文章编号:1007-1423(2018)20-0045-05 DOI:10.3969/j.issn.1007-1423.2018.20.009

# 基于朴素贝叶斯的酒店评论情感倾向性分析

焦凤

(上海海事大学信息工程学院,上海 201306)

**摘要:**
提出一种基于朴素贝叶斯的面向酒店评论领域的情感分析模型,构建酒店评论领域专用情感词典,使用机器学习的方法训练分类器进行酒店情感分类,得出最受用户关注的酒店特征,进一步挖掘出其中五大主要特征的满意程度。实验结果表明:利用朴素贝叶斯分类器与酒店专属词典、布尔权值计算相结合的方法对酒店评论有较好的分类效果。

**关键词:**
酒店;情感分类;布尔权值;朴素贝叶斯

## 0 引言

随着互联网的迅速发展,人们热衷于通过网络表达自己的态度和看法。用户发布微博、酒店评论、电影评论等来表达自己的情感。这些文本信息都包含了大量的情感信息,都是宝贵的情感语料资源,商家和消费者可以参考评论信息做出更加合理的判断。利用情感分析能从评论中获取用户的情感信息,可以准确地分析出评论中情感的倾向性。

文本情感分析技术主要分为两类:基于语义方法和基于机器学习方法。基于语义方法是指利用计算情感词语的情感值来判断文本的情感极性,杨超等人在HowNet 和 NTUSD 的基础上进行扩展,建立了一个具有倾向程度的情感词典[1],取得更加准确的分析结果。评论语料中的情感词、情感词的上下文及情感词修饰的特征词等因素都对句子情感倾向性产生影响。为了获取精确的分类结果,Meena A 等人重点考虑了连词对句子情感极性分析[2]的影响,结合短语和连词分析句子情感极性。基于机器学习的方法是指利用已经标注的语料,采用不同的特征权值[3]公式、特征选择方法对采集来的语料进行处理,利用机器学习算法进行训练得到分类器,用训练好的分类器对新文本进行识别。Fei Zhongchao 根据机器学习的方法研究了外文中体育相关评论[4]的情感分析,通过实验研究获得了有效的结果。当前研究结果表明,基于机器学习的方法分类效果优于基于语义的方法[5]。卢玲提出一种新的基于朴素贝叶斯的中文文本情感分类[6]方法,用情感短语作为文本特征,通过情感词典与否定副词相结合提取情感短语,再利用朴素贝叶斯分类器进行情感分类计算。文献[7]提出基于表情符号的中文微博情感分析系统,把微博中使用频率最高的 59 个表情符号分为 4 种情感类别。然后使用朴素贝叶斯分类器训练新浪微博的具有情感标注的微博语料,实验表明表情符号的使用提高了微博分类准确率。

由于酒店评论的口语化严重等问题,在酒店评论情感分析方面做的研究还不够。本文通过构建酒店专属词典,基于朴素贝叶斯分类器对酒店评论分类,并对酒店评论中的单因素进行分析,得出消费者对单因素的好评率。

## 1 相关工作

### 1.1 酒店评论情感分析模型

本文通过在酒店评论文本特点的基础上提出了图 1 酒店评论情感分析模型,该模型主要包括四大模块:数据抓取模块、预处理模块、分类器训练模块、分类结

果展示模块。使用爬虫程序爬取三大品牌酒店的评论数据，通过分词、去停用词、特征选择、特征权重计算等步骤后对评论文本进行逐条标注。将预处理后的训练样本训练分类器，用训练好的分类器对文本分类，得出评论文本情感倾向性。

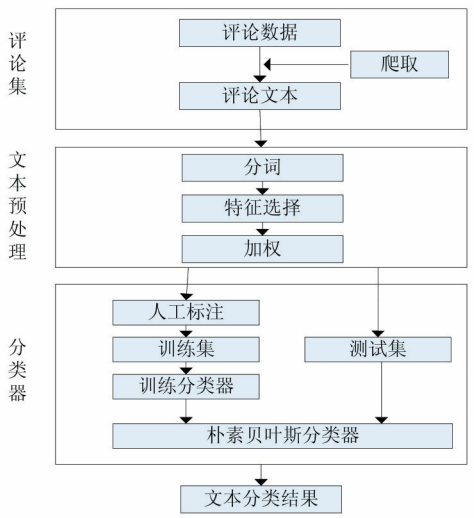


图1 酒店评论情感分析模型

### 1.2 构建情感词典

目前常用的传统情感词典是 NTUSD 词典[8]，NTUSD 情感词典包含 2810 个正面情感词汇、8276 个负面情感词汇，可以适用于各领域在线评论的情感倾向性分析。由于酒店在线评论拥有酒店领域相关的情感词语，例如“隔音”“床大”等情感词只出现在酒店评论中，在其他评论领域不会出现或者出现较少，这些词语没有收录在传统的情感词典中，基于传统情感词典不足以完成对酒店评论的分析。现将 NTUSD 词典和酒店领域专属情感词 126 个、网络评论中的网络流行词 42 个以及表情情感词 27 个合并，构成酒店评论领域专有词典。部分酒店领域专属情感词和网络流行词如表 1 所示：

表 1 新增情感词(部分)

| | |
|---|---|
| 酒店专属词 | 漏水、宽敞、实惠、网卡、地滑、床大... |
| 网络流行词 | 高大上、辣鸡、蓝瘦、香菇、醉了... |
| 表情情感词 | [撇嘴]、[微笑]、[难过]、[愉快]、[咒骂]... |

## 2 文本预处理

### 2.1 分词处理

使用 Python 语言编写爬虫程序爬取酒店评论文本，构建酒店评论语料库。例如文本“房间很干净，前台服务员非常有礼貌。”进行文本情感分类前需要对文本进行预处理，步骤如下：

(1)去除无效评论：评论中很多无效评论，很多其他商品广告之类，会干扰分类效果。

(2)分词：中文文本只有字、句、段等分界形式，无法直接获取实验需要的情感词语。实验前需要进行分词处理，本文使用中科院分词系统 ICTCLAS，将爬取得文本分词成“房间\很\干净\,\前台\服务员\非常\有\礼貌\。\”。

(3)去停用词：本文使用哈工大去停用词表去除符号、助词、动词等。

### 2.2 特征选择

常用的特征提取方法有信息增益法、互信息法、CHI 统计法等。其中 CHI 特征选择算法[9]利用了统计学中的"假设检验"的基本思想：假设特征 t 和类别 $c_i$ 之间符合 CHI 分布，CHI 统计值越大，特征与类别之间的相关性越强，对类别的贡献度越大。

实现步骤如下：

(1)统计样本集中文档总数(N)。

(2)统计特征词 t 和类别 $c_i$ 共同出现的次数(A)，特征词 t 出现但 $c_i$ 不出现的次数(B)，类别 $c_i$ 出现但 t 不出现的次数(C)，特征词 t 和 $c_i$ 都不出现的次数(D)

(3)计算每个词的卡方值，计算公式(1)如所示：

$$CHI(t,c_i)=\frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (1)$$

(4)取特征 t 的最大值作为其全局 CHI 统计量，公式(2)如下：

$$CHI_{\max}(t)=\max_{i=1}^{|c|}\{CHI(t,c_i)\} \quad (2)$$

### 2.3 特征权重

文本中的特征权重通常都基于词频来统计文本信息，赋予特征对应的权重。常用的特征权重计算方法有：布尔权重、绝对词频、TF-IDF 等多种方法。本文基于布尔权重的特征权重计算方法来计算特征在文本中的权重值。布尔权重型权值计算方法如式(3)：

$$f=\begin{cases}1, t_i\in c\\ 0, t_i\notin c\end{cases} \quad (3)$$

其中:若特征词出现在文档中，则权值为 1。否则，权值为 0。

## 3 构建分类器

朴素贝叶斯[10]是一种简单有效的分类方法，它遵守“贝叶斯假设”，即文本的特征项是相互独立的。其分类过程如下：

（1）采集并人工标记样本数据。设 $X=\{x_1,x_2,\cdots,x_n\}$ 为一条酒店评论文本数据，$x_i$ 为 $X$ 的某一特征词；类别 $Y=\{y_1,y_2,\cdots,y_n\}$。

（2）计算特征词 $x_i$ 在类别 $Y$ 中的条件概率。即 $p(y_1|X),p(y_2|X),...,p(y_n|X)$。

（3）若 $p(y_i|X)=\max\{p(y_1|X),p(y_2|X),...,p(y_n|X)\}$，其中 i<n，则 X 的类别属于 $y_i$。根据“贝叶斯假设”，利用公式（4）计算概率：

$$p(y_i|X)=\frac{p(X|y_i)p(y_i)}{p(X)} \tag{4}$$

要使得 $p(y_i|X)$ 取最大值，只需让 $p(X|y_i)p(y_i)$ 取最大值。鉴于各特征词之间相互独立，可按如下公式（5）计算：

$$p(X|y_i)=p(x_1|y_i)p(x_2|y_i)...p(x_n|y_i) \tag{5}$$

其中，$p(x_j|y_i)$ 由训练样本估值，计算公式如（6）所示。

$$p(x_j|y_i)=\frac{s_{ij}}{s_i};\ \text{其中}j=1,2,\cdots,n \tag{6}$$

类别 $y_i$ 的先验概率的计算公式如下（7）：

$$p(y_i)=\frac{s_i}{s} \tag{7}$$

公式（6）和式（7）中，$s_{ij}$ 是有特征词 $x_j$ 且属于类 $y_i$ 的样本数，$s_i$ 是属于类 $y_i$ 的样本数，$s$ 是总的样本数。

（4）经过上面训练阶段，得到文本分类器。先用特征向量 $X=\{x_1,x_2,\cdots,x_n\}$ 将待分类的微博文本表示出来，然后主要的工作就是按照上述公式计算 $p(X|y_i)$ 及 $p(y_i)$ 的值。当且仅当 $p(y_i|X)=\max\{p(y_1|X),p(y_2|X),\cdots,p(y_n|X)\}$ 时，将文本划分为 $y_i$ 类。

## 4 实验及结果分析

### 4.1 实验数据

本文通过编写爬虫程序爬取上海浦东新区的如家、汉庭和 7 天连锁三家品牌共 15 个快捷酒店的 1572 条评论作为语料库，为了保证样本具有完整性和代表性，要求每一家酒店至少有 60 条评论，剔除无效评论 63 个，剩余 1509 个有效评论。随机抽取 80%的评论作为训练集，剩下 20%评论作为测试集。

### 4.2 评价标准

为了评定分类器分类的分类效果，我们采用准确率（Precision，简记为 P）、召回率（Recall，简记为 R）及 F 值作为实验的评价指标。公式如下：

$$P=\frac{A}{A+B} \tag{8}$$

$$R=\frac{A}{A+C} \tag{9}$$

$$F=\frac{2PR}{P+R} \tag{10}$$

其中，A 表示实际好评被分类器正确判断的评论数，B 表示实际差评被误判为好评的评论数，C 表示实际好评被误判为差评的评论数，D 表示实际差评被正确分类器判断为正确的评论数。定义如表 2：

表 2

| | 实际好评 | 实际差评 |
|---|---|---|
| 判定好评 | A | B |
| 判定差评 | C | D |

### 4.3 实验结果

实验一：为了获得精准的分类评价标准，随机抽取 80%的评论作为训练集，训练集样本数为 1207 个，使用训练集对评论情感分类器进行训练。完成对分类器的训练，用测试集进行实验，实行多次训练取平均值。试验结果如表 3：

表 3 测试集测试结果

| 测试集 | 召回率 R | 准确率 P | F 值 |
|---|---|---|---|
| 第一次 | 0.8720 | 0.8555 | 0.8636 |
| 第二次 | 0.8685 | 0.8482 | 0.8582 |
| 第三次 | 0.8701 | 0.8532 | 0.8607 |
| 平均值 | 0.8702 | 0.8523 | 0.8608 |

实验二：通过筛选文本特征提取阶段有关酒店属性的特征词，使用训练好的分类器对包含产品属性特征的句子进行情感分析，得出 3 个快捷酒店品牌顾客住宿单因素好评率。对五家酒店用户评论进行词频统计，根据词频出现频率从高到底选取特征词，得到酒店用户评论特征词集合{位置、环境、服务、设施、卫生、性价比、床、房间、早餐、价格、隔音、体验……}其中，类似隔音、房间和床等特征词都可以归为设施。经过分析总结，将部分特征词合并，选取 5 个最受关注的特征词{位置、卫生、服务、设施、环境}进行研究。分类器对包含特征词的句子分类结果如表 4：

表 4 单因素情感分类结果

| | 位置 | | 卫生 | | 服务 | | 设施 | | 环境 | |
|---|---|---|---|---|---|---|---|---|---|---|
| 酒店 | 好评 | 差评 | 好评 | 差评 | 好评 | 差评 | 好评 | 差评 | 好评 | 差评 |
| 如家 | 446 | 32 | 335 | 53 | 243 | 72 | 218 | 41 | 153 | 25 |
| 汉庭 | 384 | 43 | 362 | 74 | 227 | 58 | 276 | 62 | 112 | 23 |
| 7天 | 319 | 48 | 307 | 69 | 209 | 56 | 139 | 29 | 83 | 18 |

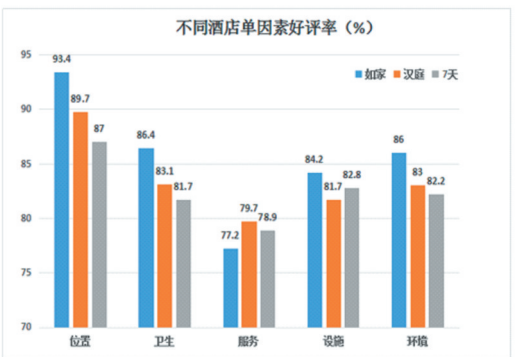


图2 单因素好评率

本文将单因素的好评率定义为单因素的好评数量与该单因素的评论总数的比值。如图 2 通过对三大品牌连锁酒店单因素方差分析结果显示，各单因素好评率有显著的差异。总体来看，消费者对如家连锁酒店的住宿好评率最高，汉庭和 7 天连锁次之。如家连锁酒店在位置、卫生、设施和环境等都优于其他两家连锁酒店。但是在服务方面，汉庭连锁酒店却优于其他两家酒店，其中 7 天连锁优于汉庭。对单因素好评率进行比较，可以发现服务的好评率最低，平均好评率为 78.6%，数据说明顾客对服务的要求越来越高，而酒店在一定程度上无法达到部分消费者的预期服务水平，提高服务水平是酒店方面需要改进和突破的重点。另外，部分酒店进驻市场时间较早，设施损耗严重，酒店也要注意对设施的维护与更新，提高用户的消费体验的满意度。

## 5 结语

本文主要研究酒店领域评论文本的情感分类问题。在已有的情感词典的基础上加入酒店专属词和表情词等，构建酒店专属情感词典。提出了一种基于朴素贝叶斯的酒店情感分类模型，实验表明，利用朴素贝叶斯分类器可以取得较好的分类效果。消费者和商家可以清晰地发现商品的优缺点，具有一定的应用价值。

鉴于酒店评论具有口语化的特点，语法错误带来的情感歧义会直接影响分类器的分类效果。下一步研究的重点主要对分类算法上进行改进，提高分类器分类的准确性，同时增强该分类算法在其他领域网络评论的适用性。

**参考文献：**

[1]杨超，冯时，王大玲，等. 基于情感词典扩展技术的网络舆情倾向性分析[J]. 小型微型计算机系统,2010,31(4):691-695.

[2]Meena A，Prabhakar T，Amati G，et al. Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis[C]. Proceedings of Advances in Information Retrieval. Berlin，Germany: Springer，2007:573-580.

[3]赵刚,徐赞.基于机器学习的商品评论情感分析模型研究[J].信息安全研究,2017,(02):166-170.

[4]FEI Zhong-chao, LIU Jian, WU Geng-feng. Sentiment Classification Using Phrase Patterns[J]. The Fourth International Conference on Computer and Information Technology，2004.

[5]Pang Bo，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)，2002：79-86.

[6]卢玲，王越，杨武等. 一种基于朴素贝叶斯的中文评论情感分类方法研究[J]. 山东大学学报(工学版)，2013，43(6):7-11.

[7]ZHAO Ji-chang, LI Dong,WU Jun-jie,et al. Mood Lens:an Emotion-based Sentiment Analysis System for Chinese Tweets in Weibo[C]. Proceedings of the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD),2012:1528-1531.

[8]杨超，冯时，王大玲，等 . 基于情感词典扩展技术的网络舆情倾向性分析[J]. 小型微型计算机系统，2010(4) : 691-695.

[9]谭松波. 高性能文本分类算法研究[D]. 北京:中国科学院计算技术研究所, 2005.

[10]贺鸣, 孙建军, 成颖. 基于朴素贝叶斯的文本分类研究综述[J]. 情报科学, 2016, 36(7): 147-154.

作者简介：
焦凤(1991-),男,安徽六安人,硕士研究生,研究方向为计算机系统应用技术
收稿日期:2018-05-22　　修稿日期:2018-07-02

# Emotional Inclination Analysis of Hotel Reviews Based on Naive Bayes

JIAO Feng

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306)

Abstract:
Presents a Naive Bayesian for comments based on sentiment analysis model, constructs the comments field special emotion dictionary, uses machine learning method to train classifier for hotel emotion classification, draws the most hotel features of user attention, further excavates the satisfaction of one of the five main features. The experimental results show that the combination of the simple Bias classifier and the exclusive Dictionary of hotels and Boolean weight calculation has a good classification effect for hotel reviews.
Keywords:
Hotel; Emotion Classification; Emotional Dictionary; Naive Bayes

(上接第 44 页)

# DCNN Brain Tumor Segmentation Method Combined with CRF

TANG Shi, WANG Fu-long

(School of Applied Mathematics, Guangdong University of Technology, Guangzhou 510520)

Abstract:
Automatic segmentation of MRI images of brain tumors has a great potential for better diagnosis, treatment and assessment of brain tumors. In recent years, many methods based on deep convolutional neural networks (DCNN) have achieved a series of successes in segmentation. Presents a fully automatic brain tumor segmentation method based on DCNN. At first, uses cascaded framework to hierarchically segment brain tumor and its substructures. Then, Atrous Spatial Pyramid Pooling helps to capture multi-scale information. Lastly, the information of relationship among pixels is integrated into the network by Conditional random field. The experiment shows that the proposed algorithm has some advantages.
Keywords:
Deep Convolutional Neural Networks; Conditional Random Field; Atrous Convolutional; Brain Tumor; Image Segmentation