Statistics and Application 统计学与应用, 2022, 11(2), 388-401
Published Online April 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.112042

Hans汉斯

基于机器学习的景点评论文本分析

郑明明，王知人，谢璐妍

燕山大学理学院，河北 秦皇岛

收稿日期：2022年3月21日；录用日期：2022年4月10日；发布日期：2022年4月20日

摘　要

使用网络爬虫技术获取了旅游网站游客在线评论作为数据源，通过Python语言对数据进行数据清洗、中文分词、文本向量化，对完成预处理的数据作了描述性统计分析；建立了朴素贝叶斯(NB)、逻辑回归(LR)两个传统机器学习文本分类模型和长短期记忆网络(LSTM)深度学习模型，利用深度学习模型LSTM进行分类的准确率为92.15%，高于传统机器学习模型中准确率最高的LR约2.6个百分点。使用LSTM模型对评论文本进行分类并对完成分类的数据构建了LDA主题聚类模型挖掘潜在主题，提取不同主题对应的特征词进行对比分析，得出结论：负面评论对山海关景区基础设施、收费管理感到不满意；正面评论对山海关景区的历史文化底蕴、体验感受、景点服务以及景点趣味性都很满意。基于从评论文本中挖掘的信息，旨在提取游客关注点与需求，为潜在消费者提供消费选择，为景点管理部门提供营销决策。

关键词

旅游大数据，机器学习，游客评论，文本分类，LDA主题聚类模型

Text Analysis of Scenic Spot Comments Based on Machine Learning

Mingming Zheng, Zhiren Wang, Luyan Xie

Science College, Yanshan University, Qinhuangdao Hebei

Received: Mar. 21st, 2022; accepted: Apr. 10th, 2022; published: Apr. 20th, 2022

Abstract

Web crawler technology is used to obtain online comments of tourists from tourist websites as data sources. Data cleaning, Chinese word segmentation and text vectorization are carried out on the data by Python language, and descriptive statistical analysis is made on the preprocessed data. Two traditional machine learning text classification models, Naive Bayes (NB) and Logistic Regression (LR), and Long-term and Short-term Memory Network (LSTM) deep learning model are established. The classification accuracy rate of LSTM is 92.15%, which is about 2.6 percentage

文章引用: 郑明明, 王知人, 谢璐妍. 基于机器学习的景点评论文本分析[J]. 统计学与应用, 2022, 11(2): 388-401.
DOI: 10.12677/sa.2022.112042

points higher than LR, the highest accuracy rate in traditional machine learning model. The LSTM model is used to classify the comment text, and LDA topic clustering model is constructed for the classified data to mine potential topics, and the feature words corresponding to different topics are extracted for comparative analysis. The conclusion is that negative comments are not satisfied with the infrastructure and charge management of Shanhaiguan scenic spot; the positive comments are very satisfied with the historical and cultural heritage, experience, scenic service and interest of Shanhaiguan scenic spot. Based on the information mined from the comment text, it aims to extract the concerns and needs of tourists, provide potential consumers with consumption choices and provide scenic spot management departments with marketing decisions.

Keywords

Tourism Big Data, Machine Learning, Tourist Comments, Text Classification, LDA Topic Clustering Model

1. 引言

大数据的迅速发展给我们带来了机遇的同时也带来了挑战，如何从海量的数据中找出有利用价值的信息是我们所关心的。就旅游景点而言，游客的在线评论数据真实客观的反应了出游时的所见、所闻、所感，是影响潜在消费者做出决策的因素之一，从而间接地影响了旅游景点的收入。但是经过实际调查研究，大多网站平台没有对其进行细致的分类，即没有明确区分开好评和差评，有的平台即使作了区分，但是区分的方法不同，甚至有的平台是直接根据综合打分这一项指标来区分好评和差评的，分类的效果很不理想。就打分来说，大部分游客会倾向于给出一个接近满分的评分，而且如果游客没有评论网站会默认满分好评，但是实际评论内容却不一定是非常满意的；就用户评论来说，有些景点会购买网络水军对其提供的旅游产品进行虚假夸赞，以吸引更多的游客。因此，使用机器学习算法对评论文本进行精确分类并对其中隐藏的信息进行深度挖掘，实现对旅游景点客观整体的评价，就显得极具意义和价值。

关于文本的情感分类和主题挖掘研究还是比较多的。基于情感词典的文本分类方法和基于机器学习的文本分类方法是目前文本分类的主要研究方法，许多研究成果表明基于机器学习的性能表现优于词典方法的(2008) [1]。因此，本文采用基于机器学习的文本分类方法训练分类器。刘志明和刘鲁(2012) [2]使用不同的特征选取算法、不同的特征项权重计算方法和不同的机器学习算法来进行组合，对微博话题进行了文本分类的研究。周咏梅等(2014) [3]通过采集新闻评论文本数据，利用已有的情感词典和从评论文本中提取的情感词构建出了新闻领域的情感词典，并用于评论文本情感倾向性分析。魏慧玲(2014) [4]在已有情感词典基础之上配合语义相似度分析实现了手机评论的情感倾向性分析。郭小芬等(2017) [5]基于贝叶斯网络和支持向量机分类算法，在实际应用中实现了对中文新闻的精确分类。丁照银(2019) [6]将机器学习分类模型和 LDA 主题模型相结合，对某品牌连锁酒店用户评论进行了研究，得出支持向量机的分类效果最好。应昊东(2021) [7]利用 LDA 主题模型对新能源汽车各车型中用户满意以及不满意维度进行主题提取，来挖掘用户的关注重点。戴维(2018) [8]讨论了逻辑回归解决文本分类问题，其不仅介绍了逻辑回归的算法原理和使用步骤还结合具体实例对算法进行了评估，其中，特征选取是基于 LDA 主题聚类模型进行操作的。通过对相关文本分类和主题模型文献的梳理与研究，发现许多文本分类研究工作针对的

是新闻、微博、电商评论，很少有学者对旅游评论文本进行研究，本文将以旅游评论文本为研究对象，旨在为文本分类在新领域的研究拓展思路。

2. 数据搜集与预处理

2.1. 数据搜集

利用网络爬虫技术爬取了携程、去哪儿网、同程、马蜂窝、途牛等多个平台上关于著名景点山海关的游客评论数据，为保证时效性，时间范围限定为 2020 年 1 月 1 日~2022 年 1 月 31 日，共计 40281 条，将数据文件保存为 csv 格式备用。每条数据内容包含 4 个指标：用户名称、评论内容、评论时间、用户评分。部分原始数据见表 1 所示：

Table 1. Raw data
表 1. 原始数据

用户名	评分	评论内容	日期
zh****28	5 分 超棒	观名胜，缅怀历史，很有收获！不错的地方！说当地人经济和素质不高，倒也不假，可这也有利有弊，看南方那些水乡吧，素质高？！可弄得太商业啦！	2020/1/30
j****og	2 分 一般	实在不怎么样，收费太贵了。明朝的东东，还没南京的城墙老呢。失望!!!!!!!!	2021/3/18

数据来源于：携程、去哪儿网等平台。

2.2. 数据处理

使用 Python 语言中的 Pandas、jieba 等库对数据进行预处理，主要包括数据清洗、文本分词、文本向量化。在数据清洗之前，为了保证数据的真实可靠性，需要人工去除网络水军以及恶意差评的评论，包括：明显是营销账号发布的评论，评论中含有 xx 网站、xx 平台、xx 酒店等；同一 ID 发布的多条评论内容，存在人工刷评论行为；评论字数过少，没有参考价值。

数据清洗前后效果对比见表 2：

Table 2. Comparison of effects before and after data cleaning
表 2. 数据清洗前后效果对比

原始文本	处理后
终于可以亲眼(●✿▽✿●)看到了山海关！古老的建筑非常壮观！还有明清时期的古炮！	终于可以亲眼看到了山海关古老的建筑非常壮观还有明清时期的古炮
好好好好好！！！！！！！！！！	好

可以看到特殊符号、重复字段都被去除，可见数据清洗效果很好，清洗之后剩余 37491 条有效数据。

2.2.1. 文本分词与去停用词

中文分词和英文分词的区别在于，中文字符不像英文单词那样有天然的空格隔开，在中文文本挖掘过程中，为了便于分析词句的特性，需要把评论语句拆分成单个的词语。本研究使用的是 jieba 分词技术来进行中文文本分词的，是使用 Python 语言实现的文本预处理软件包，其准确率较高，而且操作起来较为简便，在文本分析领域很受欢迎。

分词之后存在很多像“的”、“了”、“吧”一些并没有实际意义的语气词，“但是”、“然而”之类的转折词，还有标点符号、特殊符号等之类的这些词都称为停用词，这些对文本分析不但没有用处，还会对分析结果造成干扰，需要去除。本文采用哈工大停用词典过滤掉无用的信息，可以减少文本特征向量的维数，加快模型运算速度。

经过分词和去停用词操作后，优化分词与去停用词结果见表 3：

Table 3. Partitioning to stop words
表 3. 分词去停用词

原始文本	小张导游人美声甜有耐心，热心细致专业性，有较为丰富的带队经验。性格乐观大方，健谈和老年人，小朋友都可以打成一片。
分词去停用词	小张/导游/人美/甜/耐心/热心/细致/专业性/带队/经验/性格乐观/大方/健谈/老年人/小朋友/打成一片

可以看到许多无意义的符号、语气词等都被过滤掉且分词效果很好。

2.2.2. 文本向量化

经过分词后的文本是结构化数据，而计算机只能识别非结构化数据，文本向量化可以把经过预处理后的数据转换成计算机可以识别的结构化数据，文本向量化的主要步骤为：首先将评论文本通过 jieba 分词工具分词为 N 个词组的集合，其次利用 Word2vec 模型将分词转化为 N 个对应的词向量，最后采用 TF-IDF 算法提取每个词向量的关键词，输出最终的文本向量。

至此，数据的预处理全部完成。每条文本评论变成了固定维数的向量，这是一种数学表达，计算机是可以轻松识别的。

3. 描述性统计分析

3.1. 词云图可视化

将游客评论文本经预处理后，使用 Python 软件中 pandas 库下的 value_counts()函数，统计词频数排名前 100 词语的词频并绘制词云图。词频越高表现在词云图中的字号就越大，词频越低表现在词云图中的字号就越小。山海关景区整体词云图见图 1，可以发现：天下第一关、历史、长城、不错、值得等词语比较突出，可以得知关于山海关景区的一些特色和卖点，大部分游客对该景点还是比较满意的，可以得到很好的游玩体验。但是也有一些基于消极评论产生的关键词，例如，贵、建议、小时、停车场等词。

Figure 1. The overall word cloud map of Shanhaiguan scenic area
图 1. 山海关景区整体词云图

3.2. 语义网络图

虽然词云图可以通过词频直观地来反映游客的主要关注点，但无法反映词组在特定意义上的联系以及文本深层次的结构关系，而语义网络分析则能通过构建概念和语义关系的网络图来直观展现要素之间的关系，将评论文本 txt 文件编码由 UTF-8 转为 ANSI，将文件导入 NetDraw 软件，最终生成游客对山海关景区整体形象的语义网络图[9]。从层级结构来看，语义网络图呈现“核心－边缘”特点，箭头指向越密集，越是接近核心的词表示对文档的作用越重要，具体见图 2 中红色方块。通过图 2 可以清楚看到，山海关景区语义网络图中包含“历史”、“山海关”、“长城”、“方便”等一系列的节点。从内容要素上看，山海关景区的旅游者主要关注的是景色、知名度、历史文化、服务等旅游目的地要素，主要表现在“长城”、“城墙”、“导游”、“门票”、“方便”等节点上。我们可以初步的了解到，山海关景区购票取票服务很好，可以在网上购买也可以在现场购买，同时导游的讲解服务也被多次提到。

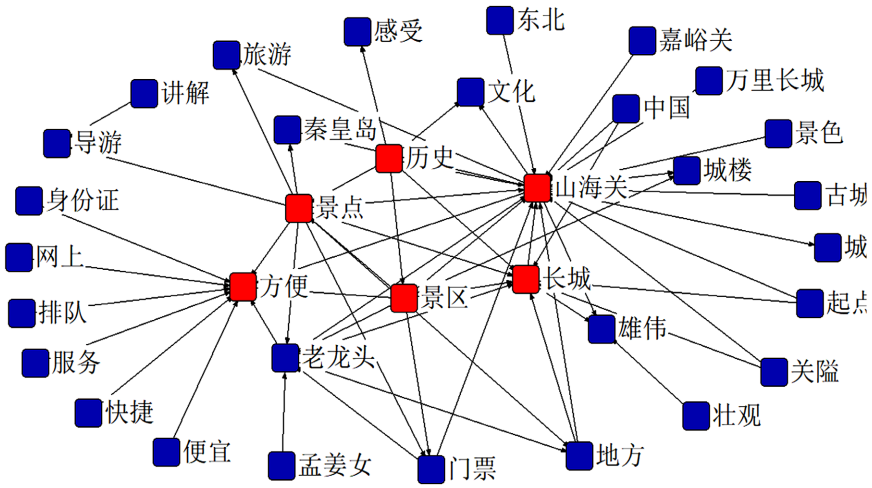

Figure 2. Semantic network diagram of Shanhaiguan scenic area
图 2. 山海关景区语义网络图

4. 分类模型

4.1. 模型训练

朴素贝叶斯、逻辑回归分类模型属于有监督的机器学习，分类器的质量很大程度上由特征属性、特征属性划分及训练样本质量决定，所以为了保证分类模型的准确率，需要让模型得到充分的训练，进而防止出现过拟合或欠拟合现象，本文通过网络收集一些人工标注的旅游景点评论组成的训练语料库，根据简单随机抽样原则随机抽取 8000 条语句进行了人工标注，进一步加大训练集的数量，共计 20,214 条语料，将处理后的评论语料库按照 8:2 的比例划分训练集与测试集，最终语料库组成为：训练数据集 16,171 条语料，测试数据集 4043 条预料，其中测试集中要包含比训练集中更新的数据。这些数据可以用于验证模型的性能，并且可以用于评估模型应用于实际环境中的表现。

语料库经过第二节预处理后，已经具备构建分类器的条件。首先在训练数据集上构建文本分类器，然后用训练出来的分类器模型在测试数据集上进行预测，根据预测情况对分类器性能进行评估。使用的是 Python 中 sklearn 机器学习库进行分类器的训练，把训练好的分类器应用到测试集中预测出分类，然后评估模型的预测效果。

4.2. 分类模型性能的评估

分类模型的性能评估是机器学习中非常重要的步骤，应该从多方面对模型进行评价，比较常见的指标就是准确率(accuracy)、精确率(precision)、召回率(recall)、F(F-measure)值、ROC 曲线和 AUC 值。

评价指标的计算公式如下：

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

对于文本分类问题，常用混淆矩阵来展示训练好的分类器在测试集中的表现。混淆矩阵是一种在机器学习分类问题中经常用到的辅助工具，可以直观地了解模型在测试集中的表现。

4.3. 模型构建及分析

4.3.1. 朴素贝叶斯模型

1) 混淆矩阵

构建基于朴素贝叶斯模型的分类模型，利用训练集训练后分类器在测试集上的表现见图 3：

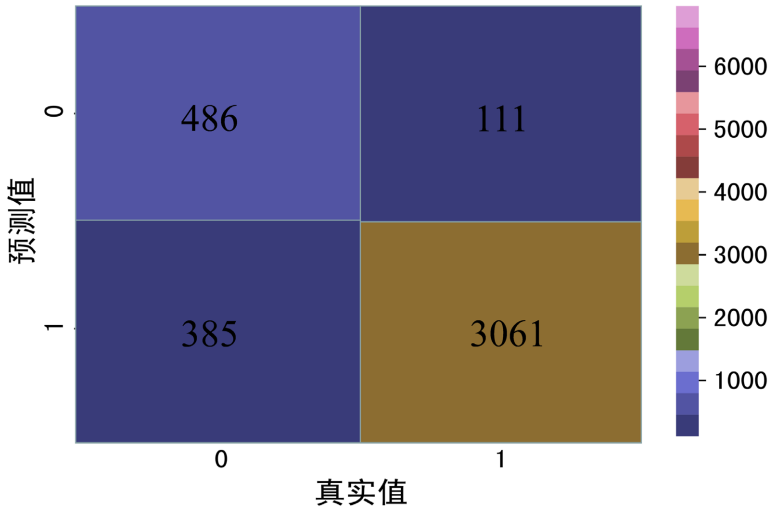

Figure 3. Naive Bayes confusion matrix
图 3. 朴素贝叶斯混淆矩阵

如上图 3，有 111 个样本原本是 1 (正例)的，却被预测成了 0 (反例)，还有 385 个样本原本是 0 的，却被预测成了 1。

具体如见表 4：

由公式(1)~(4)计算得出：准确率 = 87.73%，精确率 = 88.83%，召回率 = 96.50%，F 值 = 92.51%，都在 90%左右，说明分类器的预测效果是理想的。

Table 4. Naive Bayes confusion matrix
表 4. 朴素贝叶斯混淆矩阵

混淆矩阵		真实值	
		结果为正(P)	结果为负(N)
预测值	预测为正(P)	3061	385
	预测为负(N)	111	486

2) ROC 曲线

ROC 曲线的纵轴表示“真正例率”(true positive rate)简称 TPR，横轴表示“假正例率”(false positive rate)简称 FPR，基于表中的符号，二者公式为：

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}} \tag{5}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{6}$$

以所求得 TPR 和 FPR 的数值作为横、纵坐标，使用 Python 作图，就得到了 ROC 曲线见图 4，常用 ROC 曲线的线下面积，即 AUC (area under ROC curve)来评估分类模型的性能。

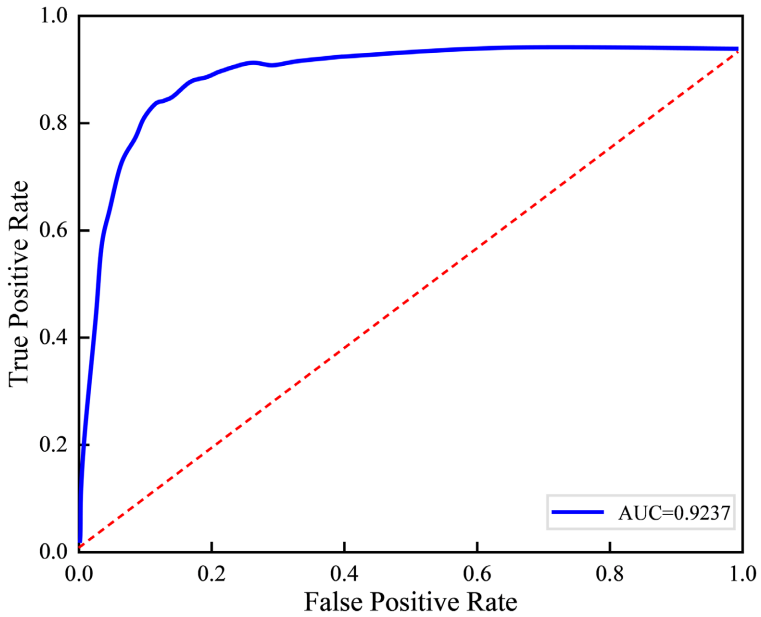

Figure 4. Naive Bayes ROC curve
图 4. 朴素贝叶斯 ROC 曲线

由图 4 可以看出，朴素贝叶斯模型 ROC 曲线占据了整个图形的左上方区域，分类器的分类效果十分理想。AUC = 0.9237 也就是 AUC 的值，这个值体现出朴素贝叶斯分类器性能良好。

4.3.2. 逻辑回归模型

1) 混淆矩阵

构建基于逻辑回归模型的分类模型，利用训练集训练后分类器在测试集上的表现见图 5：

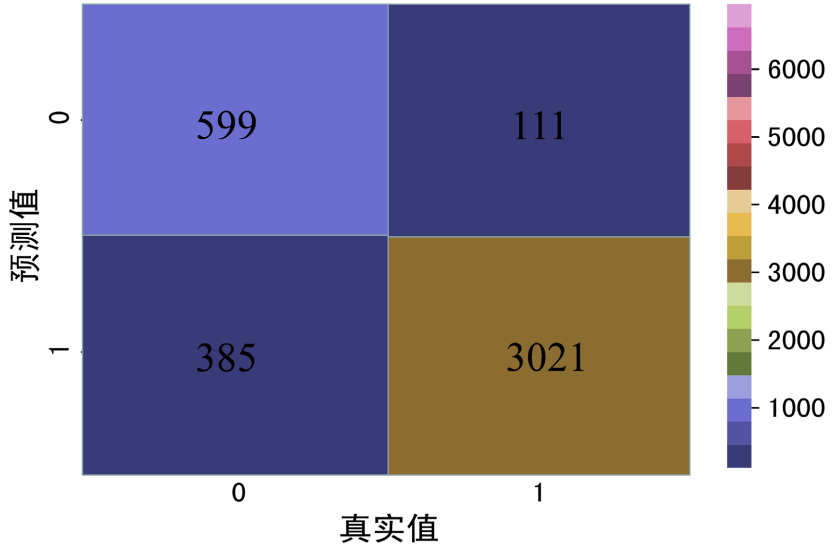

Figure 5. Logistic regression confusion matrix
图 5. 逻辑回归混淆矩阵

如上图 5，有 151 个样本原本是 1 (正例)的，却被预测成了 0 (反例)，还有 272 个，原本是 0 的，却被预测成了 1。

具体如下表 5 所示：

Table 5. Logistic regression confusion matrix
表 5. 逻辑回归混淆矩阵

混淆矩阵		真实值	
		结果为正(P)	结果为负(N)
预测值	预测为正(P)	3021	272
	预测为负(N)	151	599

由公式(1)~(4)计算得出：准确率 = 89.54%，精确率 = 91.74%，召回率 = 95.24，F 值 = 93.46%，分类器的预测效果也是很理想的。

2) ROC 曲线

由图 6 可以看出，逻辑回归模型分类器的分类效果也十分理想。AUC = 0.9338 是 AUC 的值，这个值越接近于 1 表示分类器性能越好。

4.3.3. 长短期记忆模型

神经网络模型在避免模型过拟合方面比传统机器学习模型表现好，长短期记忆网络(LSTM)是在循环神经网络的基础上改进后的模型，是一种特殊结构的 RNN。LSTM 能很好地解决长期记忆问题，也能避免出现梯度爆炸和梯度消失问题。

1) 实验与结果分析

为了防止模型过度拟合，本节采用测试集通过设初始参数进行监控，相关参数见表 6，当验证集的损失值连续 3 轮没有改善时，停止模型训练并保留最佳模型。

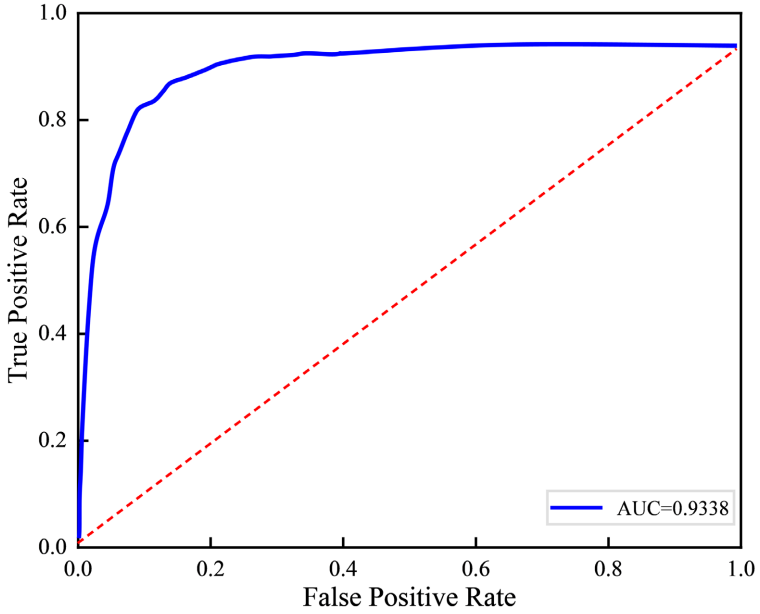

Figure 6. ROC curve of logistic regression
图 6. 逻辑回归 ROC 曲线

Table 6. Model parameter training table
表 6. 模型参数训练表

词嵌入	模型	优化器	dropout	epochs	batch_size	激活函数	损失函数
Word2Vec	LSTM	Adam	0.5	20	64	tanh	cross_entropy

准确率和损失变化见图 7、图 8，可以看出模型准确率随着训练轮数的增加不断提高，损失函数值不断下降，从第 17 轮开始基本没有变化，通过早停 earlystopping 函数停止训练。

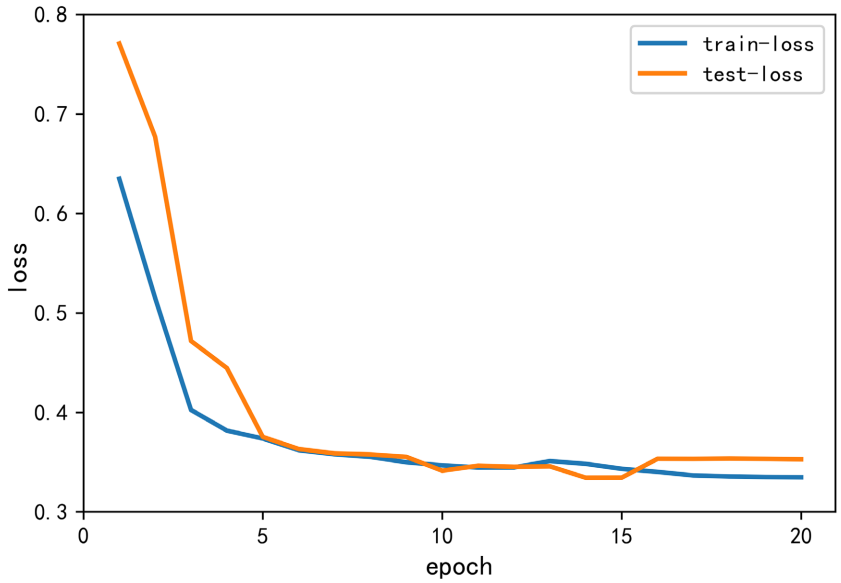

Figure 7. Changes of training loss of LSTM model
图 7. LSTM 模型训练损失变化图

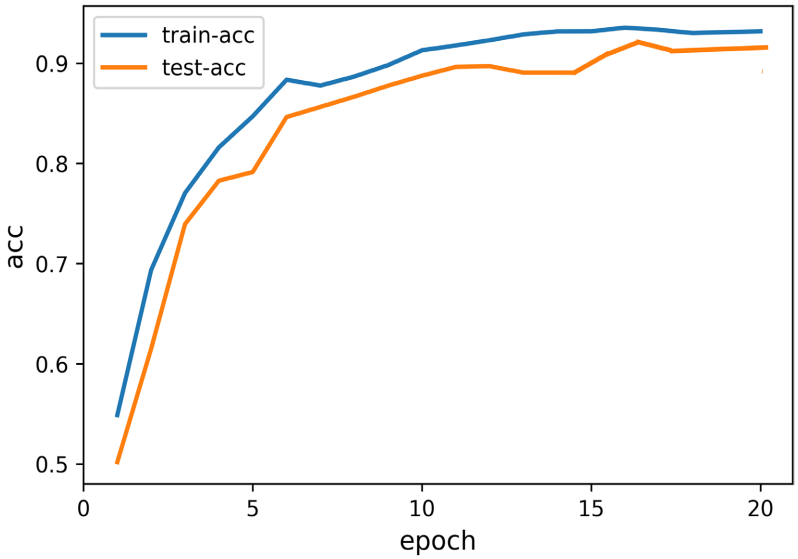

Figure 8. Changes of accuracy of LSTM model
图 8. LSTM 模型准确率变化图

混淆矩阵见图 9，通过混淆矩阵得到模型的准确率为 92.15%，精确率为 91.66%，召回率为 96.72%，F 值为 94.12%，其中模型的准确率比传统的机器学习模型最高值还高约 2.6 个百分点，展现出了 LSTM 的优良性能。

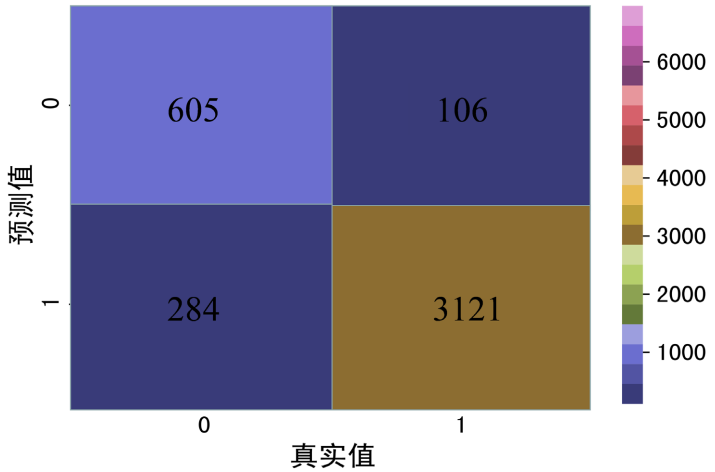

Figure 9. LSTM confusion matrix
图 9. LSTM 混淆矩阵

至此，本研究的第一个研究目标已经顺利实现。经过综合评价，LSTM 模型可以用来实现该景点游客评论文本的精确分类，可以考虑在线上推广使用。

4.4. 预测分类

在已经构建的 3 个文本分类器模型中选择性能相对来说较好的 LSTM 分类模型对山海关景区游客评论文本数据进行分类，把爬取的经过前期数据处理后的景点数据放入训练好的逻辑回归分类器中进行文本情感的分类得到正面评论和负面评论，其中游客评论正面评论占比 79.6%，负面评论占比 20.4%，即正面评论多于负面，可见山海关景区游客整体满意度很高。

5. LDA 主题聚类分析

本节运用 LDA 主题聚类模型，用以挖掘山海关旅游在线文本评论中蕴含的更深层次的信息，以期获得更有价值的内容。LDA 主题模型以文档、主题、词三层贝叶斯模型为核心结构，利用先验分布对数据进行似然估计并最终得到后验分布，模型的训练可以看成是这样一个动态链式：

$$P(\text{词}|\text{文本}) = P(\text{词}|\text{主题})P(\text{主题}|\text{文本}) \tag{7}$$

LDA 主题模型分析过程见图 10：

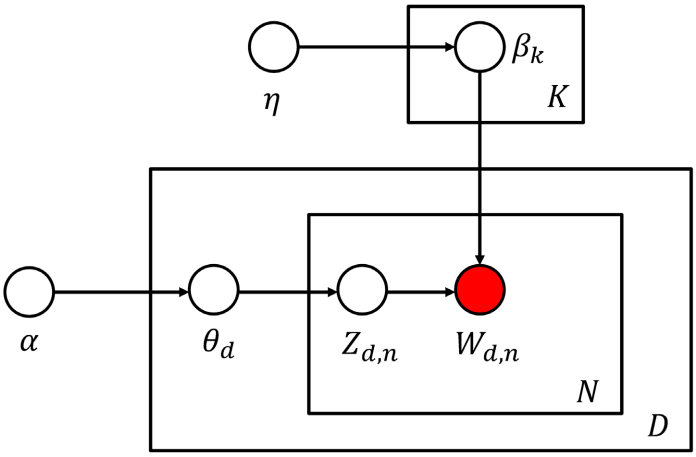

Figure 10. LDA topic probability model diagram
图 10. LDA 主题概率模型图

其中，α 和 β 分别表示主题分布 θ 和主题词分布 φ 的先验分布参数，将其先验分布视为狄利克雷分布。z 和 w 分别表示模型生成的主题及最终的主题词，M 表示文本数量，S 表示文本的词语数量。在 LDA 主题模型中，主题数 K 需要预先设定，为了得到最为合适的主题数，通过计算对应困惑度大小求得最优主题数 K。困惑度越小，说明文本聚类的效果越好。对图像来说，当困惑度下降趋势不再明显或处于拐点处时，此时的 K 值为最优主题数。

5.1. 确定主题数

在图 11 主题数－困惑度折线图中，对于游客正面评论数据集，随着主题数 K 值的增大，模型困惑度逐渐减小，并且当 $K=1$ 和 $K=4$ 的时候，存在显著的拐点：当 K 属于 $(1,4)$ 时，曲线急剧下降；当 K 属于 $(4,8)$ 时，曲线基本趋于平稳。故拐点 4 即为 K 的最佳值。同理，在游客负面评论数据集中，$K=3$ 时模型困惑度最小。故最终选择正面主题 4 个，负面主题 3 个。

接下来使用 Python 的 gensim 库进行 LDA 主题模型训练求解，分别提取正负面评论集主题的 5 个特征词以及各每个特征词的权重，更深入地挖掘景点优势和不足。

5.2. 主题展示

山海关景区负面评价的潜在主题模型聚类结果见表 7，可以看出负面评论包含三个主题且每个主题有五个主题相关特征词及其权重。主题一中包含山海关、天下第一关、知名、景区、古代五个特征词。故将主题一概括为“景点知名度”，正因为山海关景区拥有较高知名度，反而给景点带来了不小的压力，由于知名度吸引游客慕名而来，但是游客在游玩过程中的实际体验并没有和较高的知名度相匹配，旅游者在看到山海关真实的形象与自己心中想象的截然不同而产生心理落差，由此导致了负面评论的产生。

主题二中包含了景点、排队、人多、厕所、取票等五个特征词。将主题二概括为“景点基础设施”，厕所、排队、取票这几个特征词，可以看出景区的基础设施没有满足游客的需求，由此导致取票、上厕所时人过多而排队，影响游客游玩体验。主题三中包含了门票、贵、停车、收费、没有等五个特征词。将主题三概括为“景点收费管理”。主题三的特征词主要围绕收费进行展开，主要是游客对景点门票、停车的收费感觉不合理，由此导致游客负面情绪的产生。

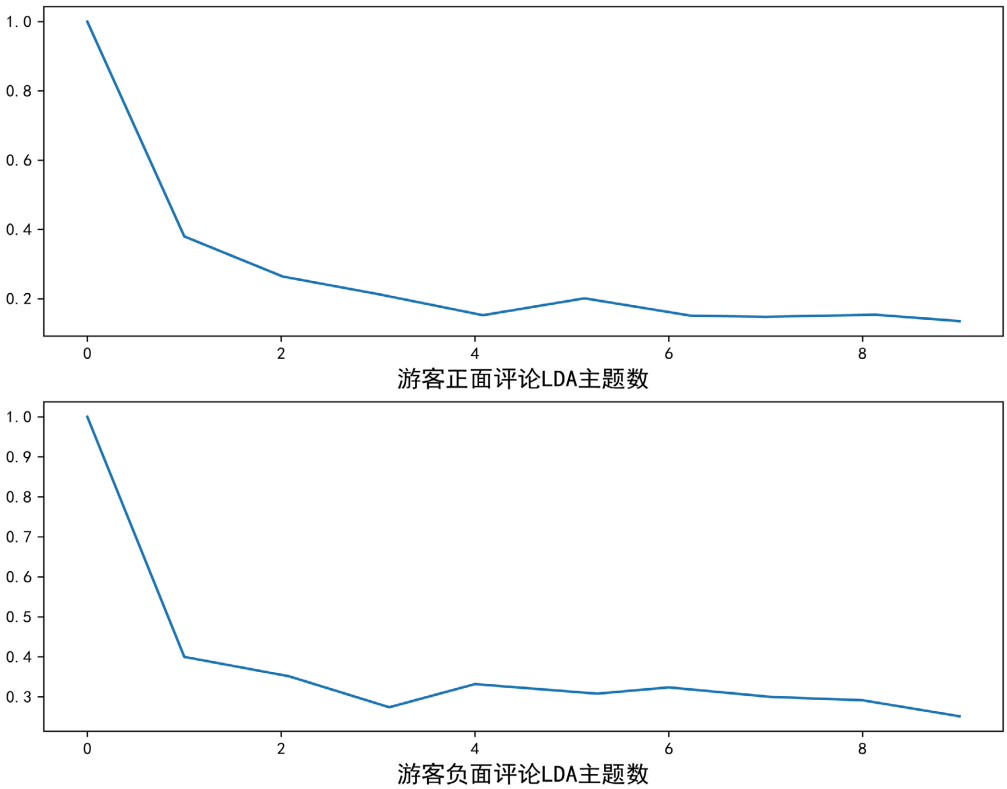

Figure 11. Topic number-confusion line chart
图 11. 主题数－困惑度折线图

Table 7. Potential topics of negative comments
表 7. 负面评论潜在主题

主题一		主题二		主题三	
山海关	0.025	景点	0.046	门票	0.061
天下第一关	0.019	排队	0.036	贵	0.044
知名	0.017	人多	0.032	停车	0.036
景区	0.016	厕所	0.025	收费	0.019
古代	0.011	取票	0.021	没有	0.018

山海关景区正面评价的潜在主题模型结果见表 8，通过表可以看出正面评论包含四个主题每个主题有五个主题相关特征词及其权重。主题一中包含历史、长城、城墙、人文、朝代等五个特征词。故将主题一概括为“景点历史文化底蕴”，正面主题一中特征词与山海关负面主题一十分接近，二者主要在围

绕山海关的历史人文、知名度来讨论。作为知名历史文化景区，山海关历史底蕴较为充足，吸引了众多游客慕名而来。主题二中包含景点、不错、感觉、雄伟、气势等五个特征词。故将主题二概括为“景点体验感受”，主题二中气势、雄伟用来形容山海关整体留给游客的印象，历经百年历史的山海关在这部分游客的眼里显得更加的雄伟、气势、壮观。主题三中包含门票、导游、身份证、网上、方便等五个特征词。故将主题三概括为“景点服务”，主要反应景点取票、进出等服务的便利性以及导游讲解服务。旅游业作为服务业，就是要以服务立身，服务质量是影响游客印象的关键所在，只有景点服务到位、贴心，才会打造良好的旅游形象，吸引更多的游客前来游玩进而增加景点收入。主题四中包含晚上、灯光、表演、游玩、值得等五个特征词。故将主题四概括为“景点趣味性”，值得注意的是，主题四中出现了体现游客情感倾向的特征词“值得”，可以看出游客对山海关景区的夜景、表演很满意，山海关景点夜景灯光秀以及传统文化表演是吸引旅游者的重要因素。

Table 8. Positive comments on potential topics
表 8. 正面评论潜在主题

主题一		主题二		主题三		主题四	
历史	0.046	景点	0.037	门票	0.019	晚上	0.281
长城	0.042	不错	0.035	导游	0.018	灯光	0.247
城墙	0.026	感觉	0.021	身份证	0.017	表演	0.224
人文	0.022	雄伟	0.018	网上	0.017	游玩	0.179
朝代	0.021	气势	0.017	方便	0.016	值得	0.120

至此，本研究第二个目标也顺利实现。

6. 总结及建议

6.1. 结论

游客评论文本的分类效果：本文构建了 LR、NB 和 LSTM 三个文本分类器模型，通过综合比较各机器学习模型，得出结论：LSTM 模型表现最佳，准确率为 92.15%，传统机器学习模型也表现不俗，其中 LR 模型准确率为 89.54%，NB 模型次之，准确率为 87.73%。LSTM 分类模型可以实现对游客评论文本的精确分类，可以考虑在线上推广使用，解决旅游平台文本分类不准确的问题。

游客评论文本的情感分析：针对该旅游景点游客评论文本，使用训练好的逻辑回归模型分析了游客对景点的情感倾向，统计得到游客积极评论占比 79.6%，消极评论占比 20.4%，说明游客对山海关景区整体感受是满意的。

游客评论文本的 LDA 主题聚类分析：对分好类的文本构建了 LDA 主题聚类模型进一步提取游客关注点，发现游客的不满意主要集中在景点收费管理、景点基础设施方面，对景点的服务、体验感受、趣味性比较满意。

6.2. 建议

6.2.1. 针对游客建议

山海关景区整体留给游客的印象多是雄伟、气势、壮观，来此地游玩游客获得了极大的满足感和幸福感；同时，部分游客在评论中表示景点的取票很方便，还着重提到了导游解说很专业，服务很到位，

使自己的游玩体验很棒，也有部分游客对于山海关景区的夜景以及灯光表演留连忘返，对于喜欢历史文化景点又注重服务体验的游客，山海关景区会是个不错的选择。

6.2.2. 针对景点建议

景区应该关注游客需求，完善基础设施，增加检票闸机，景区还需要常备一些口罩医疗物资，做好防疫措施，保证游客安全游玩，采取预约参观形式，设置阻隔带，合理安排人流。增加公共停车位，统一管理收费制度，集中整治收费乱象，保证游客游玩体验。在旅游资源的开发中应更加注重历史文化的挖掘以及历史底蕴的展现，注重打造历史文化品牌，才更好地吸引游客前来游玩，不能单纯靠“牌子”。

参考文献

[1] Tan, S.B. and Zhang, J. (2007) An Empirical Study of Sentiment Analysis for Chinese Documents. *Expert Systems with Applications*, **34**, 2622-2629. https://doi.org/10.1016/j.eswa.2007.05.028

[2] 刘志明, 刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1): 1-4.

[3] 周咏梅, 阳爱民, 杨佳能. 一种新闻评论情感词典的构建方法[J]. 计算机科学, 2014, 41(8): 67-69+80.

[4] 魏慧玲. 文本情感分析在产品评论中的应用研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2014.

[5] 郭小芬, 刘聪, 李炜. SVM 在中文广告分类中的应用[J]. 电信技术, 2017(10): 73-76.

[6] 丁照银. 基于机器学习的评论文本分析[D]: [硕士学位论文]. 芜湖: 安徽师范大学, 2019.

[7] 应昊东. 基于文本挖掘的新能源汽车评论情感分析研究及应用[D]: [硕士学位论文]. 上海: 东华大学, 2021.

[8] 戴维. 逻辑回归解决文本分类问题[J]. 通讯世界, 2018(8): 266-267.

[9] 孙晓东, 倪荣鑫. 中国邮轮游客的产品认知、情感表达与品牌形象感知——基于在线点评的内容分析[J]. 地理研究, 2018, 37(6): 1159-1180.