Geographical Science Research 地理科学研究, 2024, 13(3), 548-562
Published Online June 2024 in Hans. https://www.hanspub.org/journal/gser
https://doi.org/10.12677/gser.2024.133052

Hans汉斯

基于UGC数据的广西桂林旅游目的地评价研究

孙杰华*，王文芝，张　美，汪宏平，常书琪，杜京泽

桂林旅游学院旅游数据学院，广西 桂林

收稿日期：2024年5月21日；录用日期：2024年6月20日；发布日期：2024年6月28日

摘　要

随着网络信息技术的迅速发展与应用，更多旅游者使用网络资源平台分享旅行目的地相关资讯，而旅客利用攻略、景区评价等手段所得到的海量UGC数据，能够更有效反映其对旅行目的地的感知与评价。本文以广西桂林漓江、龙脊梯田、象鼻山、银子岩及遇龙河漂流五个景区为研究对象，运用Python、八爪鱼软件采集携程旅游平台UGC文本评论及游客的旅游信息用于构建旅游数据集并进行清洗处理。通过TF-IDF、Word2Vec及词云图等技术方法，结合ERNIE-BiLSTM-DPCNN模型，从语义特征关联分析、LDA主题模型特征分析以及语义情感词典分析研究影响广西旅游目的地游客的情感倾向的主要因素。结果表明，游客对广西桂林五个景区的整体满意度较高。对于遇龙河等消极情感比例较高的景区，相关部门应针对性地解决景区问题并定期检查改进措施的效果，以提升游客情感得分，进而推动广西桂林发展。最后，根据其中存在的问题对景区进行进一步管理、完善景区服务体系和提升游客满意度体验提出建议，为奋力打造世界级旅游城市提供理论基础与实证研究。

关键词

UGC数据，广西旅游目的地，文本评论，情感倾向

A Study on the Evaluation of Tourism Destinations in Guilin, Guangxi Province Based on UGC Data

Jiehua Sun*, Wenzhi Wang, Mei Zhang, Hongping Wang, Shuqi Chang, Jingze Du

School of Tourism Data, Guilin University of Tourism, Guilin Guangxi

Received: May 21st, 2024; accepted: Jun. 20th, 2024; published: Jun. 28th, 2024

Abstract

With the rapid development and application of network information technology, more tourists

*通讯作者。

文章引用: 孙杰华, 王文芝, 张美, 汪宏平, 常书琪, 杜京泽. 基于UGC数据的广西桂林旅游目的地评价研究[J]. 地理科学研究, 2024, 13(3): 548-562. DOI: 10.12677/gser.2024.133052

use network resource platforms to share information related to travel destinations, and the massive UGC data obtained by tourists using strategies and scenic spot evaluations can more effectively reflect their perception and evaluation of travel destinations. Taking five scenic spots of Lijiang River, Longji Terraces, Elephant Trunk Mountain, Yinziyan and Yulong River rafting in Guilin, Guangxi Province as the research objects, this paper used Python and Octopus software to collect UGC text comments and tourists' travel information on Ctrip's travel platform for construction of tourism datasets and cleaning and preprocessing. Based on TF-IDF, Word2Vec and word cloud diagrams, combined with the ERNIE-BiLSTM-DPCNN model, the main factors influencing the emotional tendency of tourists in Guangxi tourist destinations were studied from the perspective of semantic feature association analysis, LDA topic model feature analysis and semantic sentiment dictionary analysis. The results show that the overall satisfaction of tourists with the five scenic spots in Guilin, Guangxi Province is high. Finally, according to the existing problems, this paper puts forward suggestions for further management of scenic spots, improvement of scenic service system and improvement of tourist satisfaction experience, so as to provide a theoretical basis and empirical research for striving to build a world-class tourist city.

Keywords

UGC Data, Guangxi Tourist Destinations, Text Reviews, Emotional Tendencies

1. 引言

近年来，随着互联网的飞速发展，旅游信息的获取和分享方式发生了革命性的变化。旅游者在各大在线平台上积极分享攻略、游记、评价和照片等方式留下的游览足迹生成的海量数据称为 UGC (User Generated Content)数据[1] [2]。它能有效反映其对旅游目的地的整体感知与评价。

研究表明，游客在选择旅游产品时，往往会参考网络上的 UGC 数据，了解他人的经验和建议。因此，对 UGC 数据的深入研究，不仅有助于理解游客对旅游目的地的整体感知和评价，还能从中抽取出有价值的信息，为景区的设施完善和服务提升提供指导。

在旅游目的地评价领域，国外学者对城市、乡村、遗产地等旅游目的地形象进行了深入探讨[3] [4]，国内学者们则从不同角度对旅游目的地评价进行了大量的研究，池丽平[5]等基于 UGC 数据以平潭国际旅游岛为案例地，通过文本分析法剖析了国内旅游者对平潭的感知形象；张公让[6]等基于评论数据进行文本语义挖掘与情感分析，研究了游客对景区的评价，并提出了景区质量评价指标体系和游客情感倾向理论模型；王振璇[7]等利用 UGC 数据对游客行为进行分析，并评估旅游目的地的吸引力；侯成[8]通过分析 UGC 数据，研究了旅游目的地的形象感知，并探讨了如何提升其旅游吸引力。然而，尽管研究方法多样，但多数研究仍侧重于定性分析，缺乏从定量角度对游客感知进行全面评价的研究。为此，本研究将采用定性与定量相结合的方法，对游客在携程网上对桂林景区五个代表性景点的评论数据进行深入分析。

具体而言，我们将利用 Python 语言爬取相关评论数据，并运用 TF-IDF、词云图、语义特征关联分析、LDA 主题模型特征分析以及语义情感词典等方法[9]，挖掘文本数据中的有价值信息，探究游客行为

背后的动机和影响因素。通过这种方法，能够更准确地把握游客对广西旅游目的地的情感倾向和真实体验情况，识别影响游客评价和满意度的关键因素以及景区间的关联性。

基于上述研究，我们将提出一系列针对桂林景区优化和发展的策略和建议，以期提升游客满意度，促进广西旅游业的持续发展。这一研究不仅具有重要的理论价值，而且对广西桂林旅游业的实践发展具有深远的指导意义。

2. 游客评论的特征分析

2.1. 文本数据的选择与预处理

本研究聚焦于桂林漓江风景区、龙脊梯田景区、象鼻山景区、银子岩景区和遇龙河漂流景区这五个具有代表性的旅游目的地。利用携程 APP 作为数据源，收集了这五个景区的游客评论，旨在通过深入分析游客的反馈，挖掘出游客关注的核心点来更好的来完成本研究。

针对这五个桂林的热门景点评论数据，为减少杂乱信息对实验工作的影响，对所挖掘的数据进行了如下操作：去重、删除较短评论、分词、去除停用词以及词性标注。将处理好的数据格式化后，存入新的文件中，为下一步分析做准备。进而使用 Jieba 分词工具对数据集进行分词操作，同时进行词性标注和去除停用词的处理。这一步的目的是为后续的数据处理做好准备工作。通过分词、词性标注和去除停用词，可以将原始数据集进行有效地切分，并去除对分析没有意义的常用词汇，从而更好地进行后续的数据处理和分析。本文共获取了 14,821 条评论，其中无效评论有 2695 条，有效评论有 12,126 条，即 14,126 条有效评论作为语料库，具体评论数目如表 1 所示。

在数据分析阶段，借助 SPSSPRO 软件，对情感得分、好评率、总评分值和景点评论率这四个关键指标进行了详细的描述性统计分析。

由表 2 可以看出，数据中无异常值出现，可直接针对平均值进行描述分析。景区各项指标的标准差与方差较小，数据变化范围小，说明各景区在情感得分、好评率、总评分值、景点评论率方面数据波动程度低，差异小，且情感得分和总评分值的最小值与最大值之间的差异不大，未超过 1，这说明游客对每个景点的评价差异性小，各景点均未出现严重不足，但最大值并未达到 9.0，说明这些指标存在一定上限，各景区在服务质量、景观吸引力等方面均存在进一步提升的空间。

在实际应用中，评价旅游目的地分布要基于一个标准，因此，将各景区的特点进行量化，本节基于各景区已有的特征以及对应的数据，构建旅游评价综合指标，对景区进行综合评价，给每个景区赋予得分。统计学中的综合评价即基于研究对象已有的性质，确定一套量化的评估方法，对研究对象进行评价打分，为了更全面地评价各景区的综合表现，构建了一个旅游评价综合指标体系。通过变异系数法给每个指标赋予相应的权重。相关数据如表 3 所示：

Table 1. Comments
表 1. 各景点评论量

景点	评论数	有效评论	有效占比
漓江风景区	2990	2565	85.79%
龙脊梯田景区	2980	2318	77.79%
象鼻山景区	2881	2298	79.76%
银子岩景区	2979	2465	82.75%
遇龙河漂流景区	2991	2480	82.92%

Table 2. Provides descriptive statistics among indicators
表 2. 各指标间的描述性统计

名称	情感得分	好评率	总评分值	景点评论率
最小值	7.8	0.69	4.4	0.18
最大值	8.7	0.74	4.6	0.27
平均值	8.38	0.71	4.48	0.232
标准差	0.356	0.02	0.084	0.034
中位数	8.5	0.7	4.5	0.24
方差	0.127	0	0.007	0.001
25 分位数	8.05	0.695	4.4	0.2
中位数	8.5	0.7	4.5	0.24
75 分位数	8.65	0.73	4.55	0.26

Table 3. Weight determination (coefficient of variation method)
表 3. 权重确定(变异系数法)

名称	均值 95% CI(LL)	均值 95% CI(UL)	IQR	峰度系数	偏度系数	变异系数(CV)
情感得分	8.068	8.692	0.6	1.784	1.385	4.25%
好评率	0.692	0.728	0.035	0.187	0.938	2.82%
总评分值	4.407	4.553	0.15	0.612	0.512	1.87%
景点评论率	0.202	0.262	0.06	0.699	0.845	14.74%

根据变异系数法及表 3 所赋予的权重，可计算各景点综合得分值，即各景点综合得分 = 4.25% * 情感得分 + 2.82% * 好评率 + 1.87% * 总评分值 + 14.74% * 景点评论率，计算结果如表 4 所示。根据每个景点的总得分，按得分降序排列，得出遇龙河漂流景区综合得分 0.511，同时该景区情感得分最高，景区评论数最多；漓江景综合得分 0.507，该景区好评率与综合得分最高；象鼻山景区综合得分 0.497；龙脊梯田景区综合得分 0.494；银子岩景区综合得分 0.459。结果显示，遇龙河漂流景区、漓江景区、象鼻山景区和龙脊梯田景区在游客评价中表现优异，而银子岩景区则相对滞后，需要在提升服务质量、加强宣传推广等方面付出更多努力。

Table 4. Comprehensive scores of scenic spots
表 4. 各景点综合得分

景区名称	综合得分
遇龙河漂流景区	0.511
漓江景区	0.507
象鼻山景区	0.497
龙脊梯田景区	0.494
银子岩景区	0.459

2.2. 文本特征提取分析

桂林景点评论因其口语化和地域特色，富含新颖词汇和情感倾向表达，这些元素在标准情感词典中

往往缺失。然而，情感种子词的精准选择对于构建领域情感词典至关重要，它直接关系到基于该词典的情感分析准确性。

本文基于桂林景点评论数据集，首先进行了详尽的数据预处理，随后运用 TF-IDF 算法来评估词汇的重要性。该算法结合了词频与逆文档频率两个维度，从而精确地量化了词汇在特定文本及整个文本集合中的权重。TF-IDF 算法在信息检索和文本分析领域有着广泛的应用，通过它筛选出 TF-IDF 值最高的前 250 个词汇，并经过人工筛选，确定了各 50 个具有显著正面和负面情感倾向的种子词。表 5 列出了部分选定的种子词及新词。

在获得种子词后，进一步利用 SO-PMI 算法来识别并扩展情感词典。该方法通过计算候选词与种子词之间的相似度及情感分值，结合预设的阈值，筛选出与种子词情感倾向一致的新词。这些新词随后被整合到基础情感词典中，从而丰富了词典的情感表达。

值得注意的是，在筛选过程中，发现某些词汇在不同语境下可能具有不同的情感倾向。例如，“黯然失色”一词在常规语境中通常带有负面情感，但在桂林景点评论中，当它被用来形容其他景点无法与某处景点相比时，却传达了正面的赞美意味。这种一词多义的现象在构建领域情感词典时尤为重要，它提醒我们在加入新词时需充分考虑其在具体领域的特定含义。

综上所述，通过结合 TF-IDF 和 SO-PMI 算法，成功地构建了一个针对桂林景点评论的专用情感词典，它不仅包含了传统情感词典中的基础词汇，还融入了具有地域特色和领域特性的新词，从而显著提高了对桂林景点评论情感分析的准确性。

Table 5. Positive and negative emotion seed words (part) and new words (part)
表 5. 正负情感种子词(部分)及新词(部分)

情感倾向	部分种子词
正向	不虚此行、超赞、方便、服务周到、鬼斧神工、奇观、气势磅礴、热情、震撼、值得、永放光芒、圣地、好视角 新词：很棒、详细、有趣、丰富、强烈推荐、热心、专业、准时、满意、一流、汹涌澎湃、耐心、熠熠生辉、名胜、媲美
负向	堵车、后悔、不好、没什么、强制、太差、太贵、一般、不会、可惜、不值、不让、不咋地、愤怒、生硬 新词：没意思、乱、坑、不划算、混乱、黑、极差、失望、垃圾、冷漠、过分、天壤之别、离谱、没看头

2.3. 基于词云图的特征可视化分析

本文运用词云图技术对广西桂林漓江、龙脊梯田、象鼻山、银子岩及遇龙河漂流五个景区的评论数据进行了特征可视化分析。整个分析过程包含以下四个关键步骤：

数据清洗和预处理：首先，从携程网上爬取了关于这五个景区的评论数据，并进行了细致的数据清洗和预处理工作。为提高数据处理效率和减少噪声，使用 jieba 分词库对评论数据进行分词处理，移除了停用词和标点符号，同时进行了词干提取和词形还原等操作，以凸显评论中的关键信息，为后续分析提供高质量的数据基础。

计算词频或权重：在数据预处理完成后，采用了词频统计和 Word2Vec 模型来计算每个词语在文本中的出现频率或权重。词频统计直观地反映了词语在文本中的出现次数，而 Word2Vec 则通过训练模型捕捉了词语之间的语义关系，为后续的词云图生成提供了更加丰富的信息。

生成词云图：基于计算得到的词频或权重，利用 Python 中的 wordcloud 库将关键词转化为词云图中的大小和颜色，从而实现了评论数据的可视化展示。词云图中的关键词大小和颜色深浅直观地反映了它们在文本中的重要性和频率。通过比较不同景区的词云图，可以观察到游客对各个景区的不同关注点和

情感态度。

解读词云图：根据如图 1 的展示结果(顺时针方向分别为漓江、银子岩、象鼻山、遇龙河漂流、龙脊梯田景区的词云图展示)，在解读词云图时，发现“桂林”、“漓江”、“拍照”、“梯田”、“景区”、“值得”等词语在多个景区的词云图中均占据了显著位置，这些词语反映了游客对景区的主要关注点和积极评价。同时，还发现“最美”、“值得”、“不错”、“壮观”、“特别”等积极情感词汇在词云图中频繁出现，进一步印证了游客对旅游目的地服务的肯定。

同时，根据词云图识别出影响游客满意度的主要因素包括风景、体验感、旅游设施和时间。这些因素为景区管理和服务提升提供了重要参考。

基于词云图的特征可视化分析可以从文本数据中提取关键词特征，并以直观的方式展示，帮助更好地理解和解读文本数据。尽管词云图在特征可视化分析中具有直观易懂的优点，但它也存在一定的局限性。例如，词云图无法展示词语之间的语义关系和上下文信息，可能导致对某些复杂情感的误解。因此，在进行特征可视化分析时，需要结合其他分析方法和工具进行综合判断。未来，可以考虑引入基于语义网络的特征关联分析等方法，以更全面地挖掘文本数据中的信息。

Figure 1. Word cloud map of various scenic spots
图 1. 各景点词云图

2.4. 基于语义网络的特征关联分析

特征关联分析(Feature Association Analysis)是一种深入剖析数据以揭示特征之间潜在关联关系的有力工具。在旅游领域的分析中，这种方法尤其有效，能够帮助理解游客的行为偏好、兴趣焦点以及对旅

游目的地的评价。

本文基于词频分析，构建了一个以词为节点的语义网络，词与词之间的语义关系为连线，形成了一个直观、可视化的网络分布图。如图 2 所示，可以清晰地看到各个词语之间的连接强度和关联模式。此外，在 Ucinet 中引入了共词矩阵，用于进一步分析词语之间的共现关系，并识别出网络中的核心－边缘结构。经过分析，发现“十里画廊”、“路上”、“坐船”、“免费”等词语频繁出现，且与其他词语的共现频次较高，因此它们位于网络的核心区，成为中心节点。显示了游客在游览过程中对十里画廊的特别关注，在一定程度上，十里画廊能很好地吸引游客前来旅游，景区管理者可以进一步加强对十里画廊景区的建设和宣传，致力打造国际化旅游城市知名的旅游目的地，提升十里画廊自身的吸引力。除了上述中心节点外，网络中还存在一些重要的“桥接”节点，如“上山”、“龙胜”、“到达”、“特色”、“旅游”、“桂林市”、“喀斯特”等核心特征词，也被称为语义网络中的“桥”，为各高频词搭建链接，使得整个语义网络互联互通。这些核心特征词也体现了人们所关注、关心和重视的核心评论因素，在旅游环卫方面，还需要继续增加视觉吸引物，美化环境的同时，保持桂林山水“原滋原味”的魅力。与此同时，还可以注意到“特色”、“舒服”、“开心”、“甲天下”和“壮观”等一系列积极的特征词频繁出现。这些词语不仅体现了游客对旅游目的地的正面评价，也揭示了游客对自然风光、景观特色以及旅游体验的高期待。特别是“大自然”和“免费”之间的共现关系，表明游客对自然风光的热爱以及对免费资源的青睐。

结合语义网络的特征关联分析，可以为桂林的热门旅游目的地景区提出一些优化建议。首先，景区管理者应关注游客在“十里画廊”、“路上”、“坐船”、“免费”等中心词背后的需求，分析游客评论背后的原因。桂林的某些目的地，特别是阳朔境内的旅游空间有限，特别是节假日，阳朔接待旅游人数激增，人流车流混为一谈，拥堵显现特别频繁，会严重影响游客的体验感和满意度。将游客花费在路程上的时间减少，可以利用引进智慧景区管理，通过实施预测交通问题，减少交通拥堵的情况，以及景区分流，避免出现景区人员密集为游客带来极差的旅游体验；将开发免费的“坐船”项目，可以极大的吸引游客前来，在满足游客要求的前提下，将在免费的坐船游漓江项目中增强娱购体验如：特色餐饮服务、趣味体验、以

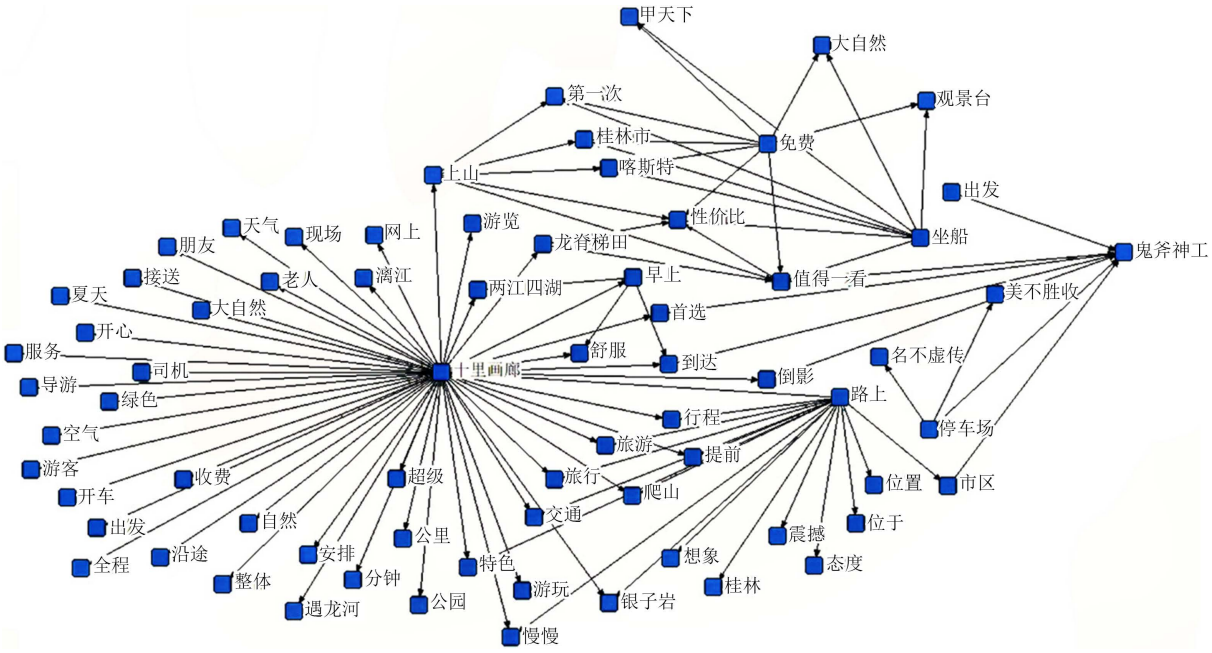

Figure 2. Semantic network diagram of co-occurrence of comment data for various scenic spots
图 2. 各景点评论数据共现语义网络图

及拍照打卡等特色服务，在免费的前提下保障有一定范围内的收益。可见，提升这些方面的服务质量和游客体验，可以极大程度的增加游客满意度。其次，对于“行程”、“服务”、“位置”、“停车场”、“观景台”等游客关注的方面，景区应该利用桂林自身的特色和优势，因地制宜，不要随意简单粗暴的克制模仿，打破“封闭式”旅游经济，用互联网+带动“开放式”地全域旅游发展，逐步的、有层次地进行全域旅游的创建。

Table 6. High-frequency words and frequency of review data for each attraction
表 6. 各景点评论数据高频词与频次

龙脊梯田高频词	频次	漓江高频词	频次	银子岩高频词	频次	象鼻山高频词	频次	遇龙河高频词	频次
梯田	3022	漓江	2360	银子	1478	桂林	1193	竹筏	1103
龙脊	1262	桂林	1207	溶洞	1013	象鼻山	1168	漂流	765
景色	765	桂林山水	787	值得	628	景区	863	景色	707
景区	639	甲天下	747	景点	531	象山	788	龙河	597
值得	549	风景	695	钟乳石	494	景点	542	风景	541
金坑	442	阳朔	583	景区	459	公园	452	体验	447
壮观	403	景色	541	景色	410	景色	443	阳朔	443
平安	402	游船	497	大自然	398	漓江	436	师傅	439
缆车	365	山水	484	桂林	374	免费	332	值得	392
风景	351	值得	389	岩洞	330	门票	324	排队	391
桂林	347	竹筏	385	鬼斧神工	324	值得	317	码头	376
时间	345	码头	342	讲解	318	桂林市	240	小时	373
景点	323	导游	290	小时	303	大象	231	小费	340
小时	296	地方	288	洞内	301	地方	222	龙桥	320
地方	283	船上	275	灯光	272	山水	212	漓江	297
观景台	267	景点	268	地方	267	风景	209	时间	276
大寨	263	小时	253	阳朔	255	打卡	196	旧县	268
建议	218	两岸	223	漂亮	246	拍照	179	好玩	261
导游	218	拍照	216	壮观	229	旅游	158	遇龙	236
季节	208	旅游	198	导游	225	携程	154	分钟	230
门票	199	美景	196	景观	202	总体	143	桂林	217
索道	191	游览	182	一去	183	山水	142	河漂流	211
山顶	188	时间	177	喀斯特	177	标志	141	拍照	209
上山	175	服务	171	时间	175	性价比	136	全程	191
山路	175	兴坪	171	缺钱	174	象鼻	133	船夫	178
开车	170	画山	167	很漂亮	172	甲天下	119	建议	166
山上	168	风光	166	拍照	168	地标	118	景点	164
体验	163	九马	158	好看	157	游客	118	山水	161
酒店	161	体验	155	门票	156	预约	117	刺激	159
推荐	141	景区	153	值得一看	139	好玩	115	人工	159

2.5. 基于 LDA 主题模型的特征分析

主题模型是一种用于揭示大量文本数据背后主题和特征的概率模型。Latent Dirichlet Allocation (LDA)作为其中一种广泛应用的模型，能够提供文本数据背后主题结构的深入理解。基于 LDA 的特征分析方法，包括三个关键步骤：数据预处理、LDA 模型训练和特征表示，这些方法共同助力解析和描述文本中的主题与特征之间的关系。

对数据预处理后的龙脊梯田景区、漓江景区、银子岩景区、象鼻山景区、遇龙河景区评论数据(部分)进行了深入的预处理和 LDA 模型分析。首先，提取了高频词，这些词汇揭示了游客对各个景区的关注点和兴趣所在。同时，也分析了这些景区之间的相关性，发现了游客关注问题的共同点和相似性。在提取高频词的过程中，过滤了与文本主题无关的词汇，如语气助词和介词等，以确保分析的准确性和有效性。最终，得出了各景区的高频词汇表，并列出了频次较高的关键词汇，如表 6 所示。

通过对这些高频特征词进行深入分析，可以得出以下结论：游客在旅游过程中非常注重服务的质量和花费的合理性，他们期望获得周到、优质的服务，并对旅游消费保持理性的态度；桂林的自然山水风光是吸引游客的重要因素，游客对桂林独特的自然资源和美景赞不绝口，充分体现出对自然景观的热爱和认可；除了自然风光外，游客对桂林的文化、美食、娱乐等旅游体验也非常重视。他们希望能够在桂林获得全方位的旅游体验，留下难忘的回忆。因此，景区管理者在开发旅游资源时，应充分考虑如何更好地保护自然环境，守好生态的底线、规划红线，合理、有序地开发各类旅游资源。同时，提供更具吸引力的免费或优惠服务，丰富旅游吸引物体系，深度挖掘桂林大自然的生态美，打造丰富的、极具地方特色的民俗文化演艺活动，以吸引更多游客前来体验。

综合以上三个主题和高频特征词的分析，可以体现来桂林旅游的游客普遍注重旅游过程中的服务感受与消费水平、特色自然旅游资源以及旅游体验。这也表明，在提供旅游服务时，景区管理者需要关注游客的这些需求，努力提升服务质量、保护自然资源、丰富旅游体验，以满足游客的期望和需求，从而赢得更高的满意度和口碑，各景区也应加大投入对景区基础设施的建设和改进力度，融合共享，持续发展。按照“数目达标、质量上乘、布局合理、使用免费”的原则，多维度地满足游客的多样化需求，进而提升桂林旅游目的地的形象。

2.6. 各指标间的相关性分析

为了更深入地理解这些关键指标之间的内在关联，利用 Python 中 numpy 库和 pandas 库对原始数据进行了细致的处理和计算。通过 Pearson 相关系数这一统计量，量化了不同特征参数与预测目标之间的相关性。Pearson 相关系数的取值范围在−1 到 1 之间，其中正值表示正相关，负值表示负相关，而 0 则表示无相关。

在数据分析过程中，设定了一个阈值：当 Pearson 系数的绝对值小于 0.2 时，将其视为极弱相关性或无相关性，并在后续建模中排除这些特征参数。这一策略不仅简化了模型的计算过程，减少了计算负担，还提高了模型的预测效率和准确性，使得模型在实际应用中能够更高效地完成预测任务。

图 3 呈现了相关性热力图，它直观地展示了各指标之间的相关程度。从图中可以看出，大部分指标之间呈现出较高的线性相关性。其中，好评率和总评分之间的相关性尤为显著，Pearson 相关系数高达 0.9，显示出极强的正相关关系，即好评率越高，总评分也相应越高。

此外，景点评论率和情感得分也呈现出较高的正相关关系，这表明当游客对某个景点产生积极的情感体验时，他们更有可能在网络上分享自己的观点和感受，从而增加该景点的评论率。这进一步证明了游客的满意度和参与度对景区声誉的重要性。

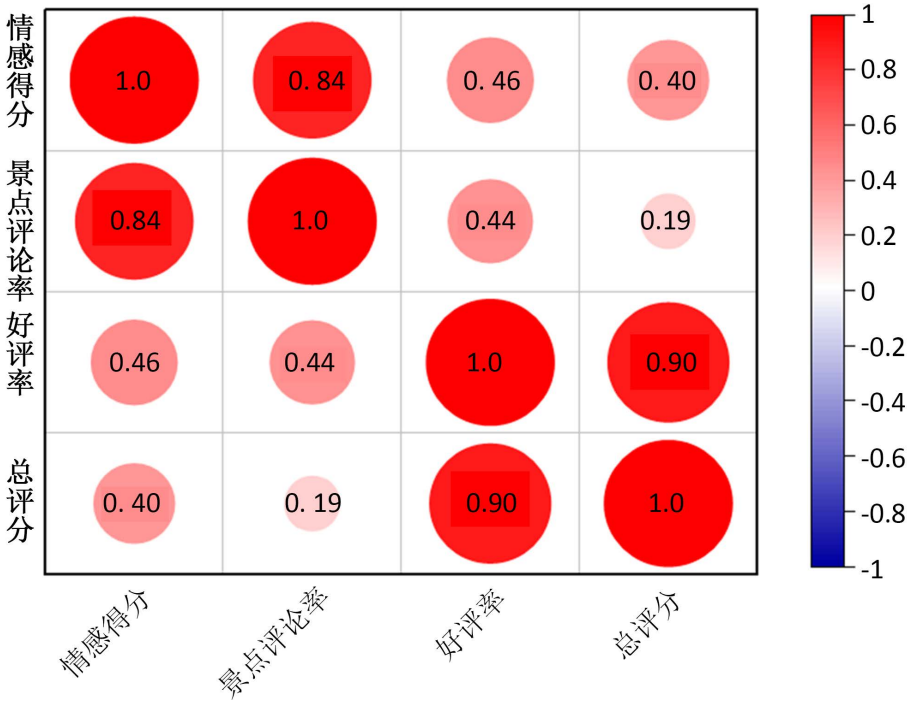

Figure 3. Heat map of correlation between various scenic spots
图 3. 各景点间的相关性热力图

尽管部分指标之间的相关系数也为正数，但数值均小于 0.5，这表示它们之间的相关性相对较弱。这些结果可能提示我们，虽然这些指标对总评分有一定影响，但它们并不是决定性的因素。

特别值得一提的是，漓江景区、龙脊梯田景区和遇龙河漂流景区的各项指标之间同样表现出较强的正相关性，这表明这些景区的整体表现较为一致，无论是从游客的好评率、总评分还是情感得分来看，都呈现出较为稳定的趋势。这为景区的持续发展和优化提供了有力的数据支持。

3. 游客评论的情感倾向分析

3.1. 情感词典的建立

本研究在深入分析广西桂林旅游目的地评价情感现状时，特别注重了桂林特色的情感词汇。桂林作为旅游胜地，其评论中常常出现与山水景色、当地文化等紧密相关的特色词汇。例如，“山水甲天下”这类独特表达，蕴含了深厚的正向情感。此外，游客的评论往往采用口语化、随意的表达方式，这增加了情感分析的复杂性。因此，构建一个专门针对桂林旅游目的地评论的领域情感词典显得尤为重要。

目前市场上已有的情感词典，如 BosonNLP、台湾大学 NTUSD 简体中文情感词典以及知网 Hownet 情感词典等，虽然具有一定的通用性，但难以完全覆盖桂林旅游领域的特色词汇和表达方式。为此，本研究以知网 Hownet 情感词典为基础，结合桂林旅游评论的实际情况，进行了情感词典的自建与扩充。情感词词典如表 7 所示，参考已有的研究和资料，本研究将积极情感词的分值设置为 2 分，消极情感词的则为−2 分。

情感词是主体对某一客体表示内在评价的词语，是情感分析的核心。将情感词典分为积极情感词和消极情感词，但这只是基础的情感判断，对于正确的情感分析来说是远远不够的。拿本研究爬取的桂林景点某评论作为例子来说，原句为“只是携程这个价格太不美丽了”，分词后为“只是/携程/这个/价格/太不/美丽了”，按照上面的算法，如果仅仅考虑“美丽”，原句的情感是正向的，显然这是不对的，主要是由于忽略了情感词“美丽”前面的程度副词“太不”。由于程度副词是对情感词情感极性的加强，

所以最终的情感值可以通过程度副词的程度值(正值表示加强，负值表示否定)乘以情感词的情感值来得到。但是，程度副词的修饰程度分为很多级，不同修饰强度的程度副词对情感词的影响程度也不同。本研究以知网程度副词词典为基础，并根据前人现有资料和本研究实际情况做出改动，将程度副词分成 5 个程度级别并赋予适当的权值，程度副词词典如表 8 所示。

Table 7. Classification and score setting of emotion words
表 7. 情感词分类及分值设置

情感词	情感词(部分)	分值设置
积极	美、好、丰富、难忘、惊叹、有趣、惬意、迷人、独特……	2
消极	恶劣、差、贵、难吃、慢、乱、不好、奇葩、崩溃……	−2

Table 8. Classification of degree adverbs and their weight distribution
表 8. 程度副词级别分类及其权值分配

程度级别	程度副词(部分)	权值大小
最(Most)	极、非常、完全、绝对、万分	2
很(Very)	多、颇、太、特	1.5
较(More)	更、还、愈加、越发、足	1.25
稍(Slightly)	略、挺、相当、有点、一些	0.5
微(insufficiently)	半点、相对、丝毫、弱、轻	0.25

同时，否定词在情感分析中也扮演着重要角色。它们虽无情感倾向，但能显著改变情感的极性。如“不美丽”这个短语中，“不”作为否定词修饰“美丽”。否定词修饰否定词即为双重否定，一般情况下不影响原来的情感倾向，对于情感程度的影响需要具体分析依存关系。如“不是不喜欢”中，第一个“不”作为否定词修饰了“是”，后面的“不喜欢”和“是”是动宾关系。否定词修饰程度词相当于对程度的否定，而非情感的否定，因此否定词修饰后，最终情感倾向未变，但是程度有所降低。如“不很喜欢”中，“不”修饰了“很”，然后才是“很”修饰“喜欢”。“很”作为程度词表现了一种情感等级的加强，但是“不”作为否定词对“很”这个程度词的否定使得最终的强度有一定的削减[10]。“不很喜欢”这个情感短语仍然表现了“喜欢”这种情感，但是在强度上强于“喜欢”这个情感词，弱于“很喜欢”这个情感短语。所以本研究通过查找相关文献和资料，找到了适用于本研究的否定词词典，并对其进行了赋值，否定词词典如表 9 所示。

Table 9. Dictionary of negative words and their scores
表 9. 否定词词典及其分值

否定词(部分)	分值大小
不	−1
没	−1
毫无	−1
非	−1
从未	−1

综上所述，本研究通过构建面向桂林旅游目的地评论的领域情感词典，结合程度副词词典和否定词

词典，为桂林旅游发展的情感分析提供了有力的工具，有助于更准确地把握游客的情感倾向和满意度，为旅游业的优化和发展提供科学依据。

3.2. 情感值分析算法的设计

在为本研究构建适用的情感词词典之后，依据不同词典的赋值标准，设计了一套情感得分的计算方法。该计算方法旨在精确评估每条评论的情感倾向，并参考了相关文献和资料，结合本研究的实际需求，形成了一套完善的情感值算法设计体系。以下是具体的实施步骤：

1) 文本预处理：首先，对广西桂林各景点的评论进行文本分句，利用 jieba 分词工具(同时去除停用词)对分句后的文本进行分词，形成可供后续分析的文本语料。

2) 情感词标记：接着，依据已构建的情感词词典，在文本语料中逐一标记出积极和消极的情感词。

3) 程度副词识别与赋值：随后，根据程度副词词典，以已标记的情感词为中心，查找并识别周围修饰情感词的程度副词。若存在程度副词，则根据赋值表对情感词进行相应程度的赋值调整。

4) 否定词识别与赋值：类似地，利用否定词词典，以情感词为核心，检查并识别周围是否存在否定词。一旦识别到否定词，将根据赋值表对情感词进行反向或调整赋值。

5) 情感符号特征化处理：在文本分析中，特别关注评论中是否包含“？”、“！”、“……”等情感符号。这些符号往往强烈表达着情感倾向的变化，因此，一旦检测到这些符号，将对情感词数目进行一定的增加，以反映这种情感强度的变化。

6) 分句情感值计算：对于每条分句，计算其情感值，具体为积极词值的总和减去消极词值的总和，从而得出该分句的情感倾向。

7) 评论总情感得分汇总：接下来，将评论中所有分句的情感得分相加，得出整条评论的总情感得分。

8) 情感倾向分类：最后，根据总情感得分的正负性，将评论分为三类：总情感得分为正数的评论被视为“好评”；得分为负数的评论被判定为“差评”；而得分为零的评论则归为“中评”。这一分类方法提供了直观了解用户对景点评价情感的途径。

3.3. 结果分析

在探索 ERNIE-BiLSTM-DPCNN 模型在情感分析任务中的有效性时，本研究针对桂林景点在线评论构建了一个专门的情感词典，并基于该词典构建了一个高质量的数据集。该数据集规模庞大，包含 15,496 条评论文本，覆盖了桂林 11 个不同的景点，并被精确标注为三个主要情感类别：负向情感评论、正向情感评论和中性情感评论。

为确保实验结果的准确性和可靠性，本研究采用了严格的数据处理和模型训练流程。首先，对评论文本进行了预处理，包括去除无关字符、停用词过滤以及词性标注等步骤，以减少噪声并提取文本中的关键信息。接着，利用构建的情感词典对预处理后的文本进行情感标注，确保数据集能够准确反映游客的真实情感体验。

在模型选择上，本研究采用了 ERNIE-BiLSTM-DPCNN 模型。这一模型结合了 ERNIE (Enhanced Representation through kNowledge Integration)的强大语义表示能力、BiLSTM (Bidirectional Long Short-Term Memory)对文本上下文信息的捕捉能力以及 DPCNN (Deep Pyramid Convolutional Neural Network)的深层特征提取能力。这种混合架构使得模型能够有效地捕捉到文本中的长距离依赖关系，并学习到丰富的特征表示。

在模型训练过程中，采用了批量处理的方式，每个批次包含 32 个样本。为了防止过拟合，设置了 Dropout 率为 0.5，并选择了 250 个过滤器来控制模型的复杂度。同时，设置了 require_improvement 参数

为1000，以确保在模型性能没有显著提高的情况下及时停止训练。整个训练过程持续了100个周期(Epoch)，以确保模型对数据集进行了充分的训练。

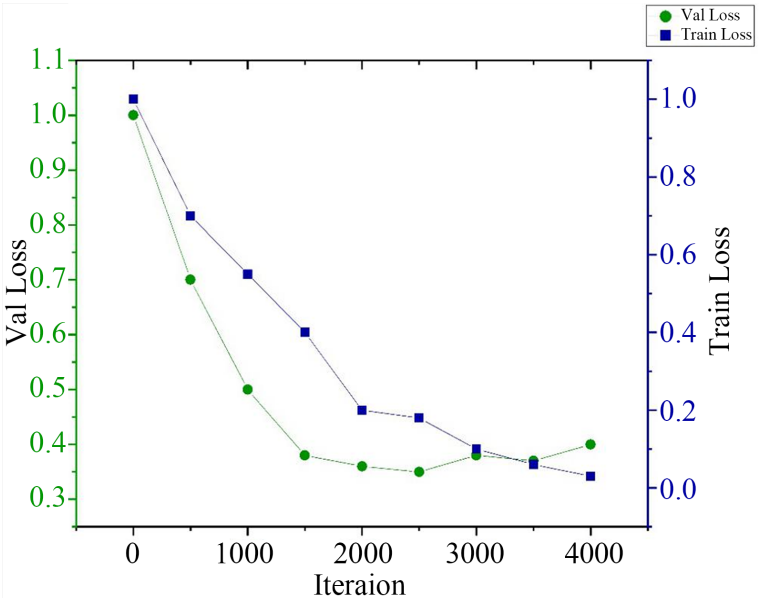

Figure 4. Loss function
图 4. 损失函数

在模型性能评估方面，采用了准确率、召回率和F1值这三个标准指标。实验结果显示，随着训练的进行，训练集和测试集的损失函数值逐渐减小，表明训练过程稳定且有效，结果如图4和图5所示。同时，随着迭代次数的增加，训练集和验证集的准确率逐渐提高，并在迭代次数达到700次时趋于稳定。最终，在测试集上，模型取得了93.23%的准确率。

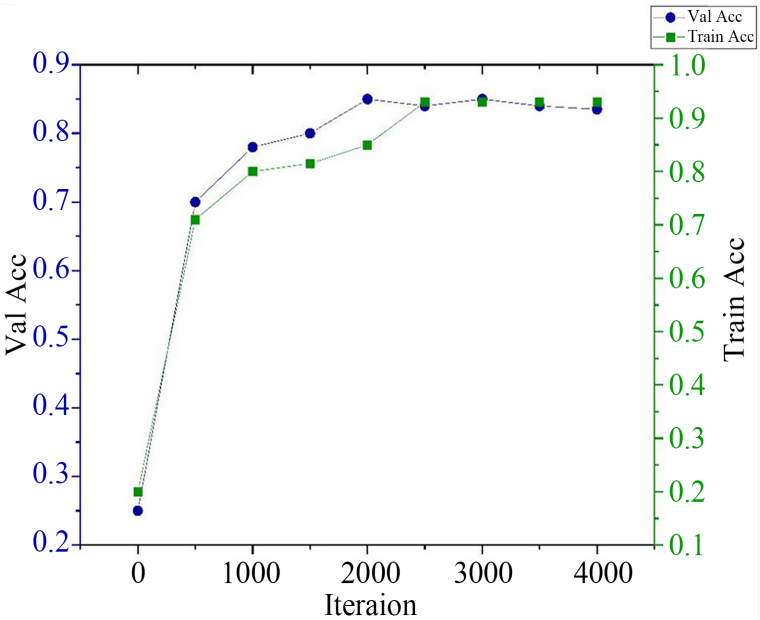

Figure 5. Accuracy
图 5. 准确率

Table 10. Performance of ERNIE-BiLSTM-DPCNN model
表 10. ERNIE-BiLSTM-DPCNN 模型性能

情感倾向	Precision(%)	Recall(%)	F1 值(%)	评论条数
积极情感	93.23%	96.83%	95.00%	2019
中性情感	73.26%	72.84%	73.05%	143
消极情感	85.70%	92.35%	88.9	837

具体来看，模型在积极情感、中性情感和消极情感三个类别上的F1值分别为95%、73.05%和88.90%。正如表10所示，这些结果表明，ERNIE-BiLSTM-DPCNN模型在桂林景点评论数据集上表现出色，能够准确地识别和分类不同情感倾向的评论。这一研究不仅验证了ERNIE-BiLSTM-DPCNN模型在情感分析任务中的有效性，也为桂林景点评论的情感分析提供了新的方法和思路。

3.4. 服务评价与相关建议

基于ERNIE-BiLSTM-DPCNN模型对游客评论的深入情感倾向分析，可以更精准地把握游客对景区服务的具体感受，并据此提出针对性的改进建议。以下是对各个景区的服务评价及建议：

积极情感评论：这类评论比例高，表明游客对景区的整体体验满意度较高。景区应继续保持并提升服务质量，同时，不应忽视积极评论中可能隐藏的改进建议。通过仔细分析这些评论，景区可以挖掘出潜在的服务优化点，如增设特色项目、提升服务细节等。

中性情感评论：中性评论通常反映出游客对景区的某些方面持保留态度。对于这类评论，景区应深入挖掘游客的犹豫和中立情绪的来源，识别游客的真实需求，并据此调整服务策略，以争取将中性情感转化为积极情感。

消极情感评论：消极评论是景区必须重视的部分，因为它们直接反映了游客的不满和问题。景区应针对这些评论进行详尽的分析，特别是那些频繁出现的负面评价，应作为改进的重点。例如，对于象鼻山和遇龙河等消极情感比例较高的景区，应立即调查游客提出的问题，如价格、设施、服务态度等，并针对性地制定改进措施，同时定期检查改进效果。

具体建议：

1) 积极情感保持：对于如漓江等积极情感比例高的景区，应继续维持并提升服务质量，通过定期回顾游客评论，确保服务始终满足或超越游客的期望。

2) 中性情感转化：对于中性情感比例较高的景区，如龙脊梯田，应通过问卷调查、游客访谈等方式深入了解游客中立情绪的原因，并据此制定策略，如提升服务质量、增加特色活动等，以争取将中性情感转化为积极情感。

3) 消极情感改善：对于消极情感比例较高的景区，应迅速响应并调查游客提出的问题，制定并实施针对性的改进措施，如调整价格策略、更新设施设备、提升员工服务水平等。同时，应定期跟踪改进效果，确保问题得到有效解决。

通过以上措施的实施，景区将能够提升游客的满意度和忠诚度，增强自身的竞争力和市场地位。同时，这也将促进旅游业的可持续发展和当地经济的繁荣。

4. 结语

本研究深入分析了携程网上广西桂林旅游代表性景点的在线旅游评论数据，旨在从游客的情感体验和满意度出发，全面评估桂林旅游目的地的吸引力与服务质量。本文综合运用了词云图、语义网络特征关联分析、LDA主题模型特征分析等方法，对游客的评论文本数据进行深入的特征分析。通过建立情感

词汇词典、程度副词词典和负面词典，能够对现有指标的好评率、情感得分、景点评论率、总得分等指标进行相关性分析，并运用变异系数法确定了相应的分值与权重。结果表明，桂林五个热门旅游目的地在携程网上受到广泛好评，游客对桂林的自然风光给予了极高的评价和认可。特别地，游客在旅游过程中非常注重体验感和对桂林山水风光的欣赏。这一结果证明了桂林作为国际旅游城市在旅游市场上仍具有巨大的竞争优势和宣传潜力。此外，在情感倾向分析中，也发现有部分游客对旅游体验感表现出消极情感。这表明，尽管桂林的旅游资源丰富，但在基础设施、服务质量等方面仍有待提升和完善。为了景区的可持续发展和桂林旅游业的整体提升，建议景区在推进商业开发的同时，加强自然资源的保护，避免过度商业化，并优化城镇、景区的道路交通，以减少节假日的拥堵和排队现象，从而提升游客满意度。

基金项目

桂林旅游学院校级科研项目(2021B04)；桂林旅游学院大学生创新创业训练计划项目(202311837061)。

参考文献

[1] 邓宁, 钟栎娜, 李宏. 基于 UGC 图片元数据的目的地形象感知: 以北京为例[J]. 旅游学刊, 2018, 33(1): 53-62.

[2] 罗秋菊, 梁思贤. 基于数字足迹的自驾车旅游客流时空特征研究: 以云南省为例[J]. 旅游学刊, 2016, 31(12): 41-50.

[3] Hollenstein, L. and Purves, R. (2010) Exploring Place through User—Generated Content: Using Flickr to Describe City Cores. *Journal of Spatial Information Science*, No. 1, 21-48.

[4] Joyner, L., Kline, C. and Oliver, J. (2018) Exploring Emotional Response to Images Used in Agritourism Destination Marketing. *Destination Marketing & Management*, **9**, 44-45. https://doi.org/10.1016/j.jdmm.2017.10.004

[5] 池丽平, 王新建. 基于 UGC 数据的滨海旅游目的地形象感知研究——以平潭国际旅游岛为例[J]. 乐山师范学院学报, 2020, 6(35): 53-58.

[6] 张公让, 鲍超, 王晓玉, 等. 基于评论数据的文本语义挖掘与情感分析[J]. 情报科学, 2021(5): 1007-7634.

[7] 王振璇. 基于 UGC 数据的游客行为分析与旅游目的地评价研究[D]: [硕士学位论文]. 临汾: 山西师范大学, 2016.

[8] 侯成. 基于 UGC 数据的南充市旅游目的地形象感知及提升研究[J]. 技术与市场, 2022, 29(5): 142-144.

[9] 杨嘉雯, 石媛媛, 闫安. 基于网络评论的文本挖掘与情感倾向分析——以北京地区博物馆为例[J]. 互联网周刊, 2023(11): 20-23.

[10] 王彬菁. 基于依存句法树方法的微博文本的情感分析研究[J]. 电脑知识与技术: 学术版, 2019, 15(24): 13-15.