项目报告

**基于情感词典与机器学习的酒店评论情感分类研究**

**摘 要**

随着电商经济的兴起,更多的人选择在线上预订酒店出行,并将自己的体验进行在线评论反馈。因此,在大数据的背景之下,如何高效地利用这些评论数据就显得尤为重要。因携程在电商旅游平台行业中处于龙头地位,本文认为携程酒店评论具有代表性并以其为例进行实践。首先,本文选取了上海地区位于推荐前四页的32家酒店，利用spyder对用户评价内容及其相关评论信息进行爬取,最终共获取2829条评论信息。然后,利用jupter notebook对获取到的文本数据进行清洗、分词、去停用词等预处理工作。最后,对处理好的数据进行情感分类,本文使用的情感分类方式主要包括情感字典和机器学习算法。首先采用情感词典分类的方式对文本数据做出大致分类的界定,并进行可视化展示，进而使用TF-IDF算法构建词向量并进行数据集划分。然后,根据相应的分类算法对训练集和测试集分别进行训练和测试,最终得出情感分类效果最好的分类器为逻辑回归模型。通过对情感极性分类结果可以看出,用户对该区域酒店体验的整体情感倾向是积极的；但实际评分和情感打分存在一定偏差，可以对评分的有用性进一步分析等。最后,上述所得结论和可视化结果，对酒店商家等主体有一定的参考价值。

关键词：酒店评论;情感词典;机器学习;自然语言处理;

**目 录**

**[一、绪论](#_Toc106549103)** [4](#_Toc106549103)

[第一节、项目背景 4](#_Toc106549104)

[第二节、项目内容 4](#_Toc106549105)

[第三节、项目组织框架 5](#_Toc106549106)

**[二、数据采集](#_Toc106549107)** [5](#_Toc106549107)

[第一节、目的 5](#_Toc106549108)

[第二节、目标网站分析 6](#_Toc106549109)

[第三节、主体代码的介绍 6](#_Toc106549110)

[第四节、数据的存储 7](#_Toc106549111)

**[三、数据集介绍与预处理](#_Toc106549112)** [7](#_Toc106549112)

[第一节、数据集介绍 8](#_Toc106549113)

[第二节、数据预处理 8](#_Toc106549114)

[（一） 总体数据清洗 8](#_Toc106549115)

[（二） 评论文本预处理 8](#_Toc106549116)

[（三） 探索性分析 9](#_Toc106549117)

**[四、实验设计与实验结果](#_Toc106549118)** [10](#_Toc106549118)

[第一节、 情感词典 10](#_Toc106549119)

[（一） 情感词典介绍 10](#_Toc106549120)

[（二） 实验设计 10](#_Toc106549121)

[（三） 实验结果 11](#_Toc106549122)

[第二节、 机器学习 13](#_Toc106549123)

[（一） 实验设计 13](#_Toc106549124)

[（二） 实验结果（预测效果） 14](#_Toc106549125)

[第三节、 本章小结 14](#_Toc106549126)

**[五、项目总结与展望](#_Toc106549127)** [14](#_Toc106549127)

**[参考文献](#_Toc106549128)** [16](#_Toc106549128)

**一、绪论**

**第一节、项目背景**

互联网的用户规模已经十分庞大,人们大多有使用互联网来表达个人想法、获取有用信息的习惯。文章、视频、图片和评论信息等等，这些信息中隐藏着用户的情感倾向,具有很高的研究价值。因此，获取用户发表在互联网上的评论,挖掘文本中隐藏的用户情感,分析影响用户情感起伏的因素,对互联网商家，媒体和政府部门等等都具有很高的参考意义。从企业的角度来说，获取互联网应用平台上消费者的评论数据相较于走访式的市场调研等更为高效和高质。但是过于庞杂的信息仍带来一定的干扰和不便：在酒店领域,各个酒店的评论留言随着时间的推移激增,从中筛选有效信息变得繁琐而费时。[1]与此同时,意见的多样性表达对文本研究具有很高的挑战性。评论文本作为一种主观文本，以个体的情感和意图为依据,在进行文本描述时,对字词的使用具有很强的随意性,甚至在句式的使用上会显得相对不规范与口语化。[2]

在文本分析中，找对重点是前进的关键。评论文本的主客观情感分析是提取用户兴趣偏好的最基本和最重要的补充。因此通过研究用户的情感倾向,并应用到相关领域的预测中,辅助决策与判断,是近年来的关注方向。而情感分类是情感分析中最重要的一个环节。情感分类是一种倾向性分类，所研究的对象是作者所表达出来的主观倾向性,分类的结果一般为褒义情感或是贬义情感,甚至是复杂多层次的情感。[1]目前有很多种分类方法，一般按照情感词典和机器学习两大类进行划分，本项目将对这两个方向展开一定探索。

**第二节、项目内容**

本项目基于携程网——国内知名在线旅游平台展开研究，首先使用网络爬虫从网络平台上获取项目分析所需要的酒店评论数据集，接着进行数据预处理，包含数据清洗、去符号字符、切词、去停用词、词性标注等步骤。然后基于《知网情感词典》开展情感分类工作，并绘制正面，负面情感词云和相关性热力图进行展示。

另一方面，开展机器学习工作。对文本数据进行TF-IDF特征提取，转换成机器看得懂的语言，并构建三种分类算法，分别是朴素贝叶斯，决策树和逻辑回归模型，实现对评论情感的分类及预测。

**第三节、项目组织框架**

本项目的组织框架如图 1.1 所示：

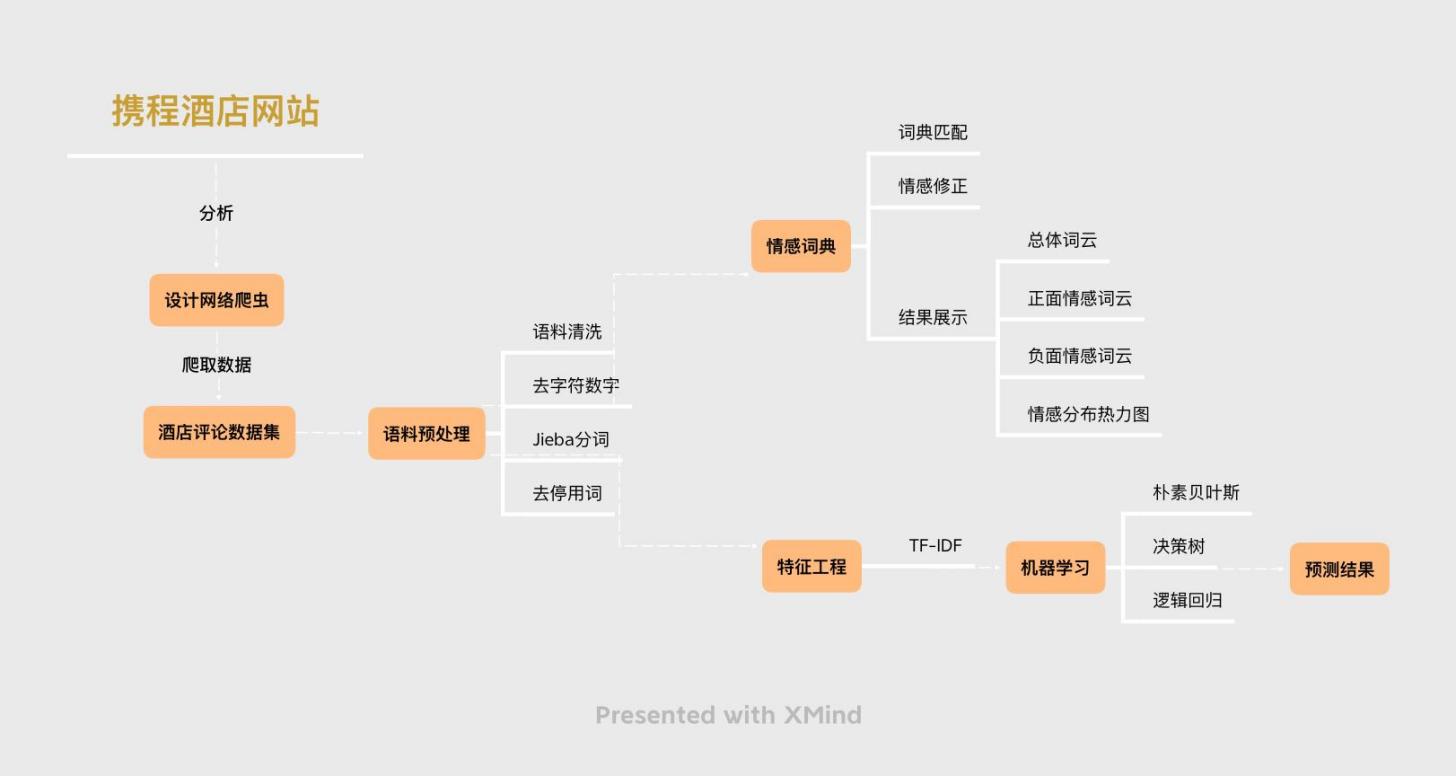


图 1.1 项目组织框架图

本报告内容将按如下顺序进行安排：

第一章为绪论，介绍本项目的项目背景以及研究内容；

第二章为数据采集，详述本项目所使用数据集的采集过程与部分爬虫代码示

例；

第三章，对本项目使用的数据集进行介绍， 描述数据的预处理过程；

第四章，详细介绍本项目的实验过程，并对实验结果进行解释；

第五章，总结项目。

**二、数据采集**

**第一节、目的**

笔者的目的是爬取携程网站上，目的地为上海，位置区域为商业区的最受欢迎的酒店以及对应的酒店评论,如图 2.1所示。



图 2.1 携程酒店图

数据包括：酒店名字；用户id；评论者；评分；入住时间；发布时间；入住目的；房间类型和评论内容，并将这些数据存入Excel中。

**第二节、目标网站分析**

该网站有一定的反爬机制，查看评论需要用户的登录。因此笔者对headers进行处理：打开开发者工具，将xhl中的headers粘贴，伪造用户登录,其中起主要作用的是user-agent和cookies。一些技术文章还提到伪造多个cookie随机更换，这里由于设置的爬取速度较慢，爬取量不大，再加上携程网站的反爬机制没有美团等网站严格，所以没有复杂化cookie。

再者，携程网站使用动态加载，因此，同样在网站的headers获取网页跳转的筛选参数，如page参数。

**第三节、主体代码的介绍**

将筛选城市等参数后的酒店首页作为start\_url,其中pageNo置为{},使用for循环翻页，爬取1-4页的酒店id和酒店名字，使用正则表达式提取，并分别放入hotelId\_list和hotelNames两个list中。

接下来爬取每一个酒店的对应评论，见图 2.2。将无关参数放在head，将所需参数放在data并初始化，将伪装user-agent放在headers。因为电脑端可以查看到的评论有限，因此将user-agent切换成移动端。

将第一个酒店的评论首页作为url,用for循环动态跳转其他评论地址。使用otherCommentList存储爬取到的所有评论，然后用for循环对每一条评论进行逐个处理，主要提取hotelName, ids, userNickName, ratingPoint, checkInDate, postDate, travelType, baseRoomName, content这九项信息。在调试中发现，有的评论没有选择房间类型，导致baseRoomName 报错keyerror，因此特地加了一个try...except...



图 2.2 携程酒店评论图

**第四节、数据的存储**

初步将数据存储成xlsx格式，提取每一条评论的上述九项信息为一行，并加上编号，数据量为2829\*10，结果如图2.2所示。由于数据预处理在jupter notebbok上进行，因此还将xlsx转换成csv格式。



图 2.2 数据存储结果

**三、数据集介绍与预处理**

本项目按照第二章所介绍的数据采集方法从携程网获取了本项目所需的数据集，记为酒店评论数据集。本章将对数据集进行介绍，并说明数据的预处理过程。

**第一节、数据集介绍**

该数据集共包含了 2829条评论，涉及了32个酒店，评分分布在1.0-5.0之间，入住目的包括家庭亲子，情侣出游，其他等6种，房间类型包括标准大床房，高级大床房等182种。其中各个字段及其示例如表 3-1 所示：



表 3-1 数据集字段及示例

**第二节、数据预处理**

数据预处理分为三部分，分别是对数据总体做缺失值重复值等处理；以及提取评论文本，做评论分词等处理；并做探索性分析。

1. **总体数据清洗**

首先，我们从采集的数据集中剔除掉不需要的字段'评论者'，并对数据重复值和缺失值做统计，删除同一用户发出的相同评论，发现没有关键数据缺失，所以不作缺失值处理。

1. **评论文本预处理**

通过人工观察数据发现，评论中夹杂着许多数字，字母等字符，对于本案例的挖掘目标而言，这类数据本身并没有实质性帮助。另外，由于该评论文本数据主要是围绕酒店质量进行评价的，其中“上海”“酒店”“房间”“携程”等词出现的频数很大，但是对分析目标并没有什么作用，因此在分词之前将这些词去除，对数据进行清洗。

接下来进行评论分词。因为电脑难以分析一个句子，但是将句子拆分成词语，就能对词语分析比对。比如通过词语本身所代表的情感，判断是好评或者是差评。通过匹配停用词词典和语气词词典，去掉一些没有意义的符合和词语。比如标点符合、停用词、语气词等等。并将词语，词性，对应评分打包成一行进行存储，作为专门的评论内容（storage-评论内容），如表 3-2 所示。



表 3-2 评论文本预处理结果

同时将处理好的评论文本按照对应索引复制一份到总体数据中存储（storage-总体结果）。

1. **探索性分析**

由于本项目主要关注评论文本，所以暂不对其他变量做描述性分析。

绘制总体评价词云，如图3.1所示。词云将针对文本中出现频率较高的‘关键词”予以视觉上的突出。首先需要对词语进行词频统计，将词频按照降序排序,选择前200个词，使用wordcloud模块中的WordCloud绘制词云，查看分词效果。



表 3-1 总体评价词云

从宏观角度来看，从上面可以明显的识别出4类关键词：

酒店服务：前台，入住、热情

酒店设施：环境、设施、干净

酒店周边：地理位置、位置、交通

其他关键字：早餐、晚上、免费

说明客户对这几方面比较重视。也可以看出正面评论居多，说明该地域的多数酒店质量符合客人预期。

**四、实验设计与实验结果**

本章将对处理好的数据进行情感分类，主要使用情感词典和机器学习两种方法。首先采用情感词典分类的方式对文本数据做出大致分类的界定,进而采取特征值提取的形式对数据集进行准确划分。然后,根据相应的分类算法对训练集和测试集分别进行训练和测试,实现对评论的情感分类和预测，并对实验结果进行分析。

1. **情感词典**
2. **情感词典介绍**

用户评价的内容往往反映情感表扬和批评的态度，情感词典的质量对情感分类的效果有直接影响。由于没有专门基于酒店语料的情感词典，因此选择了目前常用的《知网情感词典》，包括了“正面评价词语”，“正面情感词语”，“负面评价词语”，“负面情感词语”四部分。并根据酒店语料的特征，手动将“干净”，“便宜”等词汇加入到正面评价词语中，将“脏乱差”，“吵”等等词汇加入到负面评价词语中。[3]

1. **实验设计**

在情感词典的基础上，进行词典匹配。读入正负面评论情感词表，正面词语赋予初始权重1，负面词语赋予初始权重-1。并将分词结果与正负面情感词表合并，定位情感词和权重。

考虑到汉语中有双重否定的存在，会改变语义，因此对情感倾向进行一定的修正。读入否定词表，对情感值的方向进行修正：若情感词前面存在否定词，那么存在一个调整成相反的情感权重，存在两个就不调整。计算每条评论的情感得分，将评论分为正面评论和负面评论。

1. **实验结果**
2. **相关性热力图展示**

首先对分类结果进行统计：正向评论数目为：2242条 ；负向评论数目为：160条，见表4-1。

从统计结果中看出，即使评分在1.0左右，仍有正向评论，即使评分在5.0左右，仍有负向评论。其中，对于极端情况，如评分为1.0的正向评论，做异常值处理。但整体来看，反映出评分和文本情感倾向尽管存在一定相关性，但并不完全联动。

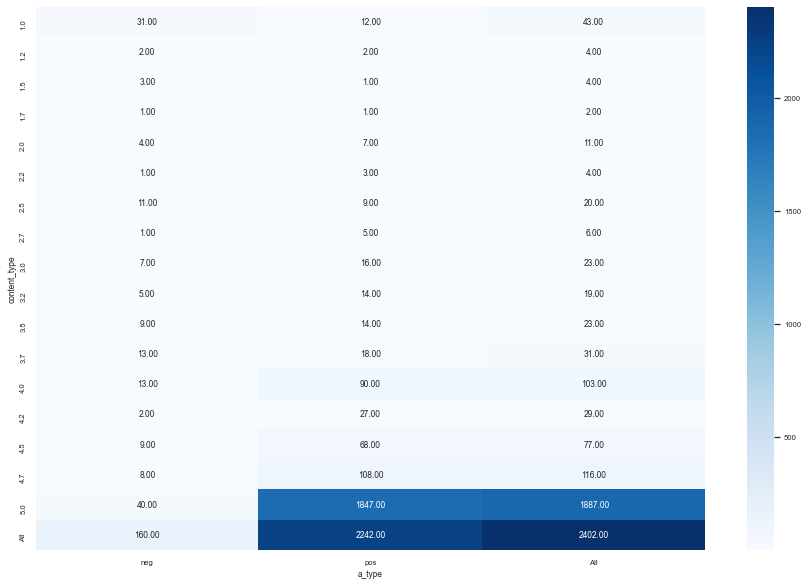


表 4-1 实际评分与文本情感倾向的相关性热力图

结合现实分析，可能的原因如下：

1. 情感词典存在缺陷。网络文本存在多样化表达，多口语化，很多热词新词不断涌现，比如“奇葩”，可能没有收录在《知网情感词典》中；同时，酒店评论有酒店的语境，比如“进口设施”是正面评价，但没有在这种通用的情感词典中体现，所以在识别过程中仍有调整空间。
2. 客人的评论文本不能完整地表达情感。例如客人尽管给出了4分以上的高分，但是在评论里着重强调了不满意的一方面，对总体的满意程度没有着重说明，所以两者很容易互相矛盾。
3. 评分虚高。一些用户可能不会特别认真地对待评分，若不选取评分便提交评价，系统会默认五星好评。
4. **词云展示**



图 4.1 正面情感词云



图 4.2 负面情感词云

由正，负面情感评论词云可知,两种情感的评论进行了很好地分离，因此可以得出基于情感词典的情感分类能较好地将情感评论抽取出来。

从分析结果看出，该区域的酒店在入住等酒店接待和环境卫生等方面总体评价较高，但在食物品质和部分服务人员的培训上还有提升空间，酒店管理可以加强这方面的重视，作为区别与其他酒店的竞争力。

若商家要进一步比较竞争优劣，需要单独对某个酒店制作正面、负面情感词云做比较分析。

**第二节、 机器学习**

1. **实验设计**

接下来，本项目将通过机器学习的模型训练，分类并预测评论的情感倾向。

首先，确定并处理自变量和因变量。以评论文本为自变量，并使用TF-TDF算法进行特征抽取，将文本信息转换成向量。其中，TF表示关键词的频率，TDF表示逆文档的频率,主要用来表示每个关键字在文档的重要性。以评分（1.0-5.0）为因变量，进行多分类。

但是由于数据集不平衡，即好评较多（4.0-5.0），差评较少（4.0以下），所以多分类效果很差，模型的测试分数基本为0.因此对模型调优，将多分类问题变成二分类问题。

|  |  |
| --- | --- |
| 评分 |  |
| 评分<=4 | 0 |
| 评分>4 | 1 |

表 4-2 因变量“评分”的参数表示

进一步，划分数据集。划分为训练集与测试集，其中训练集包含了 80%的评论，测试集包含了 20%的评论。对数据进行标准化处理。

开始模型训练。本项目采用了两类模型来进行下面的实验，分别是基于抽取的文本特征的机器学习模型，采用的机器学习模型有朴素贝叶斯分类，决策树和逻辑回归。

在实验结果的评估环节，本项目选用了分类问题常用的精确率（precision），召回率（recall）， F1-score这三项指标对预测结果进行评估。[4]

1. **实验结果（预测效果）**

根据酒店评论的文本内容预测好评差评的结果如表4-3所示，在整体上这三种机器学习模型均具有良好的分类效果和预测效果。其中，逻辑回归模型的表现是最好的，它的Precision，Recall和F1-score均为0.905，因此，进行情感分类效果最好的分类器应该选择逻辑回归分类器。

|  |  |  |  |
| --- | --- | --- | --- |
| 分类模型/指标 | Precision | Recall | F1-score |
| 朴素贝叶斯 | 0.894 | 0.894 | 0.894 |
| 决策树 | 0.867 | 0.867 | 0.867 |
| 逻辑回归 | 0.905 | 0.905 | 0.905 |

表 4-3 情感二分类预测结果

**第三节、 本章小结**

本章基于情感词典和机器学习两种路径实现了的情感的二分类。总体来说，两者的分类效果可佳。从情感词典分类方法中我们得到了三方面的结论：

1. 整体来说，客人对该区域的酒店比较满意，正面评价居多。
2. 实际评分与情感倾向存在一定的相关性，但不完全联动，并对此做出了可能性解释。
3. 从正、负面情感词云和总体词云中提取了客人较为关注的关键词和方面，有助于酒店的管理。

从机器学习分类方法中，我们了解到三种模型的二分类效果都比较好，其中逻辑回归最优。

**五、项目总结与展望**

本项目基于携程网平台，对酒店评论语料中的情感倾向展开了研究。首先，本项目根据该网站特点设计了相应的网络爬虫来获取实验所需的数据；其次，本项目对爬取的数据，尤其是评论文本数据进行了清洗与预处理，并用词云、热力图的形式对数据进行了可视化的展示；接着，本项目设计了实验，在数据集上分别进行了基于情感词典和机器学习的情感二分类，对分类及预测结果进行可视化展示，用到了词云、热力图等形式，并对实验结果做出了分析，得到了一些新的结论，对酒店商家有一定参考价值。

同时，研究过程也存在一定的不足，需要进行更深入的研究和改进。

由于酒店评论语料是中文网络文本，中文本身含义丰富并且情感多种多样,网络用语更新迅速,文本中包含的表情包也有丰富的内涵等等，因此：

1. 可以对情感词典进行更全面的扩充,关注蕴含使用者感情的网络用语,,收集各种的表情符号以及标注其对应的情感等。
2. 对文本进行多级情感分类的研究,本文只是对文本进行积极和消极的二分类,然而人的情感是多种多样的,可以把情感归类为多个类别。

再者，本项目只针对上海某区域的酒店进行整体画像，如果要对酒店商家提出更具针对性的意见，需要单独对一家酒店进行情感分类画像等等

# 参考文献

1. 石馨. 基于酒店领域情感词典的分类器研究[D].河北大学,2014.
2. 丁蔚. 基于词典和机器学习组合的情感分析[D].西安邮电大学,2017.
3. 马子睿.基于情感词典的旅游景点推荐算法[J].电脑编程技巧与维护,2022(04):20-21+37.DOI:10.16184/j.cnki.comprg.2022.04.009.
4. 郝冰川. 基于语料特征的文本分类算法研究[D].燕山大学,2019.DOI:10.27440/d.cnki.gysdu.2019.001372.