

# 基于情感分析改进的在线评论分类研究

张 姝

江南大学商学院, 江苏 无锡

收稿日期: 2022年4月7日; 录用日期: 2022年6月2日; 发布日期: 2022年6月10日

## 摘 要

面对海量的评论文本, 单纯依靠人力对其进行分类, 任务繁重且效率低下。文章提出一种基于情感分析改进的在线评论分类模型, 将灰色关联分析和朴素贝叶斯算法相结合, 考虑在线评论中的用户情感倾向, 并将灰色关联分析结果作为一项特征属性嵌入朴素贝叶斯文本分类模型中。文章以京东商城Dyson V10 Fluffy Extra手持无线吸尘器为研究对象, 对真实的在线评论数据进行挖掘, 以检验模型的分类效果。结果显示, 改进后的模型的性能明显领先于传统朴素贝叶斯分类模型, 评价指标F值提升了3.06%, 表明该方法在在线评论文本分类应用中具有一定的优势。

## 关键词

灰色关联分析, 情感分析, 朴素贝叶斯算法, 文本分类

# Research on Online Review Classification Based on Improved Sentiment Analysis

Shu Zhang

School of Business, Jiangnan University, Wuxi Jiangsu

Received: Apr. 7<sup>th</sup>, 2022; accepted: Jun. 2<sup>nd</sup>, 2022; published: Jun. 10<sup>th</sup>, 2022

## Abstract

In the face of a large number of comments, it is cumbersome and inefficient to classify them only by relying on human resources. This paper proposes an improved online comment classification model based on sentiment analysis, which combines grey correlation analysis with Naive Bayes algorithm to consider the user sentiment orientation in online comments, and embed the result of grey correlation analysis as a feature attribute into the Naive Bayes classification model. This paper takes the Dyson V10 Fluffy Extra handheld wireless vacuum cleaner of JD.com as the research object, and mines real online review data to test the classification effect of the model. The results show that the classification performance of the improved model is significantly better than that of

the traditional Naive Bayes classification model, and the evaluation index F value increases by 3.06%. This method has certain advantages in the application of online review text classification.

## Keywords

Grey Correlation Analysis, Emotion Analysis, Naive Bayes Algorithm, Text Classification

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

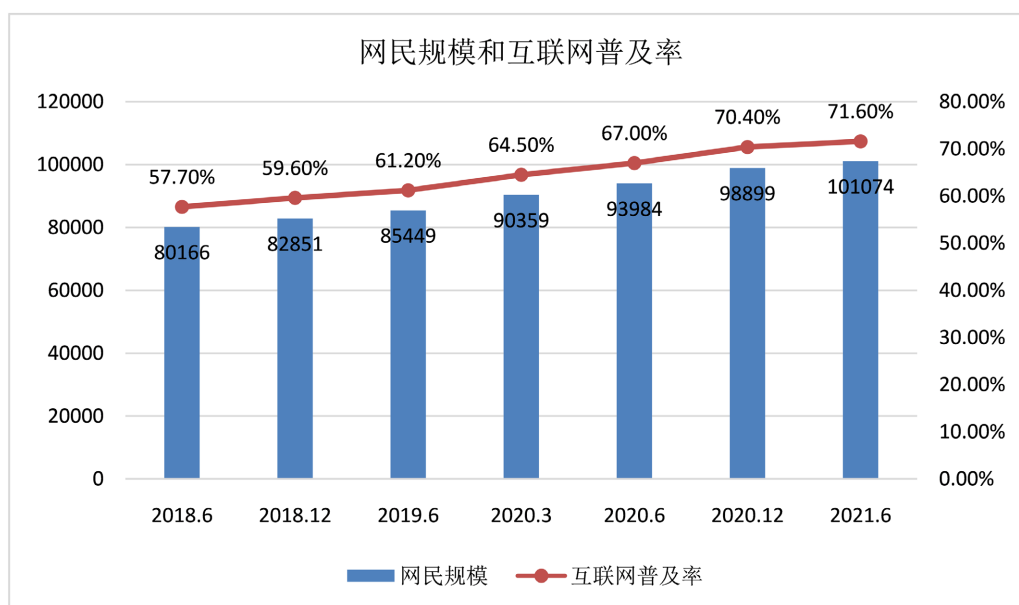
<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

网民规模壮大, 消费结构升级, 在线购物率持续上升。随着互联网行业的跨越式发展和网民规模的扩大, 开启了“十四五”数字经济发展新格局。中国互联网络信息中心(CNNIC)发布第48次《中国互联网络发展状况统计报告》显示[1], 截至2021年6月, 我国网民规模达10.11亿, 较2020年12月增长2175万, 互联网普及率达71.6%(见图1)。十亿用户接入互联网, 形成了全球最为庞大、生机勃勃的数字社会。互联网的普及和网民规模的不断壮大, 支撑了我国电子商务的巨大发展。



数据来源: 第48次《中国互联网络发展状况统计报告》

Figure 1. Internet user scale and Internet penetration rate

图1. 网民规模和互联网普及率

B2B 电商规模整体扩大, 口碑影响力日益凸显。随着社交媒体的发展, 电子商务也逐渐形成了社会商务的新发展模式, 在线用户评论也促进了电子商务发展[2]。消费者在电子商务平台扮演着非常重要的角色, 在线评论引发的口碑传播形成的社会舆论效应影响着市场决策方向[3]。通过文本挖掘技术对在线评论文本提取有价值的信息, 分析消费者对产品各个特征属性的评价及个人情感, 成为电子商务网站研究的重点[4]。

运用文本挖掘技术对在线评论进行情感分类具有十分必要的现实意义。一方面,从消费者角度,借助海量的在线评论内容自动识别获取有用的信息,了解其他消费者对产品的评价和情感倾向,为潜在消费者的购买决策行为提供参考意见。另一方面,从商家角度,通过收集消费者对产品自身的优缺点的真实评价,了解消费者对产品特征的情感倾向和用户偏好,有助于提高消费者对产品的用户体验和满意程度,为今后产品的改进优化和战略发展提供方向。

目前,对在线评论的情感分析成为热点研究课题,综合来看,文本分类技术已经突破理论上的局限,其应用领域越来越广泛。基于机器学习的文本分类方法在准确率和稳定性方面取得显著成效,其中基于朴素贝叶斯分类器的文本分类技术受到越来越多学者的关注,在算法创新方面,研究人员从不同机制不断地展开研究并取得相当大的成果。文本分类是一项复杂的课题,其中许多问题还需要深入探究,例如:文本的高维性和稀疏性、数据集类别的偏斜、自动文本分类算法的创新。

综上,本研究提出一种基于情感分析改进的在线评论分类模型,对情感分析和文本分类进行研究具有一定的创新性。将灰色关联分析和朴素贝叶斯算法相结合,考虑在线评论中的用户情感倾向。以京东商城戴森 V10 Fluffy Extra 手持无线吸尘器为例,进行在线评论挖掘。旨在从产品特征的角度对正、负面的在线评论文本进行用户的情感倾向研究,挖掘消费者对产品的意见态度,为企业提供决策支持。

## 2. 文本分类研究现状

### 2.1. 基于机器学习的情感分类方法

情感分析,又称观点挖掘,是指对用户评论文本进行情感挖掘,挖掘文本中包含的用户对商品或服务的态度和意见。口碑的概念最早是由 Arndt [5]提出,他认为口碑来源于人们日常生活中的口头传播和交流过程中产生的意见。随着电子商务的发展,“在线评论”一词出现,逐渐替代“口碑”,“在线评论”首先是由 Chatterjee [6]初次提出,他认为在线评论是一种网络交流方式,表达了消费者对产品和服务的评价。Web 2.0 使得互联网越来越以用户为中心,包含丰富的意见和情感信息的用户制作的内容在互联网上越来越多[7]。了解这种意见和情绪信息对服务和产品的提供者和用户都变得越来越重要,因为它在影响消费者的购买决定方面发挥着重要的作用[8]。

情感分析作为自然语言处理的基本内容之一,最早由 Das 和 Chen 对股票交易过程中用户情感分析时正式提出[9]。Georgiadou (2020)等将大数据分析和情感分析应用于国际谈判研究中,通过对英国-欧盟脱欧谈判案例的应用和推特用户情感的研究,表明推文的用户情感有可能成为公众对谈判结果的实时晴雨表,从而为政府决策提供信息[10]。最早的基于词典的情感研究是在 1998 年,Whissel 和 Cynthia (2009)在实验中将获取的文本和情感词典中的情感词进行匹配[11]。Oktaviani (2021)等选择印度尼西亚经常访问的电子商务应用程序,使用朴素贝叶斯分类器方法审查数据分类和文本关联,分析用户情感可用于将评论分类为正面或负面情感[12]。Sanglerdsinlapachai (2021)等利用统一医学语言系统(UMLS)中的语义类型来改进基于词典的情感分类方法,通过使用逻辑回归确定语义类型的适当极性分数[13]。基于机器学习的情感分类方法是运用计算机的分析处理能力,利用统计自然语言模型进行训练。近年来,卷积神经网络在自然语言文本领域中被广泛应用。Zhang (2021)等人提出了一种基于深度匹配和层次网络的新型跨模态语义内容相关方法,该方法弥合了图像和字幕之间的相关性。利用预训练的卷积神经网络对视觉子区域内容进行编码,并使用 GloVe 嵌入文本语义[14]。

情感分析研究领域,国内的情感倾向研究起步较晚,其中,与传统的基于词典的情感分析方法相比,机器学习方法拥有更优越的大数据集的文本处理能力,可以有效地减少人力物力成本,降低人工主观性带来的偏差。综合目前学者的研究现状,可以看出很多研究已经在传统情感分析技术基础上进行了修改和补充,然而仍然不能完全避免在实际使用中的缺陷,这使得情感分类方法仍存在局限性。

## 2.2. 基于朴素贝叶斯的分类算法

朴素贝叶斯分类算法假设词是独立的，这简化了计算机的运算过程和时间成本，在分类效果上也表现优异[15]。文本分类研究采用朴素贝叶斯假设，可以使用多项式模型，也可以使用多元伯努利模型。使用机器学习模型的文本分类用于将文档或数据组织在一个预定义的类集合中，训练后的模型将能够识别、预测和检测数据。它在 Web 内容管理、搜索引擎邮件过滤、垃圾邮件检测、意图检测等方面都非常有用[16]。Hindi (2020)等提出了一种延迟局部微调的贝叶斯方法，该算法使用查询实例的最近邻来微调 NB 使用的概率项，使用最近邻只会使独立假设更有可能是有效的，而使用微调算法来找到更准确的概率项[17]。Wei (2011)等人提出了一种辅助特征方法，针对所选的特征选择一个辅助特征对文本空间进行重新分类，调整相应的条件概率以提高分类精度[18]。Zhang (2021)等提出了一种新的改进模型——属性和实例加权朴素贝叶斯，该模型将属性加权和实例加权结合在一个统一的框架中[19]。

文本分类研究领域，综合来看，文本分类研究已经突破理论上的局限，其应用领域越来越广泛。基于机器学习的文本分类方法在精确率和分类性能方面取得显著成效[20]。目前，文本分类领域的主流研究方法包括支持向量机、朴素贝叶斯分类器、K 近邻算法等。文本分类是一项复杂的工程，其中仍存在问题需要深入探究，例如：文本信息的高维性和稀疏性、数据集类别的偏斜、自动文本分类算法的创新。

## 3. 模型构建

文章提出了一种基于情感分析改进的在线评论分类方法，这种方法将灰色关联分析和朴素贝叶斯算法相结合，考虑在线评论中的用户情感倾向，使用灰色关联分析方法计算评论文本的灰色绝对关联度，并将结果作为一项特征属性嵌入朴素贝叶斯文本分类模型中。文本分类模型框架如图 2 所示，主要功能模块如下。

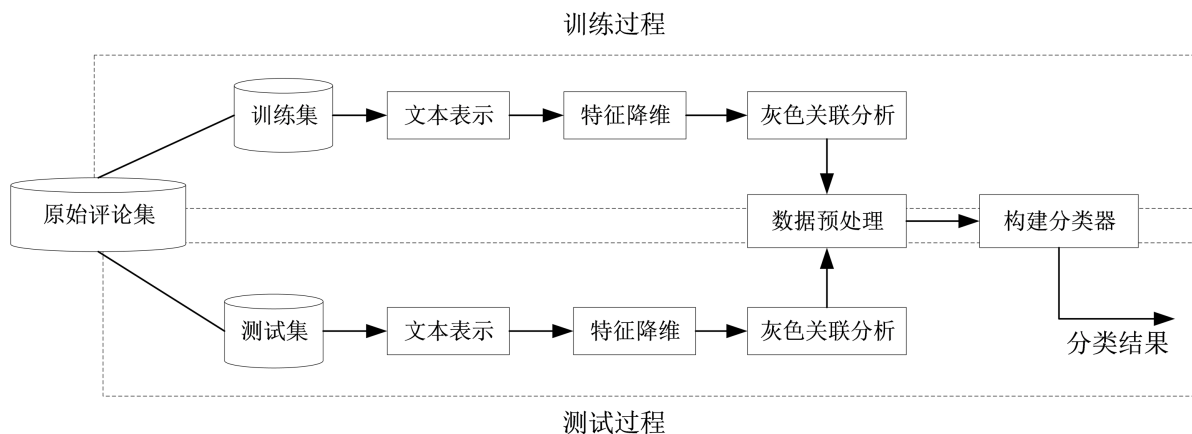


Figure 2. Review classification framework

图 2. 评论分类框架图

### 3.1. 文本表示

对采集到的评论数据进行预处理，文本预处理包括三个主要步骤：文本清洗、文本分词和去除停用词。首先，文本清洗即过滤重复文本和无用文本，以及去除文字以外的其他符号；接着，使用 JIEBA 分词方法将在线评论文本切割为一个个独立的词的集合，目的是方便计算机识别和理解；最后，结合“哈工大停用词表”以及人工添加方法，确定本文实验的停用词表，去除无意义和无突出贡献的词，保留对文本分类有价值的词，提高关键词的密度。

在文本表示前,首先需要选择文本的特征项,特征项需要具备反映文档内容和区分不同文档的能力。目前的研究中,极少有使用句子和段落作为特征项的,通常选取词作为特征项表示。采用 Word2vec 方法对文本进行向量转换后,每一个词对应一个唯一的词向量。在向量空间中,词是可以进行运算的,词之间空间距离的远近代表了词语含义的相似度。

在确定特征项后,需要选用合适的文本表示模型。本文使用 WordAVG 方法进行文本表示,将每条评论中的所有词向量计算平均,结果为这条评论的句子向量表示。

### 3.2. 特征降维

由于朴素贝叶斯基于特征条件独立性的假设过于理想化,在实际应用中会影响分类器的结果,因此研究人员通常采用特征加权的技术来抑制特征独立性的假设。本文使用 TFIDF 权重法计算特征项的权重,该方法考虑了词频和反文档频率。当特征词在某一文本中出现的频率越大,表示其对于该文本的分类作用越大;而特征词在大部分文本中出现的频率越大,则表示其对于文本的分类作用越小。具体计算公式如下:

$$IDF(t_k) = \log\left(\frac{N}{n(t_k)} + 0.01\right) \quad (1)$$

$$W_k = TF(t_k) \times IDF(t_k) = \frac{TF(t_k) \times IDF(t_k)}{\sqrt{\sum_{i=1}^m (TF(t_k) \times IDF(t_k))^2}} \quad (2)$$

其中  $TF(t_k)$  是特征  $t_k$  在训练数据集中出现的频数,  $IDF(t_k)$  表示反文档频率,  $n(t_k)$  是包含特征  $t_k$  的文档数,  $N$  是训练数据集的文档数。

### 3.3. 灰色关联分析

情感极性标注思路为:首先对文本中的情感词汇进行识别,对照情感词表找出文本中的情感词。从评论文本的第一个情感词开始搜索,若情感词前面出现修饰副词,则用该词的情感值乘以其修饰副词的权重,得到新的情感值。重复相同的步骤,直至遍历完该评论文本中所有的情感词。最后,将所有情感值加和,作为这条评论文本的情感值,对于大部分偏积极的评论而言,其情感值为正数,对于偏消极的评论而言,情感值为负数。

设评论文本集为:  $D = \{d_1, d_2, \dots, d_n\}$ , 将文本  $d_i \in D (1, 2, \dots, n)$  经过分词处理后,得到

$d_j = \{w_{i1}, w_{i2}, \dots, w_{ik}\}$ , 对照词表搜索情感词  $w_{ip}$ 、程度副词  $w_{iq}$  和否定词  $w_{ir}$ , 其中

$w_{ip}, w_{iq}, w_{ir} \in \{w_{i1}, w_{i2}, \dots, w_{ik}\}$ 。具体标注规则如下:

1) 设情感词  $w_{ip}$  的情感值为  $P_{w_{ip}}$ , 当  $w_{ip}$  为正向情感词时,令  $P_{w_{ip}} = 1$ ; 当  $w_{ip}$  为负向情感词时,令  $P_{w_{ip}} = -1$ 。

2) 程度副词  $w_{iq}$  的权重用  $P_{w_{iq}}$  表示,当情感词  $w_{ip}$  前面有程度副词  $w_{iq}$  时,此时需要更新情感值,用  $P'$  表示:

$$P' = P_{w_{ip}} \times P_{w_{iq}} \quad (3)$$

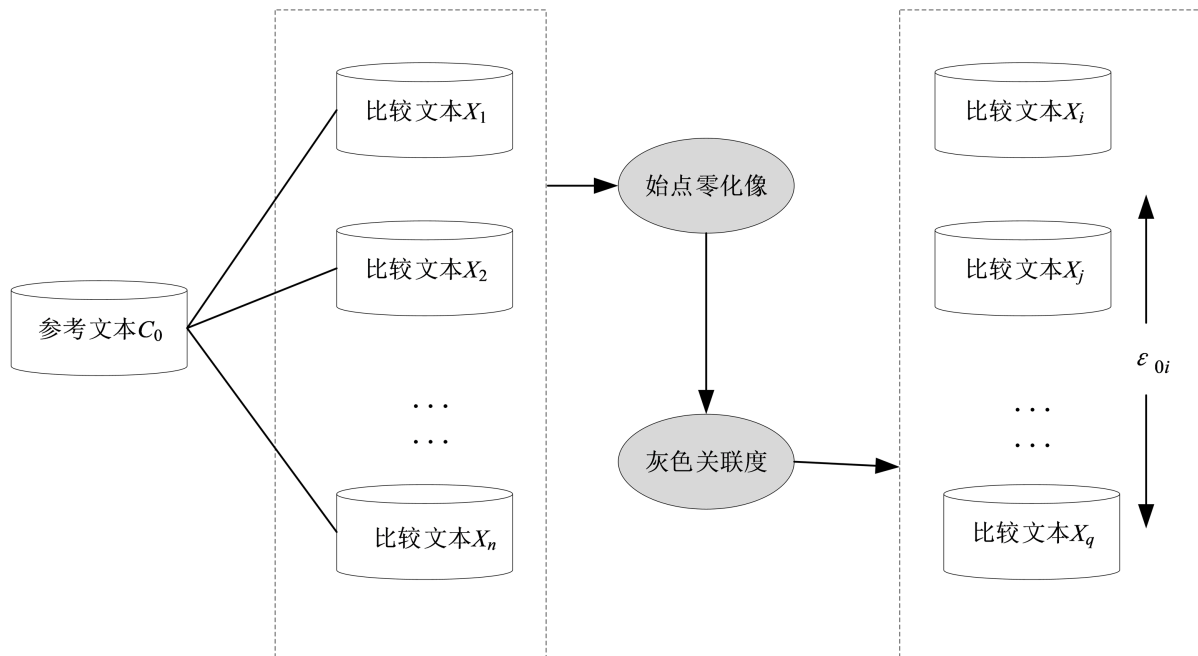
3) 否定词  $w_{ir}$  的权重用  $P_{w_{ir}}$  表示,当情感词  $w_{ip}$  前面有否定词  $w_{ir}$  时,  $P_{w_{ir}} = -1$ , 此时更新情感值,用  $P''$  表示:

$$P'' = P_{w_{ip}} \times P_{w_{ir}} \quad (4)$$

4) 情感词  $w_{ip}$  的最终情感值用  $P_{w_{ip}}$  表示,  $P_{w_{ip}} = P' + P''$ 。将评论文本  $d_i$  中所有情感值相加,得到该条评论的总体情感值  $P_{d_i}$ :

$$P_{d_i} = \sum_p P_{w_{ip}} \tag{5}$$

基于情感指数，使用灰色关联分析方法研究评论文本之间的关联关系。分别计算各评论文本和参考文本的灰色绝对关联度，灰色绝对关联度越大表示该条评论文本越接近于参考文本。灰色关联度用于比较各个序列对于同一参考序列的关联程度，反映了各个序列对于同一参考序列的优劣关系。灰色关联分析的总体框架如图 3 所示，该方法共分为 3 个步骤。



**Figure 3.** Grey correlation analysis  
**图 3.** 灰色关联分析

- 1) 构建参考序列和比较序列。由于存在  $n$  条评论， $m$  维特征向量，因此可以假设  $C_0 = (c_0(1), c_0(2), \dots, c_0(m))$  代表参考序列，而  $X_i = (x_i(1), x_i(2), \dots, x_i(m)), (i=1, 2, \dots, n)$  代表比较序列。
- 2) 求  $C_0$  和  $X_i$  的始点零化像，记为  $C_0^0$  and  $X_i^0$ 。由于不同序列的量纲不同，不可直接比较，因此需要将参考序列和比较序列的值转化为无量纲值。通过公式分别得到处理后的序列：

$$C_0^0 = (c_0^0(1), c_0^0(2), \dots, c_0^0(m)) \tag{6}$$

$$X_i^0 = (x_i^0(1), x_i^0(2), \dots, x_i^0(m)), (i=1, 2, \dots, n) \tag{7}$$

其中

$$c_0^0(k) = c_0(k) - c_0(1), k=1, 2, \dots, m \tag{8}$$

$$x_i^0(k) = x_i(k) - x_i(1), k=1, 2, \dots, m; i=1, 2, \dots, n \tag{9}$$

- 3) 计算参考序列  $C_0$  和比较序列  $X_i$  的灰色绝对关联度  $\varepsilon_{0i}$ 。本文使用公式

$$\varepsilon_{0i} = \frac{1 + |S'_0| + |S'_i|}{1 + |S'_0| + |S'_i| + |S'_i - S'_0|}, i=1, 2, \dots, n \tag{10}$$

其中

$$|S'_0| = \left| \sum_{k=2}^{m-1} c_0^0(k) + \frac{1}{2} c_0^0(m) \right| \quad (11)$$

$$|S'_i| = \left| \sum_{k=2}^{m-1} x_i^0(k) + \frac{1}{2} x_i^0(m) \right|, i = 1, 2, \dots, n \quad (12)$$

$$|S'_i - S'_0| = \left| \sum_{k=2}^{m-1} (x_i^0(k) - c_0^0(k)) + \frac{1}{2} (x_i^0(m) - c_0^0(m)) \right|, i = 1, 2, \dots, n \quad (13)$$

### 3.4. 朴素贝叶斯文本分类器改进

将灰色关联分析和朴素贝叶斯算法相结合,对传统朴素贝叶斯文本分类器进行优化,考虑在线评论中的用户情感倾向,将灰色关联度作为一项特征属性嵌入朴素贝叶斯文本分类模型中。本研究文本分类器使用朴素贝叶斯学习算法对训练集进行统计计算,计算各个类别的先验概率。模型训练阶段使用历史数据通过朴素贝叶斯算法来训练分类模型。训练数据集中的每一条评论文本  $d_i$  可以用一个  $m$  维属性向量  $d_i = \{t_1, t_2, \dots, t_m\}$  表示。类别集合为  $C = \{c_1, c_2\}$ , 表示好评类和差评类共两个类别。首先计算各个类别的先验概率:

$$P(C_j) = \frac{|C_j|}{|T_{train}|} \quad (14)$$

$$P(d_i | C_j) = P(\{t_1, t_2, \dots, t_m\} | C_j) = \prod_{k=1}^m P(t_k | C_j) \quad (15)$$

其中,  $P(d_i | C_j)$  表示文档  $d_i$  属于类别  $C_j$  的条件概率,  $P(C_j)$  表示类别  $C_j$  出现的概率,  $P(d_i)$  表示所有特征的联合概率。

将测试集中未知类别结果的文本输入预测器,利用贝叶斯公式通过先验概率求后验概率,采用极大似然估计法计算可能性最大的类别,标记该未知类别模块归于某一类别。根据贝叶斯定理:

$$P(C_j | d_i) = \frac{P(d_i | C_j)P(C_j)}{P(d_i)} \quad (16)$$

其中,  $P(C_j | d_i)$  表示样本  $d_i$  属于类别  $C_j$  的概率,是后验概率。由于分母  $P(d_i)$  对于所有类别而言是常数,则上式可以简化为只求分子  $P(d_i | C_j)P(C_j)$  最大即可。因此所求后验概率公式可转化为求解:

$$C_{map} = \max_{C_j \in C} P(C_j | d_i) = \max_{C_j \in C} P(d_i | C_j)P(C_j) = \max_{C_j \in C} P(\{t_1, t_2, \dots, t_m\} | C_j)P(C_j) \quad (17)$$

其中  $C_{map}$  表示最终分类结果。

## 4. 实证分析

本节为验证基于情感分析改进的在线评论分类模型的性能,使用实际采集评论文本数据,结合情感分析结果和文本分类预测结果,提出相关管理建议,为商家和平台提供参考。

### 4.1. 在线评论数据采集

本文选取戴森(Dyson) V10 Fluffy Extra 手持无线吸尘器作为分析对象,使用八爪鱼爬虫软件自动爬取京东网购平台该产品自 2020 年 4 月 4 日~2022 年 3 月 6 日期间的在线评论数据,共采集 2716 条评论,采集的在线评论如表 1 所示。同时,为了提高数据集类别的准确度,基于京东平台自身的归类结果,本文还使用了人工标注的方式对原始数据集分类结果进行修正。

**Table 1.** Online comment collection  
**表 1.** 在线评论采集

会员	评价内容	时间
*****	吸力大小：吸力可以，用最大档基本能将吸附到地毯猫毛吸起来；但最大档用时较短。	2022-03-06 13:37
叶****	东西是正品，感觉物有所值，吸力很大！	2022-03-06 13:29
p****	满意，操作简单方便，除尘很干净，易清洗，一机多用，噪音小！	2022-03-06 11:52
石****	超级好用，猫猫掉的毛吸的好干净，扫地机器人每周三次，卫生轻松搞定。	2022-03-06 11:52
我****	轻奢品，果然不同凡响，好用。	2022-03-06 11:43
i****	没有说明垃圾桶拆卸清洗的时候，里面过滤网如何清洗。	2022-03-06 11:06
7****	搬新家购买了这款吸尘器，满足吸尘的需要！好！	2022-03-06 10:44
懒****	用着不错，家里有宠物可以选择。用着不错，家里有宠物可以选择。用着不错，家里有宠物可以选择。	2022-03-06 09:13

## 4.2. 情感分析结果

本文以京东购物平台 Dyson 无绳吸尘器 V10 Fluffy Extra 产品为例，使用基于 Word2vec 的产品特征情感分析模型进行实验。用户情感倾向表达了消费者对产品特征的满意度，体现了该产品的市场竞争力。如果大部分消费者对产品的某个特征表达了正向的情感倾向，那么表明该产品特征的用户满意度就高。结合人工筛选和 Word2vec 词聚类方法，确定产品特征词库：“吸力”、“除尘”、“噪音”、“轻便”、“无绳”、“续航”、“重量”，表明这些产品特征属于用户重点关注的。

总体而言，通过对产品七个特征“吸力”、“除尘”、“噪音”、“轻便”、“无绳”、“续航”、“重量”进行情感分析，表 2 结果显示产品在“除尘”特征方面的用户满意度最高，达到 79.97%；其次是“吸力”、“无绳”和“噪音”，满意度分别为 77.65%、75.77%和 75.19%；满意度最低的产品特征是“重量”、“轻便”和“续航”，分别为 73.82%、72.95%和 71.56%。

**Table 2.** Identification results of “feature-point”

**表 2.** “特征 - 观点”对识别结果

特征词	“特征 - 观点”对	积极评论	占比	消极评论	占比
吸力	1168	907	77.65%	261	22.35%
除尘	2147	1717	79.97%	430	20.03%
噪音	1201	903	75.19%	298	24.81%
轻便	562	410	72.95%	152	27.05%
无绳	421	319	75.77%	102	24.23%
续航	1586	1135	71.56%	451	28.44%
重量	382	282	73.82%	100	26.18%



### 4.3. 分类结果与对比

#### 1) 评价指标

由于实验结果是判断一条评论文本为好评(正类)或是差评(负类), 是一个二分类问题。所有样本经过文本分类器分类之后, 一共会出现 4 种组合情况。

F-measure 是用来衡量分类器的综合性能的指标。AUC 表示 ROC 曲线所覆盖的区域面积值, AUC 的取值范围在 0.5 到 1 之间。AUC 可以直观地评价分类器分类性能的好坏, 值越大表示分类性能越好。

本研究使用 PD、PF、Precision、Recall、F-measure、AUC 等常用的评价参考项作为评估模型预测结果的判断依据。

$$\text{准确率(Precision): } P = \frac{TP}{TP + FP} \quad (18)$$

$$\text{召回率(Recall): } R = \frac{TP}{TP + FN} \quad (19)$$

$$\text{F-measure: } F\text{-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (20)$$

其中,  $TP$  表示属于类别“P”的样本中被归于该类的样本数量;  $FN$  表示属于类别“P”的样本被归类于“N”类的样本数量;  $TN$  表示属于类别“N”的样本中被归于该类的样本数量;  $FP$  表示属于类别“N”的样本被归于“P”类的样本数量。

#### 2) 模型分类结果

将预处理后的数据集分为两部分, 一部分用于模型训练阶段, 另一部分用于预测阶段。本研究采用 10-折交叉验证法, 模型预测结果  $TP$ 、 $FP$ 、 $P$  和  $R$  值分别为 0.909、0.091、0.909、0.909, 以及分类性能表现 F-measure 值和 AUC 值分别为 0.909 和 0.958。从实验结果看, 本文构建的基于传统朴素贝叶斯模型改进的文本分类模型具有很好的分类效果, 文章会将模型的预测效果与改进前的模型进行对比。

为了验证本文提出的文本分类模型的优势, 文章将改进前后的模型进行对比。为了确保模型比较的公平性, 采用相同的文本预处理、文本表示、朴素贝叶斯算法, 对原始数据集进行处理, 比较模型的预测效果, 并对结果进行分析。传统的基于朴素贝叶斯的文本分类方法, 经过文本分词和文本向量化表示后, 通过数据预处理步骤, 采用 10-折交叉验证法, 不考虑灰色关联分析步骤, 将原始数据集经过朴素贝叶斯文本分类器进行训练和预测。模型预测结果  $TP$ 、 $FP$ 、 $P$  和  $R$  值分别为 0.882、0.118、0.884、0.882, 以及分类性能表现 F-measure 值和 AUC 值分别为 0.882 和 0.948。改进前后的两个模型 4 个绩效评价指标值上的分类性能表现如表 3 所示, 实验结果表明本文构建的文本分类模型在准确率、召回率、F 值和 AUC 值等方面分类性能明显优于传统朴素贝叶斯分类模型, 从 4 个绩效指标可以看出, 绝对误差值分别为 2.83%、3.06%、3.06%和 1.05%, 可知改进后的模型始终优于改进前的传统朴素贝叶斯分类模型, 表明本文构建的基于情感分析改进的朴素贝叶斯分类模型在预测方面有着较好的效果, 具有一定的优势。

**Table 3.** Model comparison results

**表 3.** 模型对比结果

	$TP$	$FP$	$P$	$R$	F-measure	AUC
改进前	0.882	0.118	0.884	0.882	0.882	0.948
改进后	0.909	0.091	0.909	0.909	0.909	0.958
绝对误差	3.06%	-22.88%	2.83%	3.06%	3.06%	1.05%

## 5. 结语

本研究提出一种基于情感分析改进的在线评论分类模型,对情感分析和文本分类进行研究具有一定的创新性。本文以京东商城真实在线评论数据为例,验证模型有效性。旨在从产品特征的角度对正、负面的在线评论文本进行用户的情感倾向研究,挖掘消费者对产品的意见态度,为企业提供决策支持。文章完成了以下工作:

1) 解决了研究数据的规范问题。在预处理部分,采用准确的方法实现文本分词、去除停用词、特征降维等工作。选取京东商城用户在线评论真实数据进行了实例验证,结果显示形成的数据集是规范可用的。

2) 提高了朴素贝叶斯文本分类器的分类性能。文章考虑在线评论中的用户情感倾向,将灰色关联分析结果作为一项特征属性嵌入朴素贝叶斯文本分类模型中。结果显示,改进后的模型分类性能明显领先于传统朴素贝叶斯分类模型,评价指标 F 值提升了 3.06%,表明该方法在在线评论文本分类应用中具有一定的优势。

## 参考文献

- [1] 佚名. 第48次《中国互联网络发展状况统计报告》发布超十亿用户接入互联网我国成全球最庞大数字社会[J]. 网络传播, 2021(9): 76-81.
- [2] 祝琳琳, 李贺, 李世钰, 等. 面向电子商务网站的在线评论信息质量感知机理研究[J]. 图书情报工作, 2021, 65(6): 32-40.
- [3] 谢法举, 刘臣, 唐莉. 在线评论情感分析研究综述[J]. 软件导刊, 2018, 17(2): 1-4+7.
- [4] 赵宇晴, 阮平南, 刘晓燕, 等. 基于在线评论的用户满意度评价研究[J]. 管理评论, 2020, 32(3): 179-189.
- [5] Arndt, J. (1967) Role of Product-Related Conversations in the Diffusion of a New Product. *Journal of Marketing Research*, 4, 291-295. <https://doi.org/10.1177/002224376700400308>
- [6] Chatterjee, P. (2001) Online Reviews: Do Consumers Use Them? *Advances in Consumer Research*, 28, 129-133.
- [7] 李磊, 宋建伟, 刘继. 基于在线评论情感分析的声誉影响效应研究[J]. 管理学报, 2020, 17(4): 583-591.
- [8] 何有世, 王明. 基于多特征组合的在线产品评论情感挖掘研究[J]. 软件导刊, 2017, 16(5): 1-5.
- [9] Das, S.R. and Chen, M.Y. (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53, 1375-1388. <https://doi.org/10.1287/mnsc.1070.0704>
- [10] Georgiadou, E., Angelopoulos, S. and Drake, H. (2020) Big Data Analytics and International Negotiations: Sentiment Analysis of Brexit Negotiating Outcomes. *International Journal of Information Management*, 51, Article ID: 102048. <https://doi.org/10.1016/j.ijinfomgt.2019.102048>
- [11] Whissell, C. (2009) Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language. *Psychological Reports*, 105, 509-521. <https://doi.org/10.2466/PRO.105.2.509-521>
- [12] Wang, L., Guo, W., Yao, X., et al. (2021) Multimodal Event-Aware Network for Sentiment Analysis in Tourism. *IEEE MultiMedia*, 28, 49-58. <https://doi.org/10.1109/MMUL.2021.3079195>
- [13] Sanglerdsinlapachai, N., Plangprasopchok, A., Tu, B.H., et al. (2021) Improving Sentiment Analysis on Clinical Narratives by Exploiting UMLS Semantic Types. *Artificial Intelligence in Medicine*, 113, Article ID: 102033. <https://doi.org/10.1016/j.artmed.2021.102033>
- [14] Zhang, K., Zhu, Y., Zhang, W., et al. (2021) Cross-Modal Image Sentiment Analysis via Deep Correlation of Textual Semantic. *Knowledge-Based Systems*, 216, Article ID: 106803. <https://doi.org/10.1016/j.knosys.2021.106803>
- [15] 贾晓帆, 何利力. 融合朴素贝叶斯与决策树的用户评论分类算法[J]. 软件导刊, 2021, 20(7): 1-5.
- [16] Kollur, J. and Razia, S. (2020) Text Classification Using Nave Bayes Classifier. *Materials Today: Proceedings*, in Press. <https://doi.org/10.1016/j.matpr.2020.10.058>
- [17] Hindi, K., Alijulaidan, R.R. and Alsalman, H. (2020) Lazy Fine-Tuning Algorithms for Nave Bayesian Text Classification. *Applied Soft Computing*, 96, Article ID: 106652. <https://doi.org/10.1016/j.asoc.2020.106652>
- [18] Wei, Z. and Feng, G. (2011) An Improvement to Naive Bayes for Text Classification. *Procedia Engineering*, 15, 2160-2164. <https://doi.org/10.1016/j.proeng.2011.08.404>

- 
- [19] Zhang, H., Jiang, L. and Yu, L. (2021) Attribute and Instance Weighted Naive Bayes. *Pattern Recognition*, **111**, Article ID: 107674. <https://doi.org/10.1016/j.patcog.2020.107674>
- [20] 熊志斌, 刘冬. 朴素贝叶斯在文本分类中的应用[J]. 软件导刊, 2013, 12(2): 49-51.