东北农业科学　2020，45（3）：92-96
Journal of Northeast Agricultural Sciences　DOI：10.16423/j.cnki.1003-8701.2020.03.019

基于人参购买评论的多项特征情感倾向性分析研究

毕春光，袁　帅，皇　可，郭　海，王金龙
（吉林农业大学信息技术学院/吉林省农业物联网科技协同创新中心/吉林省智能环境研究中心/吉林省精准农业与大数据工程研究中心，长春 130118）

摘　要：在电子商务和社区论坛蓬勃发展的网络环境下，产品的评论数量激增，评论数据已成为企业提高商品质量和提升服务的重要数据源，在评论中包含用户对产品特征的关注度以及相应的情感倾向。因此，为了能够细粒度地挖掘出用户对商品多个特征的情感分布情况，文中提出了基于评论特征词构建评论分类模型以及基于朴素贝叶斯算法的特征情感分类器，从用户的评论出发，对商品的多个特征进行情感倾向性分析。

关键词：特征抽取；TextRank算法；朴素贝叶斯；分类器；情感分析
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：2096-5877（2020）03-0092-05

A Multi-feature Analysis of Emotional Tendency Based on Ginseng Purchase Reviews

BI Chunguang, YUAN Shuai, HUANG Ke, GUO Hai, WANG Jinlong
(*College of Information Technology, Jilin Agricultural University / Jilin Agricultural Internet of Things Science and Technology Co-innovation Center / Jilin Intelligent Environment Research Center / Jilin Precision Agriculture and Big Data Engineering Research Center, Changchun* 130118, China)

Abstract: With the rapid development of e-commerce and community forums, the number of product reviews has increased dramatically. Review data has become an important data source for enterprises to improve the quality of goods and services. Reviews include user's attention to product features and corresponding emotional tendencies. Therefore, in order to mine the emotion distribution of users on multiple features of commodities in a fine-grained way, a comment classification model based on comment feature words and a feature emotion classifier based on Naive Bayes algorithm were proposed in this paper. From the user's review, this paper analyzed the emotional orientation of the product's multiple features.

Key words: Feature extraction; TextRank algorithm; Naive Bayes; Classifier; Emotional analysis

互联网以它交流方便、使用快捷等特点迅速吸引了大量用户，同时也产生着大量的文本数据，这其中就包括商品购买的大量复杂评论信息[1]，里面包含了用户对商品的相关描述信息和观点看法等。越来越多网上用户的购买行为受到商品评论的影响，他们通过浏览商品的在线评论来了解商品质量和口碑，以做出正确的购买决策。此外，商品的在线评论作为一种反馈机制可以帮助商家发现商品的不足来改善服务、提升商品质量，从而在商业竞争中处于优势之地。但是，商品评论数量日益增长，动辄几千条甚至上万条评论让用户感到无所适从，且从这些评论中获取商品的准确信息变得非常困难。因此，迫切需要借助技术手段来分析这些评论[2]，获取对商品评价客观的统计数据，辅助用户的购买行为。

对商品购买评论情感分析的传统研究是针对一条评论的整体情感，从整体的层面上，无法细粒度地挖掘出用户对商品相应特征的意见，来判别用户对商品的喜恶[3]。然而一个商品包含多个特征，用户会在一条评论中表达出对不同特征的意见看法。

如图1所示，在第一条评论中，从整体的层面上感知确实是一条正面的好评，但是评论中“快递有点慢一周才到”，却是表达了用户对物流的

收稿日期：2018-12-31
基金项目：吉林省科技厅项目（20170204017NY、20170204038NY、20160623016TC）
作者简介：毕春光（1977-），女，副教授，硕士，主要从事数据挖掘、农业信息化研究。

不满;在第二条评论中,“质量确实挺好的。就是有点小贵”,表达了对“质量”“价格”这两个特征的满意程度,但是从整体的层面上分析用户对商品的看法倾向,却有些模糊。所以,产品评论通常会描述产品的多个特征属性,单条评论中所描述的多个特征可能会有不同的特征情感,不同产品特征对用户购买意愿的影响程度各不相同。因此,应从产品评论中挖掘出用户主要关注的一些商品特征,结合情感分析技术[4],分析用户对商品重要特征关注度的情感概率分布状况。

图1　购买人参的相关评论截图

1　相关研究现状

国内外在数据挖掘、电商评论、舆情分析等众多领域都已经取得了重大的研究成果。例如,Pang等讨论了机器学习算法在文本情感分析领域的应用,训练分类器对整篇电影评论进行情感极性判断[5];Ghose等使用Ling Pipe从亚马逊等购物网站爬取几种商品的评论数据,在对其进行文本预处理后,通过构造的分类器对处理后的文本进行情感分类,最后在用户进行购买时起到参考作用[6];左维松等提出基于规则和统计相结合的方法对篇章级商品评论进行情感极性判别,取得了较好的效果[7];赵丽芳提出了抽取商品属性特征进行情感极性判断,但未针对细化的商品特征属性分别设计主题词库,因而也就未能获取不同商品特征属性的情感倾向强度[8];郭捷研究了情感词典与机器学习对网络评论的情感分类,开发出文本分类的可视化应用系统[9];吴潇提出了领域情感词语的情感特征对网上购物商品评论情感倾向性研究[10],但未从某一商品具有多特征作为出发点研究等。综上所述,随着web 2.0时代的发展,情感分析已经成长为自然语言处理(NLP)中最活跃的研究领域之一,引起了越来越多学者们的注意[11]。

2　技术路线

2.1　抽取评论关键词特征

按照标点符号对评论进行分割,每条评论可以分成若干个子句,将其格式化保存。对处理后的全部评论数据采用TextRank算法,利用局部词汇关系[12]的思想,从总体的层面上进行评论观点的挖掘,找出用户对商品特征关注度的分布情况,对关注度高的商品特征进行情感分析研究。

TextRank算法是一种用于文本的基于图的排序算法,其基本思想来源于谷歌的 PageRank算法,通过把文本分割成若干组成单元(单词、句子)并建立图模型,利用投票机制对文本中的重要成分进行排序,仅利用单篇文档本身的信息即可实现关键词提取[13],其迭代计算公式如下:

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{ik}} WS(V_j) \quad \cdots\cdots (1)$$

公式(1)中,TextRank模型图由点集合V和边集合E组成,任两点V_i,V_j之间边的权重为w_{ji},$In(V_i)$是指所有指向节点i的点集合,表示节点i的入度,$Out(V_j)$表示节点j的出度,$WS(V_j)$表示节点j的权重值,d为阻尼系数,取值范围为0~1,表示从图中某一特定点指向其他任意点的概率,一般取值为0.85。

基于TextRank算法对商品评论特征关键词提取步骤如下:

(1)把给定的评论文本T按照完整句子进行分割,即$T=[S_1, S_2, \cdots, S_m]$;(2)对于每个句子$S_i \in T$,进行分词和词性标注处理,并过滤掉停用词,只保留指定词性的词语,如名词、动词、形容词,即$S_i=[t_{i,1}, t_{i,2}, \cdots, t_{i,n}]$,其中$t_{i,j} \in S_j$是保留后的候选关键词;(3)构建候选关键词图G=(V,E),由第(2)步生成的候选关键词组成,然后采用共现关系(co-occurrence)构造任两点之间的边;(4)两个节点之间存在边仅当它们对应的词汇在长度为K的窗口中共现,K表示窗口大小,即最多共现K个单词;(5)根据公式(1),迭代传播各节点的权重,直至收敛;(6)对节点权重进行倒序排序,从而得到最重要的T个词,作为候选关键词;(7)由第(6)步得到最重要的T个词,在原始文本中进行标记,若形成相邻词组,则组合成多词关键词[14]。

2.2　构建文本情感分类器

现在主流的情感分析方法分为基于词典的方法和基于机器学习的方法[15]。基于词典的情感分析方法,首先拆分文本数据的段落并解析句法,然后计算出情感值来描述文本的态度倾向,情感

词典的质量是影响情感分析准确率的关键因素，因此，情感词典的构建是一个不断积累新兴词汇与情感词粒度不断进行细化的过程。

基于机器学习的方法主要思想是先通过人工标注情感极性的方式构建文本训练语料库，然后从语料中选出部分与分类相关的特征，进行分类器的训练[16]，进而使用分类器对等待分类的文本语料进行分类，以判别等待分类的文本语料的情感倾向性。

根据提取的商品特征关键词，归纳筛选出用户主要关注的特征，并按照这些特征构造评论分类模型，对评论数据进行分类。在此基础上，基于朴素贝叶斯算法构造文本情感分类器[17]，训练分类器的数据带有标签属性，可以把75%的带有标签属性的数据作为训练数据，剩下25%的数据作为测试数据用来验证分类器的准确性，对每一个商品特征的评论数据进行情感分析。

朴素贝叶斯构建文本情感分类器[18-19]步骤如下：

（1）在商品评论数据中，设$d=\{w_1, w_2, \ldots, w_i\}$为评论文本集中的一个待分类样本，w表示d的特征项；

（2）根据类别集合$C=\{c_1, c_2, \ldots, c_j\}$，计算特征词在类比C的条件概率，即分别计算$p(c_1|d)$，$p(c_2|d)$，…，$p(c_j|d)$；

（3）根据步骤(2)，若想获取最大的$p(c_n|d)$值（c_n表示文本d所在的类别），需要根据“贝叶斯假设”，利用公式（2）计算概率：

$$p(c_n|d)=\frac{p(d|c_n)p(c_n)}{p(d)}, c_n \in C \cdots\cdots(2)$$

（4）若使得$p(c_n|d)$取得最大值，需要让$p(d|c_n)p(c_n)$取最大值，其中$p(d)$是常量，因此只需要计算$p(d|c_n)$和$p(c_n)$的值，利用公式（3）计算类别c_n的先验概率：

$$p(c_n)=\frac{t_n}{t} \cdots\cdots(3)$$

其中，t_n表示属于类别c_n的样本数，t表示总的样本数。

$p(d|c_n)$表示当属于类别c_n的前提下，文本d的概率，朴素贝叶斯采用特征之间相互独立的思想，按照公式（4）进行计算：

$$p(d|c_n)=p(w_1|c_n)\times p(w_2|c_n)\times\cdots\times p(w_i|c_n) \cdots\cdots(4)$$

其中，$p(w_i|c_n)$由训练样本估值得到，利用公式（5）计算：

$$p(w_i|c_n)=\frac{t_{in}}{t_n}, (i=1, 2, \cdots, j) \cdots\cdots(5)$$

其中，t_{in}表示有特征词w_i并且属于类别c_n的样本数。

（5）通过如上进行训练，可以得到文本情感分类器。用特征向量$d=\{w_1, w_2, \ldots, w_i\}$表示待分类的评论样本，按照前四步计算$p(c_n|d)$的最大值，即可以将样本d划分为$c_n$类。

3 实验

3.1 实验数据及环境介绍

实验均采用python 3.6语言实现，通过scrapy网络爬虫技术[20]从京东官网上获取人参购买评论，格式化处理后保存到本地；将人工标注总结的21 554条正面评论以及18 792条负面评论的电商评论作为语料库，为后续的情感分类工作做好准备。

3.2 评论关键词特征集

采用TextRank方法对评论特征关键词抽取，得出用户关于人参评论关注度较高的特征集合如下：

{京东，包装，味道，东西，购买，收到，物流，质量，效果，泡酒，速度，快递，人参，形状，参味，太小，礼盒，客服，评价，价格，产品，希望，泡水，参片，品质，信赖，实惠，老人，口感}

其中，在以上的特征集合中，类似“收到”“物流”“速度”和“快递”等特征词描述的是同一类信息，可以归为“物流”这一范围。然而，在进行实验的过程中发现，有些商品的特征在用户评论中很少涉及到，如人参的栽培方法、生产过程等[21]，说明用户对这些特征关注度低，进行情感判断的意义并不大。因此，本研究主要依据以上的特征集为主，通过筛选商品特征集合，将描述信息相同的特征词合并总结，从评论中挖掘出最受用户关注的一些人参特征：{质量，饮食，物流，品相，价格，包装}，这六个作为人参的主要特征进行情感分析研究。

3.3 构造评论分类模型

将所有人参文本评论根据标点符号构造正则表达式分割，每条评论可以分成若干个子句，并按照六大主特征进行提取和分类，会过滤掉评论中与主特征无关以及无效的评论数据，如“收到东西了”“此用户未填写评价内容”等，提高了后续情感分析工作的针对性和准确性。然而，在一些评论，如“就是有点小贵”描述的就是价格，但

表1　主特征–子特征主题词

主特征	子特征主题词
质量	质量,品质,质地,质料,产品,优劣,优质,劣质,材质,东西,东西好,东西坏,货真,货假,防伪,假冒,货品,正宗
饮食	饮食,饮食习惯,养生,养身,吃,喝,炖汤,炖菜,补气,养血,泡酒,泡水,泡茶,品尝,烹调,口味,口感,味道,参味,精力,口服,上火,菜肴,煮食,膳食,烹饪,酒水,大补
物流	物流,快递,物流业,物流配送,收到,送到,速度,太快,快速,迅捷,货运,运输,配送,发货,到货,太慢,仓库,京东快递,顺丰,圆通,申通,韵达,中通,汇通
品相	品相,形状,外形,外观,外表,尺寸,切片,片大,片小,个头,很大,有点大,很小,有点小,大点,小点,根须,太大,太小,参片,细长,小点,大点,粗壮,断须,物美,短小,碎屑,渣子
价格	价格,价钱,钱,价位,定价,消费,实惠,经济,划算,价实,售价,降价,昂贵,较贵,便宜,价廉,市价,标价,低廉,低价,要价,性价比,超值,起价,优惠
包装	包装,盒装,罐装,礼盒,精美,完美,简陋,粗陋,外盒,套装,严密,漏气

并没有出现“价格”这个主要特征词,所以在对所有分割后的人参子评论进行分类提取的过程中,需要根据上面的六大主特征通过计算机辅助,构建出与之对应的多子特征主题词库,如表1所示。

基于子特征主题词库,使用词库中词语及其搭配关系对评论子句进行匹配,从所有评论中筛选出属于各个主特征的相关评论,把所有子评论

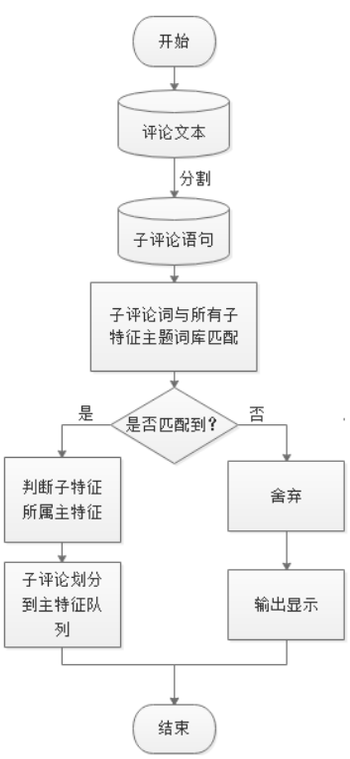

图2　评论分类流程图

按照主特征进行分类,还可以过滤掉不相关或者无效的评论,构造评论分类模型流程如图2所示。

分类的执行过程是对评论文本按照一些标点符号分割成多个评论子句;然后根据每个子句,使用包含有词语搭配关系的子特征主题词库进行匹配,评论子句所属的主特征就是与其匹配到的子特征词所属的主特征。

3.4　情感分析与评估指标

对每个类别下的所有子评论情感分析,选取语料库中75%和25%的数据分别作为情感分类器的训练数据与测试数据,为了评估分类器的分类结果,采用准确率(precision)、召回率(recall)及F_1的值作为结果的评价标准。计算公式如下:

$$precision = \frac{A}{A + B} \times 100\% \quad \cdots\cdots(6)$$

$$recall = \frac{A}{A + C} \times 100\% \quad \cdots\cdots(7)$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \quad \cdots\cdots(8)$$

其中,A表示好评被分类器正确判断出的数量,B表示分类器把差评错误判断为好评的数量,

表2　情感分类器测试结果　　%

	precision值	recall值	F_1值
第一次	80.36	79.75	80.05
第二次	79.12	78.54	78.82
第三次	81.29	79.96	80.61

C表示分类器把好评错误判断为差评的数量。为了正确验证情感分类器的分类结果,进行多次实验,结果如表2所示。

从表2中可以看出,基于朴素贝叶斯构造的分类器,在评论情感判断上准确率最低都在79%以上,召回率在78%以上,可以用于人参多个特征对应评论数据的情感分析。因此,对用户关注度较高的人参六大主特征:{质量,饮食,物流,形状,价格,包装}进行分析研究,分类器分类结果如图3所示。

人参评论情感分析的标准是每个主特征下正面子评论与该主特征下所有子评论的占比。如图

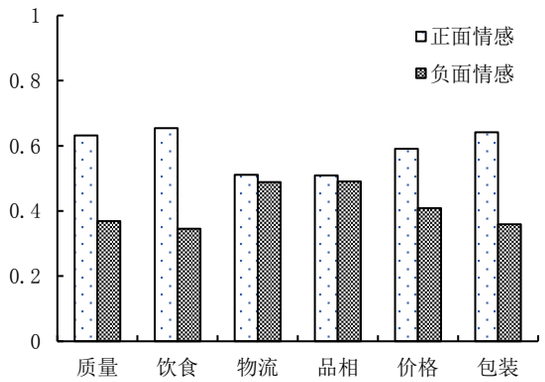

图3 情感分析结果图

3所示，人参的质量、饮食、物流、品相、价格、包装的好评占比有着明显的差异，鉴于本文情感分析的评论数据来自京东销售人参最多的商家，关于质量评论的好评占比在60%以上；在饮食、价格和包装方面，人参作为一种滋补身体的保健品，用户对人参的功效关注度较高，通常买来当作礼品赠送他人，商家的价格也相对合理，物有所值，所以用户给予的好评也就相对较多，占比在60%左右；然而在物流和品相方面，由于天气、配送方式或者商家本身的原因都会影响物流的速度，而且人参有很多根须结构，所以在物流配送的过程中可能会对人参的结构品相造成一些影响和破损，导致好评占比仅在50%左右。综上所述，商家要在保障人参优质、价格合理、包装精美等方面的同时，提高商品的物流速度，优化配送方案，保障人参在运送的过程中参体根须的完整性。

4 结 论

通过电商平台的人参评论信息，使用TextRank的思想抽取评论关键词特征，得出了用户对商品关注度较高的六个主要商品特征，并基于主特征关键词构建子特征词库，一方面扩大了对评论数据提取与分类的覆盖范围，另一方面过滤掉与主特征无关的评论数据，提高了情感分析工作中评论数据的质量。

提出了基于朴素贝叶斯构造情感分类器，采用人工标注的商品购买评论数据进行训练与测试，实验结果表明，情感分类器准确率在80%左右，对人参评论的多特征情感倾向性分析具有可行性。最终，以每个主特征下正面评论与负面评论的占比完成对人参评论数据的情感分析工作，并直观地展示了用户在人参的质量、饮食、物流、品相、价格、包装六个方面的情感倾向结果，研究成果具有一定的参考价值与应用价值。

参考文献：

[1] 陈娉婷，官 波，沈祥成，等.大数据时代开放式农业信息知识库构建研究[J].东北农业科学，2018，43(5)：60-64.
[2] 葛佳琨，刘淑霞.数字农业的发展现状及展望[J].东北农业科学，2017，42(3)：58-62.
[3] 张紫琼，叶 强，李一军.互联网商品评论情感分析研究综述[J].管理科学学报，2010，13(6)：84-96.
[4] 薛益定.中文情感分析研究综述[J].软件研发与应用，2016，5(8)：22-24.
[5] Muhammad Asif, Atiab Ishtiaq, Haseeb Ahmad, et al. Sentiment analysis of extremism in social media from textual information[J]. Telematics and Informatics, 2020, 48: 101345.
[6] Ghose A, Ipeirotis P G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(10): 1498-1512.
[7] 赵文清，侯小可，沙海虹.语义规则在微博热点话题情感分析中的应用[J].智能系统学报，2014，9(1)：121-125.
[8] 赵丽芳.基于最大熵方法的评论信息抽取研究[D].上海：上海交通大学，2009.
[9] 郭 捷.基于网络评论的情感分类技术的研究及应用[D].成都：电子科技大学，2018.
[10] 吴 潇.面向网上商城购物评论的情感倾向分析研究[D].南京：南京邮电大学，2017.
[11] 赵妍妍，秦 兵，刘 挺.文本情感分析[J].软件学报，2010，21(8)：1834-1848.
[12] 宁建飞，刘降珍.融合Word2vec与TextRank的关键词抽取研究[J].现代图书情报技术，2016(6)：20-27.
[13] 夏 天.词语位置加权TextRank的关键词抽取研究[J].现代图书情报技术，2013(9)：30-34.
[14] 宁建飞，刘降珍.融合Word2vec与TextRank的关键词抽取研究[J].现代图书情报技术，2016(6)：20-27.
[15] 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，2016，34(7)：147-154.
[16] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification[J]. Machine Learning, 2011, 85(3): 333-359.
[17] 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.
[18] Rish I. An empirical study of the naive Bayes classifier[J]. Journal of Universal Computer Science, 2001, 3(22): 41-46.
[19] 邸 鹏，段利国.一种新型朴素贝叶斯文本分类算法[J].数据采集与处理，2014，29(1)：71-75.
[20] 安子健.基于Scrapy框架的网络爬虫实现与数据抓取分析[D].长春：吉林大学，2017.
[21] 董 雪.有机蔬菜质量控制及可追溯体系研究综述[J].吉林农业科学，2010，35(3)：51-56.

(责任编辑：刘洪霞)