

基于RoBerta-BiGRU-Attention的景区评论情感分析研究

——以沈阳市为例

项瑞良, 李征宇, 孙平

沈阳建筑大学, 计算机科学与工程学院, 辽宁 沈阳

收稿日期: 2023年8月30日; 录用日期: 2023年9月30日; 发布日期: 2023年10月9日

摘要

游客在评论中所表达的意见和感受, 能够直观地反映他们对旅游目的地的评价, 同时语气鲜明、情感强烈。由此带来的巨大而动态的信息空间需要消费者和产品/服务提供者共同理解和导航。针对目前景区在线评论文本情感分类准确性不高的问题, 提出一种基于RoBerta词向量和双向门控循环单元(BiGRU)的改进模型, 使用能够表征文本丰富语义特征的Roberta模型进行词向量表示, 结合能够长期保留文本上下文关联信息的BiGRU神经网络提高模型的分类效果, 并在此基础上引入注意力(Attention)机制, 突出文本中更能表达分类结果的情感词权重, 提高情感分类的准确率。将上述模型分别在沈阳市3个景区评论数据上进行情感极性分类和预测, 实验结果表明, 该模型在各数据集上都获得了良好的性能。同时结合LDA主题模型分析, 得到游客评论的期望和诉求, 为沈阳市旅游业发展提供技术支撑以及未来发展意见。

关键词

RoBerta词向量, BiGRU, 注意力机制, 情感分析, 潜在狄利克雷分布

Research on Sentiment Analysis of Scenic Area Comments Based on RoBerta-BiGRU-Attention

—Taking Shenyang as an Example

Ruiliang Xiang, Zhengyu Li, Ping Sun

School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang Liaoning

Received: Aug. 30th, 2023; accepted: Sep. 30th, 2023; published: Oct. 9th, 2023

文章引用: 项瑞良, 李征宇, 孙平. 基于 RoBerta-BiGRU-Attention 的景区评论情感分析研究[J]. 数据挖掘, 2023, 13(4): 312-326. DOI: 10.12677/hjdm.2023.134031

Abstract

The opinions and feelings expressed by tourists in the comments can intuitively reflect their evaluation of the tourist destination, and at the same time, the tone is clear and the emotion is strong. The resulting huge and dynamic information space needs to be understood and navigated jointly by consumers and product/service providers. Aiming at the problem of low accuracy of emotion classification in online comment texts of scenic spots, an improved model based on RoBERTa word vectors and bidirectional gated recurrent units (BiGRU) is proposed. The RoBERTa model, which can represent the rich semantic features of texts, is used to represent word vectors. The BiGRU neural network, which can retain text context information for a long time, improves the classification effect of the model. On this basis, an attention mechanism is introduced to highlight the weight of emotional words in the text that can better express the classification results, and improve the accuracy of emotion classification. The above models were used to classify and predict emotional polarity on the comment data of three scenic spots in Shenyang City. The experimental results show that the model has achieved good performance on each data set. At the same time, combined with the LDA topic model analysis, the expectations and appeals of tourists' comments are obtained, and technical support and future development opinions are provided for the development of Shenyang's tourism industry.

Keywords

RoBERTa Word Vector, BiGRU, Attention Mechanism, Sentiment Analysis, Latent Dirichlet Allocation

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着 Web2.0 时代的到来, 互联网技术在旅游行业的广泛应用, 使得越来越多的游客能够在旅游平台上发布自己的个人体验和旅游评价。大量真实的、独立的、有价值的网络评论数据由此产生, 同时也带来了一个新的方向: 利用这些游客评论文本进行情感分析研究, 获取游客在旅游目的地的感受与体验, 为旅游业服务提供者提供宝贵的意见至关重要。然而, 庞大的数据源也给人工分析带来了很大的困难。自然语言处理技术(Natural Language Processing, NLP)的发展使大规模数据分析变的高效, 从而为景区在线评论情感分析研究提供技术支持。

目前情感分析领域, 主流的方法之一是基于深度学习, 在神经网络进行情感倾向分析时, 通常会采用语言表示模型以考虑不同词汇之间的关系。为此, 预训练的词嵌入模型(如 Word2Vec、GloVe 或 Bert 等)被广泛应用。这些词嵌入模型将词汇映射到高维向量空间, 使得在向量空间中具有相似含义的词具有较近的距离。这种方法有助于捕捉词汇之间的语义关联, 提取句子的特征, 并协助神经网络更好地理解文本的情感倾向。现有的深度学习模型解决了某些问题, 但忽略了其他问题。例如, Chatterjee 等人[1]利用两个预先训练好的单词嵌入和长短期记忆神经网络(Long Short Term Memory Network, LSTM)来提取情绪和语义进行情绪识别, 但他们的模型没有考虑句子不同部分重要性的差异。Rezaeinia 等人[2]改进

了预先训练好的单词嵌入,但没有考虑到过去和未来的上下文的依赖关系和具有不同重要性的单词。为改善这些模型存在的不足,本研究提出了一种新的深度学习模型,用于景区评论的情感分析,该模型考虑句内关系、捕捉过去和未来上下文的依赖关系,并通过分配权重来突出文本的重要部分,选择最重要的特征,增强对重要词汇的关注。然后,结合 LDA 主题模型(Latent Dirichlet Allocation),对沈阳市各景区的实际案例进行了研究。首先,使用自定义 Python 程序爬取游客在各旅游平台上发布的关于沈阳市景区的评论;接着,提出了主要研究框架:通过数据预处理,得到标准化文本数据;同时,运用情感分析获取游客情感倾向的相关性,进一步建立 LDA 主题模型,识别游客各特征维度的情感,研究成果可为沈阳市未来旅游业的可持续发展和创新开辟新的道路。

2. 相关工作

2.1. 情感分析研究

鉴于在线旅游网站上评论的激增以及由此产生的消费者影响,许多学者都在努力探索在线旅游评论与消费者行为之间的关系,以及评论在多大程度上影响消费者的决策和选择[3] [4]。目前倾向于评估旅游在线评论的内容质量的研究仍然稀缺。其中,衡量在线评论语言所传达的情感,特别是判断游客情绪的极性,仍处于起步阶段。

目前,情感分析方法主要分为基于统计的情感分析方法和基于深度学习的情感分析方法两大类。其中,基于深度学习的情感分析模型是通过神经网络学习的,允许网络模型根据上下文信息预测下一个单词的内容,而不依赖于人工标记的语料库,解决了传统基于统计的情感分析方法中的一个关键问题,即人工构建情感词典所需的时间和人力成本。同时,也有效地解决传统情感分析方法忽略语境语义的问题。

在众多深度网络类型中,循环神经网络(Recurrent Neural Network, RNN)在文本处理相关研究中更为常见[5]。虽然 RNN 适用于许多文本情感分析中,但当输入数据中存在长期依赖关系时,它们会出现梯度消失和梯度爆炸。而双向长短时记忆网络(BiLSTM)可以很好的解决长期依赖的关系,能够更好的提取过去和未来的上下文的依赖关系[1]。Cho 等人[6]提出了门控循环单元(Gated Recurrent Unit, GRU),GRU 相较于 LSTM 模型结构更加简单,参数更少,并且能长期保存隐藏层序列信息。尽管深度神经网络在情感分析领域有所进展,但它们仍然存在一些不足之处,不能准确关注文本的重要部分[7]。为了解决这个问题,注意力机制被应用于许多自然语言处理任务,特别是情感分析。注意力机制的灵感源自人类的视觉注意机制,试图将注意力集中在文本中更重要的部分,而非句子的所有单词。近年来,在情感分析领域,词嵌入模型和神经网络结合的方法备受关注。其中,Word2Vec 是常用的词嵌入模型,用于捕捉单词之间的相似性和词汇关系。通过卷积神经网络或长短期记忆神经网络等结构,可以有效地学习文本特征,加上注意力机制则可以提高情感分类的效果。Zhou 等人[8]则提出了一种带有注意力机制的双向 LSTM,以选择重要的特征;同时,Dragoni 等人[9]也提出了一种新的情感分析神经词嵌入方法,能够解决以前方法的主要缺点,并因此能够更好地建模长期依赖关系,获得更高的性能表现。冯兴杰等人[10]在词向量的表示上选用了 Word2vec 模型,利用卷积神经网络(Convolutional Neural Networks, CNN)不断地学习句子的特征,并且添加了注意力机制,从而对文本进行情感分类。Martín 等人[11]利用酒店相关评论进行对比实验,使用 CNN 和 LSTM 进行情感分析文本。谷歌公司于 2018 年提出 Bert (Bidirectional Encoder Representations from Transformer)模型[12],它采用双向编码器表示和 Transformer 结构,具有更强的文本表征能力。Bert 是一种深度、双向、无监督的语言表示模型,可在最新的预训练语言表示工作的基础上使用,可以避免卷积神经网络无法提取评论全局特征和 LSTM 无法并行计算的问题。因此, Bert 已成为自然语言处理任务的预训练语言模型首选。与此同时, RoBERTa (Robustly optimized BERT approach)作为 Bert 的变体[13],它是在 Bert 的基础上通过更大规模的语料库和更长的训练时间进行训练得到的, RoBERTa 使

用了更多的数据和更深层次的 Transformer 模型,对语言的理解能力更为优秀。RoBerta 采用了动态 masking 的方式来预训练模型,从而能够更好地学习上下文信息,并且通过改进训练任务, RoBerta 可以更好地处理长文本序列,增强了其文本表示能力,提高了性能表现。由于 RoBerta 在处理长文本序列和上下文理解方面更加出色,可以更好地捕捉文本中的情感信息,因此,本文选用 RoBerta 预训练语言模型作为文本表征,同时使用结构简单,参数较少的双向 GRU 网络考虑文本中过去和未来的上下文依赖关系,并对文本的不同部分给予不同的关注,从而提高情感分类的准确度。

2.2. 主题挖掘研究

主题挖掘方法是识别文本主题和挖掘用户在线意见的有效工具[14]。目前,主题挖掘有两种类型:依赖文本相似度的传统主题聚类模型[15]和概率主题模型,如 LDA 主题模型[16] [17],即潜在狄利克雷分布,由 Blei 等人在 2003 年提出[18],该模型对词汇、主题和文档的三层结构进行贝叶斯概率分布,故也称为三层贝叶斯概率模型。研究表明,LDA 模型能够提取网上关于电子商务的负面评论[19],识别社交媒体数据中的主题分布[20]。此外,LDA 模型还可以用来检测真实的旅游热点和游客在历史旅游体验中的关键信息需求,进而高效地发现短文本的主题特征[21]。因此,本研究采用 LDA 主题模型来识别旅游评论的主题。

3. 论实验流程与模型构建

3.1. 数据采集与预处理

本文调用 Python 和 XPath 技术采集携程、去哪儿、马蜂窝、同城旅行、美团、途牛、穷游、驴妈妈等旅游网站,按照景区资源类型不同爬取沈阳市 3 个 4A 级及以上的景区的游客评论信息作为本文的研究对象,分别为沈阳故宫、沈阳世博园和沈阳方特欢乐世界。通过观察爬取得到的数据发现,评论数据中有一些重复的评论、刷单评论、凑字评论、瞎乱打字的评论、系统默认的评论等对分析情感无参考意义,这些评论的存在都会对后续的文本分析工作造成不利的影响,因此采用人工与程序共同清除。同时,本文采用各平台自带的好评与差评标准,人工筛选其中评论是否归类成功,用于后续研究。接着对评论文本数据使用停用词表,去除数据中的重复词、停用词、无用词汇等,得到标准化文本数据。最后,利用 Python 的第三方库 jiaba 分词工具对数据进行分词处理,结果如表 1 所示:

Table 1. Statistics of comment data for each scenic area

表 1. 各景区评论数据统计

序号	区县	景区简称	评论条数/条
1	浑南区	沈阳世博园	14,229
2	沈河区	沈阳故宫	14,973
3	沈北新区	沈阳方特欢乐世界	10,924

3.2. 建立情感分析模型

本研究构建了一种新的情感分析模型,即基于 RoBerta 预训练模型得到的词向量和 BiGRU-Attention 的景区评论情感倾向分析模型如图 1 所示,该模型主要由 4 部分组成,分别为 RoBerta 词嵌入层、BiGRU 层,注意力层和 Softmax 分类层。模型的输入是数据预处理好的标准化在线游客评论文本,经过 RoBerta 预训练模型得到词向量,然后将其送入到下游任务 BiGRU 层提取上下文之间的信息,接着结合注意力机制得到赋予有权重的评论文本信息,最后将该层的输出使用 Softmax 函数进行情感分类。

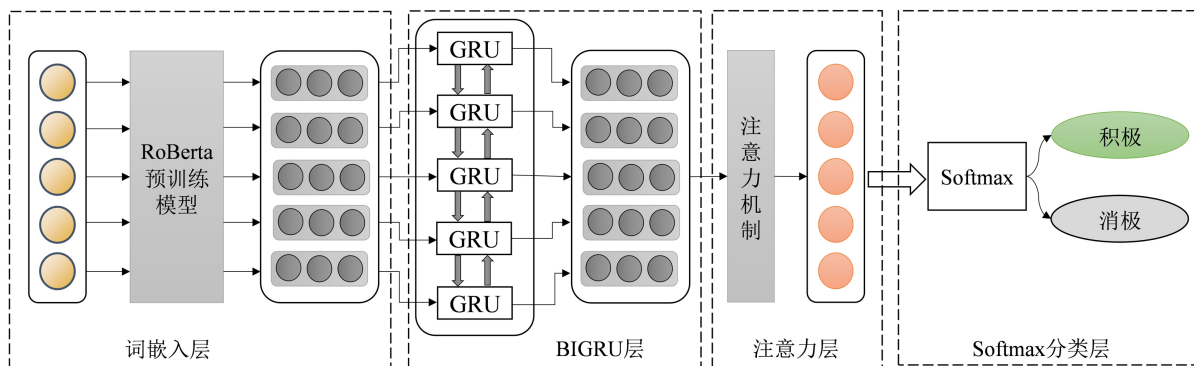


Figure 1. Emotional analysis model based on Roberta-BiGRU-Attention

图 1. 基于 RoBerta-BiGRU-Attention 情感分析模型

3.2.1. RoBerta 词嵌入层

RoBerta 是由 Facebook 人工智能研究团队开发的预训练语言模型，它基于 Bert 模型的体系结构，但包括一些改进和修改，使其在一系列自然语言处理任务中表现的更好。RoBerta 的体系结构使用双向编码器结构对文本进行编码，它也是一种双向语言模型。这意味着 RoBerta 可以同时考虑给定单词左右的上下文，以便更好地理解其含义和在句子中的使用，能够获得更全面的上下文信息，因此本文选用 RoBerta 实现词向量的表示，RoBerta 模型逻辑结构如图 2 所示。

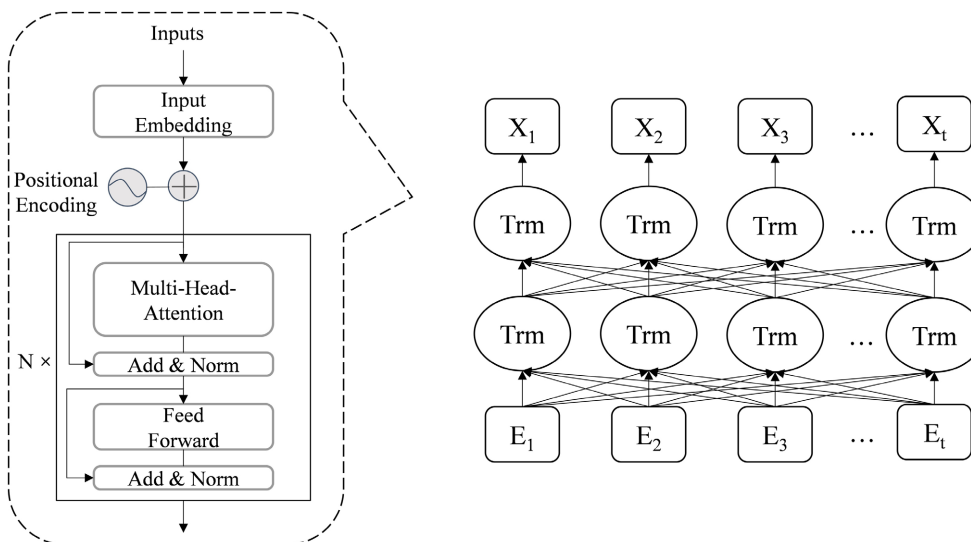


Figure 2. Logic structure diagram of RoBerta model

图 2. RoBerta 模型逻辑结构图

3.2.2. BiGRU 层

门控循环单元 GRU (Gated Recurrent Unit)是长短期记忆网络 LSTM (Long Short Term Memory)的变体，其内部结构比后者少一个遗忘门，该门的作用是控制前一时刻的记忆是否需要被遗忘，而 GRU 通过重置门和更新门的组合来实现类似的功能，因此该模型在保证网络学习效果的同时使结构更加简单，加快了训练速度。但单向 GRU 在获取文本的未来上下文信息时表现不佳，因此本文采用双向的 GRU，分别对同一个输入训练序列进行向前和向后训练提取更深层次的文本特征，使得序列中每一个节点都能完整地依赖所有上下文信息。双向 GRU 的基本结构如图 3 所示。

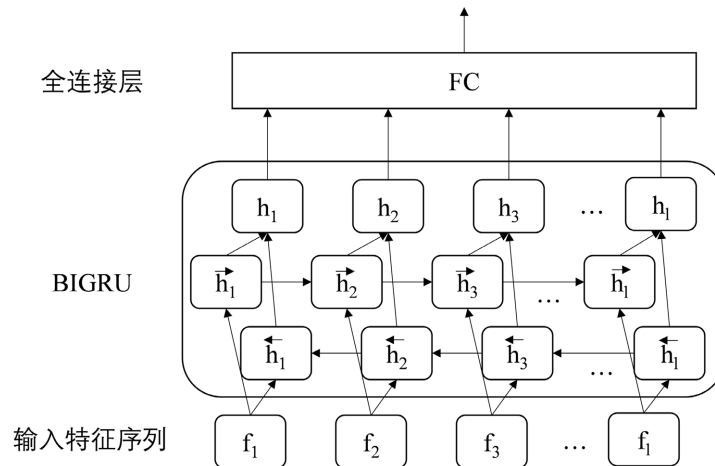


Figure 3. BiGRU model
图 3. BiGRU 模型

BiGRU 的输入为上一层 RoBERTa 模型输出的词向量，其输出由上一时刻前向隐层状态 \vec{h}_{l-1} 、反向隐层状态 \overleftarrow{h}_{l-1} 和当前时刻的输入 x_l 三部分决定。GRU 网络结构的计算过程公式如下：

$$z_l = \sigma(W_z x_l + U_z h_{l-1} + b_z) \quad (1)$$

$$r_l = \sigma(W_r x_l + U_r h_{l-1} + b_r) \quad (2)$$

$$\tilde{h}_l = \tanh(W_h x_l + U_h (r_l \cdot h_{l-1}) + b_h) \quad (3)$$

$$h_l = z_l \cdot h_{l-1} + (1 - z_l) \cdot \tilde{h}_l \quad (4)$$

式中： z_l 和 r_l 代表更新门和重置门； \tilde{h}_l 代表候选隐藏层状态； h_{l-1} 和 h_l 分别代表 $l-1$ 和 l 时刻的隐藏层态； W 和 U 为权重； b 为偏置； σ 代表 Sigmoid 函数。BiGRU 网络结构的计算过程公式如下：

$$\vec{h}_l = GRU(x_l, \vec{h}_{l-1}) \quad (5)$$

$$\overleftarrow{h}_l = GRU(x_l, \overleftarrow{h}_{l-1}) \quad (6)$$

$$h_l = F(W_{\vec{h}_l} \vec{h}_l + W_{\overleftarrow{h}_l} \overleftarrow{h}_l + b_l) \quad (7)$$

式中： \vec{h}_l 和 \overleftarrow{h}_l 分别为 l 时刻的前向和后向隐藏层的状态； F 为激活函数 $W_{\vec{h}_l}$ 和 $W_{\overleftarrow{h}_l}$ 分别为 l 时刻前向和后向隐藏层状态的权重； b_l 是 l 时刻隐藏层状态的偏置。

3.2.3. 注意力机制层

为了捕获更准确的文本信息，引入注意力机制学习词汇间的依赖关系，通过将 BiGRU 层的输出分配不同的特征权重，在关键词汇信息中给予更多的关注，进一步提高情感分类的准确度。在情感分析中，由于文本中词汇的重要程度不同，特征权重也不同，引入注意力机制可以更好地学习词汇间依赖关系，增强对重要词汇的关注度。因此，通过分配不同的权重来关注上下文的重要部分，以进一步提高分类的准确度。注意力机制根据信息的重要性为输入信息分配权重参数，称为注意力值。注意力值是通过首先确定输入信息的注意力分布来计算的，这会产生一个注意力得分(Attention Score)函数，然后使用 Softmax 函数对该函数进行归一化数值变换，得到的值根据权重系数进行加权求和。换句话说，注意力机制计算每条输入信息的重要性，并使用该信息对最终输出进行加权，具体计算公式如下：

$$u_t = v^T \tanh(W h_m + U h_t) \tag{8}$$

$$a_t = \frac{\exp(u_t)}{\sum_t \exp(u_t)} \tag{9}$$

$$V = \sum_t a_t h_m \tag{10}$$

式中： h_m 、 h_t 为时刻 m 、 t 的 BiGRU 的输出， h_m 为输入信息向量， h_t 为查找关键信息的查询向量， W 、 U 、 v 是参数向量，通过 Softmax 函数进行归一化处理。将原始计算分值转化成所有权重之和为 1 的概率分布，更加突出重要元素的权重，最后将所有输入信息的概率汇总，采用加权平均的方式求和，得到注意力值 V 。

3.2.4. Softmax 分类层

最后，我们使用密集层对得到的特征向量 h 进行降维，这一步骤有助于简化评论文本的表达，同时保留其最重要的特征。接下来，我们合并密集层的输出，并将它们通过一个全连接层，这一层有助于捕捉评论文本不同特征之间的关系，并生成更具信息性的评论文本表示。最后，我们使用 Softmax 函数来获得评论文本情感标签的概率，计算公式为：

$$y = \text{Softmax}(W_f \cdot V + b_f) \tag{11}$$

3.3. 建立 LDA 主题模型

LDA 主题模型，作为一种文档主题生成模型，该模型通过与语料库相关的 Dirichlet 先验来生成文本各段落中主题的多重分布。假设一组文档，令 D 表示文档个数， K 表示主题个数；其中每个文档 D 都是一个含有 N 个单词的序列，其中 $w_{d,n}$ 是文档 D 的第 n 个单词， $z_{d,n}$ 是文档 D 的第 n 个主题。 $\theta_d \sim \text{Dir}(\alpha)$ 表示服从超参数为 α 的 Dirichlet 概率分布， $\beta_k \sim \text{Dir}(\mu)$ 表示服从超参数为 μ 的 Dirichlet 概率分布。LDA 模型的联合概率分布函数如式(12)所示：

$$p(\beta, \theta, z, \omega) = \left(\prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) \prod_{d=1}^D p(z_{d,n} | \beta_{i,k}, z_{d,n}) \right) \left(\prod_{i=1}^K p(\beta_i | \mu) \right) \tag{12}$$

由于无法直接获得 LDA 主题的数量，本研究参考了相关文献中基于主题相似度和困惑度评价指标确定最优 LDA 主题数量的方法。假定主题取值范围为[1, 20]，计算困惑度(Perplexity)数值，计算公式如式(13)、(14)所示，该数值一般随着潜在主题数量的增加呈现递减的规律，该数值越小，该主题模型的生成能力越强。

$$\text{Perplexity}(D) = \exp \left\{ \frac{-\sum_{d=1}^M \log(p(w_d))}{\sum_{d=1}^M N_d} \right\} \tag{13}$$

$$p(w_d) = \sum_k p(k) p(w|z, g) \tag{14}$$

其中 M 为语料库的大小， N_d 为文档 D 的文本大小， k 为主题， w 为文档， g 为训练集训练的文本主题分布。模型生成能力越强， $p(w_d)$ 越小。

本文使用 Python 的数据可视化包 Matplotlib 绘制训练出来的困惑度与主题数的折线图，通过观察折线图拐点确定最优主题的数值。然后将该主题数作为 LDA 主题模型的超参数进行主题分析，我们可以得到主题特征词的概率分布，从而分析每个主题下的特征词的语义特征，最后通过人工提炼汇总每个主题下特征词的描述内容。

3.4. 实验参数设置

本文使用 Python 编程语言进行建模, 其中 RoBerta-BiGRU-Attention 模型部分使用深度学习框架 Pytorch 进行模型搭建, LDA 主题模型使用 Python 的第三方库 scikit-learn 进行建模, 具体超参数如表 2:

Table 2. Experimental parameters
表 2. 实验参数

参数名称	参数值
epochs	100
batch_size	64
学习率	1e-5
dropout	0.9
GRU 隐藏层维度	768
损失函数	Cross-Entropy
优化器	BertAdam
LDA 最大主题数	20
LDA 最大迭代次数	100
Learning_method	batch

3.5. 实验评估标准

本文使用测试集对模型进行评估, 具体使用 3 个评估指标: 加权精准率(Weighted Precision, WP)、加权召回率(Weighted Recall, WR)和加权 F1 值(Weighted F1, WF1)。假设 TP_i 表示模型预测评论为 i 极性的同时真实评论为 i 极性的数量, FP_i 表示模型预测评论为 i 极性的同时真实评论不为 i 极性的数量, FN_i 表示真实评论为 i 极性但模型预测评论不为 i 极性的数量, f_i 代表 i 极性的权重, 其中 i 为正向情感极性为 1, 反之为 0。具体计算如公式如下:

$$WP = \left(\frac{f_0 \cdot TP_0}{TP_0 + FP_0} + \frac{f_1 \cdot TP_1}{TP_1 + FP_1} \right) \cdot \left(\frac{1}{f_0 + f_1} \right) \quad (15)$$

$$WR = \left(\frac{f_0 \cdot TP_0}{TP_0 + FN_0} + \frac{f_1 \cdot TP_1}{TP_1 + FN_1} \right) \cdot \left(\frac{1}{f_0 + f_1} \right) \quad (16)$$

$$WF1 = \frac{2 \cdot WP \cdot WR}{WP + WR} \quad (17)$$

4. 实验结果与分析

4.1. 数据处理

本文使用 Python 随机抽取每个景区各 8000 条在数据预处理部分得到的标准化评论文本数据作为研究对象, 其中 5000 条数据按照 8:1:1 的比例划分为训练集、验证集、测试集, 作为 RoBerta-BiGRU-Attention 模型的训练数据, 余下的 3000 条数据进行情感倾向预测。

4.2. 情感分析模型结果

本文在训练集上训练模型，同时使用测试集进行评估。为验证本文采用模型的情感分类效果，接下来将进行与其他基线模型的对比实验。

1) RoBerta 模型：使用 RoBerta 获取词向量，直接通过全连接层(也称为分类层)进行情感倾向预测。

2) RoBerta-BiGRU 模型：使用 RoBerta 获取词向量，通过 BiGRU 对文本特征进行提取，然后通过分类层进行情感倾向预测。

为保证模型实验的客观公正性，在对比实验时使用的数据集都是各景区的在线评论，且实验的参数保持一致，三种模型的情感倾向实验结果表 3、表 4 和表 5 所示：

Table 3. Experimental results of Shenyang World Expo Park Scenic Area comment data

表 3. 沈阳世博园景区评论数据实验结果

模型类型	WP (%)	WR (%)	WF1 (%)
RoBerta	89.90	89.87	89.89
RoBerta-BiGRU	91.37	91.00	91.13
RoBerta-BiGRU-Attention	92.03	92.13	92.06

Table 4. Experimental results of Mukden Palace Scenic Area comment data

表 4. 沈阳故宫景区评论数据实验结果

模型类型	WP (%)	WR (%)	WF1 (%)
RoBerta	90.61	90.64	90.63
RoBerta-BiGRU	92.97	92.79	92.30
RoBerta-BiGRU-Attention	94.94	94.92	94.93

Table 5. Experimental results of Shenyang Fantawild Happy World Scenic Area comment data

表 5. 沈阳方特欢乐世界景区评论数据实验结果

模型类型	WP (%)	WR (%)	WF1 (%)
RoBerta	88.42	88.50	88.46
RoBerta-BiGRU	90.61	90.64	90.63
RoBerta-BiGRU-Attention	91.57	91.37	91.44

本次研究在训练集上训练本文所构建的模型，并使用测试集来进行评估。同时本文实验了其他两种情感分析模型进行对比，以此来证明本文模型的鲁棒性。3 个实验结果所示，本文提出的 RoBerta-BiGRU-Attention 模型在 3 个不同数据集上，所有评估指标上表现最好，其次是 RoBerta-BiGRU 模型，最后是 RoBerta 模型。这表明加入了双向 GRU 和注意力机制对情感倾向分析任务有益。基于预训练模型 RoBerta 在进行多次调参后，选取模型算法的最佳结果，实现的情感分析模型在各个景区表现良好，其加权平均精准率、加权平均召回率以及加权平均 F1 值在各数据集中分类效果均好，且能达到 94% 的分数，分析原因一是 RoBerta 产生的词向量是动态的，考虑了文本的上下文语义信息。原因二是注意力机制可以有效捕捉上下文甚至整体的关键信息句子，并融合到当前的词向量中，其可以给文本不同的部分分配不同的权重，进而提高情感倾向分析的准确性。

综上所述，本文将采取提出的 RoBERTa-BiGRU-Attention 模型继续对余下的 3000 条未标注的景区评论进行情感预测。同时对最后的预测结果进行人工检测，保证情感倾向分析达到最小失误。

4.3. LDA 主题挖掘结果

接着将情感倾向分析标注好的数据，利用 Python 的 Pandas 库分析正、负情感极性分布，送入到 LDA 主题模型中，输出结果为各景区评论主题。本文参考相关文献的经验[22]，将主题数 K 的取值范围设为 [1, 20]，计算不同主题数下的困惑度数值，然后使用利用 Python 的数据可视化包 Matplotlib 绘制出与主题数相关的困惑折线图。这里以沈阳故宫为例，展示困惑度分析结果如图 4 所示：

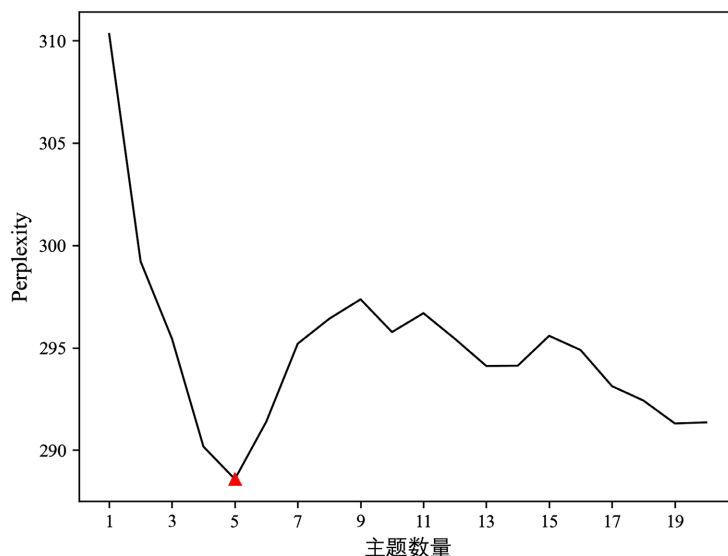


Figure 4. Perplexity line chart
图 4. 困惑度折线图

当 $K = 5$ 时，困惑度数值最小(图中红色标记处)，因此选择主题数为 5 作为 LDA 主题模型的超参数，其他景区用困惑度确定主题数的求法相同。

然后使用 Python 的 scikit-learn 库的 LatentDirichletAllocation 函数，对各景区标注好情感倾向的评论数据分别构建 LDA 主题模型。经过 LDA 主题分析后，在每个主题下生成 12 个最有可能出现的词语，随后对主题分析的结果进行评价，得到各景区评论数据的主题 - 特征词分布结果，如表 6、表 7 和表 8 所示：

1) 沈阳世博园景区

主题 1 是关于沈阳世博园景区门票和游园体验方便程度的评论，涉及到的词汇有入园、换票、买票、取票、扫码等。游客认为购票入园比较方便，排队时间较短，也提到了门票优惠信息。

主题 2 主要是针对沈阳世博园景区内部设施和景点的评价，包括一些负面评价，比如地方不好、不能进去、收费高等。同时，也有一些中性的词语，如建议、感觉等。

主题 3 是关于沈阳世博园景区适合亲子游玩的评价，包括可以孩子开心、环境适合、喜欢景色、设施齐全等词汇。游客认为沈阳世博园非常适合家庭游玩，有很多景点和游乐项目，适合各个年龄段的人群。

主题 4 主要评价沈阳世博园景区维护和季节变化的评论，包括一些词汇如现在、季节、已经、菊花、百合等。游客认为沈阳世博园景区虽然有一些缺点，但整体维护还是比较好的，不同的季节景色也各具特色。

主题 5 主要评价了沈阳世博园景区作为旅游胜地的评价，包括沈阳作为城市的特点，以及沈阳世博

园景区自身的特色，如值得一去、很大、环境优美等词汇。游客认为沈阳世博园是一个很好的旅游目的地，适合休闲游玩和观光旅游。

Table 6. Topic-characteristic words of comments on scenic area in Shenyang World Expo Park
表 6. 沈阳世博园景区评论主题 - 特征词

主题	特征词
主题 1	方便、直接、入园、换票、排队、小时、买票、便宜、景色、取票、二维码、优惠
主题 2	溜达、地方、不好、建议、小时、太大、东西、收费、项目、观光车、工作人员、感觉、一圈、
主题 3	孩子、开心、环境、适合、一起、景色、游玩、空气、家人、老人、天气、下次
主题 4	设施、感觉、现在、季节、百合、不错、已经、以前、好多、维护、郁金香、可能、没开
主题 5	值得、很大、园区、休闲、游览、环境、旅游、景色、交通、环境优美、好去处、电瓶车

综上所述，这 5 个主题分别反映了游客对沈阳世博园景区不同的维度，分别是主题一关于门票和游园方便程度的评价；主题二主要评价了景区内设施、景点和用户体验；主题三是关于适合亲子游玩的评价；主题四主要评价了季节和景区维护情况；主题五主要评价沈阳世博园作为旅游胜地的特点和特色。统计积极与消极主题评论的数量，进而剖析游客的情感倾向占比情况，其可视化结果如图 5 所示：

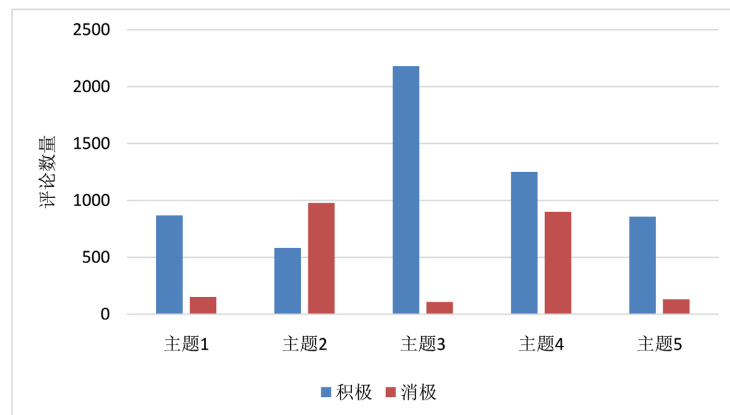


Figure 5. Statistics on the number of positive and negative comments at Shenyang World Expo Park
图 5. 沈阳世博园正负评论数量统计

由主题占比情况，我们可以得出结论，主题一、主题三、主题五的正、负情感的评论数量相差极大，说明景区的票价、交通便利情况、性价比体验等方面给游客带来的感受总体来说是很不错的，得到了多数游客的肯定。然而，在主题二中，游客的负面评论数量远超过正面评论数量，这表明一部分游客对于景区相关设施的价格和提供的价值感到不满意。他们认为景区内部的景色没有达到他们心中的期望，票价过高，给游客带来了不佳的体验。该景区服务提供者需要通过不断优化景区的各个方面，可以提升游客的整体体验和满意度。

2) 沈阳故宫景区

主题 1 中游客将沈阳故宫与北京故宫做横向对比，游客的整体评价主要关注故宫值得游览的地方、文化感受、适合孩子等方面。

主题 2 说明游客关注故宫景区的服务设施质量，评论中提到了讲解员态度和细心，同时也提到了一些景点开放问题和维修情况等建议。

主题 3 侧重关注沈阳故宫的空间大小、建筑外观、环境清洁、以及与游览体验相关的事项如售卖雪糕等。

主题 4 主要关注故宫的历史和文化底蕴方面，有故宫的宫殿、皇帝、满族等方面。

主题 5 主要关注游客对沈阳故宫的方便性、值得性、服务质量等方面，同时提出了门票价格、排队等建议。

Table 7. Topic-characteristic words of comments on Scenic Area in Mukden Palace

表 7. 沈阳故宫景区评论主题 - 特征词

主题	特征词
主题 1	故宫、历史、沈阳、看看、北京故宫、孩子、不错、值得、文化、感受、值得一看、知识
主题 2	讲解、导游、维修、建议、溜达、大政殿、开放、讲解员、凤凰殿、介绍、不让、感觉、遗憾
主题 3	里面、感觉、东西、拍照、很大、没什么、好看、雪糕、建筑、打卡、整体、环境
主题 4	宫殿、皇宫、建筑、清朝、中国、皇太极、皇家、建筑群、入关、盛京、历史、努尔哈赤
主题 5	不错、方便、值得、推荐、取票、排队、环境、人太多、景色、服务、性价比、买票

综上所述，这五个主题涵盖了游客对沈阳故宫的整体评价、导游服务、历史文化底蕴、游览体验和服务质量等方面。每个主题都有不同的评论特点和关注点，分别反映了游客们在游览过程中的不同体验和感受，然后统计积极与消极主题评论的数量，进而剖析游客主要关注沈阳故宫景区的维度，可视化结果如图 6 所示：

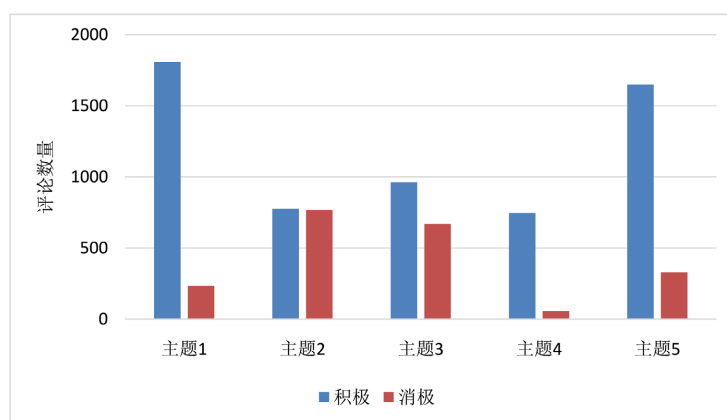


Figure 6. Statistics on the number of positive and negative comments at Mukden Palace

图 6. 沈阳故宫正负评论数量统计

根据可视化结果可知，在 5 个主题中，主题一的积极评论数量最多，说明大部分游客对沈阳故宫的整体评价是非常满意的，此外，主题二的消极评论数量较多，通过对这部分游客的评论分析可知，造成游客满意度较低的因素主要有部分游客认为讲解员的讲解不够细致，景区内部分设施在维修等。

游客对于沈阳故宫景区存在不满意的方面有：一是讲解员的服务质量不尽如人意，需要提高专业水平和服务态度；二是景区内部分设施需要加强维护管理，确保游客有更好的游览体验和感受。为此，景区服务提供者需采取针对性的措施，努力改进上述问题，提升服务水平，完善管理体系。进一步了解游客的需求和反馈，不断优化景区的服务和管理模式，为广大游客创造更好、更舒适、更安全的游览环境。

3) 沈阳方特欢乐世界景区

主题 1 描述沈阳方特欢乐世界特定人群的体验感受，小孩子表示很喜欢很开心，值得推荐。

主题 2 描述沈阳方特欢乐世界景区的整体评价，游客觉得这里非常好玩，设施特别棒，环境也很好，值得强烈推荐给大家，服务态度也很好，满意度较高。

主题 3 描述沈阳方特欢乐世界景区的夜间场景，客人们认为晚上的表演十分好看，需要注意烟花燃放时间与排队，适合大家一起来感受夜场氛围。

主题 4 描述沈阳方特欢乐世界景区的刺激项目，例如火流星、维苏威等，游客感觉这些项目非常刺激和极限，很有挑战性。

主题 5 描述沈阳方特欢乐世界景区的一些问题，例如排队时间、收费等，游客建议需要改进和提高服务质量。

Table 8. Topic-characteristic words of comments on Scenic Area in Shenyang Fantawild Happy World

表 8. 沈阳方特欢乐世界景区评论主题 - 特征词

主题	特征词
主题 1	孩子、喜欢、来方特、开心、大人、值得、十一、很漂亮、体验、梦幻、放假、无人机
主题 2	好玩、开心、推荐、设施、值得、环境、刺激、项目、满意、服务态度、体验、还会
主题 3	烟花、夜场、表演、排队、好看、时间、鱿鱼、好吃、一起、朋友、体验、漂亮
主题 4	火流星、神秘、刺激、极限、河谷、摆锤、维苏威、雨衣、旋转、恐龙危机、过山车、感觉
主题 5	项目、排队、收费、门票、人太多、花钱、设施、工作人员、东西、不要、雨衣、建议

根据可视化结果图 7 的分析，沈阳方特欢乐世界景区的游客对于前四个主题的评价普遍偏向正面，反映出游客对该景区在整体体验、游乐设施、景区内活动项目等方面的高度认同。然而，在主题五中，游客的差评数量明显高于正面评论数量，这一结果主要与景区内排队情况存在密切关联。游客反映排队等待时间过长，导致能够游玩的项目较少，游玩体验下降。同时，一些游乐设施的收费标准也存在争议，需要进行进一步的调整。因此，为了提升沈阳方特欢乐世界景区的整体竞争力，有必要加强排队管理和游乐设施的优化改善。具体而言，可以采取增加游乐设施数量、优化排队流程、引入虚拟排队等方式来解决排队问题，提升游客的满意度。除此之外，景区还可以提高服务水平，给游客提供更加细致周到的服务和关爱，增强游客的归属感。在收费方面，可以进行合理的调整，让游客感受到物有所值，从而提升游客的消费意愿。

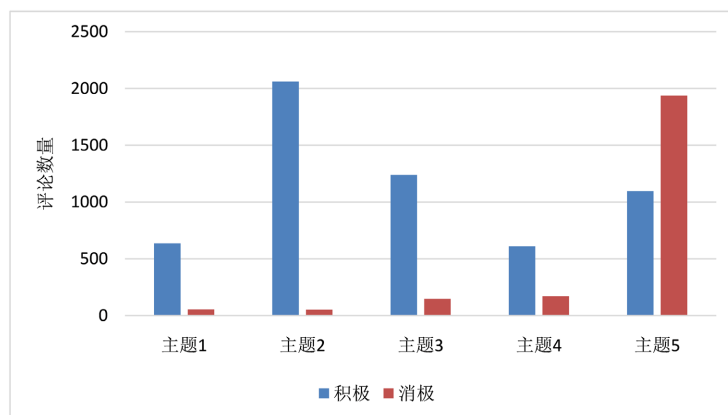


Figure 7. Statistics on the number of positive and negative comments at Shenyang Fantawild Happy World

图 7. 沈阳方特欢乐世界正负评论数量统计

5. 结语与展望

本文提出了一种名为 RoBERTa-BiGRU-Attention 的景区在线评论文本情感分类模型,它结合了 RoBERTa 预训练语言模型、BiGRU 神经网络和注意力机制。在该模型中, RoBERTa 生成词向量作为输入, BiGRU 网络对文本特征进行提取, 注意力机制对文本中更能表达情感的单词进行加权处理, 以实现文本情感分类。相较于其他模型, RoBERTa-BiGRU-Attention 能够高效地捕获文本及其上下文的总体特征, 并显著提高分类性能。实验结果显示, 该模型在沈阳市 3 个景区在线评论数据中表现良好。此外, 本文还结合 LDA 主题模型完成了游客在景区下不同主题维度情感倾向的识别分析。这一分析可以帮助景区管理者深入理解游客的评价和需求, 有利于制定更科学、合理的经营策略, 提高景区服务质量和满意度。综上所述, 本文的模型不仅有着广泛的应用前景, 而且能够对旅游业的发展产生积极的影响。

参考文献

- [1] Chatterjee, A., et al. (2019) Understanding Emotions in Text Using Deep Learning and Big Data. *Computers in Human Behavior*, **93**, 309-317. <https://doi.org/10.1016/j.chb.2018.12.029>
- [2] Rezaeina, S.M., Rahmani, R., Ghodsi, A., et al. (2018) Sentiment Analysis Based on Improved Pre-trained Word Embeddings. *Expert Systems with Applications*, **117**, 139-147. <https://doi.org/10.1016/j.eswa.2018.08.044>
- [3] Hlee, S., Lee, H. and Koo, C. (2018) Hospitality and Tourism Online Review Research: A Systematic Analysis and Heuristic-Systematic Model. *Sustainability*, **10**, Article No. 1141. <https://doi.org/10.3390/su10041141>
- [4] Liu, Y. and Hu, H. (2021) Online Review Helpfulness: The Moderating Effects of Review Comprehensiveness. *International Journal of Contemporary Hospitality Management*, **33**, 534-556. <https://doi.org/10.1108/IJCHM-08-2020-0856>
- [5] Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., et al. (2018) Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews. *Journal of Computational Science*, **27**, 386-393. <https://doi.org/10.1016/j.jocs.2017.11.006>
- [6] Cho, K., Van Merriënboer, B., Gulcehre, C., et al. (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, October 2014, 1724-1734. <https://doi.org/10.3115/v1/D14-1179>
- [7] Prabha, M.I. and Srikanth, G.U. (2019) Survey of Sentiment Analysis Using Deep Learning Techniques. 2019 *IEEE 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, Chennai, 25-26 April 2019, 1-9. <https://doi.org/10.1109/ICIICT1.2019.8741438>
- [8] Zhou, X., Wan, X. and Xiao, J. (2016) Attention-Based LSTM Network for Cross-Lingual Sentiment Classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, November 2016, 247-256. <https://doi.org/10.18653/v1/D16-1024>
- [9] Dragoni, M. and Petrucci, G. (2017) A Neural Word Embeddings Approach for Multi-Domain Sentiment Analysis. *IEEE Transactions on Affective Computing*, **8**, 457-470. <https://doi.org/10.1109/TAFFC.2017.2717879>
- [10] 冯兴杰, 张志伟, 史金钊. 基于卷积神经网络和注意力模型的文本情感分析[J]. *计算机应用研究*, 2018, 35(5): 1434-1436.
- [11] Martín, C.A., Torres, J.M., Aguilar, R.M., et al. (2018) Using Deep Learning to Predict Sentiments: Case Study in Tourism. *Complexity*, **2018**, Article ID: 7408431. <https://doi.org/10.1155/2018/7408431>
- [12] Devlin, J., Chang, M.W., Lee, K., et al. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.
- [13] Liu, Y., Ott, M., Goyal, N., et al. (2019) Roberta: A Robustly Optimized Bert Pretraining Approach. arXiv:1907.11692.
- [14] Chi, M., Pan, M. and Wang, W. (2021) Cross Platform Comparative Study of Shared Accommodation and Hotel User Comments: Thematic Social Network and Emotion Analysis Based on LDA. *Library and Information Work*, **65**, 107-116.
- [15] Xia, H.S, Li, B.G. and Yang, P. (2016) Topic Extraction in News Comments Based on Improved K-Means Clustering Algorithm. *Journal of Infection*, **35**, 55-65. http://en.cnki.com.cn/Article_en/CJFDTotal-QBXB201601007.htm
- [16] Tang, X.B. and Qiu, X. (2015) Research on Topic Oriented High Quality Comment Mining Model. *Modern Library and Information Technology*, **7**, 104-112.
- [17] Tu, H., Tang, X. and Xie, L. (2015) Research on User Requirements Mining Model Based on Online Comments. *Journal of Information Technology*, **34**, 1088-1097.

- [18] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.
- [19] Wu, Y., Na, R. and Li, H. (2021) Representative Negative Comment Extraction Based on Gaussian LDA and Spectral Clustering. *Information Sciences*, **39**, 136-142.
- [20] Ma, Y., Yue, K., Zhang, Z.-C., *et al.* (2018) Social Media Data Extraction Based on Knowledge Map and LDA Model. *Journal of East China Normal University (Natural Science Journal)*, **5**, 183-194.
<https://xbk.ecnu.edu.cn/EN/Y2018/V2018/I5/183>
- [21] Lyu, L.L. and Li, Y.T. (2017) Research on Tourism Information Recommendation—Based on Topic Mining and Presentation of Travel Notes. *Modern Intelligence*, **37**, 61-67.
http://en.cnki.com.cn/Article_en/CJFDTOTAL-XDQB201706010.htm
- [22] Pike, S. (2002) Destination Image Analysis—A Review of 142 Papers from 1973 to 2000. *Tourism Management*, **23**, 541-549. [https://doi.org/10.1016/S0261-5177\(02\)00005-5](https://doi.org/10.1016/S0261-5177(02)00005-5)