Research on Experience Evaluation and Optimization of Rural Festival Participants Using BERT Multimodal Sentiment Analysis

Man Qiao
School of Tourism and Business, Guangzhou Polytechnic University, Guangzhou 511483, GuangDong, China
Corresponding author: Man Qiao (e-mail: pp20201008@163.com)

ABSTRACT To solve the problems of separating emotional perception from textual and image commentary in rural festival activities and identifying multi-dimensional experience pain points, particularly those 'implicit' contradictions where positive textual sentiment masks negative visual cues (e.g., a visitor praising the scenery while the image reveals severe over crowding), remains challenging. This paper proposes a multimodal sentiment analysis framework that combines BERT semantic encoding and cross-modal collaborative attention. First, the textual and image data of typical rural festivals in the Yangtze River Delta region are collected and processed to align the two modalities. Second, BERT is finetuned to extract textual semantics, and ResNet-50 is employed to encode visual atmosphere. Finally, a collaborative attention mechanism is introduced to integrate text and image data, allowing the model to simultaneously process cross-modal interactions and pinpoint the specific visual sources of participant sentiment. The experimental results show that the fusion model obtains a macro-average F1 score of 91.5%, which is an improvement of 9.1% compared with the singlemodal model. Furthermore, the model can also locate the core pain points of crowdedness with the attention weight of 0.78, which provides an interpretable basis for optimizing rural festivals.

INDEX TERMS rural festival experience, multimodal sentiment analysis, bert model, cross-modal attention, smart tourism

I. INTRODUCTION

AS an important tool for the rural revitalization strategy, rural tourism has shown a significant trend of transformation from sightseeing to in-depth experience in recent years [1]. Against this backdrop, rural festivals and events have become core carriers for activating local memories and attracting off-season tourists. The Yangtze River Delta region, as a pioneer in the rural revitalization strategy, offers a diverse typology of such events—ranging from flower-viewing to traditional fishing harvests. By analyzing the Lishui Plum Blossom Festival, Digang Fishing Village Harvest Festival, and Shushan Pear Blossom Festival, this study captures representative models of seasonal, cultural, and ecological tourism that mirror the broader digital transformation challenges faced across China. However, the operation and management of current rural festivals and events still face a prominent contradiction: the means of obtaining and analyzing participants' experience evaluations are seriously lagging behind the development speed of the industry. Traditional questionnaire surveys are limited by the distribution scenario and collection cycle, making it difficult to capture the real feelings of tourists on social media platforms after the event; and facing the massive amount of text and image comment data accumulated on platforms such as Xiaohongshu and Douyin, most existing studies only perform coarse-grained sentiment classification on the text, and rarely incorporate the visual semantics such as scene atmosphere and spatial crowding implied in the images into the analysis framework [2,3]. This separation of text and image perception makes it impossible for operators to accurately identify experiential contradictions where the visual scene reveals objective issues—such as litter or congestion—that the user might not explicitly complain about in their text, thereby capturing a more comprehensive 'implicit' dissatisfaction, and it is urgent to introduce an intelligent evaluation method that can synergistically understand the language description and visual scene. Existing research on the experience evaluation of rural festivals and events mainly follows two paths. Li et al. pointed out that the image of the event and the image of the city together affect tourists' intention to revisit, indicating that the evaluation of the event experience is no longer limited to the on-

site service itself, but is closely coupled with the broader destination perception structure [4]. Hsu et al., based on the stimulus-organism-response framework, found that the perception of the authenticity and experience value of event participants significantly affects their emotional response and loyalty behavior [5]. Choo et al. further confirmed that the degree of tourist participation in local food festivals is an important antecedent variable affecting the formation of loyalty [6]. At the same time, Yen and Yu pointed out from the perspective of event brand equity that satisfaction plays a mediating role between brand awareness and loyalty formation [7]. Demir and Dalgıç, starting from the shaping effect of food festivals on destination attractiveness, emphasized the key role of event activities in the formation of tourists' perception and behavioral intentions [8]. Although the above work provides a basic framework for understanding the experience of event participants, there are still problems such as the evaluation data source still being mainly structured questionnaires, which fail to effectively release the value of unstructured social graphic data. In response to the inherent defects of the separation of graphic and textual information perception, multimodal fusion technology has shown significant advantages in the field of affective computing in recent years. For example, Das and Singh pointed out in their review of multimodal sentiment analysis that collaborative modeling of heterogeneous modalities such as text and images has become an important direction for improving the robustness of sentiment recognition and the accuracy of semantic discrimination [9]. Al-Tameemi et al. further systematically reviewed the development of visual-text sentiment analysis on social media, arguing that joint modeling of text and images can more effectively capture implicit attitudes, scene cues, and emotional triggers in user expressions [10]. Meanwhile, Arevev et al. showed that BERT models adapted for tourism corpora can significantly enhance the semantic understanding of tourism review texts, providing methodological support for more refined sentiment recognition in tourism scenarios [11]. However, the application of the above methods in tourism research is still in its early stages, especially for rural festivals, a specific scenario that combines cultural performance and spatial aggregation. No research has yet systematically integrated BERT semantic representation, residual visual encoding, and cross-modal attention fusion. Based on this, this paper constructs a multimodal sentiment analysis framework for text and image reviews of rural festivals, aiming to make up for the shortcomings of existing research in cross-modal semantic collaboration and experience pain point attribution.

The main contributions of this paper are reflected in the following three aspects: First, a multimodal sentiment analysis framework adapted to rural festival scenarios is constructed, realizing the collaborative representation of the semantics of comment text and the visual features of festival images in a high-level semantic space, breaking through the limitations of traditional single-modal evaluation methods in perceiving the correlation information between text and images. Second, through a systematic ablation and comparative experimental design, the independent contributions and interactive gains of the BERT pre-trained language model and the cross-modal attention fusion mechanism in the sentiment recognition task of festival comments are clarified, providing empirical evidence for introducing cutting-edge natural language processing technology into the rural tourism field. Third, an attention weight attribution analysis method is introduced, realizing a granular leap from "overall sentiment judgment" to "local pain point localization," providing interpretable decision support for festival operators to accurately identify the specific triggers of participants' negative experiences.

II. COLLECTION AND CROSS-MODAL ALIGNMENT PREPROCESSING OF HETEROGENEOUS TEXT AND IMAGE DATA

The data collection part was based on the following three typical rural festivals with regional influence in

Yangtze River Delta area: Lishui Plum Blossom Festival in Nanjing, Digang Fishing Village Harvest Festival in Huzhou and Shushan Pear Blossom Festival in Suzhou. To guarantee the timely and representative data collection, we limited the data collection period within two weeks before and after these three festivals. We specifically ensured that the dataset included a high density of posts from the 'peak days' of each festival to capture the most intense periods of spatial crowding, while using the pre- and post-event data to provide a baseline for environment hygiene and commercial atmosphere. An automated web crawler framework was employed to automatically collect publicly available text and image notes and short video cover frames containing event keywords from Xiaohongshu and Douyin. Collected fields included user nicknames, posting times, text content, original image URLs and metadata likes and comments. In the text part, the pure emojis, advertising and duplicate reposts were removed. In the image part, pure text posters, QR codes and blurry images with resolution lower than 224 $\times$ 224 pixels were removed.

In the cross-modal alignment part, confronted with the general situation that there are multiple images accompanying one post in users-published notes, this paper makes preliminary associations based on the native binding relationship between one comment and multiple images at the time of posting on the platform. Specifically, one sequence of comment text is used as the anchor, and all images posted by the same user at the same time will be collected into one "text-image group". And in the multiple images sample, the mean vector of residual network output for one image will be used as the visual representation of one comment in the subsequent visual feature extraction. Besides, while this study primarily focuses on the synergistic effects of multimodal data, we acknowledge that single-modality posts constitute a significant portion of social media. For the purpose of training the fusion framework, samples containing only text without effective images or only images without

text descriptions were categorized separately. In the final evaluation, we integrated these single-modality samples as a baseline to ensure the 'experience evaluation' results reflect the broader participant base rather than only the most active multimodal users. So that each record in the training set and testing set contains both complete sequence of text and at least one real image of event.

III. TEXT CHANNEL: FINE-GRAINED SEMANTIC FEATURE EXTRACTION OF BERT CONTEXT-ORIENTED COMMENT TEXT

Comment text will be cleaned and normalized (such as converting full-width to half-width) before sending comment text into encoding layer. All full-width characters are converted into half-width, and all special symbols, continuous spaces and garbled unparseable characters in comment text will be removed firstly. For internet slang, dialect homophone and station expression frequently appearing in comment text, we will directly output their original form and will not be strongly normalized to avoid the loss of colloquial expression emotion. Then, we will tokenize the cleaned comment text with our character-level segmenter, and add [CLS] at the beginning of sequence and [SEP] at the end of sequence, and uniformly truncate or padding into sequence of fixed length 128 to meet the requirement of subsequent batch processing.

The encoder uses a pre-trained Chinese BERT-base model, which contains a 12-layer Transformer structure, a hidden layer dimension of 768, and 12 self-attention heads. To address the differences in vocabulary distribution and semantic expression between texts related to rural festivals and general corpora, this study fine-tunes the model using comment texts from a constructed multimodal corpus based on pre-trained weights. During the fine-tuning phase, the parameters of the first six layers of BERT are frozen to preserve general semantic knowledge, while only the parameters of the last six layers and the classification head are updated via backpropagation. This reduces the risk of overfitting while encouraging the model to learn expression patterns specific to the festival context. Specifically, after inputting the target text sequence into the BERT encoder, the 768-dimensional vector corresponding to the [CLS] label in the last layer is taken as the contextual semantic representation of the comment. This vector is mapped to the sentiment classification space through a fully connected layer and then fused with image-side features across modalities. Furthermore, the hidden state sequence of all time steps in the last layer is retained at the text channel output for use by the subsequent cross-modal attention interaction module when calculating the corresponding weights of the text and local image regions. The parameter configuration is shown in Table 1.

TABLE 1. BERT Text Encoder Configuration Parameters

Parameter	Setting	Description
Pre-trained Model	BERT-base-Chinese	Official Chinese pre-trained weights
Transformer Layers	12	Encoder depth
Hidden Dimension	768	Output vector dimension
Self-Attention Heads	12	Parallel attention channels
Max Sequence Length	128	Input truncation/padding length
Fine-tuning Strategy	Freeze first 6 layers, update last 6	Domain adaptation approach
Optimizer	AdamW	Optimization algorithm
Learning Rate	2×10^{-5}	Parameter update step size
Dropout Rate	0.1	Random deactivation probability

IV. VISUAL CHANNELS: RESIDUAL NETWORK REPRESENTATION COMPUTATION OF FESTIVAL SCENE ATMOSPHERE FEATURES

As the name implies, image-side feature extraction maps real-life photos of rural festival activities taken by participants into visual semantic vectors that can be understood by models. Given that rural festival images contain diverse and complex scenes ranging from flower fields to folk performances, market stalls, and swarms of tourists, we choose ResNet-50 residual network pretrained on large-scale ImageNet dataset as our basic encoder. This structure well alleviates the gradient vanishing problem in deep network training through skip connections. It can extract multi-level visual features from objects' shape, spatial layout, and tonal atmosphere while keeping a relatively simple structure.

In preprocessing, all input images are resized into 224 $\times$ 224 pixel images and normalized by channel into the input distribution of pretrained model through mean and standard deviation. For a comment text that has multiple images, forward propagation is conducted on each image and the 2048-dimensional feature vector output by the global average pooling layer at the end of the network is taken as the visual semantic representation of this text and image record after arithmetic mean of all vectors element by element. The mean vector is taken as the overall visual record of the text and image record of this text and image record. This method is reasonable in aggregating visual records from multiple aspects of the same festival activity by one participant, and does not increase any parameters.

Concerning the updating strategy of ResNet50 main structure's parameters, we did not conduct full finetuning of ResNet50 main structure, but froze the convolutional part and batch normalization part's parameters pretrained on ImageNet, and only trained the newly added fully connected adaptation layer at the end. This is because scenes of festivals have certain domain-specific characteristics, but visual parts of images have strong

sharing with general natural images. Therefore, freezing the backbone network can avoid fitting on small-scale samples and accelerate the convergence of models. After dimensionality reduction by adaptation layer, the image feature vector and the BERT semantic vector on the text side remain in the same dimension. It is convenient for the later cross-modal fusion layer to conduct alignment and interactive computation between the two. In Table 2, we list the important parameters in feature extraction stage.

TABLE 2. ResNet-50 Visual Encoder Configuration Parameters

Parameter	Setting	Description
Pre-trained Dataset	ImageNet	Large-scale natural image classification dataset
Input Image Size	224 × 224 pixels	Resized uniformly
Output Feature Dimension	2048	Global average pooling layer output
Network Architecture	ResNet-50	50-layer residual network
Fine-tuning Strategy	Freeze convolutional and BN layers	Train only the final adaptation layer
Multi-image Aggregation	Element-wise arithmetic mean	Average features across multiple images
Adaptation Layer Output	768	Aligned with BERT semantic vector dimension

V. CROSS-MODAL SENTIMENT SEMANTIC FUSION COMPUTATION BASED ON COLLABORATIVE ATTENTION

Even though the BERT semantic vectors extracted from text channel and ResNet image feature vectors extracted from visual channel have gone through representation encoding in their own two channels and become an independent semantic space, these two kinds of vectors haven't formed an explicit mapping relationship between them and textual description can't determine what visual object will be projected by participants' expression in rural festival commentary texts, moreover global vectors can't reflect the contradictory information between text and crowded image. Therefore, this paper proposes collaborative attention mechanism to realize interaction between text and image in high level semantic space.

Specifically, the text-side input is the sequence of hidden states at each time step of the BERT final layer output, denoted as $T= [t_1, t_2, ..., t_n]$, where n is the sequence length, and each time step corresponds to the contextual representation of a word or phrase in the commentary text. The image input is the feature map output from the last convolutional layer of ResNet-50, denoted as $V= [v_1, v_2, ..., v_m]$, where m= 49 corresponds to a 7 × 7 spatial grid. Each grid region retains the visual semantic information within the local receptive field of the scene. Collaborative attention computation is divided into two parallel sub-branches: a text-guided image attention branch, which uses text semantics as the query condition to calculate the relevance between each spatial region of the image and the comment text content; and an image-guided text attention branch, which uses image features as the query condition to measure the importance weight of each word in the text sequence triggered by the visual scene.

In the text-guided image attention branch, BERT[CLS] vector is firstly linearly transformed to be the query matrix and the image feature map is transformed again to be the key and value matrices. The distribution of attention weight of each region of image is further acquired by computing the scaled dot product attention. Finally, the weight vector is weighted and summed with the original image feature map to obtain the image attention representation under text conditions. While in the image-guided text attention branch, symmetrically, the global average pooling vector of image is used as the query and the text hidden state sequence is used as the key and value. The attention weight of each word on current image context is computed and then weighted sum could further form the text attention representation under image conditions.

VI. EMOTIONAL POLARITY MAPPING AND ATTRIBUTION-BASED OUTPUT OF EXPERIENCE DIMENSION

The emotional semantic representation vector output from the cross-modal fusion layer is mapped to a three-dimensional space via a fully connected layer, corresponding to positive, neutral, and negative emotional labels. The Softmax function normalizes the three-dimensional output values into a probability distribution, taking the category with the highest probability as the model's predicted emotional polarity for the given image and text record. Thus, the model completes the end-to-end discrimination process from the original image and text input to emotional classification.

However, a simple emotional classification result only indicates whether the operator's comment is negative, but not the specific reason of the negative review, i.e. whether it is disappointing about the content of the event or dissatisfied with the environment of the crowed place. Thus, finding "where the problem lies" is more practically meaningful than just knowing "the problem exists" for optimizing the experience of rural festival activities. Therefore, in this study, we add an attribution-based module on the emotional discrimination. Since the image region attention weights have been calculated in the text-guided image attention branch, we extract the 7 × 7 grid matrix of attention weight for upsample the original output of text-guided image attention branch to the original image resolution with bilinear interpolation to get the heatmap with the same size as input image.

The areas with higher activation value in heatmap are the vision range that the model focuses on whether the comment has positive or negative sentiment. Then we spatially map these high-activation areas to registered event experience dimensions: if the highlighted area is in the walkways and market areas, it is attributed to crowded space; if the highlighted area is on the stall signs and the merchandise display area, it is attributed to commercial homogenization; if the highlighted area is around the trash cans or on the ground, it is attributed to environment hygiene. Regarding non-visual pain points like 'noise,' the model attributes them through the image-guided text attention branch, identifying high-frequency auditory descriptors in the text that correlate with specific spatial settings (e.g., stage areas or crowded markets) captured in the visual frame. Finally, the statistic output of the times of high-attention coverage of each category of experience dimension and the average weight of heatmap in these areas are achieved by statistical analysis and provide event operators with concrete and traceable optimization data.

VII. MODEL PERFORMANCE AND EVIDENCE CHAIN CONSTRUCTION EFFECT EVALUATION

A. ABLATION EXPERIMENT

To examine whether image-text bimodal collaboration makes a substantial contribution to sentiment recognition, this experiment set up three control groups while maintaining consistent training parameters. The first group retained only the BERT text branch and set the image input to zero; the second group retained only the ResNet visual branch, and the text-side input was uniformly filled with a fixed mask vector; the third group was the complete multimodal fusion model. All three groups were run on the same test set, and their respective sentiment classification performance was recorded.

TABLE 3. Performance Evaluation of Multi-Source Data Fusion and Clue Mining

Model Configuration	Weighted Precision	Weighted Recall	Macro F1-Score
BERT Text-Only Branch	85.6%	85.1%	85.3%
ResNet Image-Only Branch	76.2%	74.8%	75.4%
BERT-ResNet Multimodal Fusion	91.7%	91.4%	91.5%

The ablation experiment results are shown in Table 3 above. After removing the image branch, the macro-average F1 score of the single-text model dropped to 85.3%, a decrease of 6.2 percentage points compared to the complete fusion scheme; the single-image model performed the weakest, with an F1 score of only 75.4%. This numerical difference indicates that relying solely on commentary text misses the emotional cues conveyed by the visual scene, and stripping away semantic information based solely on the image is insufficient to judge the complex attitudes of participants. This confirms that cross-modal interaction between text and image has an irreplaceable beneficial effect on the accurate perception of rural festival activities.

B. COMPARATIVE EXPERIMENTS

To examine the relative advantages of the pre-trained BERT model in semantic understanding of rural festival commentary texts, this experiment kept the image branch in the multimodal fusion framework unchanged and replaced the text encoders sequentially with Word2Vec average pooling, Bi-LSTM, TextCNN, the untuned BERT-base, and the fine-tuned BERT model used in this study. All groups completed learning under the same training epochs and batch size. During the testing phase, the macro-average F1 score, weighted precision, and weighted recall for the sentiment three-class classification task were recorded.

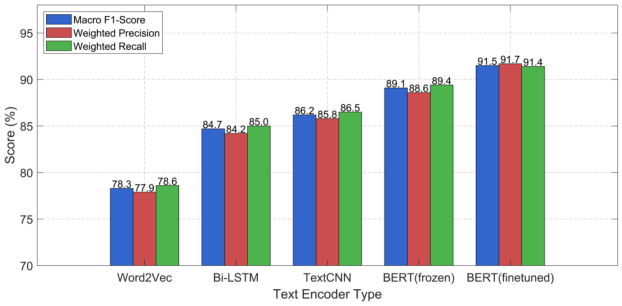

FIGURE 1. Performance Comparison in Multimodal Sentiment Classification Task

Figure 1 shows the performance differences of different text encoders under the same multimodal framework. Taking the macro-average F1 score as an example, Word2Vec static word vectors only achieved 78.3%, while Bi-LSTM and TextCNN, thanks to their sequence modeling capabilities, improved to 84.7% and 86.2%, respectively, while the untuned BERT-base reached 89.1%. After fine-tuning the event commentary corpus, the BERT model's macro-average F1 score reached 91.5%, 5.3 percentage points higher than TextCNN and 2.4 percentage points higher than the BERT with frozen parameters. This numerical gradient confirms that the combination of deep bidirectional semantic pre-training and domain-adaptive fine-tuning has an irreplaceable advantage in accurately perceiving the implicit emotional shifts and experiential evaluations in the comments of participants in rural events.

C. ATTENTION VISUALIZATION AND ATTRIBUTION EXPLANATION EXPERIMENT

This experiment sampled from the test set the sample whose emotional expression of textual reviews and image content was contradictory. Extracted was the weight distribution matrix of the cross-modal attention layer, and the attention heatmap values were mapped to the original image coordinate space. Based on the boundary of five prior labeled pain points area (crowding, commercialization, garbage dumping, etc.), the mean of normalized attention weight in each area was statistically analyzed whether the emotional judgment of the model focused on key visual clues leading to negative reviews.

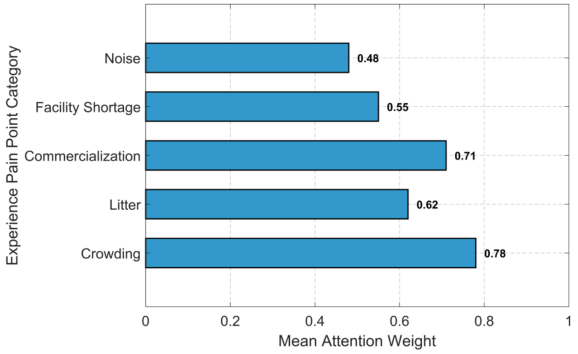

FIGURE 2. Weight Attribution Distribution on Negative Experience Factor Areas

The analysis results in Figure 2 show that the cross-modal fusion model exhibits significant differences in the attention weights given to different pain point areas. The average attention weight is highest in the crowded area (0.78), followed by the commercialized and homogenized area (0.71), then garbage dumping and facility shortage areas (0.62 and 0.55 respectively), and lowest in the noise interference area (0.48), which represents the model's identification of sound-intensive environments through cross-modal correlation rather than direct visual detection. This gradient distribution closely matches the spatial location of the manually labeled pain points, and the image elements covered by the high-weight areas precisely correspond to the frequently mentioned dissatisfaction factors in participants' comments. This confirms that the model does not blindly fit negative sentiment judgments, but rather effectively captures the visual cues that induce negative evaluations, demonstrating an interpretable attribution ability for pain points in rural festival experiences.

VIII. CONCLUSION

In response to the practical problem of extracting textual and image information in rural festival experience evaluation, this study establishes a BERT semantic encoding–cross-modal attention-based multimodal sentiment analysis framework. Experimental results demonstrate that the framework can effectively combine complementary information from comment text and festival image, correctly reflect the real emotional tendency of participants, and locate the core pain points of participants' experiences such as crowd and commercial homogenization. This work is subject to several limitations. First, the data source is limited to the rural festival type in the Yangtze River Delta region. Second, the model cannot handle dynamic video information. In the future, we will extend the geographical and activity type coverage of the samples. In addition, we will attempt to introduce temporal multimodal features to describe the dynamic evolution of participants' experiences.

AUTHOR CONTRIBUTIONS

Man Qiao designed, collected and analyzed the data, and drafted the manuscript. Man Qiao conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Man Qiao participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by,
1. Research on the Improvement Path of Rural Festival Experience Quality under the Background of Big Data (2024312351);
2. Guangzhou Polytechnic University Key Programs in Humanities and Social Sciences at School Level: A Study of Paths to Improve the Quality of Rural Festival Experiences in the Context of Big Data (2023SK01);
3. Guangdong–Hong Kong–Macao Digital Cultural Tourism Research Base, a Key Research Base for Humanities and Social Sciences of Guangdong Provincial Regular Higher Education Institutions (2023WZJD019)

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] P. Alarcón-Urbistondo, M. M. Rojas-de-Gracia, and A. Casado-Molina, "Proposal for Employing User-Generated Content as a Data Source for Measuring Tourism Destination Image," Journal of Hospitality & Tourism Research, vol. 47, no. 4, pp. 643-664, 2023, doi: 10.1177/10963480211012756.
[2] O. A. George and C. M. Q. Ramos, "Sentiment Analysis Applied to Tourism: Exploring Tourist-Generated Content in the Case of a Wellness Tourism Destination," International Journal of Spa and Wellness, vol. 7, no. 2, pp. 139-161, 2024, doi: 10.1080/24721735.2024.2352979.
[3] M. J. Sánchez-Franco and S. Rey-Tienda, "The Role of User-Generated Content in Tourism Decision-Making: An Exemplary Study of Andalusia, Spain," Management Decision, vol. 62, no. 7, pp. 2292-2328, 2024, doi: 10.1108/MD-06-2023-0966.
[4] H. Li, C.-H. Lien, S. W. Wang, T. Wang, and W. Dong, "Event and City Image: The Effect on Revisit Intention," Tourism Review, vol. 76, no. 1, pp. 212-228, 2021, doi: 10.1108/TR-10-2019-0419.
[5] F.-C. Hsu, E. Agyeiwaah, and L. I. L. Chen, "Examining Food Festival Attendees' Existential Authenticity and Experiential Value on Affective Factors and Loyalty: An Application of Stimulus-Organism-Response Paradigm," Journal of Hospitality and Tourism Management, vol. 48, no. 10, pp. 264-274, 2021, doi: 10.1016/j.jhtm.2021.06.014.
[6] H. Choo, D.-B. Park, and J. F. Petrick, "Festival Tourists' Loyalty: The Role of Involvement in Local Food Festivals," Journal of Hospitality and Tourism Management, vol. 50, no. 9, pp. 57-66, 2022, doi: 10.1016/j.jhtm.2021.12.002.
[7] I.-Y. Yen and H. A. Yu, "How Festival Brand Equity Influences Loyalty: The Mediator Effect of Satisfaction," Journal of Convention & Event Tourism, vol. 23, no. 4, pp. 343-361, 2022, doi: 10.1080/15470148.2022.2067606.

[8] M. Demir and A. Dalgıç, "Examining Gastronomy Festivals as the Attractiveness Factor for Tourism Destinations: The Case of Turkey," Journal of Convention & Event Tourism, vol. 23, no. 5, pp. 412-434, 2022, doi: 10.1080/15470148.2022.2089797.

[9] R. Das and T. D. Singh, "Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges," ACM Computing Surveys, vol. 55, no. 13s, pp. Article 270, 1-38, 2023, doi: 10.1145/3586075.

[10] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "A Comprehensive Review of Visual-Textual Sentiment Analysis from Social Media Networks," Journal of Computational Social Science, vol. 7, no. 3, pp. 2767-2838, 2024, doi: 10.1007/s42001-024-00326-y.

[11] V. Arefeva and R. Egger, "When BERT Started Traveling: TourBERT—A Natural Language Processing Model for the Travel Industry," Digital, vol. 2, no. 4, pp. 546-559, 2022, doi: 10.3390/digital2040030.