

**新疆大学课程论文（设计）、学年论文评分表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 题 目 | 基于BERT的文本情感舆情分析 | | | | |
| 作 者 | 李慧颖 | 专业年级 | 网安20-3 | 指导教师 | 程述立 |
| 指导教师评语及  评分建议 | 该生积极完成网络舆情分析课程作业并参与课程研讨，上课认真听讲，论文格式较为规范，图表清晰，文字表述通顺，用语符合专业要求，逻辑性较强，课程设计较为合理，论述过程较为详细。整体而言，论文结构科学合理，逻辑思路清晰，观点表达准确，语言流畅，论述方法清晰合理，参考文献与课程主题符合，论文内容与网络舆情分析课程内容相符。  根据网络舆情分析的课程要求，结合该生考勤和作业情况，平时成绩为： （ ）分，根据课程论文撰写质量和课程论文工作量，课程论文成绩评定为：（ ）分，综合成绩为平时成绩占30%，课程论文占70%，该生综合成绩评定为： （ ）分。  根据以上意见，建议该生课程论文最终成绩等级为：（ ）。  指导教师：  2023 年 X 月 XX 日 | | | | |
| 院  （部）  或  教  研  室  意  见 | 同意指导老师意见。  学院或教研室主任：  年 月 日 | | | | |

摘 要

针对旅游景点，旅游网站上公开的用户评论能够真实反映游客旅游后对于景点服务和游览体验的感受。本课题通过对旅游景点评论观点抽取了解游客对旅游景点的核心关注点，利用情感倾向分析了解游客对各旅游景点的喜好程度，采用聚类分类能力进行系统化的总结，最后以可视化的形式将景点按照各核心关注点的喜好程度展示。其中情感倾向分析以构建BERT模型来实现。在自建的新疆旅游评论数据集上进行测试，预测评论的情感倾向，并对新疆旅游景点各核心关注点的情感倾向进行视觉分析与呈现，从而直观地展示游客对新疆各旅游景点的喜爱程度，达到预期的目标。

**关键词：**情感分析；BERT；文本分类；深度学习

**Abstract**

Regarding tourist attractions, user reviews published on travel websites can truly reflect the tourist's feelings about the services and experience of the tourist attractions after traveling. This topic extracts the core concerns of tourists to tourist attractions through the comments of tourist attractions. The affective tendency analysis is used to understand the degree of tourists' preference for various tourist attractions. Clustering ability was used to systematize the summary. Finally, the scenic spots are displayed in the form of visualization according to the preferences of each core concern. Among them, sentiment tendency analysis is realized by constructing a BERT model. The test was carried out on the self-built data set of XinJiang tourism review. The experimental results show that the model predict the emotional orientation of commentsy. Moreover, the emotional tendency of each core focus of XinJiang tourist attractions , so as to intuitively show the degree of tourists' affection for each tourist attraction in XinJiang, and finally achieve the expected goal.

**Key Words:** sentiment analysis; BERT; text classification; deep learning

**目 录**

[第1章 绪论 1](#_Toc20053)

[1.1 研究背景及意义 1](#_Toc3944)

[1.2 国内外研究现状 2](#_Toc13835)

[1.2.1 基于情感词典的情感分析方法 2](#_Toc23807)

[1.2.2 基于机器学习的情感分析方法 3](#_Toc20026)

[1.2.3 基于深度学习的情感分析方法 3](#_Toc25654)

[1.3 课题主要研究内容 4](#_Toc20842)

[1.4 章节安排 4](#_Toc30662)

[第2章 需求分析 6](#_Toc2171)

[2.1 行业需求 6](#_Toc8689)

[2.2 市场需求 6](#_Toc11775)

[2.3 功能需求 7](#_Toc3827)

[第3章 数据获取与预处理 8](#_Toc29040)

[3.1 数据选择 8](#_Toc15330)

[3.2 数据爬取 8](#_Toc16641)

[3.3 数据清洗 9](#_Toc5554)

[3.4 文本分词 9](#_Toc30498)

[第4章 文本特征词提取 11](#_Toc20793)

[4.1 词频统计 11](#_Toc12523)

[4.2 主题词特征提取 12](#_Toc32003)

[第5章 文本情感倾向分析 14](#_Toc2551)

[5.1 标注情感标签 14](#_Toc20175)

[5.2 模型构建 14](#_Toc9374)

[5.2.1 Attention机制 14](#_Toc16349)

[5.2.2 Transformer模型 15](#_Toc15115)

[5.2.3 BERT 17](#_Toc14829)

[5.3 实验设计 18](#_Toc9844)

[5.3.1 实验数据 18](#_Toc13833)

[5.3.2 评价标准 19](#_Toc5563)

[5.3.3 参数设置 20](#_Toc24964)

[5.3.3 实验结果 20](#_Toc15096)

[第6章 数据可视化分析 21](#_Toc16353)

[6.1 数据可视化 21](#_Toc26182)

[6.2 景点改进建议 23](#_Toc22760)

[结论 25](#_Toc32767)

[参考文献 26](#_Toc15479)

第1章 绪论

1.1 研究背景及意义

随着旅游业的兴起和网络的发展，游客越来越倾向于利用社交媒体平台规划旅游出行计划，购买景点门票并分享体验感受以及对景点相关服务的看法。在数字化的当今时代，关于旅游的社交媒体平台上游客发布的评论反映了游客对景点的真情实感，带有游客的主观态度[1]。通过对社交媒体平台上的评论进行综合分析，能够直观地得到游客对旅游景点各属性的情感倾向，这将有助于游客了解先行者对某景点的情感倾向，从而优化旅游计划[2]。根据《中国旅游经济蓝皮书》显示，2015-2019年中国在线旅游市场交易规模处于持续增长状态，但增长速度略微平缓。然而，受新冠肺炎疫情影响，2020年中国在线旅游市场交易规模首次出现负增长。根据中国旅游研究院整理的数据显示，2020年国内旅游人数达到28.8亿人次，较2019年相比减少31.2亿人次，同比下降52.1%。虽然2020年旅游行业受新冠肺炎疫情影响颇深，但是景点管理者也可以利用此次疫情间隙对景点进行整改，了解游客对其服务的意见和态度，有助于识别和跟踪游客对景点服务的态度从而扬长避短，自我提升。在不久的将来，人们的生活水平越来越高，追求更加高质量的生活方式，那么旅游业的需求将会只增不减，旅游领域的发展不可估量。因此，旅游景点管理者借助社交媒体提取游客对景点及情感倾向和改进建议对景点日后的发展至关重要[3]。

然而，游客在社交媒体平台上的评论内容不仅具有简短且不规范等特征而且含有大量干扰以及冗余信息，因此需要将其按核心关注点进行属性特征分类后再情感分析，从而挖掘其中反映游客情感倾向的有价值信息，为潜在游客计划旅游提供参考[4]。仅仅依靠人工方法来挖掘信息是不切实际的，因此通过情感分析技术智能化地处理数据成为迫切需求[5]。文本情感分析是自然语言处理和文本挖掘领域中一个热点问题，主要研究内容是分析用户对产品、组织、个人以及事件及其属性的意见、情绪、评价和主观的态度感受[6]。利用意见挖掘技术，对具有情感倾向性的文本进行分析，其目的是将文本分为不同的情感类别，如积极的和消极的。

情感分析是一个极其复杂的问题，涉及到计算机科学、语言学、心理学、社会学和其他学科。文本情感分析方法旨在检测书面评论的情感，从而更好地理解公众对实体的看法。

1.2 国内外研究现状

文本情感分析是结合文本的上下文内容判断文本的整体情感倾向，是自然语言处理领域中不可或缺的分支[7]。情感分析即判断文本的观点是积极的，消极的还是中性的。从文本颗粒度来看，可以分为三个级别，分别是属性级、句子级和篇章级[8]。从情感分类粒度来看，分为主客观分类，即有无情感分类、情感倾向性分类，又称情感极性分类和细粒度情感分类，即多情感分类[9]。从技术方法来看，分为基于情感词典、基于机器学习和基于深度学习三类算法。

1.2.1 基于情感词典的情感分析方法

情感词典是传统的情感分析方法，虽然方法简单但是分类方法机械化且需要依赖人工构建情感词典，对前期建立的词典的完整性要求较高。目前，广泛应用的中文词典包括中国知网发布的HowNet情感词典、台湾大学整理的NTUSD词典以及清华大学李军中文褒贬义词典[10]。

李胜宇等人利用自建的酒店评论领域专用情感词典结合评论的句式特征以及语法特点，提出了一种面向酒店评论领域的情感分析模型，有效解决了通用情感分析模型应用在酒店评论领域情感分析不准确的问题[11]。B.Liu等人提出了多种构建情感词典的方法，通过对文本情感倾向的标注和匹配，实现了文本情感类型的计算和匹配[12]。张成功等人构建了专用于情感分析的极性词典，利用修饰词和极性词计算词句极性强度，实验证明自建的极性词典情感分析效果优于通用的情感词典[13]。王志涛等人提出了基于词典和规则集的微博情感分析方法，利用自建的微博情感词典计算微博文本的综合情感，通过实验证明了方法的有效性[14]。周咏梅等人构建了基于HowNet和SentiWordNet的中文情感词典，利用支持向量机模型构建文本情感分类器并将其应用于微博文本情感分析任务中，证明自建词典优于通用情感词典[15]。王灿伟提出了结合情感词典和感情符号计算微博情感值的方法，实现主题归类和情感分类[16]。

1.2.2 基于机器学习的情感分析方法

随着信息量的不断增长，以及新型词汇的不断出现，基于情感词典的情感分析方法存在局限性，分类效果欠佳，于是基于机器学习的文本情感分析方法应运而生。基于机器学习的方法主要有基于贝叶斯统计、最大熵、支持向量机等模型，这些方法主要是手动进行文本特征提取，对人工标注具有较强的依赖性。

Pang B等人使用朴素贝叶斯算法、最大熵分类以及支持向量机模型进行文本情感分类研究，最早地使用机器学习方法进行电影评论数据的情感分析，证明基于机器学习的情感分析方法较基于情感词典的方法效果更优[17]。徐军等人使用朴素贝叶斯模型和最大熵方法对新闻文本情感分类进行对比研究，同时发现选择具有情感语义倾向的词语作为特征项能够提高文本分类的准确率[18]。周杰等人利用支持向量机模型对网络新闻评论进行情感分类研究，发现评论中加入更多支撑评论者观点的论据词语能够提升分类性能[19]。姚娜娜等人在产品评论情感分类研究中通过提取特征项构建特征空间，并使用信息增益和互信息对特征项进行筛选，最后使用LIBSVM分类器对文本进行正负极性分类[20]。张庆庆等人运用支持向量机分类器对微博评论文本进行分类，结合了基于评价对象的特征，通过依存关系树找到影响评价对象情感表达的依存关系[21]。Bouarara H A.运用朴素贝叶斯和K临近算法对精神障碍患者在社交媒体的行为进行分析，研究社交媒体对心理健康的负面影响[22]。

1.2.3 基于深度学习的情感分析方法

传统的机器学习方法无法学习到文本中的深层次语义信息，进而导致部分情感分析任务无法完成。然而，深度学习是机器学习衍生出来的一个分支，基于深度学习的情感分析方法通过搭建神经网络模型分析文本内部信息以解决特征提取问题[23]。基于深度学习的情感分析方法不断发展更新，比如基于传统RNN、LSTM等模型。该方法能够更好地表示特征从而提高分类准确率，并且避免传统的情感分析方法带来的局限性，在对文本进行情感分析时通过学习不断优化模型。

Al-Dabet S等人提出了一种基于卷积神经网络和叠加独立长短期记忆的类别识别模型，采用堆叠的双向独立长短期记忆、位置加权机制和多关注机制层的分类模型，实现了分类准确率的提高[24]。Terra Vieira S等人利用基于卷积神经网络和带有软根符号激活函数的双向长短期记忆-递归神经网络的情感分析模型对微弱信号的主题分类准确率达到97%[25]。Pasupa K等人融合了卷积神经网络和双向长短期记忆等算法进行组合比较，证明特征组合和混合深度学习算法可以提高整体性能[26]。Lin X M等人采用长短时记忆神经网络和注意力机制相结合的实验模型，实现了以较高的准确率完成文本情感分析任务[27]。Zhang F等人提出了一种基于深度学习和机器学习的电影评论情感分析技术，其准确率达到83.13%，证明了该方法的实用性和可扩展性[28]。Muppidi S等人采用了一种结合显著特征集的监督分类器，并对模型的性能进行了比较，证明了联合CNN-LSTM深度神经网络模型准确率达到85%[29]。

1.3 课题主要研究内容

根据面向新疆旅游的社交媒体评论情感分析的研究需要，研究了文本分词、核心主题词提取、BERT预训练模型、情感分析等相关内容，给出了文本情感分析的总体设计。其中BERT是一种自然语言处理领域的预训练模型，是基于Transformer模型的改进，采用双向Transformer中的Encoder模块进行模型的搭建，利用该层对输入的评论文本进行特征提取，舍弃传统的文本处理方法，能够有效解决模型无法并行处理以及文本的长期依赖问题。

本课题基于BERT预训练模型实现高准确率分类新疆旅游评论文本信息，通过提取上下文关联信息对评论进行情感倾向分析，总结新疆旅游景点的不足之处，为景区负责人提供合理的改进建议。

1.4 章节安排

第一章：绪论 首先介绍了旅游评论情感分析的重要性和社会需求量，得出开发面向旅游的社交媒体评论情感分析的必要性。通过分析目前文本情感分析存在的问题和技术难点，以及各情感分析方法的研究现状，从而给出课题所需要研究的内容，以及所涉及的技术。

第二章：需求分析 从旅游业的角度，分析当今时代人们的消费观念，以发觉加强旅游景点建设的重要性。结合各城市景区级别数量统计，得出新疆旅游景区有待改进的相关需求。通过分析各大旅游网站的用户评论内容，确定本课题数据集来源途径及处理方法。最后，通过对比前沿的文本情感分析算法，决定采用BERT模型完成本课题的评论情感分析任务。

第三章：数据获取与预处理 介绍了本课题数据的来源以及获取方式，再对数据进行初步的预处理为接下来的文本情感分析做准备。

第四章：文本特征词提取 根据数据集的高频词统计数据归纳本课题的文本特征词和相应核心词汇，利用正则表达式将数据集按照主题词分类。

第五章：文本情感倾向分析 首先建立模型对数据集进行初步训练，然后给出情感倾向分析模型的构建以及算法原理。

第六章：数据可视化分析 针对每个景点的主题词分析游客的情感倾向，有针对性地对新疆各景点建设提出改进建议。

第2章 需求分析

2.1 行业需求

随着国民经济的发展，居民收入以及消费水平也有所提高，整体消费结构随之优化，旅游业的发展日益壮大。吴孝政等人将消费环境分为软环境和硬环境两类，提出在消费活动中，政府消费政策、消费心理、传统文化、基础设施等会对消费行为产生影响的要素也应该在消费环境中考虑到[30]。总的看来，尽管人们对消费环境的评判标准并不统一，但都是指景点中对消费者的消费行为产生重要影响的、外在的、客观的因素。通过分析游客对景点的游玩体验感受，找到景点在管理服务中存在的问题与短板，有助于景点管理者更有针对性地优化景点服务，从而加强景区建设。因此，创建良好的消费环境，是促进中国现阶段经济内循环和区域经济高质量发展必不可少的前提条件。

由于旅游业不仅能够引进人才还对其他产业有一定的推动作用，全球各地都在大力发展旅游业，旅游业已成为国民经济新的增长点。因此，本课题基于BERT模型对新疆旅游的相关评论进行情感分析，对潜在游客的旅游计划以及各景区未来发展提供建议。

2.2 市场需求

景区塑造完美的旅游形象才能在激烈的旅游业市场角逐中脱颖而出。良好的景区旅游形象不仅能够为游客们提供满意的服务体验，也能扩展整个城市的旅游市场，是最为有力的宣传依据。每个游客的关注点不同，那么旅游后也可能产生不同的旅游体验感受。因此，从不同的关注点角度分析游客发布的评论，挖掘其中蕴藏的情感能够帮助景区有针对性地提升服务。

景点在多年运营过程中也会面临游客或多或少的不满和意见，传统的方法主要是通过访谈、问卷调查等方法了解游客意见。而随着互联网的普及和发展，用户通过互联网在社交媒体平台自由发布评论的行为日益频繁且普遍，这些评论中包含着用户自由发表的评价、观点、看法、情感、立场等等。为满足游客对旅游信息服务的需求，各大综合性旅游网站应运而生，诸如携程、马蜂窝、驴妈妈、美团、途牛等。这些网站提供景区门票、度假酒店、周边游、出境游等服务，最重要的是用户能够在游玩后发表体验感受。同时，越来越多的用户也会在出行旅游前查看景点的相关评论，从而相应地更改出行计划。因此，通过分析旅游网站的用户评论能够获取用户对某一景点环境、服务、体验感等方面的喜好、态度，可以帮助景点管理人员改进景区服务态度、硬件设施等提供必要的决策依据，从而提升旅游景点的旅游品质，更好地满足游客需求，在激烈的市场竞争中取胜[31]。

2.3 功能需求

对评论的情感倾向分析光通过人工判断不仅工作量大而且分类标准无法得到统一，因此情感分析工作就需要利用相关技术提高分类准确率。现有的情感分析方法主要有基于情感词典、基于机器学习以及基于深度学习的方法，通过对比实验发现深度学习方法更加能够充分利用文本上下文内容，分类效果更佳。而基于深度学习的分类方法又细分为无监督学习、有监督学习以及半监督学习，对比发现半监督学习仅需要标记少量数据并适用于数据量大的任务，工作量较小效果最佳。本课题利用半监督学习的方法，通过协同训练和主动学习，训练实体识别模型。由于社交平台上的文本内容具有特征稀疏，错误噪声多，用语不规范等特征，因此采用BERT模型训练数据集，能够提取评论核心关注点，消除噪音信息。

第3章 数据获取与预处理

3.1 数据选择

在互联网经济飞速发展的现代社会，为满足游客对旅游信息服务的需求，各大综合性旅游网站应运而生，诸如携程、马蜂窝、驴妈妈、美团、途牛等。旅游网站下的用户评论能够真实地反应游客的游览感受，通过分析用户评论数据能够对景点的改进和潜在游客出行计划提供帮助。这些网站的用户评论数量和质量各不相同，需要从中选择合适的评论内容作为本课题的数据集。通过综合考虑网站的知名度和访问量后，本课题确定美团网中关于新疆旅游景点的用户评论作为本课题的数据集，同时将用户评论按照每个景点获取。

3.2 数据爬取

本课题以新疆旅游景点作为研究对象，利用Python爬虫技术获取美团网中新疆最受欢迎的前18个景点的用户评论数据，总共10000余条，评论数据爬取流程图如图3.1所示。



图3.1 数据爬取流程图

由于美团对数据采用了反爬虫技术，只能抓取网站页面数据，并且还需要不断变换电脑ID才可以成功获取评论数据。最后，将爬取到的评论数据以xlsx格式存储到本地如图3.2所示。

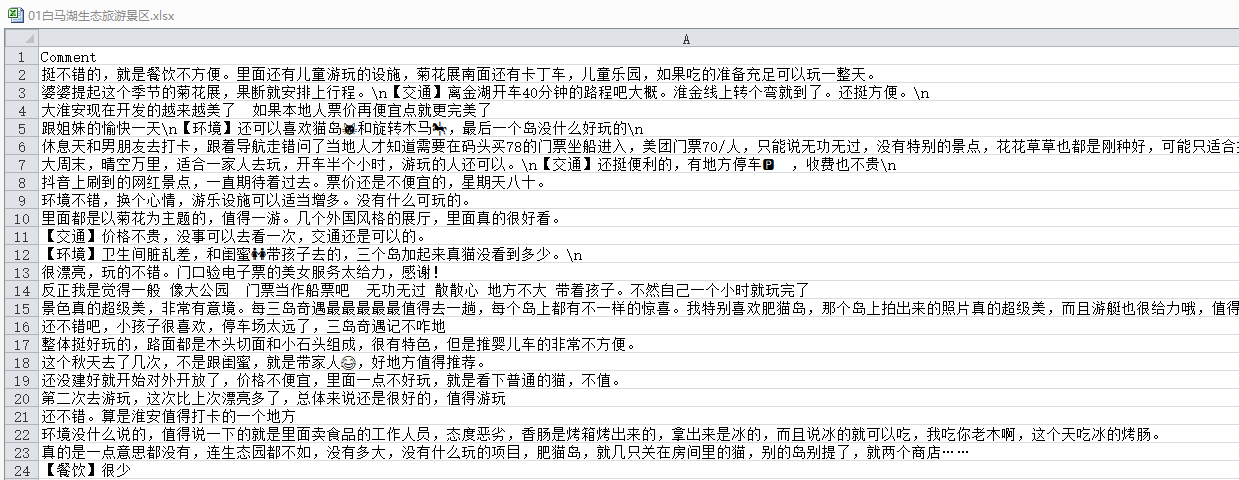


图3.2 评论数据部分截图

3.3 数据清洗

由于爬取到的评论数据格式不具规范性，其中可能存在一些重复数据、空格、表情符号和广告信息等，为了提高文本分析的可信度和准确率，需要对评论数据进行数据清洗以规范数据格式。具体清洗规则如下：

（1）删除空白评论

虽然空白评论不影响后续的文本情感倾向分析工作，但是会对最终的景点好评率结果产生影响。为了保证最终好评率的准确性，需要在数据预处理阶段将空白评论删除。

（2）删除重复评论

商家为提高好评率可能存在刷好评的行为，虽然重复评论不会对文本情感倾向分析工作产生影响，但是会影响最终的景点好评率结果。为了保证最终好评率的公平性，需要在数据预处理阶段将重复评论删除。

（3）删除广告信息

评论区出现的与景点无关的广告信息不仅给文本情感倾向分析工作带来麻烦，也影响最终的景点好评率结果。通过识别广告信息的特征，在数据预处理阶段将内容为网站链接的评论删除。

3.4 文本分词

分词是文本处理的必要步骤，为后续的文本分类任务做基础。因此，本课题使用Python中的jieba库中的精准分词函数lcut()对评论文本进行分词。同时，评论数据中也包含很多对情感分析无用的词语，诸如虚词、感叹词等，这些词语在文本数据中不仅占用空间还浪费计算时间，因此需要将这些无用的词语过滤。本课题利用哈工大停用词库对所爬取的评论数据进行进一步处理，通过观察分词结果不断更新停用词库，使得分词结果更加精准。本课题使用的停用词库如图3.3所示。

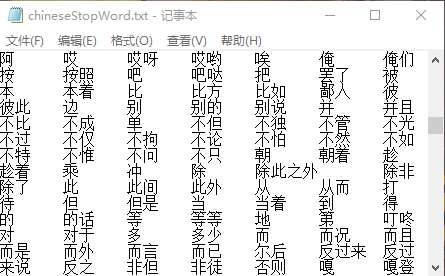


图3.3 停用词库

第4章 文本特征词提取

4.1 词频统计

统计词语在整个景点评论数据中出现的频率能够反映游客对景点评论的核心关注点，从而有针对性地改进景点服务。通过对评论文本词频统计得到如表4-1所示的高频词汇统计表。

表4-1 高频词汇统计表

|  |  |
| --- | --- |
| 词语 | 频数 |
| 玩 | 2904 |
| 环境 | 920 |
| 项目 | 833 |
| 值得 | 712 |
| 表演 | 622 |
| 设施 | 620 |
| 好玩 | 525 |
| 门票 | 492 |
| 价格 | 437 |
| 贵 | 410 |
| 服务 | 385 |
| 推荐 | 338 |
| 方便 | 276 |
| 不值 | 267 |
| 钱 | 250 |

词频统计表数据表明，游客们总体核心关注点相似。“玩”“项目”“表演”“设施”四个词语出现频数共计4979，可以看出游客对景区基础设施的娱乐性、趣味性等重视程度较高。“环境”“值得”“服务”“推荐”“不值”五个词语出现频数共计2622，可以看出游客对景区游览体验感受也十分重视。“门票”“价格”“贵”“钱”四个词语出现频数共计1339，可以看出游客对价格比较敏感，景区门票价格是游客出行的必要参考因素。因此，景区应把上述几个方面作为提升改进的重点方向。

4.2 主题词特征提取

评论内容的主题是指评论所描述的主要内容，往往反映景点的某些特征，一条评论往往涉及一个或多个主题[31]。主题词特征提取属于文本分类的核心步骤，是影响文本分类的关键任务。主题词的选择需要根据某种评价标准对数据集按照重要性排序，从而选择具有代表性的词语作为特征主题词[32]。结合TF-IDF计算评论文本内容的特征权重，根据某词语在文本中出现的次数与其在整个数据集中出现的次数相乘确定该词的重要性，从而筛选出使用频率高的词语，保留其中有意义的词语，计算公式如下：

 （1）

其中，TF-IDF值越大，表示该特征词在这个文本中越重要。TF指某关键词在该文本中出现的频率，计算公式如下：

 （2）

TF*i,j*表示关键词*j*在文本*i*中出现的频率，n*i,j*表示关键词*j*在文本*i*中出现的次数。IDF指逆文本频率，即该关键词在整个数据集中出现的次数取倒数，计算公式如下：

 （3）

IDF*i*表示词语*i*的逆文本频率，|D|表示所有文本的总数，表示出现词语i的文本总数，之所以加1是为了避免分母为0的情况。

本课题通过词频统计与分析并结合特征权重计算确定游客对于旅游景点评论的五个主题：门票、游览、交通、娱乐和饮食。通过对评论主题词分类，可以获得对景点更加丰富和全面的信息。在确定游客评论主题的基础上，本课题列出每个主题对应的核心特征词汇如表4-2所示，按照情感词典扩充的方法，得到每一个主题的主题词库。

表4-2 主题对应核心特征词汇

|  |  |
| --- | --- |
| 主题 | 核心词汇 |
| 门票 | 门票、票价、价格、性价比 |
| 游览 | 环境、服务、态度、景色、值得 |
| 交通 | 交通、停车场、路、距离、位置 |
| 娱乐 | 娱乐、项目、设施、设备、趣味性 |
| 饮食 | 好吃、难吃、卖、口味、餐饮 |

本课题采用正则表达式匹配主题词核心词汇的方式来完成评论文本的主题分类，其方法工作量小且简单易于操作。在数据表中增加type标签，规定评论主题特征若为“门票”则标记该评论type标签为“1”，若为“游览”则标记该评论type标签为“2”，若为“交通”则标记该评论type标签为“3”，若为“娱乐”则标记该评论type标签为“4”，若为“饮食”则标记该评论type标签为“5”，评论主题特征提取后如图4.1所示。

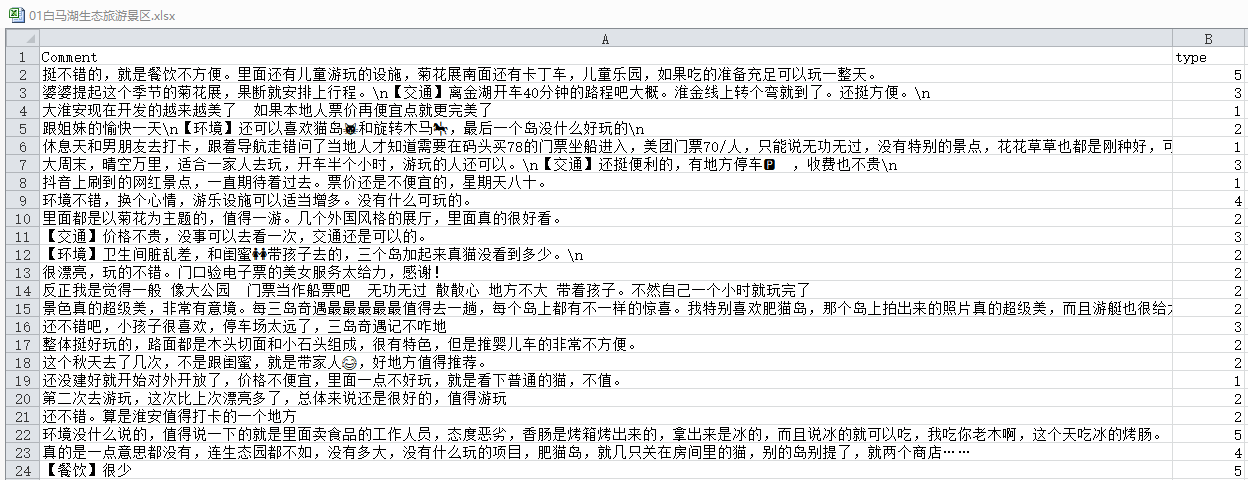


图4.1 评论主题特征提取后部分截图

第5章 文本情感倾向分析

5.1 标注情感标签

构建情感分析语言模型之前需要对数据集的情感倾向做初步的人工标注，本课题采用二分类的方式将评论分为正向评论与负向评论，增加sentiment标签，规定正向评论标注为“1”负向评论为“0”。本课题利用百度AI开放平台软件进行情感倾向的初步标注，调用百度API计算10000条评论数据的情感倾向。百度AI在进行情感分析时，无法识别带有表情字符的评论，需要跳过这类特殊字符。最后，人工检查数据集分类结果纠正错误分类，情感标注后数据集如图5.1所示。



图5.1 评论情感标注后部分截图

5.2 模型构建

5.2.1 Attention机制

Attention机制目的是从大量评论文本中筛选出对当前任务目标更关键的信息内容，即动态优化特征权重。在训练词向量过程中，Attention机制为每一个*token*分配特殊的权重，再将其应用到下面相应的任务当中[33]。通常情况下会在Encoder-Decoder模型中使用Attention机制，其数学原理如图5.2所示。



图5.2 Attention机制数学原理图

已知在Decoder上一时刻的隐藏层输入，将所有*token*输入Encoder中获得对应的词向量，接着利用F函数和Softmax函数获得相应的权重，对每个词向量的lable产生不同的影响权重，最后获得的输出值为。

5.2.2 Transformer模型

BERT的核心部分是Transformer模型，该模型是一个基于自注意力机制利用编码器和解码器的序列到序列体系结构。BERT模型利用Transformer编码器作为特征提取器，主要关注输入文本中每个词之间的关联度[34]。该编码器由多个Transformer层组合叠加而构成，包含两个子层，分别是基于自注意力机制的自注意力层和全连接前馈神经网络层，其主要结构如图5.3所示。



图5.3 Encoder结构

Encoder结构先输入评论文本的词嵌入，再计算该评论中每个词语的位置信息之和，利用自注意力层帮助Encoder查看每个词编码时对应的上下文信息。接着传输至Add&Norm层，利用Add将自注意力层的输入和输出结果相加，利用Norm对相加后的输出结果进行归一化，从而得到的向量列表传输至全连接的前馈神经网络。在该神经网络中，通过相应的Add&Norm层处理，以最终输出全新的归一化后的词向量列表。该向量列表能有效地学习每个词语的上下文信息，从而不断优化获得更好的词向量表示。

缩放点积注意组成多头自注意力模型作为其基本单位，缩放注意力计算公式如下：

（4）

输入维度为的查询向量*Q*，被查询向量*K*以及实际特征向量*V*。计算*Q*和所有的点积得到某词与其他位置上词的关联程度，值越大表示两个词越匹配，并将每个值除以以防止分子数值过大。利用Softmax函数计算该词与其他位置上词关联程度的概率值，从而得到加权融合结果，最后乘以实际特征向量*V*得到最终带有上下文信息特征的向量值。

假设对矩阵*Q*、*K*和*V*进行线性计算，将其转换成与本身不同的矩阵，需要利用缩放点积单元进行计算，将缩放点积单元计算得到的结果组合，从而能够得到更多的语义特征，计算公式如下：

（5）

由于每个缩放点积单元的输入*Q*、*K*和*V*在进行计算时的权重各不相同，因此能够获得每个文本不同空间的语义特征，以捕捉每条句子中的每个词关于上下文的信息，计算公式如下：

（6）

前馈网络层由两个大小为1的卷积核组成，采用激活函数ReLu做两次线性转换，计算公式如下：

（7）

其中和是权重，和为偏置项，通过残差网络和层归一化最终得到输出值。

Transformer编码器网络使用*n*个编码器抽取文本内容的特征，再将学习到的分布式特征使用全连接层将特征整合映射到样本标记空间，以便进行下一步的文本情感倾向性分析。

5.2.3 BERT

根据已有研究对比实验，本课题采用BERT语言模型训练数据集，通过33亿单词以及25亿维基百科和8亿文本语料的大型语料库进行预训练提高其语义提取能力。BERT是由谷歌公司提出的一种基于双向Transformer改进的通用语言表示模型，通过放缩点积注意力与多头注意力的方式直接获取语言单位的双向语义关系。BERT模型基于self-attention机制捕捉上下文约束信息条件，生成文本语义级别的词向量。通过结合文本中单词之间的联系消除词与词之间距离对文本分析的干扰，将某个词语与该句中其他词语之间的依赖关系显性表示，更好诠释文本的语义内容，同时还可以并行处理节约运行时间，BERT模型结构如图5.4所示。其中，表示词语的文本输入，经过双向Transformer编码器获取文本的向量化结果。



图5.4 BERT模型结构图

为了利用BERT模型完成评论文本的情感分析，在BERT模型中输入两个相连句子的序列，在每个句子序列首部添加一个标识符[CLS]表示句子开始，尾部添加一个标识符[SEP]表示结束，两个句子使用[SEP]隔开。对于每个词语，BERT模型进行了三种不同的嵌入操作，分别是对词语位置信息进行编码、对词语进行word2vec编码、对整个句子进行编码。将三种嵌入操作结果进行向量拼接，可以得到如图5.5所示的BERT输入序列。

图5.5 BERT模型输入序列

将注意力层的输出作为全连接层的输入进行分类，对于景点评论内容的情感倾向分析任务，将评论情感类别分为两类，使用归一化的Softmax函数计算情感倾向判别结果，其计算公式如下：

（8）

其中，表示评论的情感倾向输出概率，表示权重，*υ*表示注意力层的输出，表示偏置项。

评论情感倾向分析的目标函数为交叉熵损失函数，其计算公式如下所示：

（9）

其中，*S*表示评论数量，*M*表示所有景点情感倾向数量，表示第*s条评论*真实的情感倾向，表示第*s*条评论预测的情感倾向。

5.3 实验设计

5.3.1 实验数据

实验采用了本课题从美团网上爬取到的新疆旅游景点的用户评论数据共计10009条作为数据集，用于统计新疆各个景点各方面的游客满意度。其中包含18个景点的游客评论数据，将评论按照五种主题分类，分类占比如图5.6所示。其中，类别“1”表示“门票”占14.16%，类别“2”表示“游览”占50.7%，类别“3”表示“交通”占3.95%，类别“4”表示“娱乐”占28.53%，类别“5”表示“饮食”占2.76%。

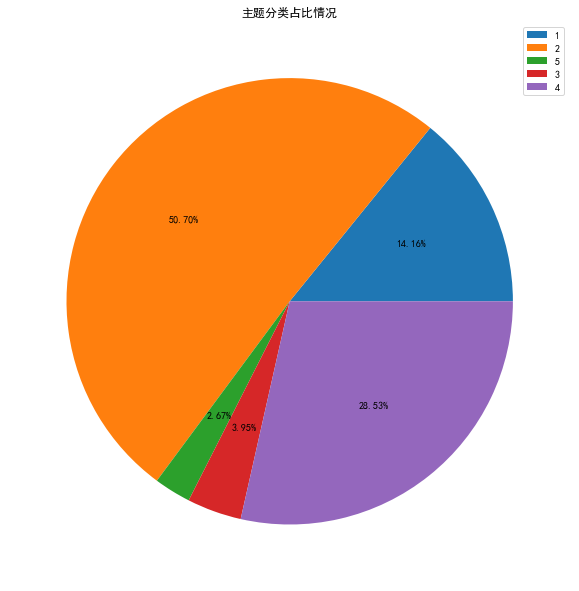


图5.6 主题分类占比情况

通过调用百度API和人工检查的方式对评论数据的情感倾向进行标注整理，得到正向评论条6455，消极评论3553条，选取8900条评论作为训练集，剩余1108条评论作为测试集。

5.3.2 评价标准

实验使用的BERT模型运用了准确率实验评价指标，用来衡量评论情感分析结果，即预测标签与实际标签吻合的评论所占比率，计算公式如下所示：

 （10）

其中，*TP*为正向评论预测为正，*FP*为负向评论预测为正，*FN*为正向评论预测为负，*TN*为负向评论预测为负。

5.3.3 参数设置

epochs=1，学习率=0.01，batch=64，词向量维度=768，句子长度=50。

5.3.3 实验结果

在准确率评价指标上得出的分类结果如图5.7所示，从实验结果可以看出，在整体准确率方面，基于预训练的迁移学习模型达到96.38%。



图5.7 模型准确率截图

第6章 数据可视化分析

6.1 数据可视化

本课题对游客在社交媒体平台发布的关于新疆旅游景点的评论进行情感分析，从门票价格、游览体验、交通、娱乐设施以及餐饮五个角度判断评论的情感倾向。利用各角度的好评率来衡量游客对该旅游景点各项服务的满意程度，帮助景点负责人有针对性地加强景点建设，优化游客游览体验感受。主题好评率计算公式如下：

 （11）

针对新疆6个热门景点，从门票价格、游览体验、交通、娱乐设施以及餐饮五个方面统计好评率。根据景点的主题好评统计数据，能够直观地了解游客对每个景点各方面的满意程度，从而更有针对性地整改景点服务。

（1）霍城

参考如图6.1所示的霍城主题好评率统计图，景区票价略高且餐饮价格制定不合理，同时景区位置偏僻交通不便，引起游客诸多不满。

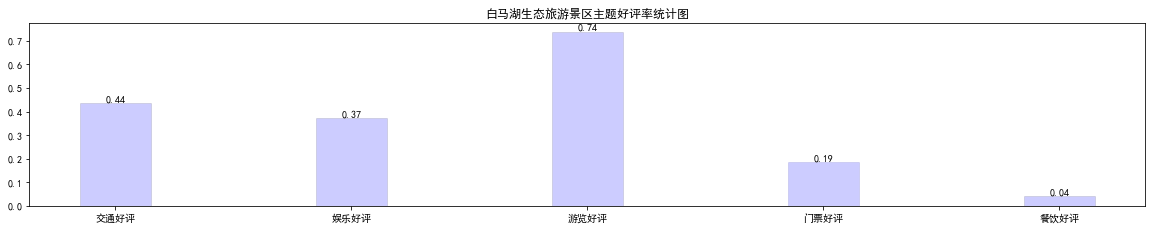


图6.1 霍城主题好评率统计图

（2）乔木滑雪场

参考如图6.2所示的乔木滑雪场主题好评率统计图，场内的餐饮得到了游客们的一致好评，但滑雪教练费用过高且服务态度恶劣使得游客极其不满，景点性价比不高。

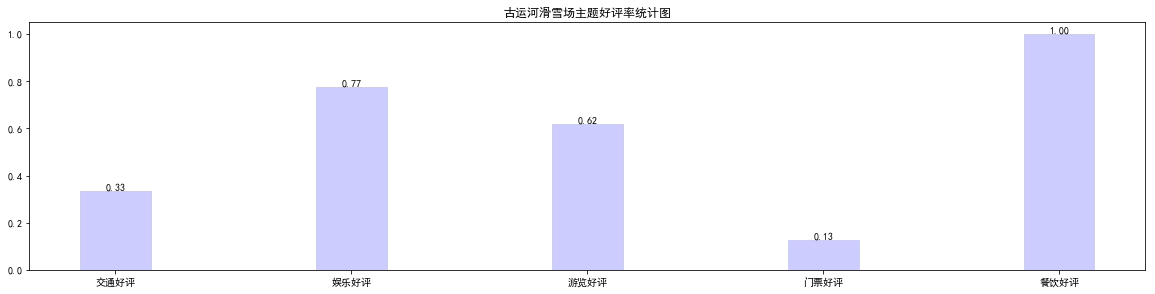


图6.2 乔木滑雪场主题好评率统计图

（3）可可托海

参考如图6.3所示的可可托海主题好评率统计图，作为历史文化景点门票价格过高会引起游客不满，同时景点还需要保证设施的完整性，保存其历史价值。

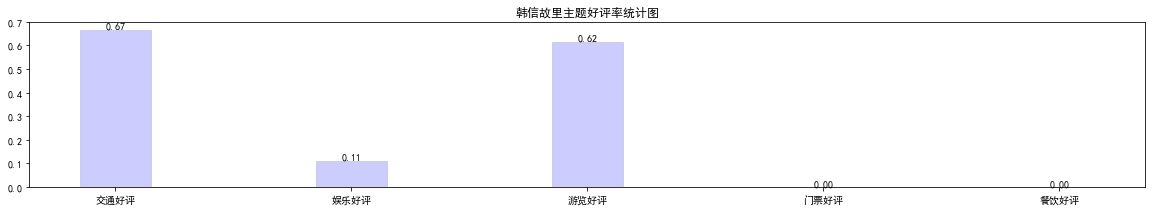


图6.3 韩信故里主题好评率统计图

（4）五彩滩景区

参考如图6.4所示的五彩滩景区主题好评率统计图，若门票再适当优惠应会吸引更多游客前来观光体验。

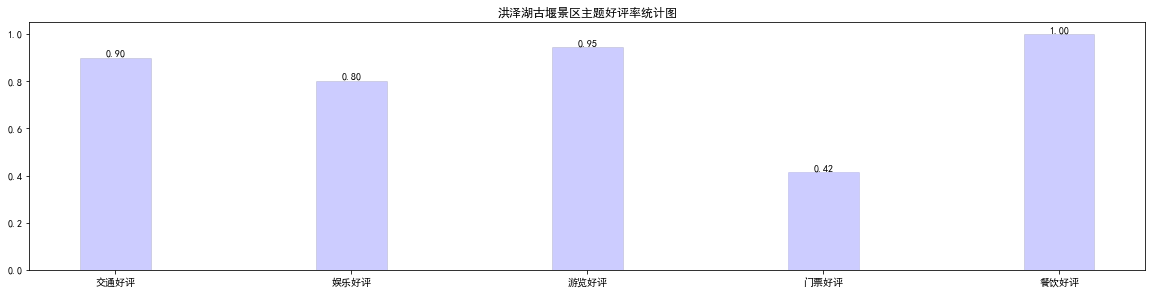


图6.4 五彩滩景区主题好评率统计图

（5）那拉提景区

参考如图6.5所示的那拉提景区主题好评率统计图，由于景区位置偏远交通不便，景色过于单调且门票性价比低，游客对景区的满意程度较低。

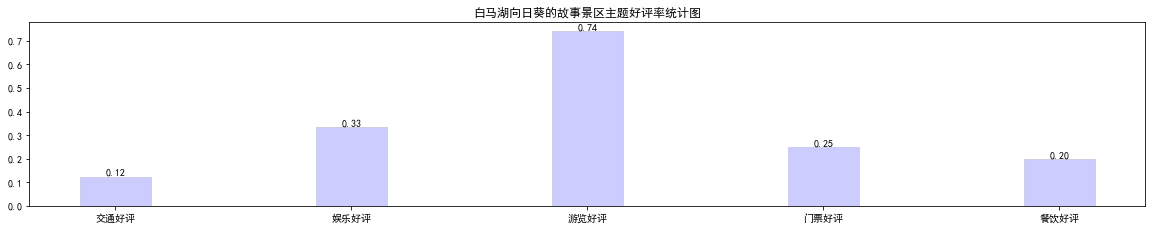


图6.5 那拉提景区主题好评率统计图

（6）博斯腾湖

参考如图6.6所示的博斯腾湖主题好评率统计图，景点门票不宜过高，停车场应尽量安排专人管理。

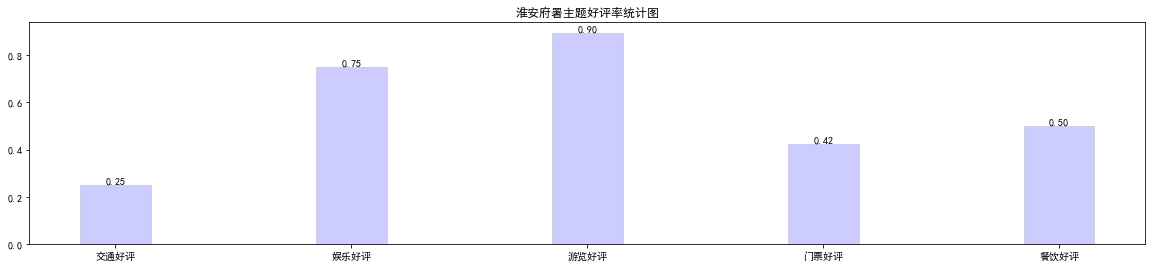


图6.6 博斯腾湖主题好评率统计图

6.2 景点改进建议

新疆旅游的综合形象是自然资源丰富，景区各具特色，历史文化底蕴浓厚。依据实验数据可视化结果，得到如表6-1所示的新疆景点主题好评率汇总表。

表6-1 景点主题好评率汇总表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 景点 | 门票好评率 | 游览好评率 | 交通好评率 | 娱乐好评率 | 餐饮好评率 |
| 霍城 | 0.19 | 0.74 | 0.44 | 0.37 | 0.04 |
| 乔木滑雪场 | 0.13 | 0.62 | 0.33 | 0.77 | 1.00 |
| 可可托海 | 0.00 | 0.62 | 0.67 | 0.11 | 0.00 |
| 五彩滩景区 | 0.42 | 0.95 | 0.90 | 0.80 | 1.00 |
| 那拉提景区 | 0.25 | 0.74 | 0.12 | 0.33 | 0.20 |
| 博斯腾湖 | 0.42 | 0.90 | 0.25 | 0.75 | 0.50 |

针对实验的6个新疆景点，分析表中数据可知，

根据好评率汇总表数据，游客对新疆大部分旅游景点的游览体验较为满意，但也有个别方面坐的不尽人意，需要有针对性地进行景区整改。表6-2根据数据分析结果为各旅游景点提供合理建议，加强景区建设。

表6-2 景点改进建议

|  |  |
| --- | --- |
| 景点 | 建议 |
| 霍城 | 合理制定票价  加强停车场管理力度  增加景区创新点  增加平价餐厅 |
| 乔木滑雪场 | 合理制定教练费用  精准导航定位 |
| 可可托海 | 合理制定票价  定期维护设施  增加商铺 |
| 五彩滩景区 | 合理制定园内项目费用 |
| 那拉提景区 | 合理制定票价  增加公交车直达路线  增加景区创新点 |
| 博斯腾湖 | 合理制定票价、加强停车场管理力度 |

结论

本文针对文本情感倾向分析的问题，提出了一套面向旅游景点评论的文本情感分析技术方案，该方案能够直观地得到旅游景点在各方面的游客满意度，挖掘旅游网站游客评论的价值信息。

相比于传统的文本情感分析模型，本文采用基于深度学习情感分析方法中的BERT模型作为预训练模型，对经过预处理的新疆旅游评论数据进行训练和预测，引入attention机制不断提高模型预测的准确率。采用开源可视化工具将实验数据嵌入到柱状图和饼状图中进行展示，同时通过调整参数实现对组件的改进。

从实验测试结果来看，本文提出的技术方案是可行的，利用相关技术能够将网上碎片化的用户生成数据转化为对潜在游客和景区负责人有用的直观信息，以辅助游客制定出行计划，帮助景区管理者了解游客意见，进而优化旅游服务加强景区建设。此外，本文的研究方案可以为旅游研究领域自动地获取游客评论提供借鉴，也可以为旅游评论观点抽取系统的构建提供参考。

参考文献

1. 马欣.电商评论情感分析及销量预测方法研究[D].大连海事大学,2019.
2. 李胜宇, 高俊波, 许莉莉. 面向酒店评论的情感分析模型[J]. 计算机系统应用, 2017, 26(1): 227-231.
3. Liu B, Zhang L. A survey of opinion mining and sentiment analysis[M]//Mining text data. Springer, Boston, MA, 2012: 415-463.
4. 张成功, 刘培玉, 朱振方, 等. 一种基于极性词典的情感分析方法[J]. 山东大学学报 (理学版), 2012, 47(3): 47-50.
5. 王志涛, 於志文, 郭斌, 等. 基于词典和规则集的中文微博情感分析[J]. 计算机工程与应用, 2015, 51(8): 218-225.
6. 周咏梅, 杨佳能, 阳爱民. 面向文本情感分析的中文情感词典构建方法[J]. 山东大学学报 (工学版), 2013, 43(6): 27-33.
7. 王灿伟. 基于主题提取的海量微博情感分析[J]. 南京大学学报 (自然科学版), 2017, 53(3): 549.
8. Pang B, Lee L, Vaithyanathan S. Thumbs up sentiment classification using machine learning techniques[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP).Stroudsburg: Association for Computational Linguistics, 2002: 79.
9. 徐军, 丁宇新, 王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报, 2007, 21(6): 95-100.
10. 周杰, 林琛, 李弼程. 基于机器学习的网络新闻评论情感分类研究[J]. 计算机应用, 2010, 30(4): 1011-1014.
11. 姚娜娜. 基于机器学习的产品评论情感分类研究 [D]. 首都师范大学, 2013.
12. 张庆庆, 刘西林. 基于机器学习的中文微博情感分类研究[J]. 未来与发展, 2015 (4): 59-63.
13. Bouarara H A. Sentiment Analysis Using Machine Learning Algorithms and Text Mining to Detect Symptoms of Mental Difficulties Over Social Media[J]. International Journal of Information Systems and Social Change, 2021, 12(2): 1-15.
14. Xu F, Zhang X, Xin Z, et al. Investigation on the Chinese text sentiment analysis based on convolutional neural networks in deep learning[J]. Comput. Mater. Continua, 2019, 58(3): 697-709.
15. Al-Dabet S, Tedmori S, Mohammad A L S. Enhancing Arabic Aspect-Based Sentiment Analysis Using Deep Learning Models[J]. Computer Speech & Language, 2021: 101224.