# 基于LDA的民宿评论情感分析研究

宋俊毅，赖昀耀，胡飞菊，朱余平

**摘　要**　以携程旅行平台上海迪士尼度假区的民宿评论数据为研究对象，利用文本挖掘技术对用户评价进行深入分析。首先，基于 TF-IDF 方法提取评论中的高频词，通过分析高频词来了解用户对民宿服务的关注点。然后，应用 LDA 主题模型对评论进行主题分类，识别出评论中的隐性主题和话题分布，进一步分析用户的不同需求和偏好。最后，结合词典情感分析技术对评论进行情感倾向性分类，并结合 TF-IDF 统计特征词的情绪值。研究结果可为提升民宿服务质量和制定市场策略提供参考。

**关键词**　LDA；TF-IDF；情感分析；民宿评论
**中图分类号**　G2　　**文献标识码**　A　　**文章编号**　1674-6708（2024）368-0111-05
DOI:10.16607/j.cnki.1674-6708.2024.23.017

## 0　引言

随着旅游业的快速发展，民宿因其较低的价格和良好的居住体验成为大量游客的首选[1]，成为旅游住宿领域的重要组成部分。互联网时代，绝大多数的游客通过线上平台挑选预订民宿，随之产生的海量用户评论蕴含了游客对民宿的意见和情感态度。但用户评论以非结构化的文本形式存在且数量极为庞大，为进一步从评论中提炼有价值的信息，本文基于携程旅行平台的民宿评论数据，使用 LDA 主题模型和情感分析技术，探究游客进行民宿预订时的关注点及情感倾向，通过对这些评论进行深入挖掘和情感分析，有助于民宿的管理者改进民宿质量、提升用户满意度和制定精确的市场策略。

## 1　相关研究

目前在关于用户评论挖掘的相关研究中，有多名学者在不同研究领域，如海洋博物馆、电商平台、高校图书馆等进行了分析研究，为本文的民宿评论挖掘研究提供了参考。

杜利明等[2]以京东商城用户评论数据集作为研究对象，运用 LDA 主题分析方法挖掘京东商城服务的影响因素。张涛等[3]融合 Word2Vec 与 LDA 模型，对我国算法治理政策进行文本分析，探索政策主题与演变趋势，为政策制定提供决策依据。王浩和方俊涛[4]通过 LDA 模型对国家海洋博物馆游客在线评论进行主题分析，揭示了游客体验与需求，助力博物馆服务优化。张文德等[5]利用 LDA 主题模型分析“双一流”大学图书馆用户评论，探讨用户关注焦点与情感态度，为图书馆建设提供了参考。高娜和东梅[6]共同运用了 Word2Vec 与 LDA 模型分析中国省级五年规划中的文化政策文本，揭示了文化政策主题的时空演变特征。

## 2　数据来源与研究方法

### 2.1　数据来源

本文利用八爪鱼采集器，对携程旅行平台上的

基金项目：江西省大学生创新创业训练计划项目“一种改进的词语相似度计算方法及其在民宿评论情感分析中的应用”（编号：S202210846038）。
作者简介：宋俊毅，江西科技学院信息工程学院，研究方向为计算机软件。
赖昀耀，江西科技学院信息工程学院，研究方向为计算机软件。
胡飞菊，讲师，江西科技学院信息工程学院，研究方向为计算机软件、大数据分析。
朱余平，江西科技学院信息工程学院，研究方向为计算机软件。

民宿评论进行数据采集。为了确保数据的时效性和代表性，本文爬取了携程旅行平台上海迪士尼度假区 2024 年 1 月至 7 月的民宿评论数据。在爬取到数据后，通过关键词过滤去除了含有广告性质的评论，手动清除了评论中的特殊字符和无意义符号，确保文本的纯净度，最终得到两千余条有效评论。

### 2.2 研究方法

TF-IDF 是一种统计方法，用以评估一个字词对于一个文件集或一个语料库的重要程度[7]。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 的主要思想是如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

LDA 主题模型是一种贝叶斯无监督学习方法，用于挖掘文本中的潜在主题信息[8]。假设文档由多个隐主题构成，其中每个主题通过不同概率生成特定的词汇。文档主题内词项的分布通过 Gibbs 抽样等方法经多次迭代后收敛得出。LDA 模型广应于用户评论和热点关注的分析。

基于词典的情感分析方法是一种常见的文本情感分析技术。该方法通过构建一个情感词典或情感词汇表，其中包含了一系列带有情感倾向的词汇和对应的情感极性（如正向、负向或中性），然后通过匹配文本中的词汇与词典中的词汇进行情感倾向的判断[9]。

本文使用八爪鱼采集携程旅行平台上海迪士尼度假区 2024 年 1 月至 7 月期间的民宿评论数据，通过 jieba 分词进行预处理，然后利用 LDA 主题模型和情感分析技术对评论数据进行主题分类和文本挖掘，深入分析了游客对民宿的感知与满意度。具体研究路线如图 1 所示。

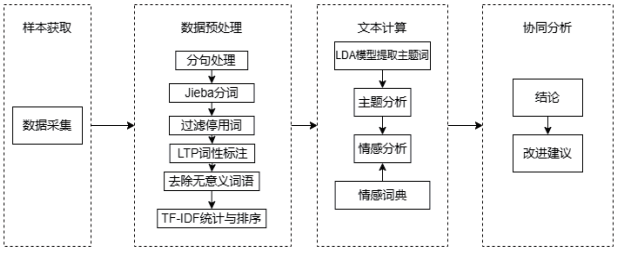


图1 研究路线图

## 3 研究过程

### 3.1 词频分析

本文通过八爪鱼采集器抓取民宿评论数据保存到文本文件中，利用 python 按照标点进行分句处理，接着使用 jieba 分词，然后自定义停用词典过滤无意义的词，其次使用 ltp 进行词性标注，去除感叹词、量词等无实际意义的词语，最后采用 TF-IDF 方法对高频词进行统计和排序，部分结果如表 1 所示。

表1 民宿评论高频词统计

| 序号 | 高频特征词 | 权重 | 序号 | 高频特征词 | 权重 |
|---|---|---|---|---|---|
| 1 | 推荐 | 0.193 | 11 | 下次 | 0.038 |
| 2 | 干净 | 0.089 | 12 | 还会 | 0.037 |
| 3 | 卫生 | 0.086 | 13 | 周到 | 0.035 |
| 4 | 异味 | 0.077 | 14 | 方便 | 0.033 |
| 5 | 省去 | 0.074 | 15 | 贴心 | 0.031 |
| 6 | 烦恼 | 0.068 | 16 | 排队 | 0.030 |
| 7 | 很近 | 0.068 | 17 | 吃饭 | 0.030 |
| 8 | 热情 | 0.049 | 18 | 很快 | 0.030 |
| 9 | 实用 | 0.042 | 19 | 很大 | 0.030 |
| 10 | 体验 | 0.038 | 20 | 价格便宜 | 0.030 |

利用 Python 绘制高频词词云图，如图 2 所示直观展示了高频词汇。从词云图可以看得出，“推荐”“干净”“卫生”等词语表明顾客对民宿卫生的高度关注，反映出卫生条件是顾客选择民宿的重要考虑因素。“异味”和烦恼“烦恼”等词提醒管理者应重视并改进顾客反馈中的负面问题。“热情”“周到”“贴心”“友好”等词汇则突出优质服务在顾客心中的重要性，表明顾客对服务人员态度和服务过程的细致程度有较高期望。“性价比”和“价格便宜”表明顾客对价格较为敏感，性价比高的民宿更容易获得顾客青睐。“温馨”“舒适”“安静”等词凸显了环境与舒适度对提升住宿体验的重要性。

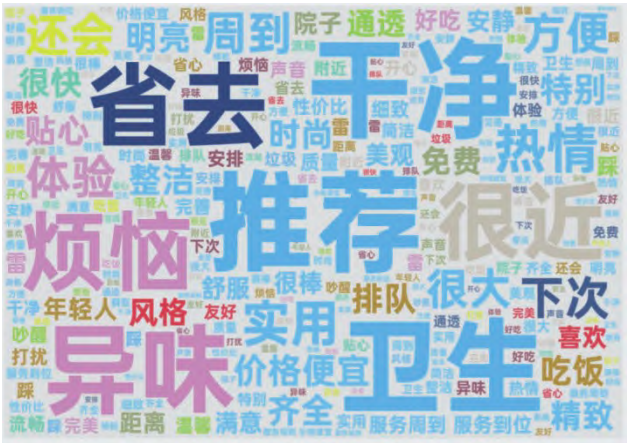


图2 民宿评论高频词云图

### 3.2 主题分析

本文使用 LDA 模型对预处理后的民宿评论文本进行主题分析。LDA 模型的效果依赖于几个关键参数的选择，包括主题数、迭代次数、α 和 β[10]。

为了优化模型的性能并提高结果的准确性，本文采用自适应算法来确定 α 参数的最佳取值。具体来说，α 值与主题数 K 值相关联，通过公式 α=50/K 动态调整。这样，在主题数较少时，每个文档可以分配到更多的主题，从而捕捉评论中的多维度体验；在主题数较多时，文档的主题分布则更加集中。与此同时，β 值设定为 0.1，确保每个主题聚焦于少数高频词，提高主题的可解释性，使模型能够更清晰地揭示出各个主题的核心内容。为了确定最佳的主题数，本文从 K 值 2 到 10 之间逐步增加，利用困惑度和主题一致性指标进行评估。实验结果显示，当 K 值为 4 时，模型表现最佳，所提取的主题能够全面覆盖游客体验、民宿位置、房间布局和民宿服务四个主要方面。这一自适应算法的使用，确保 LDA 模型在不同 K 值下都能有效捕捉评论中的关键主题。根据 LDA 模型进行主题分析后得到的四个关键主题，具体主题分布情况如图 3 所示。

主题 1 的特征词包括“度假区”“交通”“环境”“班车”和“环境优美”等。顾客对“度假区”民宿的偏好显示出他们希望在假期中获得放松和休闲的体验。“交通”和“班车”表明顾客重视民宿的交通便捷性，特别是在热门度假区，顾客期望能够轻松前往各类景点和设施。“环境”和“环境优美”则反映了顾客对周边自然景观和居住环境的期望。因此，本文将主题 1 命名为“民宿位置”。

主题 2 的特征词包括“隔音”“房间”“整体”“用品”“细节”和“体验感”等。顾客对“隔音”效果的重视表明他们希望在入住时享有一个安静、不受打扰的环境，这对位于繁忙地区的民宿尤为重要。“房间”和“整体”显示顾客对房间布局和设计的期待，他们期望房间布局合理且舒适。“用品”和“细节”则表明顾客关注房间内物品的质量和布置，他们希望房间的每一个细节都能精心设计，从而提升整体的“体验感”。因此，本文将主题 2 命名为“房间布局”。

主题 3 的特征词包括“贴心”“耐心”“热情”“主动”和“亲切”等。顾客对“贴心”服务的期待表明他们希望在住宿过程中感受到宾至如归的待遇。“耐心”和“热情”显示出顾客对服务人员态度的高度重视，他们期望民宿员工始终保持耐心和热情，及时满足他们的需求。“主动”和“亲切”则反映了顾客希望服务人员能够主动提供帮助，并以友善的态度对待每一位客人。因此，本文将主题 3 命名为“民宿服务”。

主题 4 的特征词包括“值得”“有特色”“方便的”“满意”和“舒适”等。频繁出现的“值得”和“有特色”表明顾客对那些能够提供独特体验的民宿有高度认可，尤其是那些能够满足他们需求并带来惊喜的住宿环境。同时，“方便的”“满意”和“舒适”则反映了顾客对入住期间便利性和舒适度的重视。因此，本文将主题 4 命名为“游客体验”。

### 3.3 情感分析

本文结合知网、台湾大学、大连理工大学三大情感词典对上述预处理的评论数据进行情感分析，如图 4 所示统计了各主题不同情感的特征词数量分布情况，纵坐标代表特征词个数。


图4 主题情绪占比图

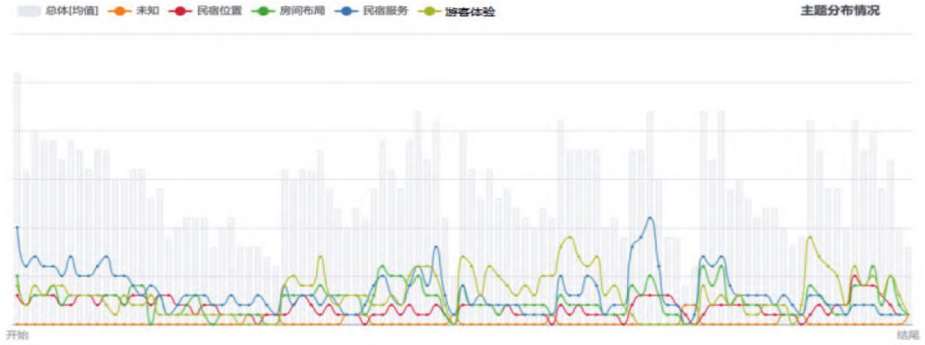


图3 主题分布情况

如图5所示是主题特征词的情绪值与数量分布的统计结果，情绪值计算由词频*TF-IDF权值*词汇情感倾向（正向为1，负向为-1）得出。从结果来看，正向情感评论的占比高于负面评论，但正向评论的平均得分偏低，其中四个箭头所指向的值为各主题情绪值的平均值。结合平均值进行分析，正向评论得分偏低的主要原因在于，多数民宿在房间布局、地理位置选择以及服务质量上存在不足，导致游客的实际体验与预期存在差距。当游客发现实际住宿情况与预期不符时，往往倾向于使用中性语言提醒和告诫其他游客。相反，当游客的实际体验优于预期时，他们可能不会对其做出评价。

由于多数评论情感趋于中性，这表明民宿管理者应对此引起足够的重视。管理者需要采取措施改善房间布局和服务态度，从而提升游客的居住体验。同时，鼓励游客给予正面评价，促进商家与游客之间的良性互动，从而推动民宿行业的健康发展。

## 4 结论与建议

### 4.1 研究结论

本文使用八爪鱼采集器获取携程旅行平台2024年1月至7月的上海迪士尼度假区民宿评论数据，通过分句、分词、去停用词和词性标注等操作进行数据预处理，然后使用TF-IDF提取数据中的高频词，接着采用LDA技术进行主题聚类分析，最后对每个主题进行情感分析，以探讨游客对民宿的关注焦点及情感倾向。研究结果显示：1）通过观察主题情绪值与数量分布情况图，发现游客对民宿的整体情感倾向偏向中性，表明游客对当前的民宿居住体验并不完全满意。游客最关注的是民宿的位置，尤其对“度假区”“景点”“车站”等词汇表现出高度关注。此外，游客的主要关注点还集中在民宿位置、设施和服务方面。尽管部分民宿地理位置优越，但在设施和服务方面仍有提升空间。2）在居住体验中，游客尤为关注民宿的室内设施布局和服务人员的态度，特别是服务是否及时响应游客需求，是否满足其期待。

### 4.2 改进建议

#### 4.2.1 提升环境质量

第一，管理者应确保清洁工作定期且彻底地进行，尤其是在客房、卫生间、厨房等易脏区域。此外，民宿应提供舒适的床上用品、家具和室内装饰，使游客在入住期间能够获得良好的休息和放松体验。

第二，管理者应采取多种措施减少噪声，包括安装隔音窗、厚重窗帘以及在房间内放置吸音材料，确保在夜间和清晨保持安静环境，使游客能够得到充分的休息。

#### 4.2.2 改进服务水平

第一，员工入职时应接受全面培训，包括服务礼仪、沟通技巧、应急处理和设备操作等，并定期进行在职培训，提升文化素养和多语言能力，以应对不同背景的游客。

第二，员工应始终保持热情、礼貌和耐心，建立以客户为中心的服务文化，主动发现和解决游客问题，提供个性化服务，满足并超出游客期望。

#### 4.2.3 加强客户反馈管理

第一，建立客户反馈机制，方便游客表达意见和建议。客房内放置反馈表，提供线上和线下反馈渠道，如网站、App、邮件和电话。退房时，主动询问游客的入住体验并收集反馈。

第二，及时处理反馈意见，特别是服务问题和投诉。指定专人负责收集和处理反馈，确保每条反馈都得到重视和回应。迅速调查和解决负面反馈，提供补救措施；对积极反馈和建议表示感谢并纳入改进计划。

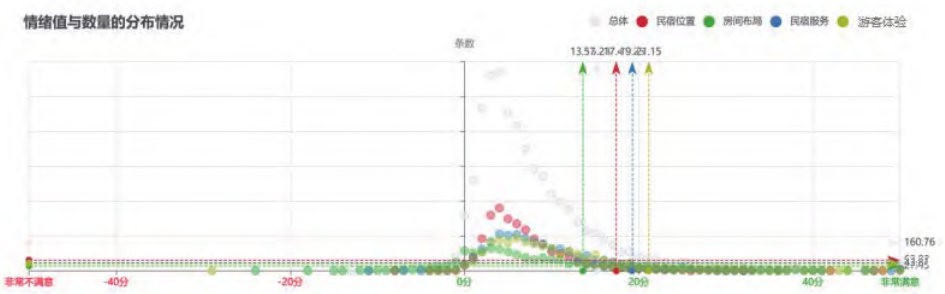


图5 主题情绪值与数量分布情况

### 4.3 讨论

本研究通过对携程旅行平台上的民宿评论数据进行分析，提出了提升环境质量、改进服务水平、加强客户反馈管理的具体建议，为民宿的服务改进提供了重要参考。然而，研究中存在数据来源单一、分析方法局限性等问题。未来研究应从扩展数据来源、改进情感分析方法等方面入手，进一步提高研究的科学性和实用性，为推动民宿行业的持续发展和繁荣助力。

### 参考文献

[1]戴其文，承忠彬，徐伟，等.多维距离对民宿游客行为意向的影响：兼论主客互动的调节效应[J].地理科学，2024（9）：1597-1608.
[2]杜利明，郭文艳，崔蕾，等.基于LDA的电商平台用户评论挖掘与情感分析研究——以京东商城App为例[J].江苏科技信息，2024（12）：125-129.
[3]张涛，王瀚功，崔文波.我国算法治理政策与科研主题协同研究——基于LDA与Word2vec融合模型[J].网络安全与数据治理，2023，42（8）：13-20.
[4]王浩，方俊涛.基于LDA模型对国家海洋博物馆游客在线评论的主题分析[J].科技和产业，2024（12）：224-230.
[5]张文德，徐子杨，赵立红，等.基于LDA主题模型的“双一流”高校图书馆用户评论文本数据挖[J].计算机工程与应用，2024（7）：120-127.
[6]高娜，东梅.基于Word2Vec和LDA主题模型的中国省级五年规划“文化政策”文本研究[J].网络安全与数据治理，2024，43（7）：47-55.
[7]英东升，翟江涛，周桥，等.一种基于eXpose和BiLSTM的DGA域名检测方法[J].东北师大学报(自然科学版)，2024，56（3）：53-61.
[8]王孟，苏进城，陈志德.基于LDA和Word2Vec模型的学位论文评阅意见主题挖掘与分析[J].福建师范大学学报(自然科学版)，2024，40（5）：41-51.
[9]CAI Y，YANG K，HUANG D，et al.A hybrid model for opinion mining based on domain sentiment dictionary[J].International Journal of Machine Learning and Cybernetics，2017.
[10]陈蛟.在线评论边际价值递减研究[D].武汉：武汉大学，2023.

(上接第110页)

布的内容塑造符合自我期待和对外展示的形象，反映了用户对自我形象建构的意识和需求，故而后期选择“公开”这一出口为他人提供反馈渠道，更加满足整饰个人印象的需要。

### 参考文献

[1]王琳，李云婧，施茜.大学生社交网络用户健康焦虑自我披露意愿的影响因素研究[J].情报科学，2024，42（3）：33-42，51.
[2]王长潇，张剑峰，张丹琨.数字原住民在视频平台中的隐私管理行为研究——基于传播隐私管理理论与隐私计算理论综合视角[J].当代传播，2024（4）：15-21.
[3]庄睿，于德山.作为情感劳动的隐私管理——中国留学生代购群体的社交媒体平台隐私管理研究[J].新闻记者，2021（1）：80-89，96.
[4]王波伟，李秋华.大数据时代微信朋友圈的隐私边界及管理规制——基于传播隐私管理的理论视角[J].情报理论与实践，2016，39（11）：37-42.
[5]李彪.双重规训与有限权利：互联网平台治理视域下的社交自我消除行为研究[J].西北师大学报(社会科学版)，2022，59（5）：65-71.
[6]董晨宇，段采薏.反向自我呈现:分手者在社交媒体中的自我消除行为研究[J].新闻记者，2020（5）：14-24.
[7]陈素白，项倩.自我表露动机与角色压力视角下的朋友圈隐私管理机制研究[J].新闻大学，2022（12）：32-45，122.
[8]陈阳，张睿丽.仅自己可见的朋友圈：社交媒体想象的互动[J].现代传播(中国传媒大学学报)，2020，42（12）：56-62.
[9]陈乾生.“日常生活中的自我呈现”：从日记到微信朋友圈[J].青年记者，2023（10）：107-109.
[10]晏庆合，操瑞青.新私人书写与公共化：社交媒体用户自我呈现中的“隐而不退”实践[J].传媒观察，2023（10）：66-76.