Hans Journal of Data Mining 数据挖掘, 2023, 13(3), 230-234
Published Online July 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.133023

Hans汉斯

基于LDA模型的商品评论情感分析研究

陈雅燕，林　耿

闽江学院数学与数据科学学院，福建 福州

收稿日期：2023年5月28日；录用日期：2023年6月28日；发布日期：2023年7月6日

摘　要

在大数据时代，商品评论情感分析可以帮助公司制定销售策略，提高产品性能，从而让消费者可以购买到优质产品。本文提出了一种基于LDA模型商品评论情感分析方法。该方法综合实际打分、预测出的评论为正面的概率、“有用”比例、是否购买、是否是会员五项指标计算出评论文本的综合情感得分。并根据以上研究结果，提出相关商品的改进建议，从而提高商品销售率。

关键词

LDA模型，TF-IDF词向量，情感分析，逻辑回归模型

Research on Sentiment Analysis of Product Reviews Based on LDA Model

Yayan Chen, Geng Lin

College of Mathematics and Data Science, Minjiang University, Fuzhou Fujian

Received: May 28th, 2023; accepted: Jun. 28th, 2023; published: Jul. 6th, 2023

Abstract

In the era of big data, emotional analysis of product reviews can help companies develop sales strategies, improve product performance, and enable consumers to purchase high-quality products. This article proposes a LDA based method for the sentiment analysis of product reviews. This method calculates the comprehensive sentiment score of the review text by integrating five indicators: actual scoring, probability of predicted positive reviews, “useful” ratio, whether to purchase, and whether to be a member. And through the above research results, suggestions for im-

文章引用: 陈雅燕, 林耿. 基于 LDA 模型的商品评论情感分析研究[J]. 数据挖掘, 2023, 13(3): 230-234.
DOI: 10.12677/hjdm.2023.133023

proving relevant products are proposed to improve the sales rate of the products.

Keywords

LDA Model, TF-IDF Word Vector, Emotional Analysis, Logical Regression Mode

1. 引言

随着数据时代的来临，对大量数据的正确处理分析可以起到促进消费的作用。计算机技术的蓬勃发展催生了大数据技术和相关挖掘信息方法的发展与应用。随着网络的发展，线上消费购物给人们带来了诸多便利的服务，足不出户就能轻松收到货物。

国外对商品评论评分的研究集中体现以下三个方面，第一：主观语言识别[1]。商品评论的文本属于自然语言，用来表达人类对这个事物的看法，褒贬不一，情感类别不一，也称为语言的倾向或极性。第二：情感倾向分类[2]。对商品评价文本处理时可以按照情感类别关键词来划分，大致可以划分为积极评价与消极评价。第三：通过挖掘商品之间的联系来分析商品间的相关性。多项研究表明，商品评论情感分析能够表达消费者对于产品的认知。但当前的研究主要是围绕商品评论和用户推荐，但少有学者综合其他变量来统计商品评价分数的变化。

因此，对商品评论的情感分析研究十分重要，要考虑综合文本数据。Saranya 等人[3]利用情绪来扩展见解，基于用户的信任偏好来分析用户的情感相互性，研究内部的潜在联系。刘永芬[4]改进了以往的支持向量机方法，提出了一种基于特征选择的多分类支持向量机方法。该方法可以有效地选取对分类有贡献的特征，提高了分类的精度和效率，在中文文章数据集中有良好的表现。曾小芹等人[5]利用 Selenium 爬虫索引，并使用 Jieba 单词分类器对评论文本进行分离和标记。在此基础上还利用 snowNLP 库进行情感分析，并将结果可视化，同时使用精确率和召回率对结果进行评估和分析。

基于此背景，本文提出：基于 LDA 模型的商品评论情感分析研究。首先进行数据清洗，再利用 LDA 模型得出主题词。其次利用 TF-IDF 词向量的方法将文本向量化，以评论正、负面作为标签，分别通过伯努利朴素贝叶斯模型、多项式朴素贝叶斯模型和逻辑回归模型进行初步文本情感分类，并利用精确率、召回率、F1 度量、ROC 曲线四种指标来评价三种模型预测评论为正面的效果，得出逻辑回归模型的预测效果最优。最后综合实际打分、预测出的评论为正面的概率、“有用”比例、是否购买、是否是会员五项指标计算出评论文本的综合情感得分。并通过以上研究提出相关商品改进建议，从而提高商品销售率。

2. 算法步骤

2.1. 数据处理

原始数据集中存在很多的无用信息，这对未来的研究挖掘有很大的影响，所以在进一步研究前首先要进行数据清洗。首先去除与数据分析无关的字段，如市场代码(marketplace)等字段，保留评论星级(star_rating)等字段。其次去除没有认证购买的商品评价，将数据标记“n”、“y”转换为“0”、“1”，将字符串数据转换成浮点数，便于下一步的统计分析和预测。最后将评价标题与正文拼接起来，去除无

关的标点符号，将缺失值填充为 0，将字母统一转换为小写，便于后续的文本分析。

2.2. 主题词提取

LDA 由 David M. Blei，Andrew Y. Ng，Jordan 于 2003 年提出[6]，用于推测文本文件围绕主题分布的情况。利用使用 LDA 方法进行主题建模。LDA 模型能够从大量的文本中挖掘出潜在的主题信息，而且这种模型具有良好的数据降维能力和模型扩展性，已经被广泛应用于各种文本分析的重要任务中。本研究找到了每种商品的最差评价(一星)和最好评价(五星)的三类主题词。

2.3. 初步情感分类

TF-IDF 是一类常用于海量信息精确检索和文本准确挖掘的加权方法[7]。其中，TF 意思是词句频率，IDF 代表逆文本内容的频率指数。该方法适用于评估某个词句对于文本或语句资料库的特殊性。其中 TF-IDF 的计算公式，如公式 2-1 所示：

$$\text{TF-IDF} = \text{TF} * \text{IDF}. \tag{2-1}$$

其中 TF 的计算公式如公式 2-2 所示：

$$\text{TF} = \frac{\text{在某一类词条S出现的次数}}{\text{该类中出现的所有词条数目}}. \tag{2-2}$$

其中 IDF 的计算公式如公式 2-3 所示：

$$\text{IDF} = \log_2 \frac{\text{语料库中的文档总数}}{\text{包含词条S的文档总数}+1}. \tag{2-3}$$

2.4. 模型构建

朴素贝叶斯算法中有一种变式被称为多项式朴素贝叶斯，其可用于处理多个分布数据，并被广泛应用于文本分类。通过研究得出多项式朴素贝叶斯模型的预测精度为 0.8307692307692308。伯努利朴素贝叶斯模型实现了针对多个伯努利分布数据的朴素贝叶斯训练和分类算法，即具有多个特征，但每个特征都假定是一个二元变量。通过研究得出伯努利朴素贝叶斯模型的预测精度为 0.8402714932126697。逻辑回归是一种广义的逻辑回归分析的实用模型，属于机器学习中的监督学习[8]。它是通过给定的 x 组数据(训练集)进行模型训练，然后对给定的一组或多组数据(测试集)进行分类。通过研究得出逻辑回归模型的预测精度为 0.8927601809954752。

通过对比伯努利朴素贝叶斯模型、多项式朴素贝叶斯模型和逻辑回归模型进行文本情感分析，可以采用精确率、召回率、F1 度量和 ROC 曲线等指标来评估各类模型的效果，通过研究可知在 ROC 曲线对比中逻辑回归模型表现最优。精确率是表示预测样本中实际为正样本的比例。根据研究可知多项式朴素贝叶斯模型预测评论为正面的精确率为 1.00；伯努利朴素贝叶斯模型预测评论为正面的精确率为 0.58；逻辑回归模型的精确率为 0.74。召回率指的是样本中的所有正样本中，有多少正样本被模型正确预测。通过研究可知多项式朴素贝叶斯模型预测评论为正面的召回率为 0.02；伯努利朴素贝叶斯模型预测评论为正面的召回率为 0.27；逻辑回归模型的召回率为 0.59。F1 度量基于精确率与召回率的调和平均定义的。通过研究可知多项式朴素贝叶斯模型预测评论为正面的 F1 度量为 0.04；伯努利朴素贝叶斯模型预测评论为正面的 F1 度量为 0.37；逻辑回归模型的 F1 度量为 0.66。

由于本研究是利用模型来预测评论为正面的概率，在挑选模型时会更加看重预测精度和 F1 度量的数值，故经过对模型预测精度、精确率、召回率、F1 度量和 ROC 曲线的综合考虑，最终选择逻辑回归模型作为最终的预测评论为正面的模型，并将 solver 设置为'lbfgs'。

3. 实验结果及分析

3.1. 实验数据集

本文的数据条数约三万余条，分别来自亚马逊购物网站的三类商品，即微波炉、吹风机、婴儿奶嘴。这些数据包括商品的名称、类目、评价内容、星级评分、评价时间等信息。

3.2. 评分标准

本文的研究文本计算是基于综合实际打分、预测出的评论为正面的概率、“有用”比例、是否购买、是否是会员的这五项重要指标来计算出文本数据综合得分。

利用训练出的逻辑回归模型去预测每条商品评价为正面评价的概率。首先定义“final_rate”为文本数据综合得分，再利用评论中的星级评定(star_rating)、预测出的评论为正面的概率(pos_prob)、“有用”比例(helpful_rate)、是否购买(verified_purchase)、是否是会员(vine)。从而利用权重公式 3-1 计算出综合得分：

$$\begin{aligned}\text{final\_rate} = &\left(\text{star\_rating}*6+\text{pos\_prob}*70\right)*\left(0.7+\text{helpful\_rate}*0.1\right.\\ &\left.+\text{verified\_purchase}*0.1+\text{vine}*0.1\right).\end{aligned} \tag{3-1}$$

3.3. 实验结果分析

本文是基于 LDA 模型研究商品评论的情感，首先进行数据处理，如去除无关字段等；再利用 LDA 模型提取三种商品的一星和五星评价主题词；在此基础上利用 TF-IDF 词向量将主题词向量化，并使用逻辑回归模型来预测商品评论为正面的概率；最后利用上文提出的权重公式计算评价综合得分。由权重公式可以看出综合得分较高的评论是可信度较高的评论。

综上，根据计算出的文本数据综合得分，关于吹风机类可以得出四点建议。建议一：设计一款便于旅行携带的小尺寸的吹风机。建议二：保证吹风机有足够大的风力以便快速吹干头发。建议三：吹风机应当设计安全装置。建议四：首先应该控制成本，制定合理的出厂价格。

根据计算出的文本数据综合得分，关于婴儿奶嘴类可以得出四点建议。建议一：保证安全的前提下设计更加可爱的造型。建议二：原材料应当是无毒无害的。建议三：提高奶嘴的安全系数。

根据计算出的文本数据综合得分，关于微波炉类可以得出四点建议。建议一：设计恰好可以放进灶台角落的微波炉。建议二：硬件质量应该升级。建议三：设计足够大的火力。建议四：微波炉的控制程序应该设计得简明清晰并且易于使用。建议五：做好售后服务。

4. 总结与展望

在线评论评分是获取用户需求的一种信息来源，而且可以帮助潜在买家做出正确的购买决策。本文提出了一种基于 LDA 主题的评论情感分析方法。该方法通过结合亚马逊网站上三种商品评论的文本、用户是否为会员等信息来分析商品评论的情感，并提出了一些建议。在未来研究中，将该情感分析方法应用于医疗服务评论情感分析中。

参考文献

[1] 张建成. 基于在线商品评论的消费者满意度和认知研究[D]: [硕士学位论文]. 宁波: 宁波大学, 2012.

[2] 王鹤琴, 王杨. 基于情感倾向和 SVM 混合极短文本分类模型[J]. 科技通报, 2018, 34(8): 149-154.

[3] Saranya, S., Veena, S., Vivek, D., *et al*. (2018) Finding Reputed Items Based on Sentimental Analysis of User Reviews and Ratings. *Journal of Computational and Theoretical Nanoscience*, **15**, 3057-3061.

https://doi.org/10.1166/jctn.2018.7591

[4] 刘永芬. 支持向量机在入侵检测中的应用[D]: [硕士学位论文]. 福州: 福建师范大学, 2010.

[5] 曾小芹, 余宏. 基于 Python 的商品评论文本情感分析[J]. 电脑知识与技术, 2020, 16(8): 181-183.

[6] 王梦宇. 基于 LDA 主题模型的在线评论聚类研究[D]: [硕士学位论文]. 兰州: 兰州大学, 2021.

[7] 刘擎权. 基于改进的 TFIDF 算法在文本分析中的应用[D]: [硕士学位论文]. 南昌: 南昌大学, 2019.

[8] 赵雅平. 基于逻辑回归模型的天津市老年人健康养老服务需求研究[D]: [硕士学位论文]. 天津: 天津财经大学, 2019.