples. The distribution of attraction types includes heritage (25.8%), commercial (25.6%), ecological (25.6%), and water-based (23.0%), ensuring balanced and diverse representation across cultural tourism categories. Figure 2 depicts the sample data images in the dataset.

Figure 2: Sample images of multimodal cultural tourism experience dataset

3.1.1 Lemmatization for standardizing textual review content

Lemmatization is a crucial preprocessing step in textual reviews that converts words into their root form, ensuring semantic consistency and reducing linguistic variability. Lemmatization was applied to normalize words by reducing them to their base or dictionary form, enhancing the consistency of textual data for model training.

3.1.2 Image resizing for uniform visual input preparation

Image resizing is a preprocessing step that standardizes the dimensions of input images to ensure uniformity across the dataset. The image resizing is used to adjust images to a uniform scale, making them suitable for model input. All raw images were resized to 124×124 pixels using high-quality interpolation techniques (figure 3 a-b).

Figure 3: outcomes of (a) original raw image and (b) resized images at 124×124 pixels

3.1.3 Normalization for stable and comparable image analysis

It standardizes pixel intensity values across tourist place images, minimizing variations caused by lighting or quality. It enhances stability of the model and improves the reliability of visual feature extraction. Z-score normalization is used to standardize data, particularly when the minimum and maximum values are unknown (equation 1).

$$w_{new} = \frac{W-v}{\sigma} = \frac{w-\min(W)}{StdDev(W)} \quad (1)$$

w_{new} is a normalized pixel value after transformation. W is the unique pixel value of the image. $W - \min(W)$ is a minimum intensity value in the image dataset. StdDev(W) is the SD (σ) of the pixel values across the cultural tourist experience dataset, used for scaling, W is a set of all pixel intensity values in the image or dataset.

3.2 Feature extraction using contrastive language–image pre-training (CLIP) framework

CLIP was used to extract high-level semantic and contextual attributes from tourist place images, rating data, and text data. It aligns information with linguistic meaning, enabling more interpretable multimodal representation. CLIP contrastively aligns the embeddings from two image and text encoders

(e_I and e_S). During each training iteration, select a mini-batch of M image-text pairs $\{(w_j, z_j)\}_{j=1}^{M}$ from the large-scale training set. Contrastive loss is defined as follows in equation (2):

$$K_{CLIP}: \frac{(K_{J\to S}+K_{S\to J})}{2} \quad (2)$$

The similarity function is denoted as $sim(.,.)$, and τ is a learnable temperature. CLIP loss K_{CLIP} averages symmetrical contrastive loss, with cross-entropy normalized along the image-to-text and text-to-image axes, which represent $(K_{J\to S} + K_{S\to J})$ respectively.

3.3 Multimodal fusion process

The intermediate fusion process was chosen over alternative fusion methods because it effectively captures nonlinear relationships between multimodal features through adaptive optimization, unlike early, late, or attention-based fusion models that struggle with imbalanced or limited datasets. First, textual and visual embeddings from the CLIP model are separately processed; the textual features are processed through a transformer-based encoder, while the visual embeddings are processed using a convolutional feature extractor. These representations are then combined through feature concatenation and refined by an attention-weighted fusion layer, which assigns adaptive importance to each modality before feeding the fused representation into the hierarchical learning module. Let the extracted embeddings be represented as equation (3) (algorithm 1).

$$T = f_{text}(x_t), V = f_{img}(x_v), R = f_{rate}(x_r) \quad (3)$$

where T, V, and R denote textual, visual, and rating feature vectors. The intermediate fusion process concatenates these features as equation (4).

$$F = [T \oplus V \oplus R] \quad (4)$$

where $\oplus$ denotes concatenation. To enhance cross-modal interaction, an attention-weighted fusion mechanism is applied as equation (5).

$$\tilde{F}_i = \alpha_t T_i + \alpha_v V_i + \alpha_r R_i \quad (5)$$

where $\alpha_t, \alpha_v, \alpha_r$ are attention weights learned via a softmax normalization as shown in equation (6).

$$\alpha_m = \frac{\exp(W_m F_m)}{\sum_{j\in\{t,v,r\}} \exp(W_j F_j)} \quad (6)$$

Algorithm 1: Multimodal Feature Process (Intermediate Fusion Process)

```
Input:
  x_t ← textual input (visitor review)
  x_v ← visual input (attraction image)
  x_r ← rating input (numerical rating)
Output:
  ŷ ← predicted sentiment or evaluation result
Begin
  Step 1: Multimodal Feature Extraction
  Step 2: Intermediate Feature Fusion
  Step 3: Adaptive Cross-Modal Weighting
  for each modality m ∈ {t, v, r} do
    α_m ← exp(W_m * F_m) / Σ_j∈{t,v,r} [exp(W_j * F_j)]
  end for
  α_m = exp(W_m F_m) / ∑_{j∈{t,v,r}} exp(W_j F_j)
  Compute weighted multimodal feature
Step 4: Deep Multimodal Interaction and Prediction
  H ← Nonlinear_Interaction_Model(F̃_i)
  ŷ ← Softmax(H)
  Output predicted sentiment or evaluation category
Return ŷ
End
```

3.4 Artificial gorilla troops optimizer-driven intelligent convolutional deep belief network (AGT-ICDBN) multimodal feature fusion and sentiment prediction

The AGT-ICDBN is designed to integrate textual and visual features into a unified multimodal framework. The AGT-ICDBN was chosen over alternative fusion methods because it effectively captures nonlinear relationships between multimodal features through adaptive optimization, unlike early, late, or attention-based fusion models that struggle with imbalanced or limited datasets. The AGT-ICDBN framework proposed utilizes an intermediate fusion strategy.

3.3.1 Intelligent convolutional deep belief network (ICDBN) for multimodal sentiment prediction

The ICDBN serves as a mechanism to fuse textual and visual features into a coherent representation. ICDBN was utilized to

extract deep hierarchical features from multimodal inputs, enabling effective representation learning for sentiment classification.

• Standard Deep Belief Network (DBN)

The standard DBN learns hierarchical feature representations by stacking several layers of Restricted Boltzmann Machines to identify complex, nonlinear relationships among multimodal data. The intention of the ICDBN is to improve predictive tasks in cultural tourism.

In equation (7), u is a state vector of visible units and g is a state vector of hidden units. Activation $g^{l}_{j,i}$of the hidden unit at position (j,i) in the $l-th$ feature map. Convolution filter X^{l} for the $l-th$ hidden feature map. a_l is a bias of the $l-th$ hidden feature map.

$$-\log O(u,g) \propto F(u,g) = -\sum_{l=1}^{L}\sum_{i=1}^{M_G}\sum_{t=1}^{M_x} g^{l}_{j,i} X^{l}_{q,s} v_{j+q,i+t-1} - \sum_{l=1}^{L} a_l \sum_{j,i=1}^{M_G} g^{l}_{j,i} - D\sum_{j,i=1}^{M_U} u_{j,i} \quad (7)$$

The D represents the bias of visible layer units. $O(u,g)$ is a Output probability, $F(u,g)$ is the energy function of the CRBM. l denotes a layer index in the DBN. Total number of layers is depicted as L in the DBN. j,i are the indices of hidden feature maps. M_G is the number of hidden feature maps. The bias term a_l for the $l-th$ hidden feature map. Dimensionality M_x of the visible input along one axis. $v_{j+q,i+t-1}$ The visible unit value at position shifted by $(q,t-1)$ Part of the input patch that the convolution filter processes. M_U is a number of visible units. $X^{l}_{q,s}$ represents local receptive field indices. t is the index over the visible/input dimension. After some computational steps, the energy function can be converted into the following equation (8).

$$F(u,g) = -\sum_{l}\sum_{j,i} g^{l}_{j,i}.(\tilde{X}^{l} * u)_{ji} - \sum_{l} a_l \sum_{j,i} g^{l}_{j,i} - d\sum_{j,i} u_{j,i} \quad (8)$$

The standard CRBM's conditional probabilities can be calculated using one-step Gibbs sampling in equation (9).

$$O\left(g^{l}_{j,i}=1\middle|U\right) = \sigma((\tilde{X}^{l} * u)_{ji} + a_o) \quad (9)$$

$O\left(g^{l}_{j,i}=1\middle|U\right)$ is the probability that the position of hidden unit at (j,i) in $l-th$ feature map is activated given the input U. $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid activation function, which was measured in equation (10), which maps input values to a probability between 0 and 1. $\tilde{X}^{l}$is is a convolution filter corresponding to the $l-th$ hidden feature map. U is an input matrix, and a_o is the bias term for the hidden units in this layer.

$$O\left(u_{j,i}=1\middle|g\right) = \sigma(\sum_{l,j,i}(X^{l} * g^{l})_{ji} + d) \quad (10)$$

$O\left(u_{j,i}=1\middle|g\right)$ is the probability that the visible unit at position (j,i) is activated, given the hidden layer activations g. Sigmoid activation function is expressed as $\sigma(x) = \frac{1}{1+e^{-x}}$, mapping inputs to a probability between 0 and 1. X^{l} is a convolution filter associated with the $l-th$ hidden feature map. Hidden feature g^{l} map of l. d is a bias term of the visible layer units.

• Convolutional restricted Boltzmann machine (CRBM) model

The CRBM is designed to efficiently capture local dependencies in input data, reduce dimensionality through shared weights, and provide robust hierarchical feature representations for multimodal sentiment analysis. The CRBM architecture is composed of three main layers: the visible layer (U), the hidden layer (G), and the pooling layer (O). The input layer U is characterized as a dual matrix of size $M_U \times M_U$. When the CRBM employs l convolutional kernels, each of size $M_X \times M_X$, the hidden layer produces K-dimensional feature maps with dimensions $M_G \times M_G$ where, $(M_G = M_U - M_X + 1)$. The pooling layer further reduces the dimensionality of the hidden representation into an $M_O \times M_O$ format. Using a pooling window of size D, each feature element $D \times D$ in the pooling layer corresponds to a receptive region α of size $D \times D$ in the hidden layer. Figure 4 illustrates the overall CRBM structure.

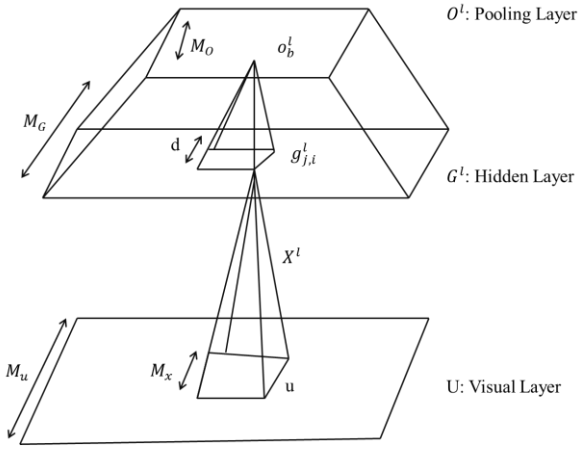

Figure 4: The architecture of CRBM.

$$O\left(g_{ji}^{l}=1\middle|U,\theta\right)=\sigma((\widehat{X}^{l}*U)_{ji}+a_{l}) \quad (11)$$

In equation (11), $O\left(g_{ji}^{l}=1\middle|U,\theta\right)$ is the probability that the hidden unit at position (j,i) in the $l-th$ feature map is activated, given the input U and model parameters θ. a_l is a bias term for the $l-th$ hidden feature map(Equation 12).

$$O\left(U_{ji}^{l}=1\middle|g,\theta\right)=\sigma((\sum_{L}X^{l}*g^{l})_{ji}+b) \quad (12)$$

$O\left(U_{ji}^{l}=1\middle|g,\theta\right)$ Probability that the visible unit at position in the $l-th$ channel is activated, given hidden activations g and model parameters θ (Equation 11). X^l is the convolution kernel corresponding to the $l-th$ hidden feature map. g^l Hidden feature map l. $*$ is a convolution operation. $(\sum_{L}X^{l}*g^{l})_{ji}$ is a convolution result summed over all hidden feature maps at position (j,i). b is a bias of the visible layer. $*$ indicates a convolution operation. The parameter a represents the bias of the visible layer. $a=(a_1,a_2,\dots a_n)^S$ represents the hidden layer bias, while $\theta=(W,a,b)$ is the CRBM model parameter. The activation chances for every characteristic point in the layer is calculated using equation (13).

$$O\left(g_{ji}^{l}=1\middle|U,\theta\right)=\frac{f^{J}(g_{ji}^{l})}{1+\sum_{n,m}f^{J}(g_{ji}^{l})} \quad (13)$$

The probability of the hidden unit at position (j,i) in the $l-th$ hidden feature map being activated, given input U and CRBM parameters θ. Net activation function $f^J(g_{ji}^l)$ applied to the hidden unit (g_{ji}^l), often including convolution outputs plus bias, $\sum_{n,m}f^J(g_{ji}^l)$ is a Sum of the activations. U represents the input matrix. Where f denotes the exponential operation, that was calculated in equation (14),

$$O\left(o_{\alpha}^{l}=0\middle|U,\theta\right)=\frac{1}{1+\sum_{(j,i)\in A_{\alpha}}f^{J}(g_{ji}^{l})} \quad (14)$$

The o_{α}^{l} One output unit. The input data U. θ The model's settings (weights, biases). The group of hidden units A_α connected to this output. g_{ji}^{l} The activity level of each hidden unit in that group. A function f^J that changes the hidden unit activity in some way. The ICDBN overcomes the limitations of standard DBN by integrating convolutional layers with optimization strategies, enhancing feature extraction and reducing overfitting. In this research, the CRBM is pre-trained. the robustness and diversity of the learned convolutional features (equation 15).

$$\Delta a_{l}^{sparisity}=\text{olog}\left(\frac{1}{M}\sum_{m=1}^{M}r_{l}^{m}\right)+91-\text{olog}\,(1-\frac{1}{M}\sum_{m=1}^{M}r_{l}^{m}) \quad (15)$$

The hidden layer's average activation probability value r_l^m in output corresponds to the $l-th$ convolution kernel of the $m-th$ sample, M is the each batch's sample size, $\Delta a_l^{sparisity}$ is the $l-th$ bias penalty term, and o is the sparse coefficient, which was calculated in equation (16).

$$r_{l}^{m}=\frac{1}{M_G\times M_G}\sum_{j=1}^{M_G}\sum_{i=1}^{M}o(g_{ji}^{l}\,|U) \quad (16)$$

MG is the number of hidden feature maps, g_{ji}^{l} is the state of the hidden unit at position (j,i) in the $l-th$ feature map. U is the input layer. Equation (17) signifies that if every time of crbm parameters is changed, the consequence period is appended to the gradient of Δa_l.

$$\Delta a_{l}=\frac{1}{M_G\times M_G}\sum_{j,i}(g_{ji}^{0}-g_{ji}^{l})+\Delta a_{l}^{sparisity} \quad (17)$$

g_{ji}^{l} is the state of the hidden unit at position (j,i) in the $l-th$ feature map. g_{ji}^{0} is the initial (or target/expected) state of the hidden unit before update.

3.3.2 Artificial gorilla troops (AGT) optimization for feature selection and model optimization

The AGT optimization is employed to optimize the selection of multimodal features and fine-tune learning parameters. The AGT algorithm fine-tunes the key parameters of the ICDBN, such as the learning rate, sparsity coefficient and hidden layer weights, through a local adaptive search in the multimodal space of the features to reduce the reconstruction loss and improve the efficacy of feature selection. Task specific tuning also improves the convergence stability and robustness of the model and, it gives better predictive power for the AGT-ICDBN. Figure 5 demonstrates the overall process of AGT optimization.

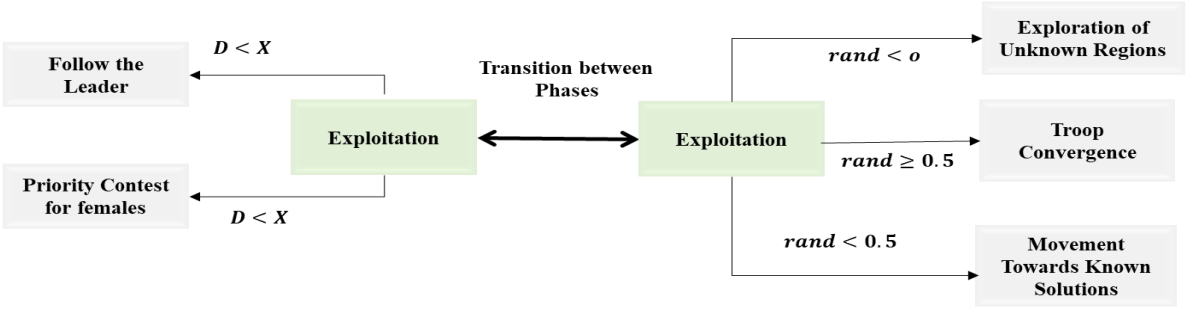

Figure 5: Comprehensive process of AGT optimization

• Exploration phase

Gorilla behavior suggests that they live in groups led by silverbacks, which must be followed for sentiment prediction. Gorillas occasionally split off from their groups and migrate to new areas. The silverback gorilla is the greatest candidate solution in the AGT algorithm, with a random parameter O determining gorilla migration patterns. Equation (18). s, update the gorilla's candidate position HW in every iteration.

$$HW(s+1) = \begin{cases} (VA - KA) \times q_1 + KA & rand < O \\ (q_2 - D) \times W_q(s) + K \times rand \geq 0.5 \\ W(j) - K \times \left(K \times \left(W(s) - W_{(s)} - HW_q(s)\right) + q_3 \times \left(W(s) - HW_q(s)\right)\right) & rand < 0.5 \end{cases} \quad (18)$$

Updated position $HW(s+1)$ of the solution at iteration $(s+1)$. $HW_q(s)$ Position of another solution (gorilla) chosen randomly at iterations. $W(s)$ is a current solution position at iteration s. $W_q(s)$ is another gorilla's position vector at iteration s. $W(j)$ represents a randomly selected solution/position from the population. Variable VA represents a vector adjustment term. A control factor KA for updating positions. Balancing coefficient K, which scales exploration/exploitation. q_1, q_2, q_3 are random numbers $\in [0,1]$ that inject randomness into movement. Random number $rand \in [0,1]$ used to decide which case applies. The threshold value O is used to compare with rand for choosing the movement type. Dimension D of the problem. Where $W(s)$is the gorilla's current position and $HW(s+1)$ is its position vector in the subsequent iteration; VA and KA show the variable's upper and lower values; as a result, rand, q_1, q_2, and q_3 are random variables in the [0-1] range. A randomly chosen gorilla and its position vector are denoted by W_q and HW_q. Here is how D, K and H are computed. Equation (19) defines a control parameter that decreases over iterations.

$$D = E \times (1 - \frac{It}{MaxIt}) \quad (19)$$

The control parameter D decreases over time. Initial value (E) of the control parameter. Current iteration number It. Maximum number of iterations, $MaxIt$, allowed. Equation (20) generates a dynamic oscillating factor (between 0 and 2) to control updates.

$$E = Cos(2 \times q_4) + 1 \quad (20)$$

Control factor E is used in other equations. Random number $q_4 \in [0,1]$ is used to introduce randomness into the control factor. $Cos()$ is a cosine function that ensures E oscillates smoothly. Equation (21) calculates a scaled random coefficient from D for balancing movement.

$$K = D \times k \quad (21)$$

In equation, K is a Scaling/adjustment parameter used in position update rules. Control parameter D decreases over time to shift from exploration to exploitation. Random number $k \in [0,1]$ introduces stochasticity into K. Equation (22) scales the current solution's position by a random factor Y.

$$G = Y \times W(s) \quad (22)$$

G is an updated value, A coefficient or scaling factor, Y. $W(s)$ is the current position of the gorilla (solution) at iteration s. Equation (23) defines the random range of Y based on the control parameter D.

$$Y = [-D, D] \quad (23)$$

A coefficient Y. Control parameter D decreases over time.

• Exploitation phase

The AGT algorithm's two exploitation phases enhance convergence, where gorillas follow or compete with the Silverback for adaptive multimodal optimization.

• Track the silverback

Young gorillas, guided by the Silverback, efficiently update their positions for sentiment prediction using Equation (24), which adjusts movement toward the Silverback with random scaling factors K and N to balance exploration and exploitation.

$$HW(s+1) = K \times N \times (W(s) - W_{Silverback}) + W(s) \quad (24)$$

$HW(s+1)$ is a new position of the gorilla at iteration $s+1$. The $W(s)$ signifies the current position of the gorilla at iteration s. $W_{Silverback}$ depicts the position of the Silverback gorilla. The scaling parameter K balances exploration and exploitation. Random number (N), which introduces variability into the step size. Where $W_{Silverback}$ represents the best solution so far, and K is calculated using equation (25).

$$N = (|\frac{1}{M} \sum_{j=1}^{M} HW_j(s)|^h)^{1|h} \quad (25)$$

Coefficient N was used in position updates. M is the total number of gorillas (population size). The $HW_j(s)$ is a position of the $j-th$ gorilla at iteration s. Power parameter

(h), $1_{|h}$ represents the fractional power. equation (26) defines exponent h as an exponential function of the scaling factor K.

$$h = 2^{K} \tag{26}$$

Exponent parameter h is used in the formula for N. Scaling coefficient K. M denotes the total number of gorillas, whereas $HW_j(s)$ represents each gorilla's position vector throughout iteration.

• Competition for adult females

The equation (27) updates a gorilla's position by competing with the Silverback's influence.

$$HW(j) = W_{Silverback} - (W_{Silverback} \times R - W_s \times R) \times B \tag{27}$$

Updated position HW(j) of the $j-th$ gorilla. $W_{Silverback}$ is a position of the Silverback gorilla. Current position W_s of a gorilla at iterations. R is a random coefficient ($\in [0,1]$) that introduces stochasticity into the update. B is a coefficient vector. equation (28) generates a random coefficient R in the range $[-1,1]$.

$$R = 2 \times q_5 - 1 \tag{28}$$

Random coefficient R is used in gorilla position updates. Random number $q_5 \in [0,1]$. equation (29) defines Silverback's influence coefficient B as a product of constant β and adjustment factor F.

$$B = \beta \times F \tag{29}$$

Predefined constant β represents the scale that scales the effect. F is a threshold/adjustment factor. equation (30) sets the threshold factor F conditionally based on a random number.

$$F = \begin{cases} M_1 \text{ rand } \leq 0.5 \\ M_2 \text{ rand } > 0.5 \end{cases} \tag{30}$$

Threshold/adjustment factor F used in other equations. M_1 is a constant or parameter chosen when the random value is ≤ 0.5. M_2 is constant or a parameter chosen when the random value is > 0.5. Random number $rand \in [0,1]$ that decides which case applies.

• Time complexity of AGT

The AGT algorithm's time complexity depends on M agents, S iterations, and feature dimension C, covering initialization, evaluation, and adaptive update phases. Equation (31) splits the optimizer's total computational cost into three phases.

$$P(AGT) = P(\text{Initialization}) + P(\text{Evaluation}) + P(\text{Update Positions}) \tag{31}$$

The initialization phase has a time complexity of $P(M)$. Each iteration evaluates all gorillas, resulting in $P(S \times M)$. The updated positions use $P(S \times M \times C)$ for both exploitation and exploration. Therefore, the AGT's time complexity was calculated in equation (32).

$$P(AGT) = P(M) + P(S \times M) + P(S \times M \times C) \times 2 = P(M \times (1 + S + SC) \times 2) \tag{32}$$

P(AGT) represents the overall probability of the AGT optimizer. M is the number of gorillas. Search dimension S or the scaling factor related to exploration space. C is a parameter related to competition behavior among gorillas. Algorithm 2 denotes the AGT-ICDBN Multimodal framework for Feature Fusion and Sentiment Prediction. Table 2 depicts the Hyperparameter Tuning of the AGT-ICDBN Framework.

Table 2: Hyperparameter Tuning of the AGT-ICDBN Framework

Parameter	Search Range	Optimized Value
Learning Rate (η)	0.0001 – 0.01	0.0012
Batch Size	16, 32, 64, 128	64
Epochs	5 – 30	10
Number of Hidden Layers	2 – 5	3
Neurons per Layer	64 – 512	256
Sparsity Coefficient (λ)	0.01 – 0.05	0.03
Dropout Rate	0.1 – 0.5	0.3
AGT Population Size	10 – 50	30
Maximum Iterations	20 – 100	60
Momentum (μ)	0.7 – 0.99	0.85
Activation Function	ReLU, LeakyReLU, Sigmoid	LeakyReLU

Algorithm 2: AGT-ICDBN for Multimodal Feature Fusion and Sentiment Prediction

```
Input: Text features Xt, Image features Xi
Output: Predicted sentiment score S
1. Initialize parameters:
    text_dim, image_dim, hidden_dim, output_dim
2. Define ICDBN feature extractor:
    Text_Feature  = ReLU(Wt * Xt)
    Image_Feature = ReLU(Wi * Xi)
    Fused_Feature = Concatenate(Text_Feature, Image_Feature)
3. Apply AGT optimization on fused features:
    Initialize gorilla population (G1…Gn) around Fused_Feature
    For each iteration t:
      For each gorilla Gi:
        If random() < 0.5 → Exploration:
            Gi = Gi + random_noise()
        Else → Exploitation:
            Gi = Gi + (Best_G - Gi) * random()
        Update Best_G if fitness(Gi) > fitness(Best_G)
    Optimized_Feature = Best_G
4. Define sentiment classifier:
    S = Sigmoid(Ws * Optimized_Feature)
5. Output predicted sentiment score S
```

4 Performance analysis

The goal of this research is to create a DL -driven multimodal framework that combines textual, visual, and rating data to analyze and evaluate sentiment towards cultural tourism attractions. The AGT-ICDBN was chosen over alternative fusion methods because it effectively captures nonlinear relationships between multimodal features through adaptive optimization, unlike early, late, or attention-based fusion models that struggle with imbalanced or limited datasets. Table 3 lists the hardware, software, and dataset specifications used for implementing and evaluating the multimodal framework.

Table 3: Experimental specifications for multimodal sentiment prediction

Category	Specification
Hardware Components	
Processor (CPU)	Intel Core i9-13900K (24 cores, 3.0 GHz)
GPU	NVIDIA RTX 3080 (10 GB VRAM) and NVIDIA A100 (40 GB VRAM)
Memory	64 GB DDR5 RAM
Storage	1 TB NVMe SSD
Software Components	
Software Environment	Windows 11 Pro (Build 23H2), Python 3.9, PyTorch 2.1, CUDA 12.2, cuDNN 8.9
Average Training Time per Epoch	2.6 minutes (RTX 3080) / 1.2 minutes (A100)
Total Convergence Time	45–50 minutes (RTX 3080) / 20–25 minutes (A100)
Peak GPU Memory Usage	8.7 GB
System RAM Utilization	≤ 35 GB during multimodal fusion and AGT-based optimization
Optimizer	Artificial Gorilla Troops (AGT) metaheuristic implementation (custom in PyTorch)
Frameworks	Jupyter Notebook / Google Colab / Kaggle for prototyping and training

4.1 Multimodal prediction and sentiment evaluation of cultural tourism attractions

The analysis is to evaluate how accurately the proposed multimodal framework predicts visitor ratings across different cultural tourism attractions. Table 4 present a comparative analysis of predicted and actual ratings alongside dominant sentiments across different cultural tourism attraction types.

Table 4: Comparative analysis of predicted and actual ratings with sentiments

Attraction Type	Avg. Predicted Rating	Avg. Actual Rating	Dominant Sentiment
Heritage Sites	4.6	4.7	Positive
Eco-Towns	4.3	4.4	Positive
Water-based	4.5	4.5	Positive
Commercial Towns	4.0	4.1	Neutral–Positive

By employing the proposed AGT-ICDBN model, heritage sites achieved an average predicted rating of 4.6 compared to an actual rating of 4.7, with the dominant sentiment being positive. Table 5 shows sample multimodal reviews with their predicted sentiment, emotion label, and model confidence scores.

Table 5: Sentiment and emotion prediction from multimodal cultural tourism reviews

Review + Image	Predicted Sentiment	Emotion Label	Confidence (%)
I did not enjoy the guided tour; it was boring.	Negative	Anger	85
Highly positive visit experience, strongly satisfactory service quality.	Positive	Joy	93
Visually appealing environment, though affected by high visitor density.	Neutral	Sadness	78
Architecturally remarkable site with strong aesthetic appeal	Positive	Joy	90
Museum visit yielded positive impressions; staff assistance rated high.	Positive	Joy	92
Poor service, will not return.	Negative	Disgust	88
The local cuisine was delicious and the ambiance perfect.	Positive	Joy	91
Average experience, nothing special.	Neutral	Surprise	75
The place was dirty, and the entry fee was too high.	Negative	Anger	87
Positive perception of cultural performances and heritage representation.	Positive	Joy	89

It purposes to identify the unique strengths of each modality across semantic, aesthetic, and evaluative feature sets. Figure 6 illustrates the relative strengths of Textual, Visual, and Rating data across various features. Overall, Rating Data offers robust quantitative insights, Textual Data enhances interpretability through rich sentiment and semantic features, and Visual Data provides contextual and aesthetic richness.

4.2 Multidimensional evaluation of visitor satisfaction in cultural tourism sites

The analysis examines the relationships among visitor ratings across key dimensions, including overall experience, service, authenticity, and atmosphere. Figure 7 illustrates the correlation patterns among visitor ratings for key aspects of cultural tourism attractions.

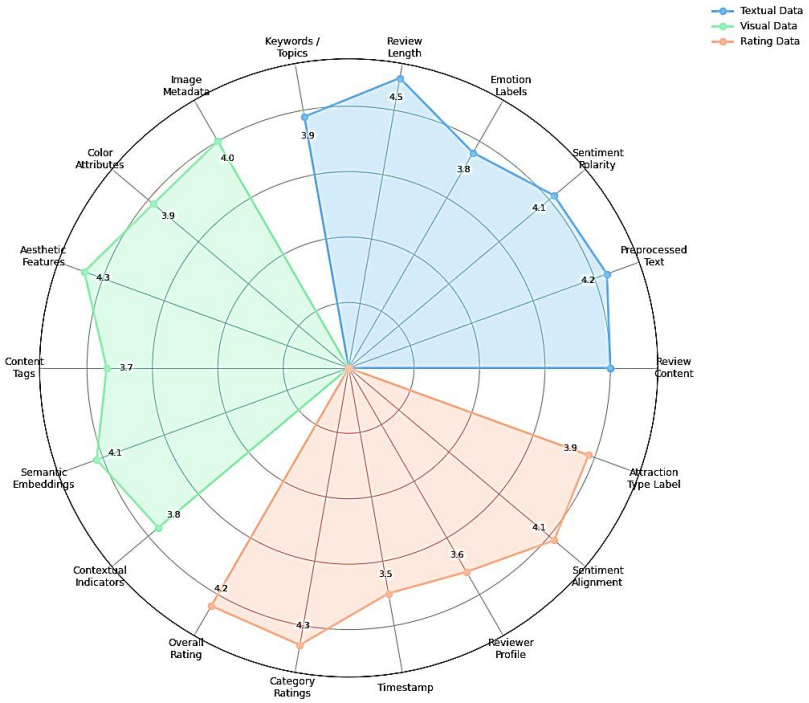

Figure 6: graphical analysis of data Types with those features

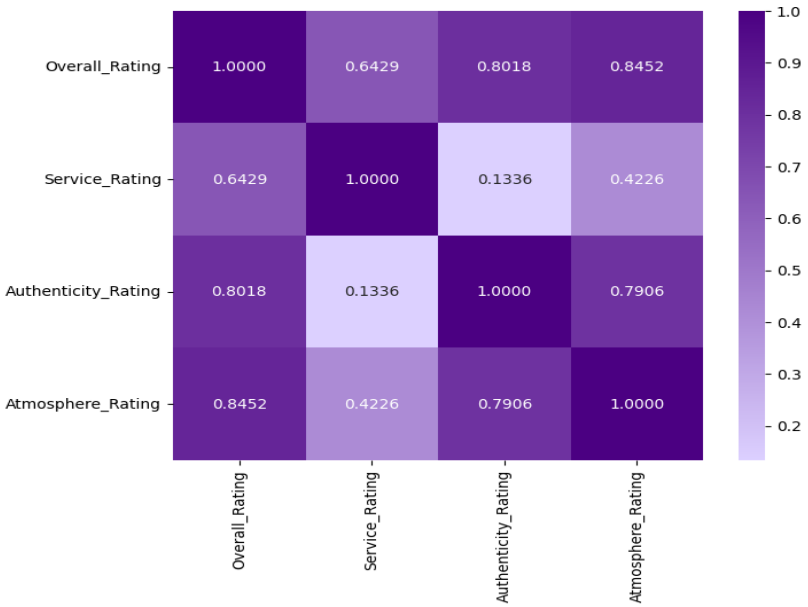

Figure 7: Correlation analysis of visitor ratings in cultural tourism attractions

These observations underscore the significance of combining various facets of visitor experience service, ambiance, and authenticity in multimodal sentiment evaluation predictive models.

4.3 Comparative analysis of attraction categories through transformed feature curves

The intention of this research was to investigate visitor rating patterns between different cultural tourism categories. Figure 8 refers to converted visitor rating

patterns between ecological, heritage, water-based, and commercial attractions, revealing uniform functional trends in evaluation.

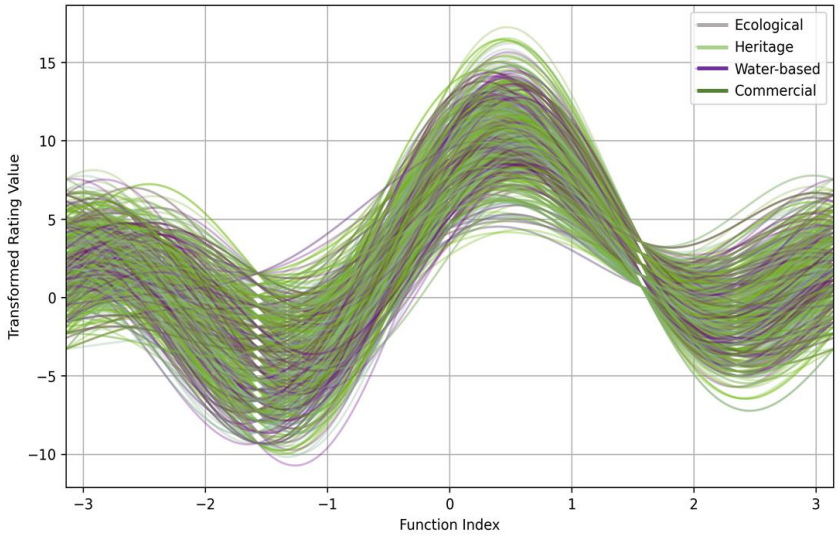

Figure 8: Graphical results of Multidimensional tourist experience patterns

The shifted rating curves showed uniform functional patterns across ecological, heritage, water-based, and commercial attractions, with ecological and water-based sites peaking slightly higher, demonstrating that the proposed model AGT-ICDBN effectively captures cross-category sentiment–evaluation relationships. It aims to deliver a holistic interpretation of visitor attitudes across varied categories of attractions. Figure 9 presents the graphical results of (a) Multimodal sentiment analysis and (b) cultural tourism attractions visitor flow.

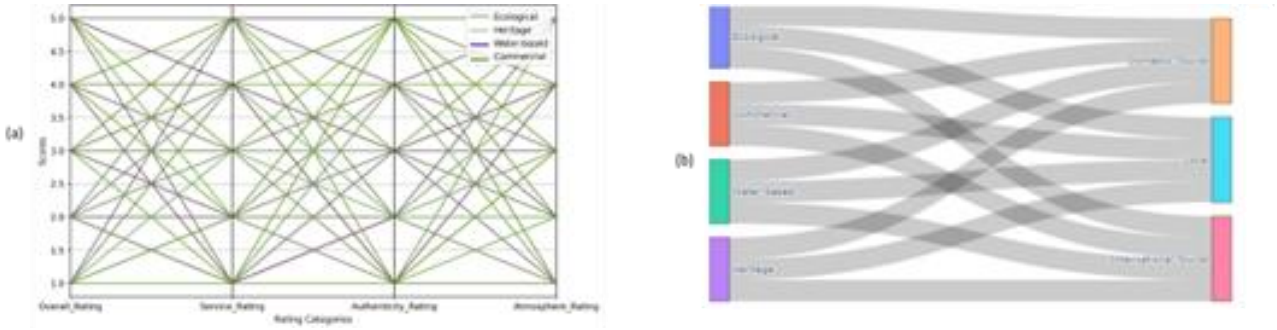

Figure 9: findings of (a) Multimodal sentiment evaluation and (b) visitor flow of cultural tourism attractions

The results presented in Figure 9(a) reveal variations in visitor evaluations across cultural tourism attractions, reflecting diverse priorities such as authenticity, service, and atmosphere. Figure 9(b) demonstrates how these attractions correspond to different visitor groups, highlighting distinct engagement patterns among local, domestic, and international tourists. It seeks to reveal how perceptions differ across attraction categories and visitor groups, providing actionable insights for destination management, cultural preservation, and strategic marketing. Figure 10 represents the multimodal sentiment evaluation of cultural tourism attractions.

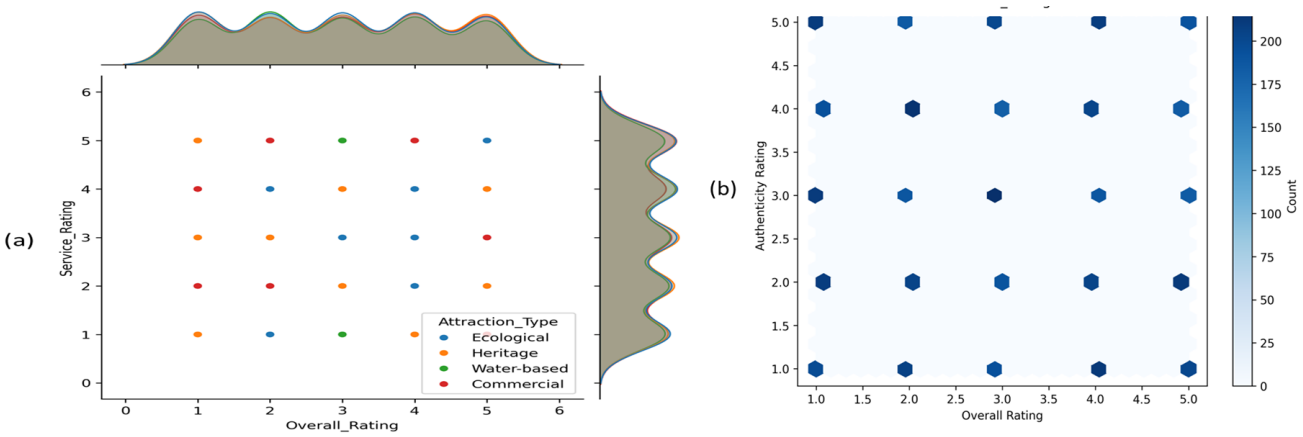

Figure 10: outcomes of (a) Visitor evaluations across cultural tourism attraction categories and (b) Visitor group distribution across attraction types

Figure 10(a) shows that water-based attractions achieved the highest overall visitor rating of 4.6, but lowest in commercial towns (3.8). Figure 10(b) indicates that international tourists dominate visits to cultural–religious sites. These values confirm the framework's ability to uncover meaningful differences in perception and engagement.

SHapley Additive exPlanations (SHAP)

SHAP is an explainable AI method that explains in which way the features of an individual instance influence the predictions of a model by giving each a Shapley value according to cooperative game theory. It serves as an interpreter of the AGT-ICDBN model's decision process for sentiment prediction of cultural tourism reviews, thus, assisting in discovering which features like sentiment polarity, emotion labels, and review content dominate the output as given by Figure 11.

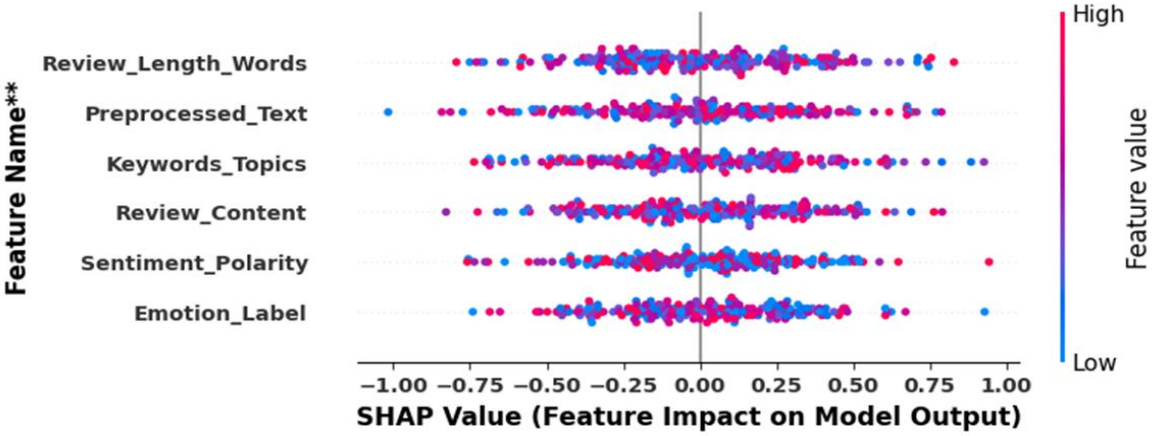

Figure 11: SHAP Beeswarm Plot illustrating the impact of key multimodal textual and emotional features

Features such as sentiment polarity and emotion label have a broader SHAP range which implies that they have a great influence on the model's output whereas review length and preprocessed text have a lesser impact.

Confusion Matrix

A confusion matrix is a table that displays the comparison between a model's predicted and actual classification results. Figure 12 depicts the confusion Matrix for Multiclass Sentiment Prediction using AGT-ICDBN.

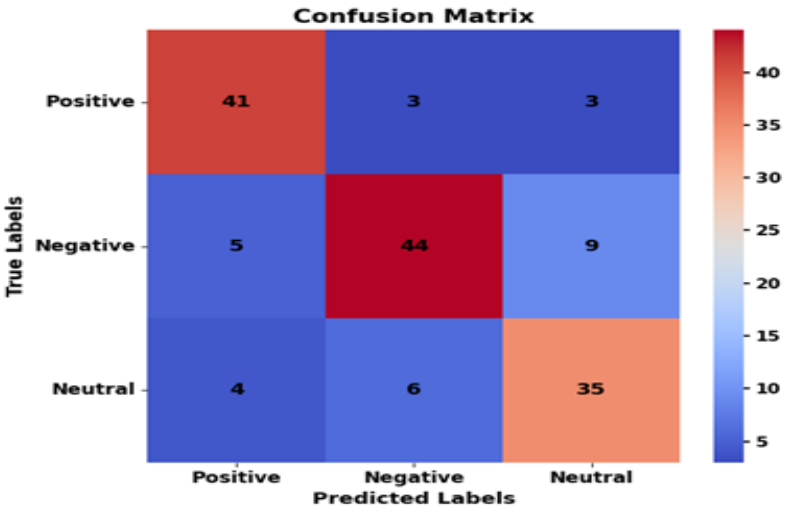

Figure 12: Confusion matrix for multiclass sentiment prediction using AGT-ICDBN

Out of the total samples, the model correctly identified 41 positive, 44 negative, and 35 neutral sentiments, with minor misclassifications observed between neighboring categories These results confirm the model's robustness in handling class overlap and its effectiveness in capturing sentiment nuances across multimodal features.

ROC Curve

The ROC curve compares the true positive rate with the false positive rate at different classification threshold levels. Figure 13 depicts the ROC Curve for Sentiment Classification on the Multimodal Cultural Tourism Experience Dataset.

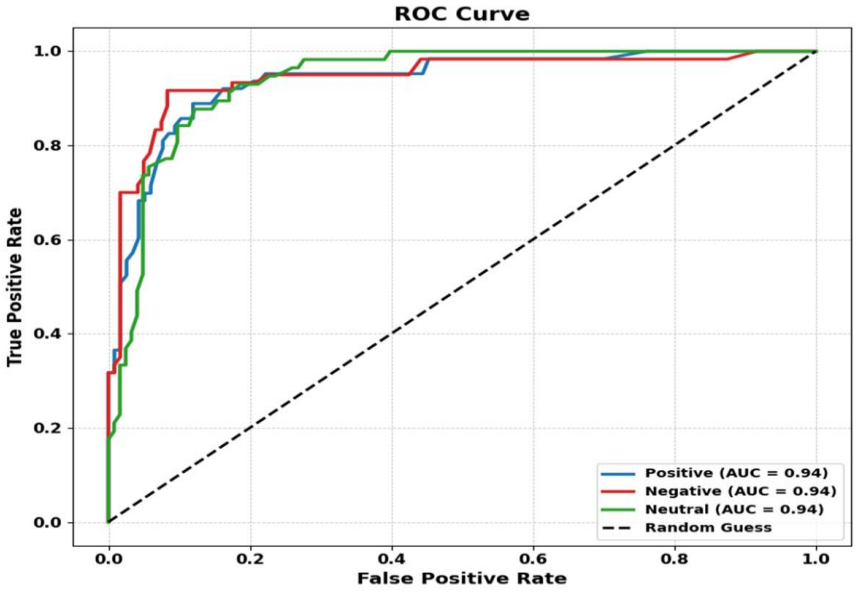

Figure 13: ROC Curve for sentiment classification on the multimodal cultural tourism experience Dataset

The model achieved a consistent AUC value of 0.94 for all three sentiment categories such as positive, negative, and neutral. It signifies superior discriminative ability and balanced classification across all classes compared to random guessing. Table 6 depicts the Ablation Study Results of the AGT-ICDBN Framework.

Table 6: Ablation study results of the AGT-ICDBN Framework

Model Variant	Precision	Recall	F1-score	AUC	CA	RMSE	SRC	MAE
ICDBN	0.820	0.757	0.819	0.9678	0.8384	0.322	0.2956	0.2809
AGT	0.869	0.780	0.828	0.9770	0.852	0.315	0.2710	0.2717
AGT-ICDBN [Proposed]	0.9895	0.9895	0.872	0.9853	0.9895	0.2912	0.3012	0.2784

These results confirm that the AGT-ICDBN enhances both accuracy and robustness, enabling more reliable sentiment and evaluation predictions in multimodal cultural tourism analysis.

4.4 Comparative analysis of proposed and traditional models

The research evaluated existing models for both classification and regression tasks in the context of sentiment prediction and the evaluation of the prediction of cultural tourism attractions. For classification-based sentiment prediction, models such as Linear Regression (LR) [24], Neural Networks (NN) [24], Support Vector Machines (SVM) [24], LR [25], SVM with Radial Regression [25], Decision Tree (DT) [25], Random Forest Regression (RFR) [25], Gradient Boosting Machine (GBM) [25], XGBoost Tree [25], CharCNN [28], LSTM [28], BiLSTM [28] and BERT [28] were assessed using metrics, which are defined in table 7 and figure 14 (Equations 33-40).

Table 7: compared metrics and their explanations

No. of. Equations	Metrics	Equation	Definition
(33)	Area Under Curve (AUC)	$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\text{d}(\text{FPR})$	Measures overall classification ability. TPR = True Positive Rate, FPR = False Positive Rate
(34)	Classification Accuracy (CA)	$\text{CA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$	The proportion of correctly predicted samples. TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives
(35)	F1	$\text{F1 Score} = 2.\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of Precision and Recall
(36)	Precision	$\text{precision} = \frac{\text{T}_\text{P}}{\text{T}_\text{P} + \text{F}_\text{P}}$	Proportion of predicted positives that are actually positive
(37)	Recall	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$	Proportion of actual positives correctly predicted
(38)	Root Mean Squared Error (RMSE)	$\text{RMSE} = \sqrt{\frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}}(\text{y}_\text{i} - \hat{y}_\text{i})^2}$	Measures the average squared difference between predicted ($\hat{y}_\text{i}$) and actual (y_i) values; n = total samples
(39)	Mean Absolute Error (MAE)	$\text{MAE} = \frac{\sum_{\text{i}=1}^{\text{n}} \lvert\text{y}_\text{i} - \text{x}_{\text{i}\rvert}}{\text{n}}$	It calculates the average absolute difference between predicted (y_i) and actual ($\text{x}_{\text{i}\rvert}$) values; n = total samples
(40)	Spearman's Rank Correlation Coefficient (SRC)	$\text{SRC} = 1 - \frac{6\sum_{\text{i}=1}^{\text{n}} \text{d}_\text{i}^2}{\text{n}(\text{n}^2 - 1)}$	Spearman Rank Correlation; d_i difference between ranks of predicted and actual values measures a monotonic relationship

Table 8 presents a comparison of different models for sentiment prediction, highlighting their recall and F1-score to show which model performs best. Figure 11 shows the numerical results of multimodal evaluation.

Table 8: Comparative performance of models for sentiment prediction

Approaches	AUC	CA	F1	Precision	Recall
LR [24]	0.967	0.850	0.850	0.850	0.800
NN [24]	0.967	0.840	0.839	0.840	0.787
SVM [24]	0.973	0.840	0.843	0.859	0.791
AGT-ICDBN [Proposed model]	0.9853	0.9895	0.9877	0.9876	0.9893

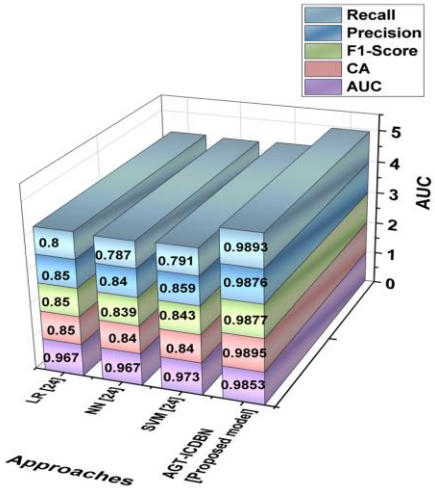

Figure 14: Graphical representation of sentiment prediction in metrics values

The AGT-ICDBN model was better than other techniques for sentiment forecasting. It had the highest AUC (0.9942), CA (0.898), F1-score (0.872), precision (0.882), and recall (0.842) in comparison to LR, NN, and SVM. Table 9 depicts the numerical results of multi-model sentiment analysis.

Table 9: Numerical results of multi-model sentiment analysis

Approaches	SRC	MAE	RMSE
LR [25]	0.1439	0.2919	0.3433
SVM Radial Regression [25]	0.1506	0.2854	0.3504
DT [25]	0.0904	0.2915	0.3467
RFR [25]	0.1937	0.2885	0.3373
GBM [25]	0.2866	0.2809	0.3316
XGBoost Tree [25]	0.2790	0.2817	0.3331
AGT-ICDBN [Proposed model]	0.3012	0.2784	0.2912

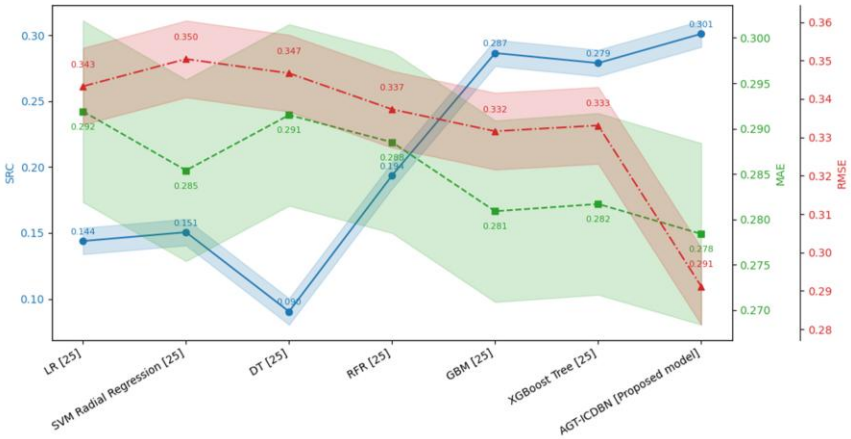

Figure 15: Graphical depiction of compared metrics for multi-model sentiment analysis

The AGT-ICDBN model performed better than conventional regression models in forecasting evaluation scores for cultural tourism sites, as shown in Table 9 and Figure 15. It had the lowest RMSE (0.2912) and MAE (0.2784) but the highest SRC (0.3012) among LR, SVM, DT, RFR, GBM, and XGBoost.Table 10 depicts the outcomes of the proposed method for sentimental analysis when compared with existing method.

Table 10: Outcomes of the proposed method for sentimental analysis when compared with existing method

Methods	CA (%)	Precision (%)	Recall (%)	F1-score (%)
CharCNN [28]	79.89	78.60	75.63	76.12
LSTM [28]	82.46	80.47	81.75	83.76
BiLSTM [28]	86.51	84.35	82.73	85.36
BERT [28]	92.78	92.26	90.23	91.65
AGT-ICDBN [Proposed]	0.9895	0.9876	0.9893	0.9877

The proposed AGT-ICDBN model performed effectively with an accuracy of 0.9895, precision of 0.9876, recall of 0.9893, and F1-score of 0.9877, indicating its ability to handle multimodal inputs. Table 11 depicts the outcomes of the proposed method while compared with existing dataset.

Table 11: Outcomes of the proposed method while compared with existing dataset

Dataset	CA (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
Tourist Review Sentiment Analysis [29]	0.9678	0.9626	0.9623	0.9565	0.9601
Multimodal cultural tourism experience Dataset [Proposed]	0.9895	0.9876	0.9893	0.9877	0.9853

The Tourist Review Sentiment Analysis Dataset [29] has shown great performance with classification accuracy (CA) demonstrating a robust sense of sentiment classification using text. The proposed Multimodal Cultural Tourism Experience Dataset shows substantially better results with a CA of0.9895%, precision of 0.9876%, recall of 98.93%, F1-score of 0.9877%, and AUC of 0.9853%, ranking higher in all evaluation assess for performance.

ANOVA statistical significance analysis

One-way Analysis of Variance (ANOVA) is used to statistically validate performance comparisons between models. Table 12 depicts the One-Way ANOVA Results for Model Performance.

Table 12: One-Way ANOVA results for model performance

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-Value	p-Value
Between Groups (Models)	0.0246	5	0.00492	11.37	0.0038
Within Groups (Error)	0.0087	24	0.00036	–	–
Total	0.0333	29	–	–	–

It shows a between-group sum of squares (SS) of 0.0246, a within-group SS of 0.0087, and a total SS of 0.0333, with degrees of freedom (df) of 5, 24, and 29, respectively, confirming that the AGT-ICDBN model's superior outcomes

are not the result of chance, but rather genuine performance improvements over the baseline models.

5 Discussion

Tourism is a vital sector for economic development and cultural preservation, and understanding visitor experiences is crucial for effective destination management.

5.1 Comparative discussion with state-of-the-art models

Existing models such as Logistic Regression (LR) [24, 25, 28, 29], Neural Networks (NN) [24], Support Vector Machines (SVM) [24, 25], SVM with Radial Regression [25], Decision Trees (DT) [25], Random Forest Regression (RFR) [25], Gradient Boosting Machines (GBM) [25], and XGBoost Trees [25] face significant challenges in multimodal sentiment analysis. LR [24, 25] assumes linearity and fails to capture nonlinear cross-modal relationships; NN [24] requires large labeled datasets and risks overfitting; SVM [24, 25] struggles with kernel tuning and scalability; DT and RFR [25] are prone to overfitting and lack semantic understanding; while GBM and XGBoost [25] need complex hyperparameter optimization and show limited interpretability. To address these challenges, this research develops a DL driven multimodal framework that integrates textual, visual, and rating data to perform sentiment analysis and evaluation prediction for cultural tourism attractions.

AGT-ICDBN achieves better results in terms of accuracy, interpretability and stability than existing state-of-the-art models. Its dynamic parameter tuning improves feature interaction across text, image, and rating modalities, leading to greater accuracy and faster convergence than existing models. It offers interpretable results connecting visual features and textural signals to visitor sentiment, allowing for informed cultural tourism management decisions.

6 Conclusion

The aim is to implement a DL-based multimodal framework that incorporates textual, visual, and rating information to classify sentiment and assess cultural tourism destinations. Image, text, and numerical ratings of reviews were gathered from the Multimodal Cultural Tourism Experience Dataset, reflecting various opinions regarding heritage sites, ecological towns, commercial places, and aquatic-based attractions. Text reviews were preprocessed using tokenization, removal of stop-words, and lemmatization, whereas images were resized to 124×124 pixels and normalized by Z-score transformation to make them stable and comparable. The CLIP framework was utilized to extract semantic visual features, and textual along with visual features were mapped to multimodal embeddings. The AGT-ICDBN was employed to fuse multimodal features, optimize learning parameters, and predict visitor sentiment with improved robustness and interpretability. The AGT-ICDBN achieved superior performance over traditional models, reaching an AUC of 0.9853, CA of 0.9895, F1-score of 0.9877, precision of 0.9876, and recall of 0.9893 for sentiment prediction, while attaining the lowest RMSE of 0.2912 and MAE of 0.2784 with the highest SRC of 0.3012 for evaluation prediction. The multimodal framework successfully transformed visitor feedback into interpretable and accurate sentiment and evaluation predictions, providing actionable insights for destination management, cultural heritage preservation, and tourism strategy.

Limitations and future scope

While the AGT-ICDBN framework reflects high performance, it still has some limitations. The generalization of the model to large-scale real-world tourism platforms has not been confirmed yet, so there is a need for more tests with different user-generated data. The scalability in computational of the AGT optimizer may decrease the efficiency for a large multimodal data for its iterative tuning procedure. Also, the lack of external ground-truth sequences limits the evaluation of model generalization to new geographic regions and types of attractions. Future research can expand the dataset to multiple regions and cultural contexts, integrate additional modalities such as audio and geolocation data, and optimize the framework for real-time applications in smart tourism systems.

Funding

This work was supported by Project at the school level of Xinzhou Normal University, applied for in 2021, led by, researching the formation mechanism of tourism and cultural spaces along the Xinzhou section of the Great Wall (No. CCYJ202105) and Mount Wutai Cultural Ecology Collaborative Innovation Center of "1331 Project" of Xinzhou Normal University: Research on high-quality development path of Mount Wutai cultural tourism integration based on Taixin integration, applied and presided in 2022 (No. WYSYJ202222)

Author contributions

Yuanhui Gong writing original draft preparation & methodology, Yuanhui Gong investigation & writing review and editing.

References

[1] Mteti, S.H., Mpambije, C.J. & Manyerere, D.J., 2025. Unlocking cultural tourism: Local community awareness and perceptions of cultural heritage resources in Katavi Region in the southern circuit of Tanzania. Social Sciences & Humanities Open, 11, p.101295. https://doi.org/10.1016/j.ssaho.2025.101295

[2] Chen, X. & Yu, S., 2024. Synergizing culture and tourism talents: Empowering tourism enterprises for success. Journal of the Knowledge Economy, 15(3),

pp.12439–12471. https://doi.org/10.1007/s13132-023-01598-x

[3] Ye, J., Qin, Y. & Wu, H., 2024. Cultural heritage and sustainable tourism: unveiling the positive correlations and economic impacts. Current Psychology, 43(47), pp.36393–36415. https://doi.org/10.1007/s12144-024-07070-6

[4] Zhao, X., Elahi, E., Wang, F., Xing, H. & Khalid, Z., 2024. Sustainable tourism development for traditional Chinese drama's intangible cultural heritage. Heliyon, 10(3), p.e24560.https://doi.org/10.1016/j.heliyon.2024.e25483

[5] Pai, C.H., Zhang, Y., Wang, Y.L., Li, K. & Shang, Y., 2025. Current challenges and opportunities in cultural heritage preservation through sustainable tourism practices. Current Issues in Tourism, pp.1–19. https://doi.org/10.1080/13683500.2024.2443776

[6] Ma, C., Somrak, T., Manajit, S. & Gao, C., 2024. Exploring the potential synergy between disruptive technology and historical/cultural heritage in Thailand's tourism industry for achieving sustainable development in the future. International Journal of Tourism Research, 26(5), p.e2759. https://doi.org/10.1002/jtr.2759

[7] Zubiaga, M., Sopelana, A., Gandini, A., Aliaga, H.M. & Kalvet, T., 2024. Sustainable cultural tourism: Proposal for a comparative indicator-based framework in European destinations. Sustainability, 16(5), p.2062. https://doi.org/10.3390/su16052062

[8] Moura, A., Eusébio, C. & Devile, E., 2023. The "and what for" participation in tourism activities: Travel motivations of people with disabilities. Current Issues in Tourism, 26(6), pp.941–957. https://doi.org/10.1080/13683500.2022.2044292

[9] Soltani Nejad, N., Rastegar, R. & Jahanshahi, M., 2024. Tourist engagement with mobile apps of E-leisure: A combined model of self-determination theory and technology acceptance model. Tourism Recreation Research, 49(4), pp.714–725. https://doi.org/10.1080/02508281.2022.2100194

[10] Sun, T., Li, Y. & Tai, H., 2023. Different cultures, different images: A comparison between historic conservation area destination image choices of Chinese and Western tourists. Journal of Tourism and Cultural Change, 21(1), pp.110–127. https://doi.org/10.1080/14766825.2021.1962894

[11] Lao, Y., Zhu, J. & Liu, J., 2023. Tourism destinations and tourist behavior based on community interaction models of film-enabled tourism destinations. Frontiers in Psychology, 13, p.1108812. https://doi.org/10.3389/fpsyg.2022.1108812

[12] Ariyani, N. & Fauzi, A., 2024. Unlocking sustainable rural tourism to support rural development: A Bayesian approach to managing water-based destinations in Indonesia. Sustainability, 16(13), p.5506. https://doi.org/10.3390/su16135506

[13] Zhang, S., Lin, Z. & Yhang, W.J., 2025. Forecasting international tourist arrivals in South Korea: A deep learning approach. Journal of Hospitality and Tourism Technology, 16(2), pp.247–268. https://doi.org/10.1108/JHTT-03-2024-0176

[14] Bi, J.W., Han, T.Y. & Yao, Y., 2024. Collaborative forecasting of tourism demand for multiple tourist attractions with spatial dependence: A combined deep learning model. Tourism Economics, 30(2), pp.361–388. https://doi.org/10.1177/13548166231153908

[15] Jiang, R. & Dai, B., 2024. Cultural tourism attraction recommendation model based on optimized weighted association rule algorithm. Systems and Soft Computing, 6, p.200094. https://doi.org/10.1016/j.sasc.2024.200094

[16] Zhang, X. & Wang, X., 2025. Rural tourist attractions recommendation model based on multi-feature fusion graph neural networks. International Journal of Computational Intelligence and Applications, 24(1), p.2450027. https://doi.org/10.1142/S1469026824500275

[17] Jeribi, F., Perumal, U. & Alhameed, M.H., 2024. Recommendation system for sustainable day and night-time cultural tourism using the mean signed error-centric recurrent neural network for Riyadh historical sites. Sustainability, 16(13), p.5566. https://doi.org/10.3390/su16135566

[18] Liang, X., Li, X., Shu, L., Wang, X. & Luo, P., 2025. Tourism demand forecasting using graph neural network. Current Issues in Tourism, 28(6), pp.982–1001. https://doi.org/10.1080/13683500.2024.2320851

[19] Bozkurt, A. & Şeker, F., 2023. Harmonizing heritage and artificial neural networks: The role of sustainable tourism in UNESCO world heritage sites. Sustainability, 15(17), p.13031. https://doi.org/10.3390/su151713031

[20] Paudel, T., Li, W. & Dhakal, T., 2024. Forecasting tourist arrivals in Nepal: A comparative analysis of seasonal models and implications. Journal of Statistical Theory and Applications, 23(3), pp.206–223. https://doi.org/10.1007/s44199-024-00079-7

[21] Tian, Y. & Tang, X., 2025. The use of artificial neural network algorithms to enhance tourism economic efficiency under information and communication technology. Scientific Reports, 15(1), p.8988. https://doi.org/10.1038/s41598-025-94268-8

[22] Zhang, Y., Tan, W.H. & Zeng, Z., 2025. Tourism demand forecasting based on a hybrid temporal neural network model for sustainable tourism. Sustainability, 17(5), p.2210. https://doi.org/10.3390/su17052210

[23] Nanjappa, Y. et al., 2024. Improving migration forecasting for transitory foreign tourists using an

ensemble DNN-LSTM model. Entertainment Computing, 50, p.100665. https://doi.org/10.1016/j.entcom.2024.100665

[24] Calderón-Fajardo, V., Rodríguez-Rodríguez, I. & Puig-Cabrera, M., 2025. From words to visuals: A transformer-based multi-modal framework for emotion-driven tourism analytics. Information Technology & Tourism, pp.1–41. https://doi.org/10.1007/s40558-025-00334-2

[25] Gupta, V., Jung, K. & Yoo, S.C., 2020. Exploring the power of multimodal features for predicting the popularity of social media images in a tourist destination. Multimodal Technologies and Interaction, 4(3), p.64. https://doi.org/10.3390/mti4030064

[26] Puh, K., & Bagić Babac, M. (2023). Predicting sentiment and rating of tourist reviews using machine learning. Journal of hospitality and tourism insights, 6(3), 1188-1204. https://doi.org/10.1108/JHTI-02-2022-0078

[27] Erdoğan, D., Kayakuş, M., Çelik Çaylak, P., Ekşili, N., Moiceanu, G., Kabas, O., & Ichimov, M. A. M. (2025). Developing a Deep Learning-Based Sentiment Analysis System of Hotel Customer Reviews for Sustainable Tourism. Sustainability, 17(13), 5756. https://doi.org/10.3390/su17135756

[28] Cao, Z., Xu, H., & Teo, B. S. X. (2023). Sentiment of chinese tourists towards malaysia cultural heritage based on online travel reviews. Sustainability, 15(4), 3478. https://doi.org/10.3390/su15043478

[29] https://www.kaggle.com/datasets/sangitapokhrel/tourist-review-sentiment-analysis