# MSAQE: A Large-scale Dataset for Multi-view Scenic Areas Quality Evaluation

Yajie Wang*[1][0009−0008−5981−5761], Meiling Li*[1][0000−0002−4224−8528], Gaozhi Liu[1][0009−0008−2813−0452], Xinpeng Zhang[1][0000−0001−5867−1315], Sun Yunlong[2][0000−0002−7358−4575], and Zhenxing Qian✉[1][0000−0002−5224−6374]

[1] Fudan University, School of Computer Science, Yangpu 200433, China
{yjwang22,gzliu24}@m.fudan.edu.cn,{mlli20,zhangxinpeng,zxqian}@fudan.edu.cn
[2] Fudan University, Department of Tourism, Yangpu 200433, China
yunlongsun@fudan.edu.cn

**Abstract.** Addressing the research gap in comprehensive scenic area quality evaluation is crucial in today's evolving tourism industry. Existing approaches rely heavily on sentiment analysis, capturing general visitor sentiment but failing to reflect fine-grained aspects. To address this limitation, we introduce the first large-scale, multi-view dataset for Multi-view Scenic Areas Quality Evaluation (MSAQE), consisting of 291,714 comments labeled across eight specific quality aspects: business management, excursions, hygiene, post and telecommunications, tourism transportation, travel safety, travel shopping, as well as resources and environmental protection. Meanwhile, we propose an innovative data-driven framework integrating Reference-Based Sentiment Analysis (RBSA) and Global and Local Ensemble Encoding (GLEE) based multi-label classification. RBSA employs a fine-tuned ALBERT model to generate sentiment scores for comments, followed by similarity calculations with a reference comment set, resulting in more accurate sentiment evaluations. Since sentiment analysis alone cannot fully assess quality, we integrate GLEE-based multi-label classification to evaluate eight specific quality aspects more comprehensively. Experimental results confirm our framework's superiority over existing sentiment analysis and multi-label classification benchmarks. The dataset and code are available at https://github.com/Yajie-good/MSAQE.

**Keywords:** Multi-view Scenic Areas Quality Evaluation · Sentiment Analysis · Multi-label Classification.

## 1 Introduction

The quality of scenic areas is crucial for attracting and retaining tourists, which in turn drives the local economy and promotes cultural exchange. In the contemporary interconnected world, social networks have emerged as a crucial medium for expressing and disseminating tourist experiences. Many Online Travel Agency (OTA) platforms such as Weibo, TripAdvisor, and Ctrip have democratized the

[1] *These authors contributed equally to this work.

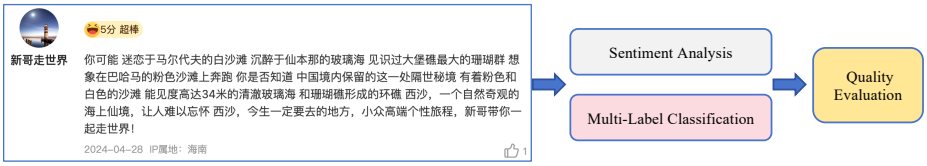


Fig. 1: User-generated content from a typical OTA and a quality evaluation schematic diagram.

collection of tourist opinions, thereby providing a scalable substitute for traditional surveys [4; 17; 18]. The comments posted by visitors on these platforms substantially influence the perception and evaluation of scenic areas. Nevertheless, the majority of analyses primarily concentrate on the general sentiment polarity (e.g., positive, negative, or neutral) [14; 27; 29], seldom delving into the fine-grained quality aspects that are fundamental to comprehensively understanding scenic areas. A data-driven, fine-grained comprehension of these quality aspects can enhance tourist satisfaction and foster the sustainable development of the tourism industry.

Despite the rich volume of user-generated content, current methods in scenic area evaluation often lack the detailed, aspect-specific evaluation necessary for actionable insights. In recent years, researchers have engaged in understanding and enhancing the quality of scenic areas. Traditional methods for evaluating scenic area quality have primarily relied on small-scale surveys and manual analysis [9; 32]. Though insightful, these methods are constrained by limited sample sizes and subjective biases, providing only a partial view of visitor perceptions. Recent approaches leverage unsupervised topic modeling techniques [19; 23; 39] to categorize feedback, but these methods typically yield suboptimal performance and require extensive human curation.

There exist several challenges for comprehensive scenic area quality evaluation. First, there is a lack of large-scale datasets for quality evaluation from multiple perspectives. Second, there is a lack of systematic evaluation methods that can both reflect the degree of tourists' attention to various aspects of the scenic spot and the fine-grained emotional evaluation of tourists for each dimension. To address these issues, we first construct a Multi-view Scenic Areas Quality Evaluation (MSAQE) dataset, the first large-scale dataset designed for fine-grained evaluation of scenic area quality. The MSAQE dataset consists of 291,714 comments gathered from major OTA platforms, namely Ctrip[3] and Qunar[4], and it has been meticulously annotated concerning eight quality aspects. This dataset offers actionable insights to researchers, policymakers, and tourism operators, filling the gaps in data-driven and detailed evaluations for real-world applications. Afterward, we propose a framework for a comprehensive evaluation of scenic area quality, which obtains the feedback from semantic-level and sentiment-level of user comments. The framework mainly contains a reference-

[3] http://www.ctrip.com/
[4] https://www.qunar.com/

Table 1: Summary of existing datasets related to tourism. MSAQE is the first large-scale, multi-view dataset designed for scenic areas fine-grained quality evaluation through semantic and sentiment analysis.

| Datasets | Language | Task | Scenario | Data Scale |
|---|---|---|---|---|
| SPACE [3] | English | Unsupervised Opinion Summarization | Hotel reviews from TripAdvisor | 1k |
| SubjQA [6] | English | Question Answering | Books, movies, grocery, electronics, hotels, and restaurants | 10k |
| NoReC [35] | Norwegian | Sentiment Analysis, Opinion Mining | Literature, movies, video games, restaurants, music, and theater | 43k |
| ReviewQA [13] | English | Question Answering | Hotel reviews | 100k |
| Tripadvisor Reviews [7] | English | Review Rating, Topic Modeling on Reviews | Hotel reviews from Tripadvisor | 20k |
| 515K Hotel Reviews[24] | English | Sentiment analysis, Reviews clustering | Hotels reviews | 515k |
| Trip Advisor Reviews [2] | English | Question Answering, Review Rating, Topic Modeling | Hotel Reviews from Tripadvisor | 20k |
| MSAQE | Chinese | Fine-grained quality evaluation | Scenic areas comments from Ctrip and Qunar | 291k |

based sentiment analysis module and a multi-label classification module, with the former module focusing on the user's affective inclination on each quality aspect of the scenic area, and the latter reflecting the user's degree of attention on the specific aspect by categorizing comments across multiple quality aspects. The framework provides a more fine-grained evaluation, enabling deeper insight into various quality dimensions. Fig. 1 depicts a typical OTA application scenario, demonstrating how user-generated content enriches our comprehension of scenic area quality. Our main contributions are summarized as follows:

- We present MSAQE, a large-scale, multi-view dataset designed for scenic area quality evaluation across eight fine-grained aspects, filling a critical gap in the comprehensive dataset for tourism service quality evaluation.
- We propose an integrated data-driven framework for scenic area quality evaluation that combines Reference-Based Sentiment Analysis (RBSA) and Global and Local Ensemble Encoding (GLEE) based multi-label classification. RBSA utilizes a reference-based approach for more accurate sentiment scoring, while GLEE captures detailed quality aspects.
- Extensive experimental results demonstrate that our framework significantly outperforms baselines, offering a more fine-grained and comprehensive approach to scenic area quality evaluation.

## 2 Related Works

This section reviews existing tourism datasets and tourism quality evaluation approaches, highlighting the need for a fine-grained evaluation method.

### 2.1 Tourism-related Datasets

Current tourism-related datasets are mainly designed for tasks like sentiment analysis [2; 24], opinion summarization [3], question answering [6; 13], review rating [7], and opinion mining [35], etc. Despite their contributions, these datasets neglect multiple aspects of the comments, failing to meet fine-grained scenic area quality evaluation requirements. In response, MSAQE bridges this gap by offering extensive multi-aspect annotations. As displayed in Table 1, with 291,714 comments labeled in eight aspects, MSAQE sets a new standard in fine-grained scenic area quality evaluation and provides a valuable resource for tourism-related research.

Table 2: A comment example from the MSAQE dataset. The text segments are marked with the same color as the corresponding labels.

| Comment | Label |
|---|---|
| 小岛不大，娱乐项目有远海潜水、海底观光艇、海洋馆等。潜水能看到美丽的珊瑚和漂亮的热带鱼，空气清新，环境保护得比较好。潜水教练也很不错，是由景区统一管理的。不要去玩海底观光艇，不值得。如果是跟团的话，导游会大力推荐去美狮厨餐厅吃自助，饮食倒是干净卫生。小岛没什么好玩的,就是距离市区近, 交通比较方便, 坐公交就可以直达。<br><br>The island is not big, and the entertainment items include high-sea diving, underwater sightseeing boats, aquariums, etc. Diving can see beautiful corals and beautiful tropical fish, the air is fresh and the environment is well protected. The diving instructor is also very good and is managed by the scenic spot. Don't go on a submarine sightseeing boat, it's not worth it. If you are in a group, the tour guide will strongly recommend going to the Meishi Kitchen Restaurant for a buffet, and the food is clean and hygienic. Nothing is interesting about the island, but it is close to the urban area, and the transportation is relatively convenient. You can go directly to it by bus. | BM., Exc., REP., Hyg., TT. |

### 2.2 Approaches for Tourism Quality Evaluation

Service quality evaluation in tourism has evolved from traditional survey-based assessments to sophisticated analytical methodologies leveraging modern statistical and computational techniques. Early research such as [21] utilizes factor analysis and structural equation modeling to demonstrate the positive impacts of perceived service quality on tourist satisfaction. Similarly, Wu and Dong [36] apply the SERVQUAL model to explore the critical role of service attributes in shaping tourist experiences. Recent advancements introduce more complex tools. Hou [16] explores fuzzy clustering analysis for dynamic service quality evaluation, showcasing how fuzzy logic adeptly handles the subjective nuances of tourist experiences. Qi et al. [28] employ a back propagation neural network to analyze tourist satisfaction in agro-tourism, highlighting machine learning's effectiveness in discerning complex patterns from large datasets. Furthermore, Li et al. [22] use natural language processing to analyze tourist perceptions of mountain scenic spots, enhancing understanding of tourist behavior through cognitive-emotion theory. While these methods capture overall trends in tourist satisfaction, they generally address single dimensions of sentiment or specific service aspects and frequently overlook the detailed, multi-dimensional nature of the quality that tourists experience in scenic areas. For example, current research on sentiment analysis typically focuses on general polarity (positive, negative, neutral) and lacks the granularity required to assess specific quality aspects. As a result, they fall short of delivering the comprehensive, fine-grained evaluation needed for actionable insights into scenic area management.

Unlike the above approaches, our approach combines reference-based sentiment analysis with multi-label classification, filling this gap by capturing multiple quality dimensions within a unified framework. Leveraging the MSAQE dataset, the integrated method allows for a more comprehensive and fine-grained evaluation, advancing beyond the capabilities of current fine-grained evaluation approaches for scenic areas' service quality.

Table 3: Ratio of comments with different numbers of labels.

| Label Numbers | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total |
|---|---|---|---|---|---|---|---|---|
| Count | 75395 | 47355 | 45839 | 74489 | 39339 | 9247 | 50 | 291714 |
| Ratio (%) | (25.8%) | (16.2%) | (15.7%) | (25.5%) | (13.5%) | (3.2%) | (0.1%) | (100%) |

Table 4: Ratio of each label in the MSAQE dataset.

| Label | BM. | Exc. | Hyg. | PT. | TT. | TSA. | TSH. | REP. |
|---|---|---|---|---|---|---|---|---|
| Count | 230499 | 204846 | 164656 | 218 | 53262 | 10478 | 5246 | 188900 |
| Ratio (%) | (79.0%) | (70.2%) | (56.4%) | (0.1%) | (18.2%) | (3.6%) | (1.8%) | (64.8%) |

## 3 MSAQE Dataset

This section elaborates on the creation of the MSAQE dataset, from data collection and annotation to dataset analysis.

**Data Collection** We collected $6,052,100$ raw comments from Ctrip and Qunar, two major OTA platforms in China. To ensure data quality, we conducted several pre-processing operations. First, we retained comments with more than five characters and high semantic representation. Second, we removed comments with URLs, inappropriate content, or duplicates (with text similarity over 70%).

**Dataset Annotation** We selected 8 primary aspects from the 12 outlined in China's Standard of Rating for Quality of Tourist Attractions (GB/T17775—2003) [1] for quality evaluation. This choice was guided by two main considerations. First, these 8 aspects—business management (BM.), excursions (Exc.), hygiene (Hyg.), post and telecommunications (PT.), tourism transportation (TT.), travel safety (TSA.), travel shopping (TSH.), and resources and environmental protection (REP.)—are most frequently referenced in user comments, capturing core areas of visitor concern. Second, these aspects provide greater operational relevance and improvement potential within scenic areas, each supported by detailed definitions to ensure consistent annotation (see appendix). Table 2 illustrates an example comment annotated with multiple labels.

**Dataset Analysis** Table 3 and Table 4 provide the statistic of various labels in MSAQE dataset. This underscores the dataset's richness and multi-dimensional nature, highlighting that many comments span multiple quality aspects simultaneously. To better understand comment representations with different labels within the dataset, we selected $1,000$ comments for each label and applied ALBERT to generate their semantic embeddings respectively. As shown in Fig. 2, comments with different labels exhibit distinct overlapping clusters, reflecting complex interrelationships between labels, and highlighting the challenges inherent in multi-label classification.

To further explore label co-occurrence, we calculate the joint distribution of labels. Give label $L_i$ and $L_j$, we can get the correlation between $L_i$ and $L_j$:

$$C_{ij} = \frac{N_{ij} \times 100}{min\{|L_i|, |L_j|\}}, \tag{1}$$

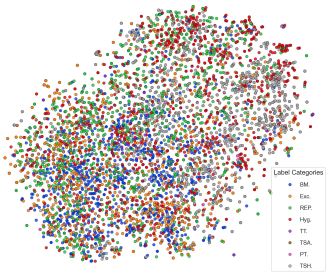


Fig. 2: The t-SNE visualization of single-label comment embeddings in the MSAQE dataset. Different colors represent different labels.

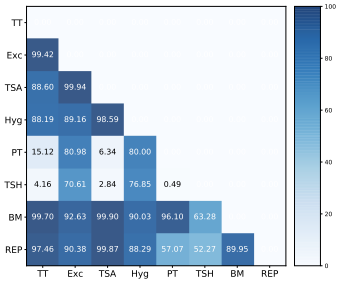


Fig. 3: Joint distribution of different labels.

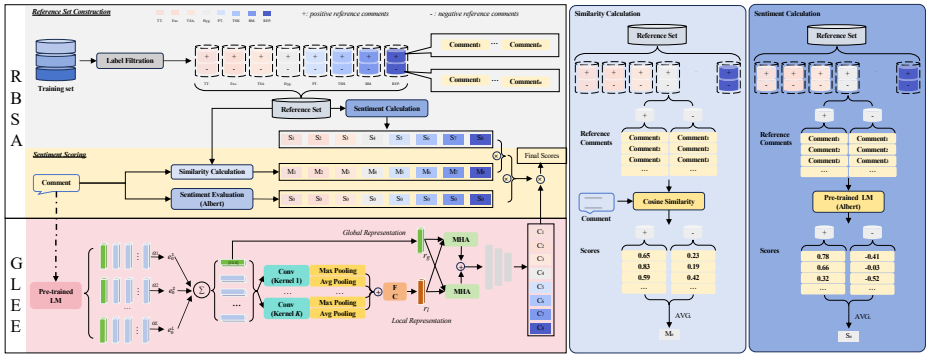


Fig. 4: The overall framework of the proposed approach.

where $N_{ij}$ represents comments labeled with both $L_i$ and $L_j$, and $min|L_i|, |L_j|$ is the minimum number of comments for either $L_i$ or $L_j$. Fig. 3 reveals frequent co-occurrences; for example, 'TT.' often appears with 'BM.', 'REP.', and 'Exc.', while 'Hyg.' commonly overlaps with 'TSA.', 'BM.', and 'Exc.'.

## 4 Methodology

While generic pre-trained language models (PLMs) offer substantial capabilities, they often lack domain-specific knowledge and adaptation for specialized applications like scenic area quality evaluation, where fine-grained insights are essential. To address these limitations, we propose a data-driven integrated framework tailored to our setting, incorporating domain-relevant features and enabling a comprehensive, fine-grained analysis of tourism-related sentiment and quality dimensions. Fig. 4 illustrates the overall framework of the proposed approach, which consists of Reference-based Sentiment Analysis (RBSA) and Global and Local Ensemble Encoding (GLEE) based multi-label classification. Specifically,

**Algorithm 1** Reference-based Sentiment Analysis (RBSA)

**Input:** Comment $X$, ALBERT model for sentiment scoring, reference dataset $D_p^c$, $D_n^c$ for each quality aspect $c$

**Output:** Final sentiment score $S_X^c$ for comment $X$ on each aspect $c$

**for** each quality aspect $c \in \{1, \ldots, C\}$ **do**
  **Step 1: Albert Sentiment Score**
  Compute $S_{direct}^c \leftarrow$ ALBERT_sentiment_score($X$)
  **Step 2: Similarity-weighted Sentiment Score**
  Initialize score_positive $\leftarrow 0$, score_negative $\leftarrow 0$
  **for** each reference comment $X_p^c \in D_p^c$ **do**
    Compute similarity $SIM_p^c \leftarrow \text{CosineSim}(X, X_p^c)$
    $Score_p \mathrel{+}= SIM_p^c \cdot S_p^c$
  **end for**
  **for** each reference comment $X_n^c \in D_n^c$ **do**
    Compute similarity $SIM_n^c \leftarrow \text{CosineSim}(X, X_n^c)$
    $Score_n \mathrel{+}= SIM_n^c \cdot S_n^c$
  **end for**
  Compute $Score_{ref}^c \leftarrow \frac{Score_p + Score_n}{|D_p^c| + |D_n^c|}$
  **Step 3: Final Sentiment Score**
  $S_X^c \leftarrow \delta \cdot S_{direct}^c + \omega \cdot Score_{ref}^c$
**end for**
**return** $S_X^c$ for each aspect $c$

Table 5: Scale of the positive and negative reference sets for each quality aspect.

| Sentiment | BM. | Exc. | Hyg. | PT. | TT. | TSA. | TSH. | REP. |
|---|---|---|---|---|---|---|---|---|
| Positive | 89399 | 77846 | 58805 | 88 | 20700 | 3753 | 1596 | 68767 |
| Negative | 38001 | 33525 | 24500 | 32 | 11421 | 2021 | 126 | 29810 |

the RBSA module evaluates the sentiment score of test comments concerning each quality aspect label. Meanwhile, the GLEE-based multi-label classification provides aspect-specific labeling, allowing us to measure the degree of attention paid to each quality dimension within users' feedback.

### 4.1 Reference-Based Sentiment Analysis

We first constructed a reference set including eight comment sets that correspond to eight quality aspects $c \in \{1, \ldots, C\}$, each consisting of a positive subset $D_p^c$ and a negative subset $D_n^c$. Table 5 summarizes the distribution within each subset, illustrating the imbalance across various aspects. We use an ALBERT model to score the sentiment of each comment[5]. The reference set supports more fine-grained sentiment scoring, enhancing the relevance of sentiment analysis in scenic area quality evaluation.

[5] Sentiment scoring tool: https://huggingface.co/voidful/albert_chinese_small_sentiment. Comments with scores above 0.2 are retained for reliable reference.

Given a test comment $X$, the RBSA module evaluates the sentiment score of $X$ concerning each quality aspect label. Using these subsets, the sentiment score for comment $X$ on label $c$ is calculated as follows:

$$Score^c_{ref} = \frac{\sum_{X^c_p \in D^c_p} CosSim(X, X^c_p) \cdot S^c_p + \sum_{X^c_n \in D^c_n} CosSim(X, X^c_n) \cdot S^c_n}{|D^c_p| + |D^c_n|}, \tag{2}$$

where $X^c_p$ and $X^c_n$ are comments in the positive subset $D^c_p$ and negative subset $D^c_n$, respectively, with sentiment scores $S^c_p$ and $S^c_n$. The $CosSim(\cdot)$ function calculates the cosine similarity between the representation of comment $X$ and each reference comment in $D^c_p$ and $D^c_n$.

Moreover, to achieve more precise sentiment evaluation, we combine the reference-based score $Score^c_{ref}$ with ALBERT-based score $S^c_{direct}$. Then we obtain the final sentiment score $Score^c_X$:

$$Score^c_X = \delta \cdot S^c_{direct} + \omega \cdot Score^c_{ref}, \tag{3}$$

where $\delta$ and $\omega$ are hyper-parameters. This dual approach ensures that each comment' s sentiment aligns with established sentiment benchmarks, enhancing the clarity and relevance of sentiment evaluation. The process for RBSA is summarized in Algorithm 1.

### 4.2 Multi-label Classification

The multi-label classification module captures the level of focus each comment places on different quality aspects by classifying each comment into relevant quality dimensions. Here, we propose a Global and Local Embedding Ensemble (GLEE) approach to achieve multi-label classification.

**Global Encoding** We take the power of PLMs to capture the global semantic representation of comments. We extract the hidden states of the `[CLS]` token output by the PLMs and compute their weighted sum as the global representation $r_g$, *i.e.*,

$$r_g = \sum_{l=1}^{L} \alpha_l \cdot e^l_0[\texttt{CLS}], \tag{4}$$

where $\alpha_l$ denotes the layer weights, and $e^l_0$ represents the hidden states from the $l$-th layer of the transformer.

**Local Encoding** Local encoding highlights key phrases and dependencies within the comment, capturing fine-grained details. We apply Convolutional Neural Network (CNN) to the PLMs outputs with convolutional filters of varying sizes. For each kernel size, both max-pooling and avg-pooling are performed, and the features are concatenated as the local representation $r_l$, *i.e.*,

$$\begin{aligned} r_l &= \mathrm{ReLU}(W \cdot f_{\mathrm{concat}} + b), \\ f_{\mathrm{concat}} &= \mathrm{concat}(f^1_{max}, f^1_{avg}, \ldots, f^K_{max}, f^K_{avg}), \end{aligned} \tag{5}$$

where $f_{\max}^k$ and $f_{\text{avg}}^k$ ($k \in \{1, \ldots, K\}$) represent the max-pooled and avg-pooled features for each convolutional kernel $k$, $W$ and $b$ are trainable parameters, and $K$ is the number of kernel sizes.

**Multi-Head Attention** To combine global and local representations effectively, we employ a bidirectional Multi-Head Attention (MHA) mechanism that selectively enhances salient features from both contexts:

$$\text{MHA}(r_g, r_l) = \text{concat}(\text{Att}(r_g W_g^Q, r_l W_l^K, r_l W_l^V), \text{Att}(r_l W_l^Q, r_g W_g^K, r_g W_g^V)), \tag{6}$$

where $r_g$ and $r_l$ are the global and local representations, respectively. The matrices $W^Q$, $W^K$, and $W^V$ are used to compute the query, key, and value for the attention mechanism. The MHA mechanism enhances the model's understanding of context and interaction between global and local features.

**Label Prediction** After fusing the global and local features, we apply a linear layer followed by normalization and ReLU activation. The output probability $P_X^c$ for each quality aspect label $c$ indicates the level of attention each comment places on specific quality aspects.

**Loss Function** To effectively address the issue of class imbalance in multi-label classification, we adopt the Class Balanced Focal Loss (CBFL) [11] as the objective function, which combines a class balancing term and a focusing mechanism. Unlike traditional cross-entropy loss, CBFL enhances the model's learning capability for under-represented classes by modulating the loss according to class frequencies, thus focusing more on rare or difficult-to-predict labels. The CBFL loss function is given by:

$$L_{CB} = \begin{cases} -r_{CB}\left(1 - p_i^c\right)^\gamma \log\left(p_i^c\right) & \text{if } y_i^c = 1 \\ -r_{CB}\left(p_i^c\right)^\gamma \log\left(1 - p_i^c\right) & \text{otherwise,} \end{cases}, r_{CB} = \frac{1-\beta}{1-\beta^{n_i}}, \tag{7}$$

where $p_i^c$ is the predicted probability for the $c$-th class of instance $i$, $y_i^c$ is the ground truth label, and $\gamma$ is the focusing parameter. This focusing term reduces the loss contribution of well-classified examples, thereby prioritizing difficult, misclassified samples. The balancing factor $r_{CB}$ adjusts the loss based on the number of samples $n_i$ for each class $i$, which allows CBFL to dynamically modulate the impact of each class in the loss function. The CBFL facilitates a balanced and fine-grained classification across all labels, which is critical for complex, multi-label quality evaluation tasks in scenic areas. Algorithm 2 presents the multi-label classification process.

### 4.3 Overall Scenic Quality Evaluation

To derive a comprehensive and fine-grained quality evaluation for scenic areas, we combine the results from the Reference-based Sentiment Analysis and the Multi-label Classification modules. The RBSA module provides the sentiment score $Score_X^c$ for each quality aspect $c$, while the GLEE module supplies the relevance probability $Prob_X^c$, indicating the degree of focus on each quality aspect

**Algorithm 2** Multi-label Classification with GLEE

**Input:** Comment $X$, pre-trained language model (PLM), convolutional kernel sizes $K$, attention heads $h$, quality aspect labels $\{c_1, c_2, \ldots, c_C\}$
**Output:** Predicted probability $Prob_X^c$ for aspect $c$

**Step 1: Global Encoding** ▷ Extract global features using PLM
Tokenize and pass $X$ through PLM to get hidden states
Compute global feature $r_g \leftarrow \sum_{l=1}^{L} \alpha_l \cdot e_0^l[\texttt{CLS}]$ ▷ Weighted sum of [CLS] token
**Step 2: Local Encoding** ▷ Extract local features using CNN
Permute PLM outputs to match convolution input dimensions
**for** each kernel size $k \in K$ **do**
    Apply convolution filter to get local features
    Perform max-pooling and avg-pooling on each feature map
**end for**
Concatenate pooled features from each kernel to form $f_{\text{concat}}$
Compute local feature $r_l \leftarrow \text{ReLU}(W \cdot f_{\text{concat}} + b)$
**Step 3: Multi-Head Attention Mechanism** ▷ Enhance context understanding
Apply multi-head attention to combine $r_g$ and $r_l$
Compute fused feature $r_{\text{MHA}} \leftarrow \text{MHA}(r_g, r_l)$
**Step 4: Label Prediction** ▷ Multi-label probabilities
Apply a linear layer to $r_{\text{MHA}}$, followed by ReLU activation
Regularize with a dropout layer
Output label probabilities $Prob_X^c = \text{sigmoid}(\text{Linear}(r_{\text{MHA}}))$
**return** $Prob_X^c$ for each aspect label $c$

in the comment. Therefore, we can get the final quality evaluation score for each quality aspect $Output_c$, *i.e.*,

$$Output_c = Score_X^c \cdot Prob_X^c. \tag{8}$$

## 5 Experiments and Analysis

### 5.1 Experimental Settings

**Baseline Selection and Experimental Design** The choice of baselines in our experiments ensures a rigorous benchmark by comparing our approach to state-of-the-art methods in both sentiment analysis and multi-label classification tasks, thus validating the effectiveness of our model for scenic area quality evaluation. Given PLMs' notable advancements in downstream tasks, our experimental setup emphasizes these models, which excel in capturing complex linguistic features. However, for sentiment analysis, we also include traditional models, given their established effectiveness in sentiment tasks.
**Baseline Selection** For Sentiment Analysis, we consider both traditional models (*i.e.*, RNN [30], LSTM [15], BiLSTM [15], and TextCNN [8]) and transformer-based models (*i.e.*, XLNet [37], ERNIE 3.0 [34], RoBERTa [25], and ALBERT [20]). For Multi-label Classification, we choose baseline models including BERT [12], RoBERTa [25], XLNet [37], ALBERT [20], RoFormer_V2 [33], BigBird [38], and BLOOM [31].

Table 6: Comparison of model performance on sentiment analysis.

| Model | Accuracy | Consistency | Interpretability | Granularity | Fleiss' Kappa |
|---|---|---|---|---|---|
| RNN [30] | 6.7 | 7 | 7.5 | 6.5 | 0.2584 |
| LSTM [15] | 6.8 | 7.1 | 7.6 | 6.6 | 0.2319 |
| BiLSTM [15] | 6.5 | 7.2 | 7.4 | 6.2 | 0.2777 |
| TextCNN [8] | 5.9 | 7.3 | 7 | 8.5 | 0.4591 |
| XLNet [37] | 7.3 | 7.4 | 8.5 | 7.1 | 0.3158 |
| ERNIE 3.0 [34] | 7.5 | 7.5 | 8 | 7.8 | 0.2691 |
| RoBERTa [25] | 8.4 | 6.7 | 9 | 7.8 | 0.3594 |
| ALBERT[20] | 8.6 | 7.4 | 9.5 | 8.9 | 0.2885 |
| Ours (RBSA) | 9.3+.7 | 8.7+1.3 | 9.7+.2 | 9.4+.5 | 0.3787 |

The value of Fleiss' Kappa represents *fair agreement* when it is during $0.21 \sim 0.40$; *moderate agreement* during $0.41 \sim 0.60$.

Table 7: Comparison of multi-label classification performance.

| Model | Kernel Size | F1(macro) | F1(micro) | F1(weighted) | Jaccard | Accuracy | Hamming Loss |
|---|---|---|---|---|---|---|---|
| BERT (bert-base-uncased) [12] | - | .9036 | .9498 | .9497 | .9052 | .8167 | .0372 |
| BERT (hfl/chinese-bert-wwm) [10; 12] | - | .9342 | .9723 | .9723 | .9466 | .8786 | .0204 |
| BERT (hfl/chinese-bert-wwm-ext) [10; 12] | - | .9324 | .9720 | .9719 | .9458 | .8776 | .0206 |
| RoBERTa (roberta-base) [25] | - | .9104 | .9553 | .9552 | .9154 | .8220 | .0331 |
| XLNet (xlnet-base-cased) [37] | - | .5882 | .8342 | .8300 | .7179 | .4220 | .1239 |
| XLNet (hfl/chinese-xlnet-base) [10; 37] | - | .9288 | .9681 | .9681 | .9386 | .8656 | .0234 |
| ALBERT (albert-base-v2) [20] | - | .6926 | .8764 | .8747 | .7813 | .6086 | .0909 |
| ALBERT (uer/albert-base-chinese-cluecorpussmall) [20; 40] | - | .9378 | .9713 | .9712 | .9445 | .8749 | .0217 |
| BLOOM (bigscience/bloom-560m) [5; 31] | - | .9116 | .9581 | .9578 | .9200 | .8323 | .0309 |
| RoBERTa (hfl/chinese-roberta-wwm-ext)[10] (Global) | - | .9354 | .9728 | .9727 | .9474 | .8789 | .0201 |
| RoBERTa (Local) | k=2,3 | .9351 | .9721 | .9720 | .9460 | .8757 | .0206 |
| Ours (RoBERTa+GLEE) | k=2, 3 | .9368+.0014 | .9732+.0004 | .9731+.0004 | .9480+.0006 | .8794+.0005 | .0199+.0002 |
| Ours (RoBERTa+GLEE+MHA) | k=1, 2 | .9377+.0023 | .9734+.0006 | .9733+.0006 | .9484+.0010 | .8816+.0027 | .0197+.0004 |
| RoFormer_V2 (junnyu/roformer_v2_chinese_char_base)[33] (Global) | - | .9305 | .9718 | .9717 | .9454 | .8785 | .0208 |
| RoFormer_V2 (Local) | k=2,3 | .9324 | .9717 | .9716 | .9451 | .8795 | .0208 |
| Ours (RoFormer_V2+GLEE) | k=2, 3 | .9372+.0048 | .9722+.0004 | .9721+.0004 | .9461+.0007 | .8796+.0001 | .0205+.0003 |
| Ours (RoFormer_V2+GLEE+MHA) | k=1, 3 | .9326+.0002 | .9725+.0007 | .9724+.0007 | .9467+.0013 | .8798+.0003 | .0203+.0005 |
| Bigbird (Lowin/chinese-bigbird-base-4096)[26] [38](Global) | - | .9284 | .9724 | .9724 | .9466 | .8785 | .0204 |
| Bigbird (Local) | k=2,3 | .9323 | .9723 | .9723 | .9466 | .8780 | .0205 |
| Ours (Bigbird+GLEE) | k=2, 3 | .9367+.0083 | .9733+.0009 | .9733+.0009 | .9483+.0017 | .8798+.0013 | .0198+.0006 |
| Ours (Bigbird+GLEE+MHA) | k=2, 3 | .9327+.0043 | .9733+.0009 | .9733+.0009 | .9483+.0017 | .8816+.0031 | .0198+.0006 |

**Evaluation metrics** For Sentiment Analysis, we employed 30 evaluators to measure the performance of different models on accuracy, consistency, interpretability, and granularity. The score ranges from 1 to 10, with higher scores indicating that the evaluator better agrees with the model's sentiment scores. Besides, we applied Fleiss' Kappa to measure the consistency of different evaluators. For Multi-Label Classification, the evaluation metrics include F1 Score (Macro, Micro, and Weighted), Jaccard Index, Accuracy, and Hamming Loss.

**Training Details** We conducted experiments in a Linux environment with eight NVIDIA GeForce RTX 4090 GPUs, CUDA version 11.6, and PyTorch version 2.0.1+cu117. The dataset was divided into training, validation, and testing sets with an $8:1:1$ ratio. We trained models for 30 epochs with a batch size of 4 and the learning rate of $2 \times 10^{-5}$. The value of $\delta$ and $\omega$ is fixed as 0.5.

### 5.2 Experimental Results

**Sentiment Analysis Performance** As shown in Table 6, ALBERT achieved the highest accuracy (8.6), interpretability (9.5) and granularity (8.9) score among baselines. The RBSA method outperformed all baselines, achieving top scores in all metrics, thus setting new benchmarks. Additionally, RBSA reached a Fleiss' Kappa of 0.3787, indicating fair agreement with human evaluators. Fig.

Table 8: Ablation studies of multi-label classification on MSAQE dataset.

| Model (+GLEE) | MHA | Kernel Size (KS) | F1(macro) | F1(micro) | F1(weighted) | Jaccard | Accuracy | Hamming Loss |
|---|---|---|---|---|---|---|---|---|
| RoBERTa | × | 2,3 | .93676 | .97315 | .97310 | .94795 | .87937 | .01988 |
| | ✓ | 2,3 | .92851 | .97312 | .97308 | .94795 | .88030 | .01992 |
| | ✓ | 1,2 | .93765 | **.97338** | **.97332** | **.94837** | **.88160** | **.01970** |
| | ✓ | 1,3 | **.93956** | .97299 | .97298 | .94772 | .88088 | .02006 |
| | ✓ | 1,4 | .93299 | .97276 | .97273 | .94727 | .87834 | .02023 |
| | ✓ | 1,5 | .92038 | .97139 | .97131 | .94464 | .87365 | .02120 |
| | ✓ | 4,5 | .93586 | .97268 | .97264 | .94707 | .87961 | .02024 |
| | ✓ | 2,3,4,5 | .91792 | .97092 | .97093 | .94392 | .87265 | .02158 |
| RoFormer_V2 | × | 2,3 | **.93717** | .97217 | .97207 | .94606 | .87964 | .02057 |
| | ✓ | 2,3 | .93236 | .97241 | .97237 | .94662 | .87978 | .02043 |
| | ✓ | 1,2 | .93357 | .97213 | .97206 | .94605 | .87954 | .02058 |
| | ✓ | 1,3 | .93255 | **.97248** | **.97239** | **.94666** | .87978 | **.02034** |
| | ✓ | 1,4 | .92717 | .97224 | .97216 | .94623 | **.88078** | .02046 |
| | ✓ | 1,5 | .93554 | .97167 | .97159 | .94514 | .87666 | .02101 |
| | ✓ | 4,5 | .92922 | .97065 | .97054 | .94320 | .87481 | .02158 |
| | ✓ | 2,3,4,5 | .93370 | .97104 | .97095 | .94393 | .87429 | .02140 |
| Bigbird | × | 2,3 | .93667 | .97332 | .97327 | .94829 | .87975 | .01977 |
| | ✓ | 2,3 | .93274 | **.97333** | **.97330** | **.94834** | **.88156** | **.01976** |
| | ✓ | 1,2 | .93665 | .97286 | .97286 | .94752 | .87940 | .02014 |
| | ✓ | 1,3 | .92033 | .97217 | .97219 | .94635 | .87560 | .02065 |
| | ✓ | 1,4 | .93184 | .97248 | .97245 | .94674 | .87790 | .02039 |
| | ✓ | 1,5 | .93660 | .97275 | .97274 | .94725 | .87940 | .02021 |
| | ✓ | 4,5 | **.93742** | .97227 | .97224 | .94632 | .87731 | .02052 |
| | ✓ | 2,3,4,5 | .93561 | .97299 | .97295 | .94769 | .87975 | .02004 |

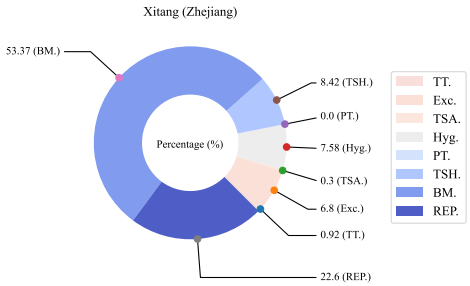


Fig. 5: Scenic areas quality evaluation label attention level of Xitang in Zhejiang.

6 illustrates the average sentiment scores across quality aspects for Xitang in Zhejiang, validating RBSA's effectiveness in aligning with human evaluations. This detailed sentiment analysis provides an in-depth understanding of how specific quality aspects are perceived by visitors. Notably, 'BM', 'TSH' and 'REP' received relatively higher average sentiment scores, suggesting that these aspects are well-regarded by visitors. In contrast, 'PT' 'TSA' and 'TT' had lower sentiment scores, indicating areas where visitor satisfaction may be improved. This distribution underscores the capacity of the RBSA to provide clear insight into visitor perceptions, effectively identifying both strengths and areas for potential enhancement in scenic quality.

**Multi-label Classification Performance** As shown in Table 7, the integration of GLEE and MHA yields significant overall performance improvements, with MHA enhancing contextual feature capture, which supports precise predictions on scenic quality labels. Table 8 displays the impact of kernel sizes on perfor-

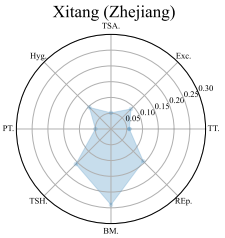


Fig. 6: The average sentiment score of overall scenic areas quality evaluation label on Xitang in Zhejiang.

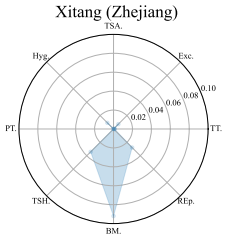


Fig. 7: Overall scenic areas quality evaluation label's average weighted sentiment score of Xitang in Zhejiang.

mance. Optimal configurations, such as KS=1,2 for RoBERTa+GLEE+MHA, KS=1,3 for RoFormer_V2+GLEE+MHA, and KS=2,3 for Bigbird+GLEE+MHA, demonstrated enhanced feature extraction capabilities. Fig. 5 illustrates our refined RoBERTa+GLEE+MHA model in action with KS=1,2, displaying the distribution of label attention for different quality aspects in Xitang, Zhejiang. As seen, 'BM' and 'REP' are the most emphasized aspects of visitors, confirming the model's fine-grained capacity to accurately assess scenic quality aspects.

**Integrated Analysis** We evaluated scenic areas by averaging sentiment scores of all comments per location. Fig. 7 shows Xitang's weighted sentiment scores, combining RBSA sentiment outputs with multi-label classification probabilities. This integration reflects an overlay of sentiment scores and attention levels for each quality dimension, which enables holistic quality assessment, highlighting its potential for multi-dimensional quality improvements. As shown, aspects like 'BM,' 'TSH,' and 'REP' exhibit stronger sentiment, suggesting stronger visitor focus, though overall neutral scores (0-0.1) reflect balanced positive/negative evaluations. A more in-depth analysis could focus on comments with higher absolute sentiment values, revealing specific areas for improvement more objectively by highlighting comments representing positive or negative aspects distinctly.

## 6 Conclusion

In this work, we introduce the MSAQE dataset, a large-scale, fine-grained dataset for scenic area quality evaluation, marking a pioneering effort in data-driven analysis of tourism services. Our data-driven integrated framework combines Reference-Based Sentiment Analysis (RBSA) with Global and Local Ensemble Encoding (GLEE) for multi-label classification, effectively capturing both sentiment and quality dimensions. Extensive experiments demonstrate that the proposed approach enhances feature extraction and achieves robust performance across scenic area quality evaluation tasks, setting new benchmarks in this field.

**Acknowledgments.** This study was funded by the Natural Key Research and Development Program of China (grant number 2023YFF0905000).

# Bibliography


[1] Standard of rating for quality tourist attractions. GB/T 17775-2003 (May 2003), [Chinese National Standard (Recommended)]

[2] Alam, M.H., Ryu, W.J., Lee, S.: Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Information Sciences **339**, 206–223 (2016)

[3] Angelidis, S., Amplayo, R.K., Suhara, Y., Wang, X., Lapata, M.: Extractive opinion summarization in quantized transformer spaces. Transactions of the Association for Computational Linguistics **9**, 277–293 (2021)

[4] Arenas-Márquez, F.J., Martinez-Torres, R., Toral, S.: Convolutional neural encoding of online reviews for the identification of travel group type topics on tripadvisor. Information Processing & Management **58**(5), 102645 (2021)

[5] BigScience: Bigscience language open-science open-access multilingual (bloom) language model (2022), international, May 2021-May 2022

[6] Bjerva, J., Bhutani, N., Golahn, B., Tan, W.C., Augenstein, I.: Subjqa: A dataset for subjectivity and review comprehension. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (November 2020)

[7] Chaki, A.: Tripadvisor reviews 2023. https://www.kaggle.com/datasets/arnabchaki/tripadvisor-reviews-2023 (2023), accessed: yyyy-mm-dd

[8] Chen, Y.: Convolutional neural network for sentence classification. Master's thesis, University of Waterloo (2015)

[9] Colladon, A.F., Guardabascio, B., Innarella, R.: Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. Decision Support Systems **123**, 113075 (2019)

[10] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for Chinese natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. pp. 657–668. Association for Computational Linguistics, Online (Nov 2020), https://www.aclweb.org/anthology/2020.findings-emnlp.58

[11] Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)

[12] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. **abs/1810.04805** (2019)

[13] Grail, Q., Perez, J.: Reviewqa: a relational aspect-based opinion reading dataset. arXiv preprint arXiv:1810.12196 (2018)

[14] HnestDs: Hotel reviews and listing. https://www.kaggle.com/hamzafarooq50/hotel-listings-and-reviews (2020), [Dataset]

[15] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[16] Hou, J.: Service quality evaluation algorithm of tourist scenic spots based on fuzzy cluster analysis. In: 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC). pp. 1–6. IEEE (2022)
[17] Hou, Z., Cui, F., Meng, Y., Lian, T., Yu, C.: Opinion mining from online travel reviews: A comparative analysis of chinese major otas using semantic association analysis. Tourism Management **74**, 276–289 (2019)
[18] Kim, S.E., Lee, K.Y., Shin, S.I., Yang, S.B.: Effects of tourism information quality in social media on destination image formation: The case of sina weibo. Information & management **54**(6), 687–702 (2017)
[19] Kirilenko, A.P., Stepchenkova, S.O., Dai, X.: Automated topic modeling of tourist reviews: Does the anna karenina principle apply? Tourism Management **83**, 104241 (2021)
[20] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=H1eA7AEtvS
[21] Lee, H.C., Pan, H.L., Chung, C.C.: The study of destination image, service quality, satisfaction and behavioral intention–an example of dapeng bay national scenic area. International Journal of Organizational Innovation (Online) **11**(3), 25 (2019)
[22] Li, F.j., Liao, X., Liu, J.m., Jiang, L.l., Wang, M.d., Liu, J.f.: Investigating the tourism image of mountain scenic spots in china through the lens of tourist perception. Journal of Mountain Science **20**(8), 2298–2314 (2023)
[23] Li, Z., Xiong, G., Wei, Z., Zhang, Y., Zheng, M., Liu, X., Tarkoma, S., Huang, M., Lv, Y., Wu, C.: Trip purposes mining from mobile signaling data. IEEE Transactions on Intelligent Transportation Systems **23**(8), 13190–13202 (2021)
[24] Liu, J.: 515k hotel reviews data in europe. https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe/data (2017), accessed: yyyy-mm-dd
[25] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
[26] LowinLi: Lowin: chinese-bigbird-base-4096 (2022)
[27] Olery: Olery - worldwide hospitality data | hospitality, travel & tourism data | destination api | sentiment analysis. https://datarade.ai/data-products/olery-destination-api-worldwide-hospitality-data-olery (2023), [Dataset]
[28] Qi, C., Zhang, Y., Luo, Q.: Research on service quality improvement of agro-tourism integration scenic spots based on bp neural network. In: 2022 4th International Conference on Applied Machine Learning (ICAML). pp. 1–5. IEEE (2022)
[29] Quốc, H.B.: 515k hotel reviews data in europe. https://www.kaggle.com/huypui/data-515k-rating-hotel (2022), [Dataset]

[30] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)
[31] Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
[32] Shakibaei, S., De Jong, G.C., Alpkökin, P., Rashidi, T.H.: Impact of the covid-19 pandemic on travel behavior in istanbul: A panel data analysis. Sustainable cities and society **65**, 102619 (2021)
[33] Su, J., Pan, S., Wen, B., Liu, Y.: Roformerv2: A faster and better roformer - zhuiyiai. Tech. rep. (2022), https://github.com/ZhuiyiTechnology/roformer-v2
[34] Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al.: Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137 (2021)
[35] Velldal, E., Øvrelid, L., Bergem, E.A., Stadsnes, C., Touileb, S., Jørgensen, F.: NoReC: The Norwegian Review Corpus. In: Proceedings of the 11th edition of the Language Resources and Evaluation Conference. pp. 4186–4191. Miyazaki, Japan (2018)
[36] Wu, H., Dong, H.: Research on service quality improvement of tianmu lake scenic spot based on servqual model. In: 2022 2nd International Conference on Management Science and Software Engineering (ICMSSE 2022). pp. 608–612. Atlantis Press (2022)
[37] Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 5754–5764 (2019)
[38] Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al.: Big bird: Transformers for longer sequences. Advances in neural information processing systems **33**, 17283–17297 (2020)
[39] Zhang, L., Xu, J., Gong, Y., Yu, L., Zhang, J., Shen, J.: Unsupervised image and text fusion for travel information enhancement. IEEE Transactions on Multimedia **24**, 1415–1425 (2021)
[40] Zhao, Z., Li, Y., Hou, C., Zhao, J., et al.: Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. ACL 2023 p. 217 (2023)