信息安全研究 第3卷 第2期 2017年2月
Journal of Information Security Research Vol.3 No.2 Feb. 2017

基于机器学习的商品评论情感分析模型研究

赵 刚　　徐 赞
(北京信息科技大学信息管理学院信息安全系　北京　100192)
(zhaogang@bistu.edu.cn)

Research on the Sentiment Analysis Model of Product Reviews Based on Machine Learning

Zhao Gang and Xu Zan
(*Information Security Faculty, School of Information Management, Beijing Information Science and Technology University, Beijing* 100192)

Abstract Online product reviews have become the primary means to enable people to explain their own views on a particular commodity. And, the research on the sentiment analysis model owns values in both business and academic areas. Discussing on several machine learning models for sentiments analysis, using enlarged emotional dictionaries, and describing full machine learning procedures, this paper proposes a set of sentiment analysis model for the sentiment analysis on the catering industry. Then, this paper discusses some classify algorithms, such as Naive Bayes and C4.5, and gives detailed discussions about effects of different models based on various evaluation methods. The experimental results show that the proposed model gives full play to emotion dictionary efficiency, and is more suited to judge customer emotional tendencies.

Key words online product reviews; sentiment analysis; emotion dictionary; machine learning; model evaluation

摘　要　在线商品评论已成为对商品阐述看法的主要手段,对商品评论的情感分析研究具有学术及商业价值.研究情感分析领域若干机器学习模型,通过扩充情感词典,运用机器学习方法,设计餐饮领域网上评论情感分析模型.深入探讨朴素贝叶斯、C4.5等分类算法,利用多种性能评价方法,详细讨论不同模型的分析效果,结果表明所设计模型发挥出情感词典的有效性,更加适合于判断客户情感倾向.

关键词　网络商品评论;情感分析;情感词典;机器学习;模型评价

中图法分类号　TP309

收稿日期:2016-12-29
基金项目:国家自然科学基金项目(61272513);北京市科委重大项目(D151100004215003)

随着互联网的发展，越来越多的用户习惯于对商品进行在线评论，它不仅是用户在商品使用之后一种反馈的有效方式，同时对于商户以及浏览者都有着不可替代的参考性. 但是目前绝大多数评论没有被有效利用，在浩如烟海的评论中，人们无法直观得出结论. 因而如何有效挖掘处理这些评论是一件非常有意义的事情. 与此同时，由于互联网评论与传统的文本有着极大的差别，对于在线评论的情感分析比传统文本将更加困难[1-2]. 而对于情感分析的研究有着重大的学术价值以及直观的广阔的商业价值，目前在世界各国的计算机领域中，情感分析都成为了科学研究的热点方向[3-5].

情感分析是充分体现出多学科交叉的任务，本领域的研究者们将越来越多的其他专业领域的研究方法引入了文本情感分析这一领域，创新了许多优秀的方法，并促进了本领域的发展[6]. 而对向量空间模型(vector space model，VSM)在文本处理上的应用，让许多其他领域内的算法在情感分析领域得到应用成为可能[7-8]. 这使得各种机器学习算法能够方便地应用到文本情感分析中来，其中常用的分析算法有朴素贝叶斯(Naive bayes，NB)、Logistic 回归(Logistic regression，LR)、决策树(decision tree，DT)、支持向量机(support vector machine，SVM)、K-近邻算法、LDA(linear discriminant analysis)、QDA(quadratic discriminant analysis)等[9-10].

本文深入探讨了情感词典在在线商品评论情感分析中的作用，对典型的机器学习方法在餐饮评论情感分析效果进行了比较，在此基础上针对在线商品评论情感分析的隐含属性发觉、商品间关联性发现、客户情感倾向判断提出基于扩充情感词典方法的行之有效的机器学习情感分析模型.

1 商品评论情感分析系统

为得出基于网站评论的全面的餐厅评论情感分析，本文依据自然语言处理方法，结合机器学习方法对在线商品评论进行挖掘预处理以及情感分析. 首先选定大众点评上的餐厅获取语料，其次对语料实施预处理，运用改进的情感词典构造方法完成餐厅情感分析所需要的情感词典，利用机器学习算法，实现餐饮商户评论的情感分析. 本文的核心在于全面地挖掘商品评论以及基于对用户评论的分析处理而得到对餐厅评论情感分析的整体描述，其商品评论情感分析总体流程如图 1 所示：

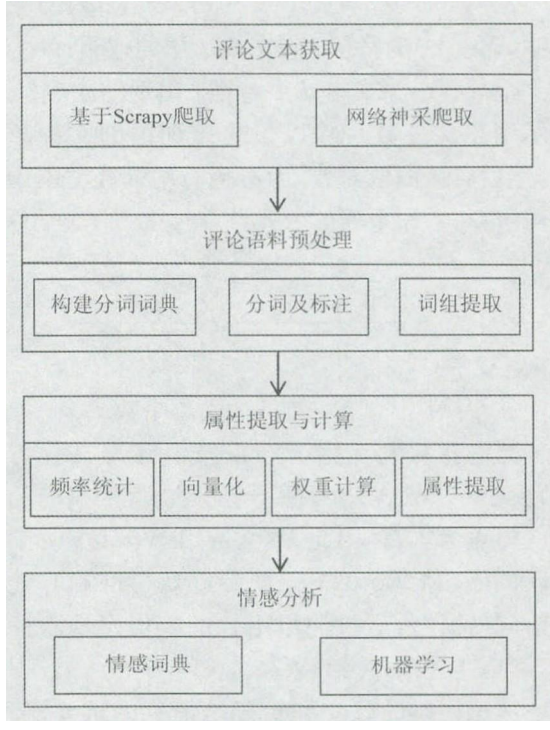


图 1 商品评论情感分析流程示意图

步骤 1. 语料获取，通过预处理选择单词乃至词组并提取特征；

步骤 2. 筛选特征，筛选关联词语以及发现隐性属性，构造扩充情感词典；

步骤 3. 文本特征化，对特征化的文本确定类标签；

步骤 4. 运用机器学习算法进行训练，构建情感分析模型；

步骤 5. 通过测试获得机器学习方法正确率、错误率、准确率、召回率、以及 ROC 曲线(receiver operating characteristic curve)等判断指标，分析确定符合餐饮评论情感分析的机器学习方法.

2 特征处理技术

为构建机器学习模型，首先需要计算权重，提取属性词语，筛选特征. 本文应用 TF-IDF(term frequency-inverse document frequency)作为计算权重的依据. TF-IDF 是基于统计学方法的文本词

语加权方法，用于评估文本向量在整篇文档集中的重要程度. TF-IDF 的核心思想在于影响权重的2个因素，即词频 TF 和逆向文档频率 IDF：

$$W(t,d) = tf_{ij} \times idf_i = tf_{ij} \times \mathrm{lb}(\frac{N}{idf_i}), \quad (1)$$

其中，$W(t,d)$为特征 t 在文本 d 中的权重；N 为总文本个数；tf_{ij} 为文本 d 中特征 t 出现的次数；idf_i 表示用总文件数目除以包含该词语的文件数的商，然后对该商取对数. 为避免有些词在文本中不出现，本文采用下面的公式计算权重：

$$W(t,d) = \frac{(1+\mathrm{lb}(tf_{ij})) \times \mathrm{lb}(N/n_t)}{\sqrt{\sum_{t \in d}[(1+\mathrm{lb}(tf_{ij})) \times \mathrm{lb}(N/n_t)]^2}}, \quad (2)$$

其中 n_t 为 N 中含有 t 特征的文本个数.

与潜在狄利克雷分布以及潜在语义分析相比，由于 Word2Vec 考虑了上下文，其所携带的语义变得更为丰富. 因此，本文选用 Word2Vec 筛选关联词语，对 Word2Vec 调参以获得准确有效的词语向量模型，通过词组组合的权值发现隐性属性，筛选非直接列出的菜品.

情感词典的建立和扩充是情感的重要流程. 在极性词典构建方面，本文将所获得的语料中所有词语做成字典，用以过滤普通极性字典中的词语，过滤没有在评论文档中出现的词语，从而提高词典利用效率. 利用 Word2Vec 扩展已有程度词典中的类似词语，达到扩充程度词典的目的，提高程度词典利用效率. 根据 TF-IDF 标注有情感倾向的词语，同样利用 Word2Vec 筛选相关联的近义词，扩充领域词典.

3 机器学习方法

针对商品评论情感分析问题，本文构建并测试了以下几种典型的机器学习算法，包括 Ada Boosting、Bagging、Bayes Network、Decision Tree、C4.5 分类树、Naive Bayes 分类器、Multinomial Naive Bayes 以及 Ripper 等算法. 这里就典型的且在测试中取得较好效果的朴素贝叶斯分类算法和 C4.5 分类树加以说明.

朴素贝叶斯分类算法作为典型概率模型算法，根据贝叶斯公式计算得出待分类文本分类的概率值，取这些概率中最大值完成分类. 假定文本集中每个文本样本可用一个 n 维特征向量表示，基于贝叶斯类条件概率方法计算待定新文本 $\boldsymbol{d}$ 的后验概率用 $p(\boldsymbol{c}|\boldsymbol{d})$表示：

$$p(\boldsymbol{c} \mid \boldsymbol{d}) = \frac{p(\boldsymbol{c})p(\boldsymbol{d} \mid \boldsymbol{c})}{p(\boldsymbol{d})}, \quad (3)$$

其中，$\boldsymbol{c}$ 表示类别向量. 使用朴素贝叶斯方法时，各特征之间视为完全相互独立，于是：

$$p(\boldsymbol{d} \mid \boldsymbol{c}) = \frac{p(\boldsymbol{d})}{p(\boldsymbol{c})} \prod p(t_k \mid \boldsymbol{c}). \quad (4)$$

在多项式模型中，设某个文本 $\boldsymbol{d}=(t_1, t_2, \cdots, t_k)$，$t_k$ 是该文本中出现过的单词，可以重复出现，则先验概率 $p(\boldsymbol{c})$为类 $\boldsymbol{c}$ 下单词总数/整个训练文本的单词总数，类条件概率 $p(t_k|\boldsymbol{c})$=(类 c_i 单词 t_k 在各个文本中出现过的次数之和+1)/(类 c_i 下单词总数+$|v|$)，$\boldsymbol{v}$ 是训练文本的单词表向量，$|\boldsymbol{v}|$则表示训练文本包含多少种单词，$p(t_k|\boldsymbol{c})$可以看作是单词 t_k 在证明 $\boldsymbol{d}$ 属于类 c_i 上提供了多大的证据，而 $p(\boldsymbol{c})$则可以认为是类别 $\boldsymbol{c}$ 在整体上占多大比例.

C4.5 算法是一种典型的决策树，C4.5 算法在构建决策树时同样采用自上而下的方法，在每一步选择一个最佳的特征来分裂. 这里最佳的特征通常是能够使得子节点中的训练集尽量的纯，在 C4.5 算法中最常用的衡量训练集纯度的指标是信息增益率. 信息增益率的计算如下：

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)}, \quad (5)$$

其中，$Gain(S,A)$表示信息增益，划分信息 $SplitInfo(S,A)$代表按特征 A 划分样本集 S 的广度和均衡性. 信息增益率越大代表训练集越不纯，所以在构建 C4.5 分类树的过程中就需要计算出每一个特征划分的信息增益率.

C4.5 分类树分别计算每个变量的各种切分和组合情况的信息增益率，找出该变量的最佳值组合和切分点，再比较各个变量的最佳值组合和切分点，最终找出最佳变量和该变量的最佳值组合和切分点.

C4.5 建树步骤如下：

1) 判断当前样本集是否满足终止条件，若没有，计算当前样本集每个特征的信息增益率；

2) 选择信息增益率最大的特征作为划分特征；

3) 将该特征划分出来的2部分样本集分别作为新样本集，重复步骤 1)～3).

4 测试分析

针对餐厅评论情感分析问题，利用筛选的1 900条特征，运用多种机器学习算法构建分类器，通过十折交叉测试的结果验证算法性能指标，确定有效的机器学习算法.

首先是各种算法正确区分积极、消极2种情绪的概率，如图2所示：

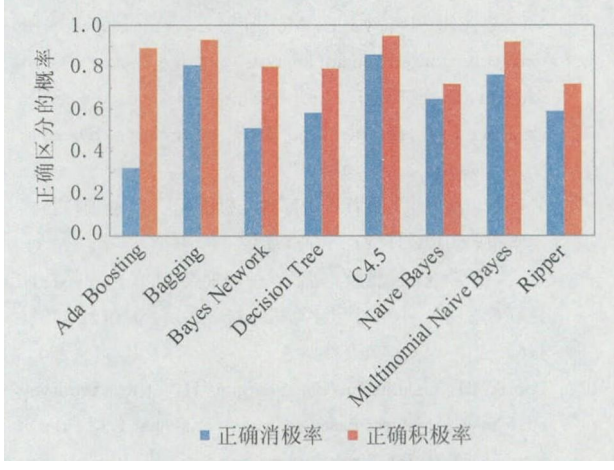


图2 正确区分2种情绪的概率

图2所示的是正确区分出积极情绪及正确区分出消极情绪的概率. 从测试结果可以看出，C4.5算法在2个方面表现优异. Ada Boosting在分辨积极情绪上表现优异，但在分辨消极情绪上结果十分不理想. 一般情况下，在分辨情绪领域中消极性的分析要难于积极性的分析. 可以看出较为优秀的算法是C4.5，Bagging，Multinomial Naive Bayes.

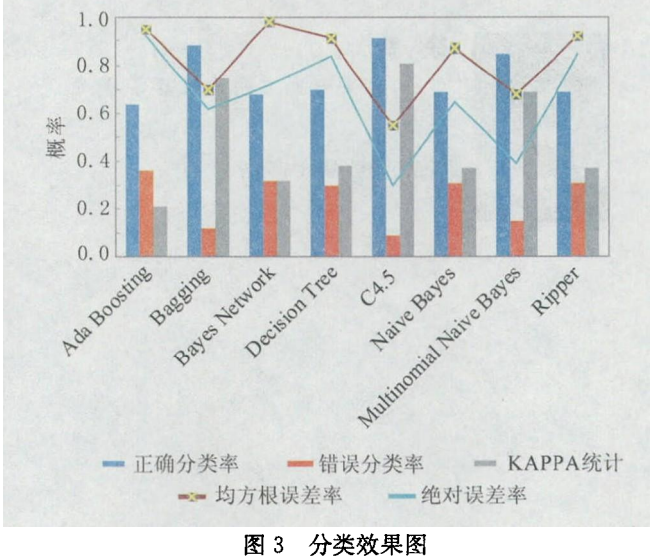


图3 分类效果图

在图3分类效果图中，绝对误差和均方根误差都是预测和实际的差值，因此越小越好，同样，误差率也是越低越好.

而KAPPA统计是分类器统计和真实分类的相对情况，其范围在[－1，1]之间，当取值靠近1时表示和真实分类越近似，当取值接近0时表示和随机分布越相似，当越靠近－1时说明和真实分类越相反. 从图3的性能指标结果可以看出C4.5，Bagging，Multinomial Naive Bayes是比较优秀的算法.

图4和图5的机器学习性能指标测试结果中，精确率表示正确识别的评论占总评论的多少，召回率表示正确识别的评论占所有应该正确识别的评论的多少. 精确查全综合指数是指综合了召回率和精确率的一个指数. 相关系数取值在[－1，1]

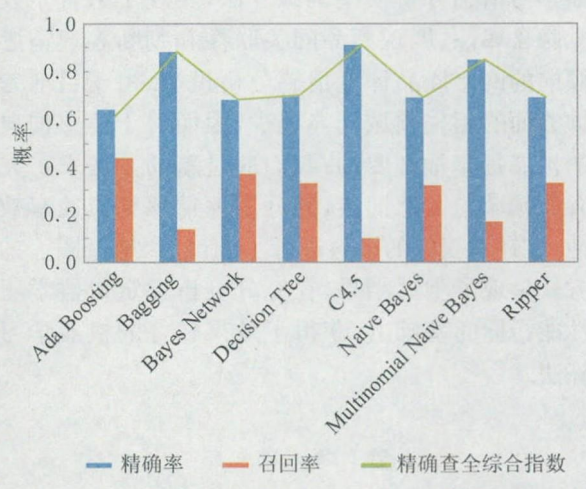


图4 分类属性图

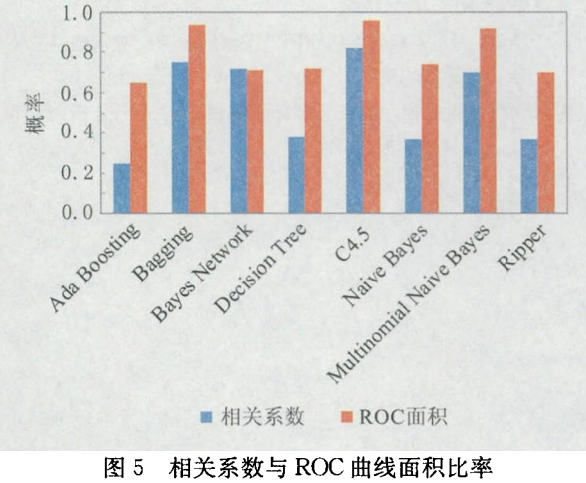


图5 相关系数与ROC曲线面积比率

之间,1代表完全等同于实际预测,0代表和随机预测效果是一样的,-1代表预测结果与实际结果完全相反.相关系数指标与KAPPA相比,更适用于类标签文本数量不平衡的语料.ROC曲线的面积范围在[0,1]之间,等于1时说明效果最好.

从图标中可以得出,C4.5,Bagging,Multinomial Naive Bayes算法的效果比较好.

5 结　论

本文介绍了自然语言处理和网络在线商品评论,详细说明了评论挖掘以及情感分析的常用方法.针对客户对餐厅商品的在线评论,本文在描述情感分析总体流程的基础上,进一步讨论了语料获取、特征提取、特征筛选、扩充情感词典构造、类标签确定的设计实现方法,详细分析了核心的机器学习情感分析模型训练算法,实现了适合于发觉隐含属性、展现商品间关联性和判断客户情感倾向的网上商品评论情感分析模型.由于目前餐饮方面的相关领域词典很少,因而对于情感词典的扩展是非常必要的.我们通过新的方法来扩充餐饮领域的情感词典,充分发挥情感词典的有效性.在针对获取的语料进行了测试后,通过图表的方式呈现了测试结果,在充分分析评定机器学习性能指标的基础上,获得了效果较好的机器学习算法.

参 考 文 献

[1] 崔连超. 互联网评论文本情感分析研究[D]. 济南:山东大学,2015

[2] 李阳,吕欣. 社交网络空间的安全问题与应对策略[J]. 信息安全研究,2015,1(2):126-130

[3] 张磊,陈贞翔,杨波. 社交网络用户的人格分析与预测[J]. 计算机学报,2014,37(8):1877-1894

[4] 蓝天广. 电子商务产品在线评论的细粒度情感强度分析[D]. 北京:北京邮电大学,2015

[5] Khan A Z H, Atique M, Thakare V M. Combining lexicon-based and learning-based methods for Twitter sentiment analysis [J]. International Journal of Electronics, Communication and Soft Computing Science & Engineering: Special Issue of National Conf on Advanced Technologies in Computing and Networking, 2015, 4(1): 89-91

[6] 吴燕波,向大为,麦永浩. 大数据时代的空间舆情管理与引导研究[J]. 信息安全研究,2016,2(4):356-360

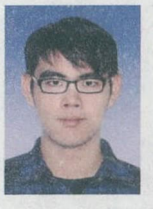
[7] Manek A S, Shenoy P D, Mohan M C, et al. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier [J]. Internet and Web Information Systems, 2016, 19(1): 1-20

[8] 朱郁筱,吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报,2012,41(2):163-175

[9] 彭云,万常选,江腾蛟,等. 一种词聚类LDA的商品特征提取算法[J]. 小型微型计算机系统,2015,36(7):1458-1463

[10] Pontiki M, Galanis D, Papageorgiou H, et al. SemEval-2016 task 5: Aspect based sentiment analysis [C] //Proc of SemEval 2016 Int Workshop on Semantic Evaluation. San Diego: Association for Computational Linguistics, 2016: 19-30

赵　刚
博士,教授,主要研究方向为人工智能与信息安全.
zhaogang@bistu.edu.cn



徐　赞
学士,主要研究方向为机器学习、信息安全.
xuzan@bistu.edu.cn