

文章编号: 1672-7010(2018)05-0027-06

## 基于词典的商品评论情感分析

常 丹<sup>1</sup>, 王玉珍<sup>2</sup>

(1. 兰州财经大学 信息工程学院, 甘肃 兰州, 730020;  
2. 兰州财经大学 丝绸之路经济研究院, 甘肃 兰州, 730020)

**摘 要:** 为了挖掘商品评论中有价值的信息, 有效分析用户对商品的直接反馈, 文章基于通用的情感词典, 以国内某电商平台上的商品评论为研究对象, 根据词频共现方法拓展情感词典, 基于构建的情感词典对评论进行情感极性分析。根据分析结果, 不仅能够帮助商家优化商品属性, 更好地满足消费者的需求, 还可以指导用户制定合理的购买决策。

**关键词:** 词典; 商品评论; 情感分析

中图分类号: TP391

文献标志码: A

## Product Reviews Sentiment Analysis Based on a Dictionary

CHANG Dan<sup>1</sup>, WANG Yuzhen<sup>2</sup>

(1. School of Information Engineering, Lanzhou University of Finance and Economics, Lanzhou 730020, China;  
2. Silkway Economic Research Institute, Lanzhou University of Finance and Economics, Lanzhou 730020, China)

**Abstract:** In order to mine valuable information in commodity reviews and effectively analyze users' direct feedback on the products, based on a general emotional dictionary, this paper uses commodity reviews on an e-commerce platform in China as the research object, and expands the emotional dictionary by the word-frequency co-occurrence method, then carries out the sentiment polarity analysis of the reviews based on the set-up emotion dictionary. According to the analysis results, it can not only help merchants to optimize product attributes, better satisfy consumer needs, but also guide users to formulate reasonable purchase decisions.

**Key words:** dictionary; commodity reviews; sentiment analysis

收稿日期: 2018-06-19

基金项目: 兰州财经大学丝绸之路经济研究院项目(JYYY201704); 兰州财经大学科研项目(Lzufe2018B-04)

作者简介: 常丹, 女, 硕士研究生, 从事数据挖掘、电子商务研究

通信作者: 王玉珍, 女, 教授, 硕士, 从事数据挖掘、电子商务研究

互联网技术的发展使得电子商务行业快速崛起,伴随着各类购物网站用户的持续增长,人们已从过去简单的信息获取转变为向互联网传输自己的观点看法,从而每天产生数以万计的商品评论。这些呈指数级增长的评论信息虽然数据量庞大,难以梳理,但是很有研究的价值。文中将以京东和淘宝网站销售 OPPO R11s 手机为例,通过拓展的情感词典对商品评论进行研究,分析用户对产品不同特征的情感倾向,从而指导用户的消费行为,帮助商家改进商品上的不足。

近年来,交互式资源的出现,使得情感分析成为热门、前沿的研究领域。如:龚安,费凡<sup>[1]</sup>以酒店评论数据集为语料,在改进的情感规则中融合一元词、句法等特征,利用信息增益进行特征选择,提高了文本分类的性能和精度;周锦峰<sup>[2]</sup>等人针对文本局部语义特征,提出一种深度卷积神经网络模型,并表明这种模型不仅在训练效率和预测速度上有所提升,而且情感分类的准确率有所提高;冯兴杰<sup>[3]</sup>等人提出基于卷积神经网络和注意力模型相结合的方法对数据集进行分析,指出这种方法的实验结果相较于传统的机器学习方法有明显提高;范炜昊<sup>[4]</sup>等人在情感分析的基础上构建了用户痛点分析模型,提出用户痛点指数计算公式,对用户痛点进行量化;李桃迎<sup>[5]</sup>等人以服务类网购评语为研究对象,构建高频词共现网络,为网络评语挖掘研究领域提供了新的研究思路;赵刚<sup>[6]</sup>等人运用机器学习方法,设计餐饮领域的情感分析模型,通过实验,这种模型更加适用于客户的情感倾向分析;胡朝举<sup>[7]</sup>等人为有效解决传统的基于注意力的深度学习模型所存在的问题,构建了融合主题特征的深层注意力的 LSTM 模型,实验表明,该模型在情感分析的准确率和稳定性上都有一定的提高;赵志滨<sup>[8]</sup>等人通过抽取中文产品评论中的维度信息,计算各维度的维度情感,并提出维度权值计算方法,综合维度情感和维度权值计算评论的整体情感。可见,由于商品评论信息研究的重要性,情感分析受到各界学者的广泛关注。但是随着研究方法的不断改进,基于词典的情感分析方法逐渐不适用于大量文本的分析,主要是因为词典构建的难度较大,基准词典不适用于所有的文本情感分析。因此文中通过拓展手机电子设备领域的情感词典,并将其应用于时下销量较好的手机型号做情感分析。

## 1 情感分析概述

情感是人对客观事物是否满足自己的需要而产生的态度体验,可以表现为人们对于某个事物的评价、情绪或观点<sup>[9]</sup>。在商品评论情感分析中,情感被认为等同于用户对产品的观点、评价,因此,情感分析又称意见挖掘,是自然语言处理的范畴之一,涉及到机器学习、数据挖掘、信息检索等多个研究领域。情感分析主要是对语料集进行情感极性分析和情感极性强度分析,文章将主要对商品评论的情感极性进行分析,即用户对商品是否表示积极、消极或中性的情绪。除此之外,情感分析还可根据文本的不同分为新闻评论分析、产品评论分析、电影评论分析等类型。对这些主观性评论文本进行集成和分析,并根据上下文语境,计算文本的情感值,分析产品评论的情感倾向,对商家改进产品和用户选购商品都具有重要意义。

## 2 基于词典的情感值计算方法

目前,常用的文本情感分析主要是根据手工标注好的词典,基于传统的机器学习方法对文本进行分类。由于文本情感分析领域是极其敏感的,对于不同领域的产品评论,都有其适应的情感分析方法。因此文章基于知网发布的中文情感词典,构建适合手机评论的用户词典对商品评论进行分析,即建立情感词典,提取情感单元,计算各评论文本的情感倾向值。

### 2.1 情感词典的构建

由于文中以手机评论数据集为语料进行分析,因此,基于手机评论的词典,不仅要包含通用的情感词典,还要从产品属性层面上细化,补充网络流行用语,拓展相关词典,这样才能提高情感分类的准确度。

情感词典构建方法:

第一步:以 How Net 2007 版情感词典提供的正面情感词、负面情感词、正面评价词、负面评价词等为基础拓展词典;

第二步:通过经验和对数据集的观察,人工采集规则,其中主要的几种规则有“名词+形容词”及“名词+副词+形容词”、“动名词+形容词”及“动词+名词”等;

第三步:对语料集进行分句处理,即以“。”“!”“?”等对评论文本进行切分;

第四步:利用分词系统对评论文本进行分词和词性标注,依据收集到的规则找出新的情感词;

第五步:根据两个词的共现程度  $PMI^{[10]}$  判断情感词的情感倾向,当两个词与基准情感词的  $PMI$  差值大于某一阈值,则较大的一方的情感极性为新词的情感极性;

$$PMI(W_1, W_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \quad (1)$$

其中  $w_1, w_2$  表示两个不同的词语。

第六步:将新词录入情感词典中。

### 2.2 基于词典的文本情感值计算

对于商品评论文本来说,影响评论情感倾向的不仅是情感词,还包括程度副词、否定词等,因此,文中使用情感值计算函数<sup>[11]</sup>(公式2),通过情感词和修饰词的综合计算结果,计算商品评论的情感值,从而分析该属性的情感倾向。

$$F = \sum_{i=1}^n [f(x_i) \prod_{j=1}^m a_{ij}] \quad (2)$$

其中  $F$  表示情感极值,  $f(x_i)$  表示第  $i$  个情感词极值,  $a_{ij}$  表示情感词  $x_i$  对应的程度副词的权值。

## 3 基于商品评论的情感分析

### 3.1 数据来源

文中分析的数据主要来源于京东和淘宝的网站销售的 OPPO R11s 手机的评论数据。由于 OPPO R11s 手机上市以来,广受欢迎,但对于这个型号的手机属性还存在争论,因此从

京东、淘宝这两家访问量较大的国内购物网站上,通过八爪鱼数据采集器收集了 13883 条评论数据,通过分析用户评论信息,挖掘用户对手机各项特征的情感倾向。

### 3.2 数据处理

商品评论是用户对产品的主观性评价,在表达上具有随意性,数据容易产生噪声。因此,文章首先对评论文本进行预处理,即去除噪声和重复性文本。如在原始数据中,由于用户未对商品进行评价,系统会默认给出“此用户未及时评价系统默认为好评”,这类评价信息研究的意义不大,因此在数据预处理中去除这一类型的文本。然后进行分词、词性标注和词频统计等工作,目前常用的分词工具主要有 NLPIR 中文分词系统、哈工大的 LTP 语言平台等,文中主要是通过结巴分析系统对评论文本进行分词和词性标注,主要是因为结巴分词系统可以直接在 python 中导入结巴分词模块,并且分词速度较快。最后,对分词结果进行词频统计,部分结果如表 1 所示:

表 1 手机部分词频统计结果

Table 1 Frequency statistics of mobile phones

关键词	词频	词性	关键词	词频	词性	关键词	词频	词性
手机	4387	n	拍照	1132	v	手感	733	n
不错	2721	a	满意	1108	v	东西	668	n
喜欢	1971	v	很快	994	d	感觉	663	n
非常	1508	d	速度	846	n	清晰	626	a
收到	1239	v	正品	808	n	可以	620	c

说明: n-名词, a-形容词, v-动词, d-副词, c-连词

### 3.3 情感值计算

通过以上数据预处理,将分词后的词频统计结果按由大到小的顺序进行排序,并选择排名前 14 的手机特征指标词,对手机指标进行分类,得到五个类别指标,见表 2。

表 2 手机特征指标分类表

Table 2 Cellular feature index classification

指标类别	指标词
拍照	拍照、像素、照相
续航	充电、电池
外观	屏幕、外观
性能	功能、性能、系统、音质
性价比	性价比

对数据进行特征提取以及情感极性计算,为了更精确的计算用户对不同特征的情感倾向,文章将程度副词考虑在情感值计算中,给不同程度的副词赋予相应的权值,文章参考已有的研究成果,具有权值大小见表 3,情感分析所有程序通过 python 3.6 实现。

表3 程度副词权值设置表

Table 3 Degree adverbs weight setting table

程度级别	权值大小
most	2.00
very	1.50
more	1.25
ish	0.50
insufficiently	0.25
inverse	-1.00

其中,most 表示程度级别最大,代表词有“超级,过于”等,very 代表词有“很”,“非常”等,more 代表词有“比较”,“还”等,ish 代表词有“稍微”,“略微”,insufficiently 代表词“不怎么”,inverse 是对整体的文本评论的情感倾向做了反转,因此赋予负值。

### 3.4 结果分析

对预处理后的 9997 条有效数据进行分析,得到的手机特征词情感分析结果如图 1 所示。

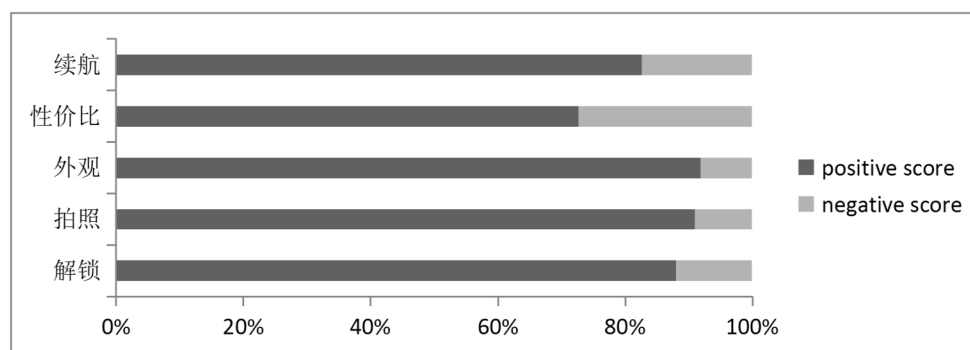


图1 特征词正负平均情感占比

Fig.1 Positive and negative average emotion ratios of characteristic words

由图 1 可知,对于 OPPO R11s 这款手机来说,用户对手机的总体评价较好,尤其是手机的外观和拍照功能,积极评论占比超过 85%,得到了用户的普遍好评。而手机的续航,即充电和电池性能方面,以及手机的性价比相对来说评价较弱,这些特征更应该引起商家的足够重视。通过进一步分析用户评论,能够发现对于 oppo R11s 手机的续航这一特征来说,主要是由于手机电池不耐用,并且在充电的过程中手机会发热,导致用户产生消极的情感。而性价比方面,部分用户给出了“手机性价比不高”,“不是性价比之选”等消极评论。主要是由于这款手机在上市时,价格较高,除了拍照功能较好,外观漂亮外,其他性能并不出众。而且通过词频统计,发现这款手机的主要用户为女士居多,可见对于男士来说这款手机并不是最好的选择。

通过上述分析,建议商家能够延续在外观和拍照上的优势,优化手机的电池,解决手机电池的耗电以及充电时手机的发热问题,针对不同的用户群体制定相应的精准营销策略,提高客户的满意度,进而使得商家获得更高的收益。

## 4 结论

对商品评论进行情感分析和意见挖掘,不仅能指导用户的购买行为,而且能够帮助商家更好地了解用户需求,从而改进产品。文中基于拓展的情感词典对手机评论进行情感分析,这种情感词典的构建和情感值计算方法不仅适用于这种型号的手机评论情感分析,对于同类型的其他商品来说,同样也具有参考价值。

### 参考文献:

- [1] 龚安, 费凡. 基于多特征融合的评论文本情感分析[J]. 计算机技术与发展, 2018, 4(08): 91-95.
- [2] 周锦峰, 叶施仁, 王晖. 基于速度优化卷积神经网络的文本情感分类[J/OL]. 计算机工程, 2018: 1-9. [2018-03-01]. <http://kns.cnki.net/kcms/detail/31.1289.TP.20180301.1442.001.html>.
- [3] 冯兴杰, 张志伟, 史金钊. 基于卷积神经网络和注意力模型的文本情感分析[J]. 计算机应用研究, 2018, 35(05): 1434-1436.
- [4] 范炜昊, 徐健. 基于网络用户评论情感计算的用户痛点分析——以手机评论为例[J]. 情报理论与实践, 2018, 41(01): 94-99.
- [5] 李桃迎, 吕晓宁, 李峰. 网购评语高频词共现网络的结构特征分析[J/OL]. 计算机应用研究, 2019, (01): 1-6. [2018-02-08]. <http://kns.cnki.net/kcms/detail/51.1196.TP.20180208.1713.022.html>.
- [6] 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, 3(02): 166-170.
- [7] 胡朝举, 赵晓伟. 基于词向量技术和混合神经网络的情感分析[J/OL]. 计算机应用研究, 2018, 4(12): 1-6. [2017-12-12]. <http://kns.cnki.net/kcms/detail/51.1196.TP.20171212.1833.018.html>.
- [8] 赵志滨, 刘欢, 姚兰, 等. 中文产品评论的维度挖掘及情感分析技术研究[J]. 计算机科学与探索, 2018, 12(03): 341-349.
- [9] 唐晓波, 刘广超. 细粒度情感分析研究综述[J]. 图书情报工作, 2017, 61(05): 132-140.
- [10] 丁卜建. 基于情感词典拓展和词向量的中文情感分析技术的研究[D]. 沈阳: 辽宁大学, 2016.
- [11] 周红伟. 商品评价信息的中文情感分析[D]. 杭州: 浙江工商大学, 2015.