doi:10.3772/j.issn.2095-915x.2023.01.002

基于网评数据的游客印象挖掘与情感分析

完颜兵[1] 张超群[2,3] 王大睿[1] 李晓翔[1] 郝小芳[1]

1. 广西民族大学电子信息学院 南宁 530006；
2. 广西民族大学人工智能学院 南宁 530006；
3. 广西混杂计算与集成电路设计分析重点实验室 南宁 530006

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 从国内游客在线评论文本中分析挖掘出游客对目的地的潜在印象，有助于相关部门和企业了解游客的真正需求，从而科学规划目的地的发展，提升目的地的美誉度。[方法 / 过程] 通过构建词云图进行热词分析；通过 DBSCAN 密度聚类、构建语义网络关系图完成特色分析；通过提出一种基于 Bert 的双路神经网络融合的文本情感分类模型 TNNFMB(Two-way Neural Network Fusion Model Base on BERT) 结合迁移学习实现情感分析，以此挖掘游客的潜在印象。[结果 / 结论] 通过实验，总体挖掘分析出游客高度关注目的地的服务、环境、设施、性价比、位置、景点景色、景区项目，并验证了 TNNFMB 模型在分类准确率上比基线模型至少提升 3.06%，取得了更好的分类效果。

关键词：在线评论；DBSCAN；迁移学习；神经网络；情感分析；印象挖掘

中图分类号：G35

Tourist Impression Mining and sentiment Analysis Based on Online Review Data

WAN Yanbing[1] ZHANG Chaoqun[2,3] WANG Darui[1] LI Xiaoxiang[1] HAO Xiaofang[1]

1. College of Electronic information, Guangxi Minzu University, Nanning 530006, China;
2. College of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China;
3. Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Nanning 530006, China

Abstract: [Objective/Significance] Analysis and mining of tourists' potential impressions of the destination from the online comment texts of domestic tourists will help relevant departments and enterprises to understand the real needs of tourists, so

基金项目 国家自然科学基金项目“基于数据流的微分代数事件结构及其层次化理论研究”（62062011）；广西民族大学研究生科研创新项目“基于 Spark 技术的旅游景点推荐系统”（gxun-chxps202088）；广西民族大学研究生科研创新项目“基于知识图谱的旅游推荐系统半自动化构建”（gxun-chxs2021066）。

作者简介 完颜兵（1997-），硕士研究生，研究方向为自然语言处理、大数据技术与应用；张超群（1974-），博士，副教授，硕士生导师，研究方向为大数据技术与应用、智能计算，E-mail: chaozi_0771@163.com；王大睿（1996-），硕士研究生，研究方向为大数据技术与应用；李晓翔（1996-），硕士研究生，研究方向为大数据技术与应用。郝小芳（1997-），硕士研究生，研究方向为知识图谱。

引用格式 完颜兵 , 张超群 , 王大睿 , 等 . 基于网评数据的游客印象挖掘与情感分析 [J]. 情报工程 , 2023, 9(1): 15-29.

doi:10.3772/j.issn.2095-915x.2023.01.002

as to plan the development of the destination scientifically and enhance the reputation of the destination. [Methods/Process] By constructing a word cloud map for hot word analysis; by constructing DBSCAN density clustering and semantic network relationship map to complete feature analysis; by proposing a two-way neural network fusion model base on Bert for text sentiment classification TNNFMB (Two-way Neural Network Fusion Model Base on BERT) combined with migration learning to achieve sentiment analysis as a way to explore the potential impressions of tourists.[Results/Conclusions] Through the experiment, the overall excavation and analysis show that tourists pay high attention to the service, environment, facilities, cost performance, location, scenic spots, scenic spots and projects of the destination, and verify that the classification accuracy of TNNFMB model is at at least 3.06% higher than that of the baseline model, achieving a better classification effect.

Keywords: online reviews; DBSCAN; transfer learning; neural networks; sentiment analysis; impression mining

引言

随着国内旅游业的发展，外出旅游已成为人们生活的一部分，十四五规划提出要着力推进旅游为民、发挥旅游带动作用[1]。当下旅游景点数不胜数，如何提升景区等旅游目的地美誉度是各地文化和旅游主管部门、旅游企业非常重视的工作，其涉及提高竞争优势、吸引游客和稳定客源等内容。

旅游目的地一般涉及景区和周边酒店等，景区的美誉度与游客访问量有着直接的关系。在快速发展的社会，人们倾向于根据景区的美誉度等因素选择旅行，以此来获得更好的体验感和提高生活质量。换言之，旅游目的地口碑越好、网络热度越高、名人聚集越多、平台互动性越好、旅游形象映射品质越高，就越容易促成游客出游[2]。此外，由于人们出行通常会住在酒店，酒店的服务等因素在一定程度上会影响人们对目的地的选择和体验，不同满意度的游客体验感存在显著差异[3]。

1 相关研究

随着互联网技术的发展，网络文本已成为目的地形象研究的重要数据来源[4]。网络在线评论数据能更加真实、准确地表达顾客的客观感受，成为商家和消费者等情报的重要来源，相关数据的获取与分析的方法也层出不穷。例如，陈天琪等[5]以携程网抓取的网络评论文本为研究素材利用 ROST Content Mining 软件，挖掘游客评论数据；蒋建洪等[6]利用网络爬虫技术爬取个人游记、评论等信息作为初始文本，将复杂网络理论应用于文本挖掘；赵志杰等[7]使用 TF-IDF 算法获取不同类型酒店客户评论特征权值并分析；Yuan Zhai 等[8]采用 Word2vec 对携程网关于莫干山周边在线评论进行特征提取并分析。

情感分析是一个可以从在线旅游评论文本中提取游客对旅游目的地的情感的过程。情感分析的结果构成旅游决策的重要依据[9]。情感分析方法分为基于词典、基于机器学习以及将

两者结合等方法[10]。Jardim S 等[11]面向旅游服务通过建立情感字典，利用旅游平台用户评论作情感极性判断。Mostafa 等[12]利用机器学习中支持向量机、朴素贝叶斯和决策树对旅行者评论进行情绪分类。

上述及其他相关研究存在两个问题：（1）在数据挖掘中对文本缺乏进一步的有效性处理；（2）在情感分析中，传统的基于词典和基于机器学习的方法耗时、费力，需要大量人工特征，直接影响模型性能[13]，而深度学习模型能减少特征设计的复杂性，且优于机器学习等方法[14]。本文主要利用具有噪声的基于密度的空间聚类方法 DBSCAN[15] 对文本进行有效性处理，利用迁移学习结合所提出的 TNNFM-B(Two-way Neural Network Fusion Model Base on BERT) 深度学习模型实现情感分类和情感分析。本文主要以热词分析、特色分析、情感分析三个部分来完成对游客印象的挖掘。

2 目的地热词分析

2.1 数据选择

本文待分析的数据来源于2021年第九届“泰迪杯”全国数据挖掘挑战赛官网 (https://www.tipdm.org:10010/#/competition/1354705811842195456/question) 所提供的50家景区和50家酒店的真实评论数据，对于不同的目的地其分析策略相同。本文选取所有景区和酒店的评论作为实验数据，其中，景区共有59106条评论数据，酒店共有25225条评论数据。对这些评论数据进行分析，初步挖掘游客对目的地的印象。

2.2 文本处理与热词提取

热词分析主要是对游客评论数据进行高频词统计、结果可视化与分析。首先加载游客网评数据后对该数据进行脱敏处理，接着分词和根据常用的多个停用词表来过滤语气词、标点符号等无用词，最后进行词频统计，并初步分析结果，其词频排名在前20的词如表1所示。

表1 50家景区与50家酒店排名前20的词

序号	景区	酒店	序号	景区	酒店
1	不错	酒店	11	项目	位置
2	方便	服务	12	感觉	设施
3	值得	不错	13	景色	卫生
4	取票	房间	14	动物	下次
5	好玩	早餐	15	排队	特别
6	表演	环境	16	门票	服务态度
7	地方	方便	17	风景	热情
8	景区	前台	18	时间	交通
9	开心	入住	19	便宜	性价比
10	景点	干净	20	环境	感觉

如表1所示，在景区排名前20的热词中，其中“不错”“方便”“值得”“好玩”“开心”“便宜”这些词表达游客对于景区的个人情感感知，在数量比例上占据景区排序前20词的30%，并且这些感知都是积极正面的，一方面可能是值得游玩的景区占据多数，另一方面可能是游客倾向于表达正面的个人情感。“表演”“景点”“景色”“动物”“风景”，这些偏向于观赏性感知的词占据了30%，“取票”“感觉”“排队”“时间”“环境”这些偏向于体验性感知的词占据了25%，其他词占据15%。从游客评论中提取

doi:10.3772/j.issn.2095-915x.2023.01.002

的热词能说明游客倾向于对景区观赏性和体验性对象的情感表达，这些目的地对象在游客游玩后在其感知记忆中占据主要位置。

同理，酒店排名前 20 的热词中“不错”“方便”“干净”“热情”这些词表达游客对目的地的个人情感感知，这些情感也全是正面积极的。不同的是，“服务”“房间”“早餐”“环境”“位置”“设施”“卫生”“服务态度”“性价比”，这些体验性的感知占据 50%，而观赏性的感知几乎没有，这也不难理解，酒店主要是以给游客带来良好的体验作为关键点，也说明游客在以酒店为目的地时，倾向于表达对酒店入住后的个人体验感知。

2.3 基于图的可视化与分析

将游客的网络评论词可视化为词云图，目的是了解游客对目的地的整体印象，所构建的词云图分别如图 1、图 2 所示。对表 1 中景区排名前 20 的评论热词采用饼状图展示 , 如图 3 所示；将酒店排名前 20 的评论热词采用柱状图表示，如图 4 所示。结合图 1—4 的结果进行游客印象挖掘，初步分析游客对目的地的关注点，为特色分析及进一步挖掘游客印象点做准备。

图 1　50 家景区评论词云图

图 2　50 家酒店评论词云图

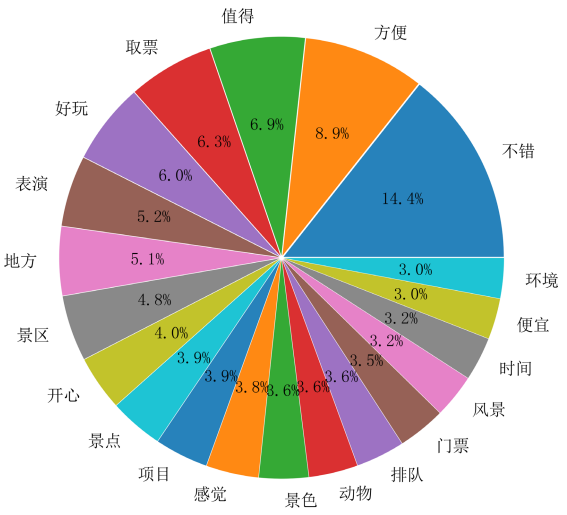

图 3　50 家景区评论热词的饼状图

由图 1 和图 3，游客对所游玩的景区总体印象偏好，其中，“不错”一词所占印象比最高，其次“取票”“好玩”两个词反映游客对进入景区前的取票过程和进入景区后的游玩印象深刻。根据其他热词及联系上下文可推测游客对于景区内项目表演、景区环境、景点景色等同样难以忘怀。同时，由于景区游玩的人多等因素，游客会关心排队时间等问题。经上述分析可以初步挖掘出游客的关注点偏向于景区的一些项目、景色、环境、性价比等。

根据图 2 和图 4，游客对酒店的整体印象也是偏好的，具中“不错”一词排名第一，“方便”“干净”“卫生”等高频词所体现的正是使游客印象偏好的原因之一。联系上下文的所有热词，再对酒店自身进行分析，挖掘分析出具备卫生干净、性价比高、环境好等因素的酒店会给游客带来深刻的正面印象。此外，大多数游客会考虑目的地交通是否便利，地理位置是否优越等问题。从“前台”和“服务”、“房间”、“环境”和“干净”、“性价比”和“位置”等词初步分析出游客的关注点偏向于服务、设施、卫生、性价比、位置。

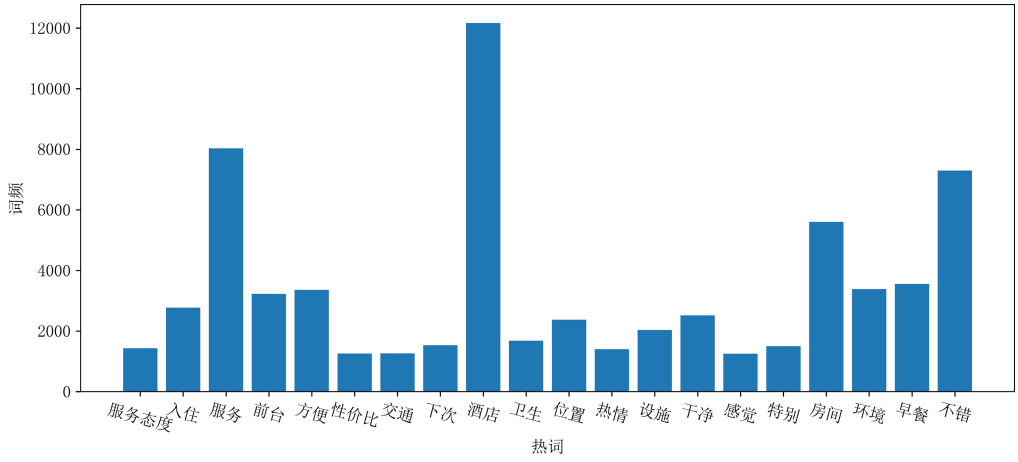

图 4　50 家酒店评论热词的柱状图

通过对提取的热词进行构建图的可视化分析，研究发现游客会对一些目的地的不同方面产生不同程度的个人印象，初步反映游客的偏好，但还需进一步验证分析。

3　目的地特色分析

3.1　数据选择

目的地特色分析主要是从游客评论所抽取

doi:10.3772/j.issn.2095-915x.2023.01.002

的关键词中分析出目的地特色以及分析游客对目的地的主要关注点。与高频词不同，关键词可以代表文本重要内容，代表文本所表达的核心思想，并且在对文本进行有效性处理的基础上进行关键词抽取，既能筛掉无效评论文本对关键词抽取的干扰，又能挖掘出文本重要含义，是对游客进一步的印象挖掘分析。

由于景区众多，根据数据中每个景区和酒店的综合评分，选取景区和酒店各 4 个作为目的地分别进行单独分析。这些目的地的综合评分与总平均分很接近，目的是合理地挖掘游客潜在印象，避免评分高或评分低的目的地带来游客印象单一问题。最后对 50 家景区和 50 家酒店所有数据分别当作一个文本来挖掘主要印象点，目的是从宏观上进行分析。由于游客评论以目的地为单位，因此利用 DBSCAN 密度聚类，分别对每个目的地进行文本有效性处理。

3.2 DBSCAN密度聚类及关键词提取

DBSCAN 是一种基于密度的聚类算法，这类密度聚类算法一般假定类别可以由样本分布的紧密程度决定，同一类别的样本是紧密相连的。与 K-means[16] 等聚类算法相比，DBSCAN 擅长找到离群点 [17]，对于本文来说离群点就是要保留的有效数据，因此 DBSCAN 适用于本文的文本处理。以 A01 景区评论数据为实验代表，聚类测试过程如图 5、图 6 所示。

图 5 中，X 轴代表密度聚类半径，Y 轴代表不同聚类半径所对应的聚类簇数，橙色标记“×”代表聚类簇数的最大值。图 6 中，Y 轴代表不同聚类半径所对应的游离个数，通常用 -1 表示游离簇，橙色标记“×”代表最大簇数半径下的游离个数，其聚类效果为最优，因此选取对应节点参数进行实验，其部分聚类结果如表 2 所示。

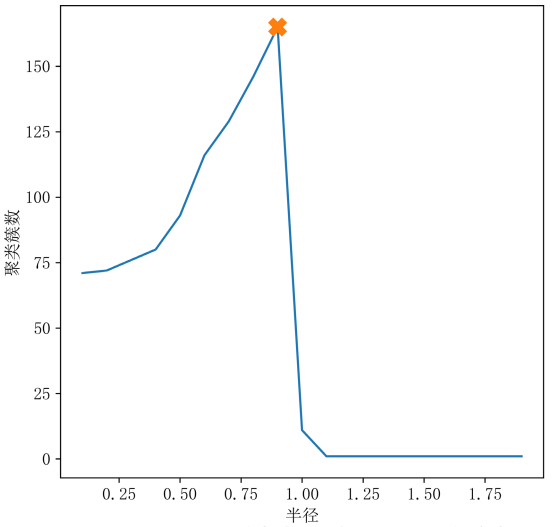

图 6　A01 景区评论数据聚类测试图 - 游离数

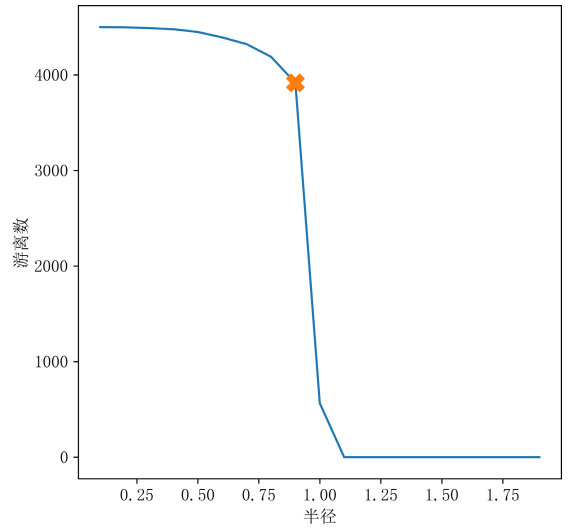

图 5　A01 景区评论数据聚类测试图 - 聚类簇数

在表 2 中，索引指内容在原数据中的位置，簇代表数据所属第几类。根据表 2 所展示的聚类结果，说明文本中存在随意评论和刷评论的游客，其中完全相同、非常相似、简单修改的评论会聚到其所属的簇中，这些游客会对一些已有的评论进行复制，有的进行稍加修改，有的甚至直接拿来评论，而这些评论并不能代表

游客的本身真实印象，且会对实验及分析结果产生影响，因此需要对数据进行有效性处理。经处理后，景区、酒店分别剩余 51 193、20 270 条有效评论数据。对处理后的评论内容根据计算的 TF-IDF 值提取关键词，取排名前 15 的关键词列于表 3。

表 2　A01 景区部分评论文本

索引	内容	簇
292	演出非常精彩。就是等级要求太恶心了。	24
294	非常精彩的演出。就是等级要求太恶心了。	24
335	小孩觉的还是很好玩的，就是暑假有点贵多人	29
514	小孩觉的还是很好玩的，就是暑假有点贵好多人	29
386	还是要换票的！还是要换票的！还是要换票的！ 无须换票不真实。	36
4577	还是要换票的！还是要换票的！还是要换票的！ 无须换票不真实。	36
436	服务很好 房间干净。设施一应俱全！ 好评！	43
4590	服务很好 房间干净。设施一应俱全！ 好评！	43
745	身体抱恙没法去也不给退，很过分	61
4627	身体抱恙没法去也不给退，很过分	61
1367	好不错方便迅速快捷折扣力度再大点就更好	95
1680	不错，比较方便迅速及时，折扣力度再大点就更好了	95

表 3　部分景区和酒店及各 50 家汇总数据的关键词

目的地	A24	A29	A34	A46	景区汇总	H13	H29	H38	H48	酒店汇总
1	温泉	蝴蝶谷	好玩	栈道	不错	服务	早餐	早餐	服务	服务
2	不错	景点	不错	玻璃	取票	入住	服务	沙滩	不错	房间
3	水上	门票	项目	景点	方便	房间	前台	房间	房间	不错
4	泡温泉	不错	开心	不错	好玩	早餐	不错	设施	方便	早餐
5	环境	值得	门票	天道	值得	贴心	干净	不错	卫生	前台
6	乐园	景色	设施	缆车	表演	经理	房间	泳池	前台	入住
7	设施	空气	小孩	值得	景点	港珠澳	入住	入住	位置	方便
8	池子	爬山	方便	景色	门票	下次	服务态度	环境	干净	干净
9	大观园	肇庆	汕头	取票	排队	大桥	清馨	服务	服务态度	环境
10	水果	庆云寺	取票	风景	开心	不错	小姐姐	海滩	地理位置	设施
11	值得	风景	值得	观光车	景色	海景房	下次	适合	性价比	服务态度
12	服务	地方	游客	门票	游玩	海景	卫生	游泳池	交通	下次
13	开心	鼎湖	小朋友	方便	风景	前台	舒服	前台	便利	位置
14	小孩	负离子	孩子	好玩	感觉	升级	清灵	海景房	环境	卫生
15	下次	空气清新	刺激	丹霞风貌	地方	长隆	热情	老旧	下次	性价比

doi:10.3772/j.issn.2095-915x.2023.01.002

表3体现目的地的相关属性和一些视觉性、体验性感知以及游客围绕这些感知印象所表达的个人情感感知。其中，个人情感感知包含正面情感（如“开心”“好玩”）和负面情感（如“老旧”）。结合一些目的地内部因素可以将情感表达分为对整体游玩后的情感表达、对景点景色的情感表达、对项目设施的情感表达等，另外不同目的地所出现的关键词存在重复现象，说明游客潜意识中会去抓住这些目的地的关键点作为主要印象。

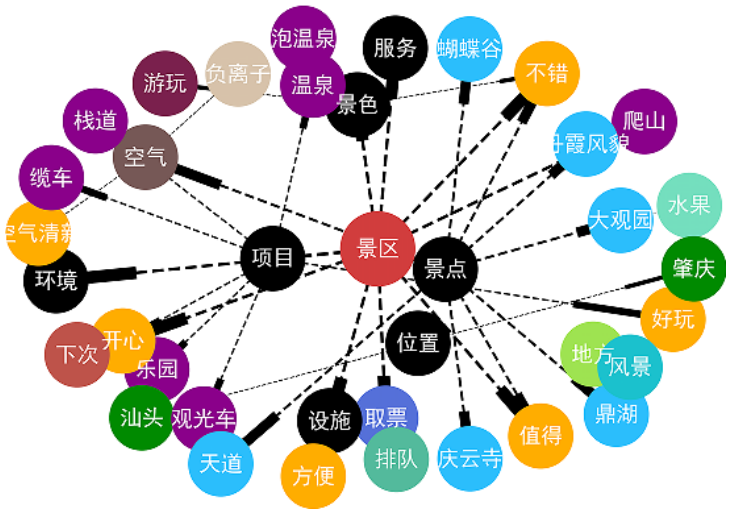

图 7　表 3 中景区关键词的语义网络关系图

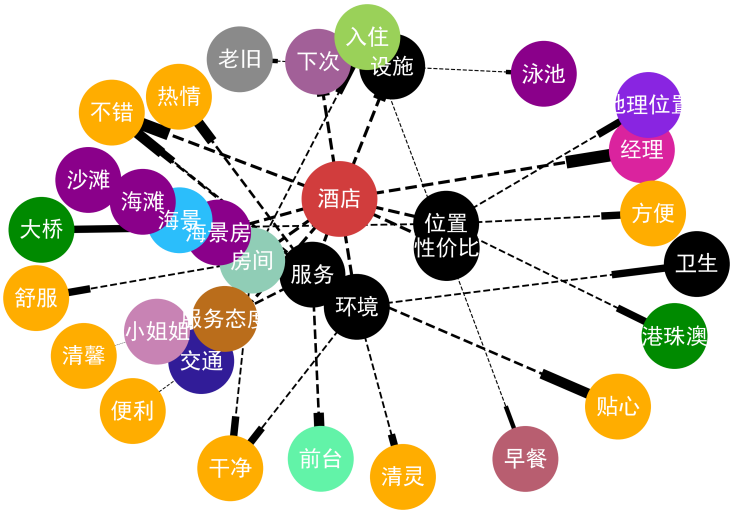

图 8　表 3 中酒店关键词的语义网络关系图

3.3　语义网络关系图的构建

经过计算提取的关键词可以反映出游客对目的地的印象，不仅如此，由于关键词具有一定的独特性[18]，因此分析这些关键词之间的关系，可以提取出目的地自身所具有的特色。本文根据表 3 提取的关键词利用 Python 编程生成如图 7、图 8 所示的语义网络关系图，其语义关系主要指的是词之间的总分关系、修饰关系

等。例如，“景区”与“景点”存在总分关系，“景区”与“不错”存在修饰关系等。对表3中的关键词进行适当分类，图7、图8中黑色节点是指与目的地具有直接包含关系的词，蓝色指景点风景，橙色代表游客积极感知，灰色代表消极感知、紫色代表一些可玩项目、地点等。

3.4 结果分析

依据表3和图7进行分析，首先从A24景区排名最高的“温泉”、A29景区的“蝴蝶谷”以及A46景区的“栈道”等词可分析出，游客对于景区部分首要的关注点是目的地的景点及景色和好玩的项目，其次从关键词“环境”“服务”“设施”可反映出环境、服务、设施也是重要关注点。根据各目的地所有上下文的关键词，从其各自特色角度来分析，可大致分析出A24景区具有温泉、乐园、大观园等特色项目和景点，而且景区内环境好，服务好；A29景区有蝴蝶谷、庆云寺、鼎湖等特色景点，且景色优美，空气质量好；A34景区具有一些特色项目；A46景区具有栈道、缆车等一些特色项目和丹霞风貌等景点。总体来看，游客对景区的印象点主要在于景区内的景点景色、项目、环境、服务、设施等。

由表3和图8可见，从H13的“港珠澳”“大桥”可看出该酒店位置独特，在地理位置上给游客留下深刻印象，其中“海景房”“海景”更是让游客体会到与酒店不同的特色；“服务”一词权值第一，结合“贴心”说明好的服务最能让游客记住；H29的“干净”“卫生”“清灵”这些关于环境的正面描述是该目的地的显然特色，“服务态度”“热情”更是让游客记忆深刻；H38的“沙滩”“海滩”“海景房”给游客留下不一样的印象，让游客感到“适合”“入住”，但是“设施”“老旧”给游客留下不好的印象，应该及时维护；H48的“方便”“交通”“便利”给游客带来不同的体验，“性价比”也是游客印象点，“服务”“服务态度”“环境”关注居多。结合汇总来看，游客主要关注点为服务、环境、设施、性价比、位置。

经过上述4家景区和4家酒店的单独分析及总结，再结合50家景区汇总和50家酒店汇总数据得出的关键词以及基于热词的初步分析，最后实验表明游客对目的地主要关注点在于服务、环境、设施、性价比、位置五个方面，其中，游客对景区主要关注前三个方面，除此之外，景点景色以及景区内的项目也同等重要，对于酒店的关注则完全包含上述五个方面。相关部门和旅游企业应该着重做好前四个方面以及景区内景点的美化和项目的部署，以此吸引游客，再结合前面所有的游客目的地印象分析，应当对当地旅游目的地根据真实印象因地制宜，扬长避短。

4 目的地情感分析

基于深度学习的情感分析方法，其分类效果优于基于情感词典的方法和基于机器学习的方法[19]，由于原数据集存在质量参差不齐、类别间失衡、缺乏标注等问题，因此很难把原数据直接通过深度学习的方式进行情感分析。本文将引入迁移学习解决上述问题，并建立深度学习模型进行情感分类。

doi:10.3772/j.issn.2095-915x.2023.01.002

4.1 迁移过程

采用迁移学习的方式，将原数据作为目标域数据，所采用的另一份带有标签的相似情感分析数据集作为源域数据。将源域数据与目标域数据域进行语义对齐，训练源域数据作用于目标域数据。

4.1.1 源域数据

源域数据来自 AI Challenger 2018 竞赛中情感分析赛道数据 (https://github.com/AI-Challenger/AI_Challenger_2018)，其本身来源于顾客对餐馆的网评，数据中涉及对服务、位置、环境、性价比、设施、总体等主题的情感判断，与目标域数据高度吻合，详情如图 9 所示。红色部分代表需要迁移的主题，其中总体情感代表顾客对餐馆上述几个方面的整体情感倾向，更适合用于迁移到目标域数据做情感分析。

4.1.2 语义对齐

为了消除训练时源域数据中餐馆相关的特有词的语义干扰，通过提取目标域高频词对源域数据进行词筛选，目的是使两份数据语义高度对齐，其示例过程如图 10 所示。

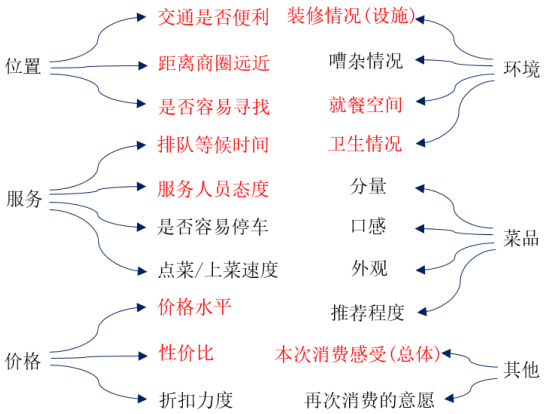

图 9 源域数据主题分布

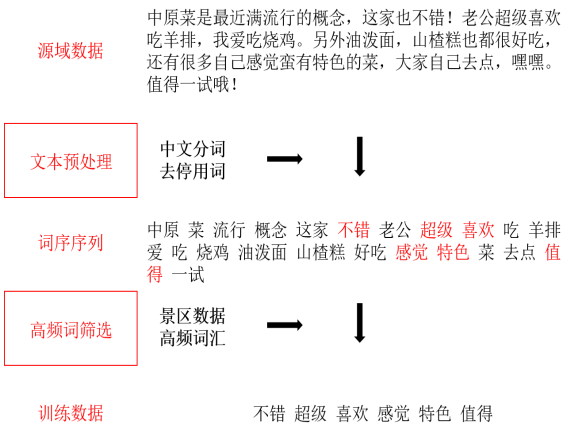

图 10 语义对齐示例过程图

4.2 TNNFMB模型

为了对目的地进行情感分析，我们提出一种基于 Bert 的双路神经网络融合的文本情感分类模型 TNNFMB，其整体模型如图 11 所示。首先对源数据域采用所设计的双路融合神经网络进行训练，将训练的模型作用于目标数据域进行情感分类，从而达成目标。

4.2.1 模型介绍

(1) 输入层和文本表示层

文本表示是字词表示为计算机可以处理的向量或矩阵。在深度学习中，常用的词嵌入模型有 Word2Vec、Glove 等，但这些模型很难解决汉语词汇的多义性和汉字的多意性问题，尽管添加了段落信息的 Doc2Vec 也很难解决，而通过上下文和字向量训练的 Bert 模型，能较好地改善以上问题 [20]，另外模型会依据语境信息自动调整向量。

(2) 特征抽取层与特征融合层

基于 TextCNN 的关键词的特征抽取，通过不同大小的卷积窗口来捕捉不同的 N-Gram 特征。一段文本除了具有整体的全局语义特征外，

其局部文本还包含大量重要的语义特征。因此，利用 TextCNN 来提取文本的局部特征 [21]，有助于分类效果。具体以 BERT 模型输出结果作为 TextCNN 的输入，TextCNN 中设置三个不同大小的卷积核，在其卷积及最大池化后进行结果的拼接，即为 TextCNN 层的输出结果 [22]。

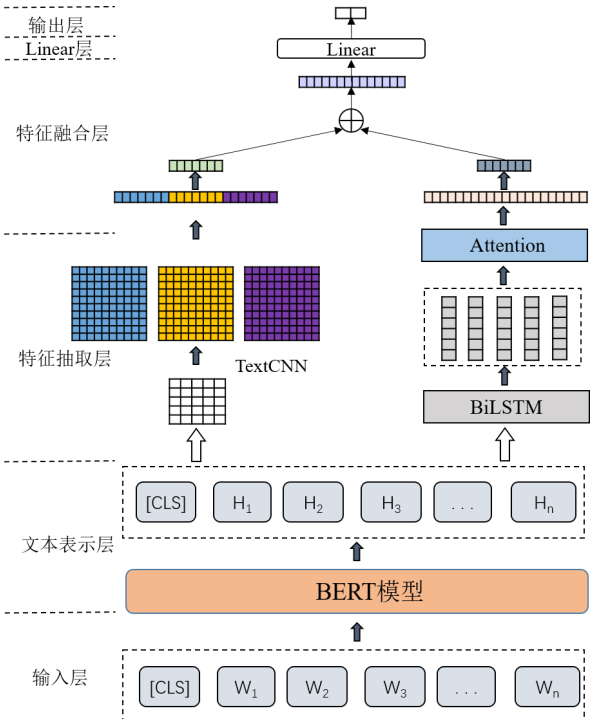

图 11　TNNFMB 模型

基于 BiLSTM-Attention 的上下文特征提取可以弥补多通道 TextCNN 无法表示多信息点的时序特征的缺陷 [23]。对于单向 LSTM 仅能保留过去的信息，而 BiLSTM 使用两个隐藏状态组合能在任何时间点保存过去和未来的信息。BiLSTM 的结构如图 12 所示。

在图 12 中，将 BiLSTM 的输入（即文本表示层的输出）表示为 H_i(i=1，…，n)，其中 LSTM 与 LSTM 分别代表方向相反的 LSTM 网络单元；将输出连接后的结果表示为 h_t。文本表示层的输出经过双向 LSTM 编码，把方向相反的 LSTM 得到的结果进行连接作为输出。将左右相反的第 t 时隐藏层状态分别记 h_t^{lr} 与 h_t^{rl}，则得出 h_t 的公式如 (4) 所示。

$$h_t = [h_t^{lr}, h_t^{rl}] \tag{4}$$

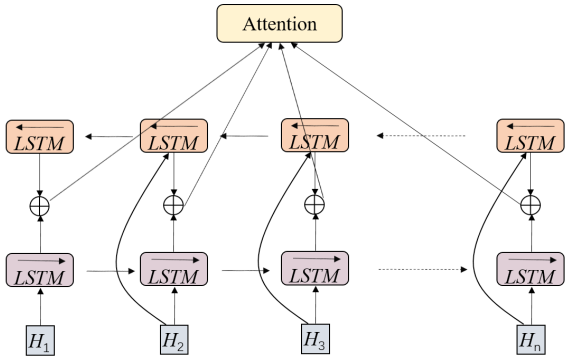

图 12　含有注意力的双向长短时记忆网络

为了突出关键词对语句语义表达的影响，将 Attention 机制引入在 BiLSTM 输出之前，目标是更新各维度的权重（即 BiLSTM 的输出），本质操作为加权平均计算，其 Attention 机制的具体公式如下：

$$u_t = \tanh\left(w \times h_t + b\right) \tag{5}$$

$$a_t = \frac{e\left(u_t\right)^T \times u_w}{\sum e\left(u_t\right)^T \times u_w} \tag{6}$$

$$s = \sum\left(a_t \times h_t\right) \tag{7}$$

由公式 (5) 可知，双向 LSTM 输出值的权重为 w，h_t 表示时间 t 时的双向 LSTM 输出值，偏移项为 b，时间为 t 时每个词的权重为 u_t；如公式 (6) 所示，转置词的权数用 u_t^T 表示，Attention 矩阵经随机赋值后表示为 u_w，t 时的归一化权数用 a_t 表示；公式 (7) 指将本质为含有 Attention 的语义编码表示为 s，也代表着双向 LSTM 的输出值与 t 时的权数加权后相加的结果。

doi:10.3772/j.issn.2095-915x.2023.01.002

(3) 基于双路特征融合的上下文特征提取

双路特征融合部分旨在将各路输出的最终向量进行首尾连接，然后通过 Linear 层做线性变换转化为所预测的最终向量维度。它主要包括三部分，具体为：①各路的全连接层；②各路向量的前后连接；③最终结果的全连接层。通过各路的全连接层可以将各路的输出进行降维，这样既能灵活调节各路输出结果的占比，又能使模型更具有可训练性，从而大大提高训练效果。

4.3 实验准备

源域数据集共有训练集 105 000 条，包含四种标签，其中，“-1”代表负向，“-2”代表未提及，“1”代表正向，“0”代表中立，涉及总体的正向评论共 70 070 条，负向为 9 384 条，未提及为 2 110 条，中立为 23 436 条。我们将中立与正向进行合并作为正向评论数据；负向与未提及合并作为训练集负向评论数据，共 11 494 条；将未提及数据中的正向词，使用目标数据域高频负向特征词进行随机替换，形成负向评论。从正向评论中随机抽取 11 494 条数据作为训练集中的正向评论。总的训练集共有 22 988 条，验证集为 7 662 条，测试集为 11 492 条，正负样本比例均为 1:1。模型训练时的超参数设置如表 4 所示。

4.3.1 实验结果与分析

模型实验效果对比如表 5，其中“p”代表正向情感，“n”代表负向情感，分别在整体准确率 (Accuracy)，各类别的召回率 (Recall) 与 F1 值上进行对比，TNNFMB 模型在准确率上比 BERT+BiLSTM+Attention 模型提升 3.06%，比 BERT+TextCNN 模型提升 4.0%，其正负类别在召回率与 F1 值上也都优于其他基线模型，证明 TNNFMB 模型在分类效果上具有一定的优越性。

表 4 TNNFMB 模型参数设置

参数描述	参数值
嵌入维度	768
卷积核大小	(2,3,4)
每种卷积核的数量	256
隐藏层单元个数	256
Dropout随机失活率	0.1
优化器	Adam
学习率	0.0005
训练批次大小	64
Transforme编码器层数	12

4.4 情感分类与分析

我们选取景区 51 193 条评论数据，经再处理后剩余 51 154 条，对这些数据进行情感分类，最终得到负面评论为 6 742 条，正面评论为 44 412 条。游客的负面评论对于景区的完善发展具有重要意义，对所有负面评论进行挖掘分析，经过筛选后的词频大于 100 的负面高频中心词如图 13 所示。

表 5 基线模型对比

模型	Accuracy	Recall(p)	Recall(n)	F1(p)	F1(n)
BERT+BiLSTM[24]	87.28%	88.57%	85.99%	87.44%	87.11%
BERT+BiLSTM+MaxPooling[25]	86.93%	91.73%	82.13%	87.53%	86.27%
BERT+BiLSTM+Attention[26]	89.18%	91.56%	86.81%	89.43%	88.92%
BERT+TextCNN[27]	88.24%	93.24%	83.23%	89.11%	87.21%
TNNFMB	**92.24%**	96.65%	87.83%	93.56%	90.89%

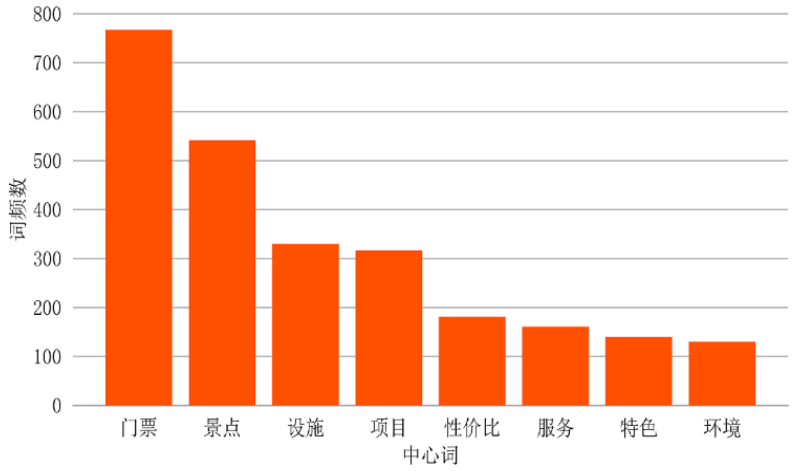

图 13　景区负面高频中心词

经过裁剪后的部分消极评论结果如表 6 所示，该表中“总体负”代表负面情感特征值，“总体正”为正面情感特征值。结合图 13 进行分析，其中，景区“门票”的票价贵成为游客最多的吐槽点，“景点”的观赏度差位居吐槽第二，其次景区“设施”少、“项目”无聊、景区“性价比”低、“服务”差、景区内部“特色”缺乏，“环境”差等都成为游客主要的消极印象点。虽然这些负面印象点大多都是目的地的部分建设所引来的问题，但这些问题却给整个目的地带来不好的影响，这些消极评论不仅影响游客自身的消费感受，并且这些负面评论所产生的消极印象会影响其他潜在的旅游消费者，严重影响景区美誉度，间接造成景区经济损失。因此相关景区部门人员不能忽视这些负面评论，应根据实际情况尽可能解决。

根据上述分析提出四点建议，(1) 景区建设收费要以人为本，坚决杜绝乱收费，重复收费现象；(2) 景点等方面必须有自己特色，如果没有特色将很难吸引游客；(3) 做好服务区的建设，完善服务机制；(4) 加强景区文化、项目设施、环境等的建设。总之要给游客带来更好的体验，才能消除更多的负面评论。

表 6　部分消极评论

景区	评论内容	总体负	总体正	情感
A01	门票贵啊，也就自驾的一段有点意思。	5.6263	-6.7471	消极
A07	缩小版的景点看起来没感觉。	5.7792	-6.8160	消极
A13	1分，太垃圾太不好玩设施太少。	5.9708	-6.6820	消极
A03	老旧 项目无聊 别相信网评！！	5.8376	-6.8567	消极
A19	性价比真的很低，门票不知为何还挺贵。	5.6515	-6.8394	消极
A27	服务差到无盆友。	6.0164	-6.8279	消极
A38	温泉没什么特色。	5.4703	-7.0155	消极
A48	索道周边环境差。	5.6651	-6.6506	消极

doi:10.3772/j.issn.2095-915x.2023.01.002

5 结语

本文通过热词分析、特色分析、情感分析来挖掘游客对目的地的高度印象点，其中，在热词分析中通过构建词云图初步挖掘游客印象；在特色分析中，利用 DBSCAN 善于发现离群点的特点进行文本的有效性处理；在情感分析中，提出一种基于 Bert 的双路神经网络融合的文本情感分类模型 TNNFMB，并结合迁移学习来对目标数据进行情感分类。通过实验，总体挖掘分析出游客高度关注目的地的服务、环境、设施、性价比、位置、景点景色、景区项目，结果表明了 TNNFMB 模型在分类准确率上比基线模型至少提升 3.06%，说明该模型具有一定的优越性。挖掘分析的结果有助于目的地因地制宜，扬长避短，科学规划和长远发展。

未来的改进方向：(1) 在文本挖掘过程中，可以考虑将词性、词的相关性等因素加入词频统计，以提高热词提取的准确性；(2) 虽然已剔除简单复制和修改的网评数据，但应加强剔除其中不相关的内容；(3)TNNFMB 模型中双路特征融合部分，采用的方法是把各路输出的向量进行直接前后连接，比较传统，其模型采用的 Attention 机制设定较为一般，可进一步改进，以提升最终的分类效果。

参 考 文 献

[1] 文化和旅游部 . “十四五”文化和旅游发展规划 [EB/OL]. 2021-4-29.http://www.gov.cn/zhengce/zhengceku/2021-06/03/content_5615106.htm.

[2] 王双 , 邱守明 . 新媒体语境下游客旅游目的地决策影响因素研究 [J]. 西南林业大学学报 (社会科学), 2021, 5(4): 35-41.

[3] 李春晓 , 李辉 , 刘艳筝 , 等 . 多彩华夏 : 大数据视角的入境游客体验感知差异深描 [J]. 南开管理评论 , 2020, 23(1): 28-39.

[4] 徐菲菲 , 剌利青 , Ye Feng. 基于网络数据文本分析的目的地形象维度分异研究——以南京为例 [J]. 资源科学 , 2018, 40(7): 1483-1493.

[5] 陈天琪 , 张建春 . 基于文本挖掘的景区旅游形象感知研究——以杭州西溪国家湿地公园为例 [J]. 资源开发与市场 , 2021, 37(6): 741-746.

[6] 蒋建洪 , 马瑞云 . 基于文本挖掘的个性化旅游偏好特征属性分析 [J]. 企业经济 , 2017, 36(12): 129-133.

[7] 赵志杰 , 刘岩 , 张艳荣 , 等 . 基于 Lasso-LDA 的酒店用户偏好模型 [J]. 计算机应用与软件 , 2021, 38(2): 19-26.

[8] Zhai Y, Chen P. Sentiment analysis on tourist satisfaction with rural homestayinns based on reviews from the website of online travel agency[J]. International Journal of Sustainable Development and Planning, 2020, 15(5): 705-712

[9] Chen W, Xu Z, Zheng X, et al. Research on sentiment classification of online travel review text[J]. Applied Sciences, 2020, 10(15): 5275.

[10] Sánchez-Rada J F, Iglesias C A. Social context in sentiment analysis:formal definition, overview of current trends and frameworkfor comparison[J]. Information Fusion, 2019, 52:344-356.

[11] Jardim S, Mora C.Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning[J]. Procedia Computer Science, 2022, 196:199-206.

[12] Mostafa L.Machine learning-based sentiment analysis for analyzing the travelers reviews on egyptian hotels[C]//The International Conference on Artificial Intelligence a nd Computer Vision.Springer, Cham, 2020:405-413.

[13] Lxa B, Yun X B, Hua W C, et al. Exploring fine-grained syntactic information for aspect-based sentiment classification with dual graph neural networks[J]. Neurocomputing, 2022, 471:48-59.

[14] Yadav A, Vishwakarma D K. Sentiment analysis using deep learning architectures:a review[J].

Artificial Intelligence Review, 2020, 53(6): 4335-4385.

[15] Sander J, Ester M, Kriegel H P, et al. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 169-194.

[16] Hartigan J A, Wong M A. Algorithm as 136:a K-means clustering algorithm[J]. Journal of the Royal Statistical Society.Series C(Applied Statistics), 1979, 28(1): 100-108.

[17] 王帅 , 纪雪梅 . 基于在线健康社区用户画像的情感表达特征研究 [J]. 情报理论与实践 , 2022, 45(6): 179-187.

[18] 徐坤 , 毕强 . 关键词分类判定及领域热点特征分析 [J]. 情报理论与实践 , 2019, 42(4): 96-100.

[19] 王春东 , 张卉 , 莫秀良 , 等 . 微博情感分析综述 [J]. 计算机工程与科学 , 2022, 44(1): 165-175.

[20] 李悦晨 , 钱玲飞 , 马静 . 基于 BERT-RCNN 模型的微博谣言早期检测研究 [J]. 情报理论与实践 , 2021, 44(7): 173-177+151.

[21] 范昊 , 何灏 . 融合上下文特征和 BERT 词嵌入的新闻标题分类研究 [J]. 情报科学 , 2022, 40(6): 90-97.

[22] Guo B, Zhang C, Liu J, et al. Improving text classification with weighted word embeddings via a multi-channel textCNN model[J]. Neurocomputing, 2019, 363: 366-374.

[23] Li H. Deep learning for natural languageprocessing: advantages and challenges[J]. National Science Review, 2018, 5(1): 24-26.

[24] Cai R, Qin B, Chen Y, et al.Sentiment analysis about investors and consumers in energy market based on bert-bilstm[J]. IEEE Access, 2020, 8: 171408-171415.

[25] Jiang Q, Zhang H, Shang J, et al. Densely connected bidirectional LSTM with max-pooling of CNN network for text classification[C]//International Conference on Advanced DataMining and Applications.Springer, Cham, 2020: 98-113.

[26] Lee L H, Lu Y, Chen P H, et al. Ncuee at medical 2019:medical text inference using ensemble bert-bilstm-attention model[C]//Proceedings of the 18th BioNLP Workshop and Shared Task. 2019: 528-532.

[27] Yang L, Huang X S, Wang J Y, et al. Identifying subtypes of clinical trial diseases with bert-textcnn[J]. Data Analysis and Knowledge Discovery, 2022, 6(4): 69-81.