文章编号: 1007-1423(2024)21-0141-05 DOI: 10.3969/j.issn.1007-1423.2024.21.026

京东手机用户评论的情感分析及聚类分析

苏舒菲，蔺 聪*

（广东财经大学统计与数学学院，广州 510320）

摘要： 为了帮助商家了解消费者对商品的需求偏好以及消费者群体构成，构建了基于词典划分的情感分析和K-means聚类来识别在线评论中产品需求偏好以及客户群组模型。通过爬取京东平台华为Mate60系列手机在线评论并对其进行处理；采用LDA主题模型确定消费者关注的主题并利用HowNet词典结合自定义词典的情感分析来评分。最后基于词向量利用K-means聚类算法得到消费者细分构成，帮助商家根据不同聚类群组的特点制定明确的产品定位和特色以满足消费者的需求。

关键词： 在线评论；情感分析；K-means聚类；主题挖掘；文本预处理

0 引言

目前在电子数码领域，对用户评论的情感分析和聚类分析已有一些研究，大多集中在亚马逊、淘宝、天猫平台，对京东的在线评论进行情感和聚类分析的研究较为有限。林伟振等[1]基于亚马逊平台健康监测穿戴产品的在线评论，使用LDA等方法确定用户对产品的喜恶，深度挖掘其影响，从而对商家优化产品和提升服务给出建议。李宗敏等[2]介绍基于情感分析和聚类分析微博在线评论的探索，涵盖情感分析的基本概念、情感词典构建方法等；还介绍聚类分析在微博评论中的应用。AL-Sharuee等[3]通过情感分析技术提取商品评论的情感信息，并使用聚类分析方法将相似情感评论聚集在一起，实现对不同方面观点和消费者情感倾向的准确识别。本文在现有研究的基础上，运用自然语言技术对京东平台上华为Mate60系列在线评论进行情感分析和聚类分析。

1 数据获取与预处理

本文主要使用主题网络[4]的爬取技术来获取相关数据。本文预设主题是华为系列四款手机的名称，使用相关Python库，如requests、pandas等来对京东的在线评论进行爬取，共爬取66个ID下17834条评论，为了提高数据质量进行以下数据预处理工作：

（1）去除小于等于10字符的评论：如“哈哈哈”“66”等评论提供的信息并没有太大的价值，使用pandas库将这些长度在10字符（含10）以下的评论去除。经过处理剩下16231条在线评论。

（2）去除重复评论：针对存在相似或完全相同的评论进行去重处理，剩下8065条独立评论。

（3）Jieba精确分词处理：将长度超过10字符的评论进行切分，使评论转化为更易处理和分析的词语序列。

（4）哈工大停用词处理：对“我”“华为”这类无意义词语进行停用词去除处理。

收稿日期： 2024-04-16 **修稿日期：** 2024-07-21

基金项目： 广东省普通高校特色创新项目（自然科学）（2022KTSCX041）；广州市海珠区科技计划项目（海科工商信计2022-45）；2022年度广东财经大学一流本科教学质量与教学改革工程项目（粤财大〔2022〕132号）：面向互联网+的《非结构化数据挖掘》混合式教学改革探索；2024年广东财经大学统计与数学学院课程建设项目（培育项目）：《非结构化数据挖掘》案例建设

作者简介： 苏舒菲（2001—），女，广东湛江人，在读本科生，研究方向为数据挖掘；*通信作者：蔺聪（1979—），男，河北邯郸人，博士，讲师，硕导，研究方向为信息安全、数据挖掘，E-mail：lotuslin2005@126.com

2 算法模型构建

2.1 LDA主题模型挖掘主题-特征词

本文使用网格搜索法结合5折交叉验证选择最佳主题数，并根据主题一致性得分和困惑度[5](计算如式(1))评估，观察主题-困惑度折线图寻找困惑度变化“肘形”点，将其作为最优的主题数目[6]。

$$P(D)=\exp\frac{\sum_{d=1}^{M}\log P(W_d)}{\sum_{d=1}^{M}N_d} \tag{1}$$

式中：D表示数据集所有词集合；W_d表示文档d中的词；$P(W_d)$表示文档d中的词出现的概率；M表示文档总数量；N_d表示每个文档中d的词数。

在此之前考虑到特征词词义相近、重复出现等问题，通过Word2Vec训练有效的在线评论，计算相似度来对在线评论数据文本之间相似程度进行衡量，以此识别具有相似语义的文本词语，从而降低重复率[7]。对数据集采用Word2Vec中的Skip-gram方法进行训练：首先通过梯度下降法对目标函数[8](如式(2))进行梯度下降；提取形容词、名词作为输入使用Word2Vec训练；使用式(3)预测周围词的总概率对的数值；最后利用式(4)来计算词语之间的相似特征。

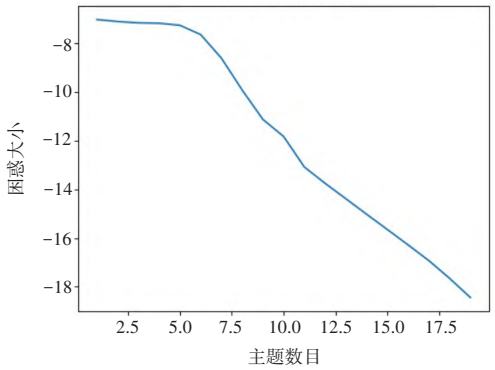
$$\frac{1}{N}\sum_{n=1-c\le i\le c,i\ne 0}^{N}\sum_{i\ne 0}\log P\left(\boldsymbol{\omega}_{n+j}\middle|\boldsymbol{\omega}_n\right) \tag{2}$$

$$P\left(\boldsymbol{\omega}_n\in C_n\middle|\boldsymbol{\omega}_n\right)=\sigma(W_n^N W_k)=(1+e^{-(W_n^N W_k)})^{-1} \tag{3}$$

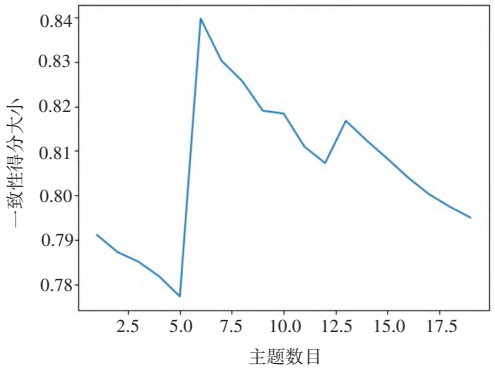
$$\sin(\boldsymbol{\omega}_n,\boldsymbol{\omega}_k)=\cos(W_n^N,W_k)=\frac{W_n^N W_k}{\|W_n\|\|W_k\|} \tag{4}$$

式中：变量N是语料库词的总个数；$\boldsymbol{\omega}_n$是当前词的词向量；$\boldsymbol{\omega}_k$是任意词语的词向量。

如图1(a)所示，主题数超过6时，困惑度下降，说明模型此时具有较高的复杂性，且一直上升，导致出现过拟合，故在主题数为1至6的范围内确定最终主题数。结合一致性得分变化，如图1(b)，当主题数从1增加到5时，一致性得分呈现下降趋势；当主题数增加到6时，一致性得分上升；随后随主题数增加，一致性得分整体上呈现下降趋势。故确定主题数为6，此时一致性得分最高，为0.8390。



(a) 主题-困惑变化情况



(b) 主题-一致性得分变化情况

图1 主题-困惑度/一致性得分变化

通过对评论文本数据进行主题挖掘，得到了六个主题组，主题命名结合了华为Mate60发布会背景，确保主题与手机的关键属性相一致，最终主题命名为性价优质、功能体验、性能配置、外观设计、物流服务和拍照质量。结果见表1。

表1 LDA主题-特征词及主题分类

主题类别	主题特征词
性价优质	不错、正品、国货、国产、价格、好用、推荐、购买、支持、值得
功能体验	体验、喜欢、流畅、效果、强大、满意、音效、运行、待机时间、完美、功能
性能配置	遥遥领先、待机时间、质量、电池、续航、屏幕、系统、信号、性能、质感、充电
外观设计	款式、颜色、外观、好看、漂亮、大气、外形、颜值
物流服务	物流、速度、快递、服务、很快、包装、发货、商家
拍照质量	拍照、照相、相机、清晰、像素、摄像头、照片、特别、高、出色

2.2　基于HowNet情感词典情感分析

采用HowNet[9]作为情感词典，基于LDA[10]主题-特征词挖掘，将主题属性与情感词典进行关联，以确定每个主题属性的情感值。

首先，根据主题特征词选择在线评论进行预处理。将在线评论按标点符号如“。”“!”等进行切割，得到独立在线评论，接着进行Jieba精确模式分词处理，确保每个在线评论只包含一个主题的特征词。

其次，使用完善后的情感词典对分词后的在线评论按以下方法进行情感标记和计算：对于正面情感在线评论得分加1；对于消极情感在线评论得分减1；考虑否定词的影响，有偶数个否定词情感倾向不变；反之情感倾向乘以“-1”来改变方向。

接下来将构建HowNet程度副词词典。对于每个程度副词，将赋予相应的权重以反映其在情感表达中的强度，本文将程度副词分为低量、中量、高量、极量四个等级。其中“中量”是基础等级，强度值设为1；“低量”强度被模糊量化50%，设为0.5；以此类推，“高量”为1.5，“极量”为2[11-12]。文本单元情感得分计算如式(5)。

$$score(d_m) = (-1)^i \sum_{i=1}^{k} S_l \prod_{j=1}^{n} D_j \tag{5}$$

最后计算各主题情感得分。特征词集合$W_{ij} = \{w_{11}, w_{12}, \cdots, w_{ij}\}$，将其与在线评论文本单元进行匹配，确定在评论数据出现的次数a_{ij}，根据式(6)、(7)分别计算每个特征词的情感得分、各主题的情感得分。

$$score(w_{ij}) = \frac{\sum score(d_m)}{a_{ij}}, w_{ij} \in d_m \tag{6}$$

$$score(z_k) = \frac{\sum_{i=k} score(w_{ij}) a_{ij}}{\sum_{i=k} a_{ij}} \tag{7}$$

从情感得分情况(见表2)来看，华为Mate60系列手机用户对于各个主题的关注度从高到低依次是性价优质、功能体验、拍照质量、外观设计、性能配置和物流服务。

表2　主题特征情感得分情况

主题	占比/%	情感得分	积极	消极	中性
性价优质	18.4	1.1714	1.1337	0.0370	0.0006
功能体验	15.8	1.1225	1.0861	0.0355	0.0009
性能配置	16.3	1.0135	0.9632	0.0494	0.0009
外观设计	15.8	1.0735	1.0341	0.0389	0.0005
物流服务	16.3	0.7767	0.7394	0.0357	0.0016
拍照质量	17.4	1.1365	1.0890	0.0477	0.0002

2.3　轮廓系数确定K值的K-means算法

基于对每个主题特征的情感得分，将在线评论转换为空间向量，通过比较不同K值下的轮廓系数，选择具有较高轮廓系数的K值作为最佳聚类个数，根据样本i的簇内不相似度a_i和簇间不相似度b_i，定义样本i轮廓系数，如式(8)[13]。

$$\begin{cases} s(i) = \dfrac{b(i) - a(i)}{\max\{a(i), b(i)\}} \\ b(i) = \min\{b_{i1}, b_{i2}, \cdots, b_{ik}\} \end{cases} \tag{8}$$

式中：$s(i)$取值在[-1，1]之间，当$s(i)$接近1时，样本i与其所属簇内的其他样本紧密相连，与其他簇距离较远，表示聚类结果良好。所有样本$s(i)$的均值称为聚类的轮廓系数，定义为S。S值越大，聚类分析的效应越强[14]。

$$S = \frac{\sum s(i)}{n} \tag{9}$$

根据实验结果(如图2所示)，当聚类个数从1逐渐增加到6时，轮廓系数呈现上升的趋势，表明聚类效果逐渐提升；当聚类个数超过6时，轮廓系数下降，这意味着聚类结果变得不太理想。

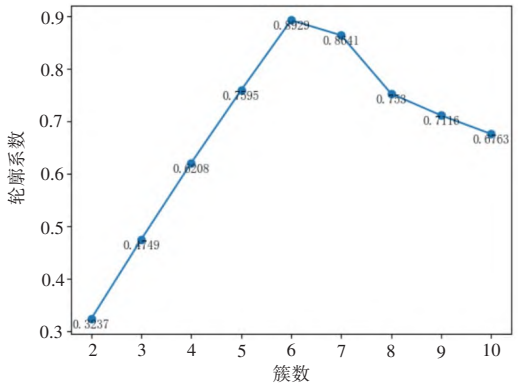


图2　K-means聚类轮廓系数-簇数变化折线图

在聚类个数为6时，轮廓系数为0.8929，接近于1，表明聚类结果较为合理。经过K-means聚类分析，本文得到以下结果(见表3和图3)：

(1) 误差平方和为31.2906，较小的*SSE*值表示该点与其所属聚类中心的距离较近，聚类结果较准确；

(2) 组内方差均小于0.04，意味着每个聚类内样本相似度较高，样本之间的差异较小；

(3) 单个样本的轮廓系数在0.95左右，表明大多数样本在自己的聚类中具有较高的相似度，此外平均轮廓系数*S*为0.9565，表明每个聚类内部样本具有较高的相似度，*K*值为6合适。

表3 K-means聚类质量评估

评估指标	聚类一	聚类二	聚类三	聚类四	聚类五	聚类六
组内方差	0.0372	0.0381	0.0378	0.0375	0.0370	0.0372
SSE	31.2906					
$s(i)$	0.9547	0.9528	0.9517	0.9602	0.9610	0.9589
S	0.9565					

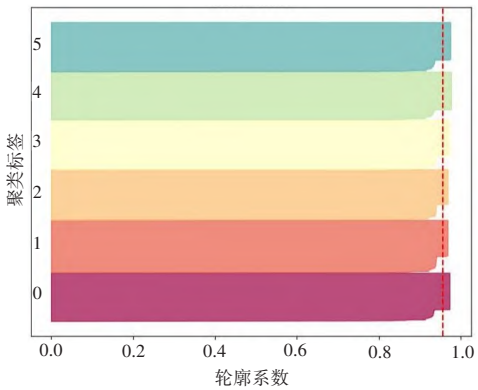


图3 单个主题轮廓系数值图

因此，本次聚类分析具有较好的效果。根据前文的情感分析得分，计算各聚类中各主题的情感分数，进一步了解用户评论在不同聚类中对特定特征的偏向程度，聚类结果见表4。

表4 K-means聚类情感倾向结果

主题	聚类一	聚类二	聚类三	聚类四	聚类五	聚类六
性价优质	125.8	19.48	334.95	119.13	44.72	28.6
功能体验	166.16	55.74	137.55	839.41	162.39	599.62
性能配置	122.76	57.64	55.68	972.28	110.59	227.08
外观设计	65.02	30.19	39.89	195.17	26.01	267.88
物流服务	39.91	16.44	49.85	207.97	321.06	230.86
拍照质量	395.56	49.51	97.80	413.04	61.95	149.82

3 情感分析与聚类分析综合讨论

不同聚类中的用户对各个主题的关注程度存在差异。功能体验主题在多个聚类中得到了高情感分数，说明用户普遍重视手机的功能和使用体验，而外观设计和物流服务主题的情感分数普遍较低，这可能意味着用户对手机外观设计和物流服务方面的体验不满意。根据其偏好和关注的不同，可以分为六个聚类群组。

聚类一是拍照偏好组，消费者看重手机的拍照质量，如主摄参数，华为Mate60Pro的超高像素和超微距长焦镜头实现全场景的全焦段覆盖，为使用者带来高质感成像效果，制造商可以再深一步提升拍照功能。

聚类二是差评组，消费者对手机的各主题都没有太高关注度，可能对商品持有负面评价，商家可以通过改进产品的性价比、功能、性能和外观等方面来提升产品质量。

聚类三是性价比组，消费者最注重手机的性价比，其他方面如拍照、性能等的关注度较均衡，商家可以提供性价比更高的产品并注重各个方面的平衡与优化。

聚类四是品质功能组，消费者看重华为Mate60系列的功能体验和性能配置，其他方面关注度也很高。商家可以持续提升产品的性能配置、功能体验，并确保产品的整体品质和可靠性。

聚类五是购物体验组，消费者信赖京东平台的物流服务，如商家态度、物流速度，但是对手机的其他方面满意度较低，京东平台可以优化物流服务、加强售后支持来提升消费者购物满意度。

聚类六是综合性能组，消费者除了性价优质方面关注度不高，其他方面满意度都很高，商家可以提供更具性价比产品以及在功能体验、拍照质量方面持续改进。

4 结语

本文采用实证分析方法对京东平台华为Mate60系列手机的在线评论进行情感和聚类分析，探究不同主题下的情感和不同聚类下的主题得分，从消费者的视角来反映满意度和关注度。具体包括Word2Vec词向量化、LDA主题挖

掘、HowNet词典主题评分、K-means算法细分消费者。这种方法不仅适用于电商平台，还可以应用于不同领域平台，比如在政务平台上，对用户进行满意度测评的分析，能帮助政府部门了解公众对政策、服务等方面的态度，从而对政务工作起到优化的作用。

参考文献：

[1] 林伟振,刘洪伟,陈燕君,等.基于在线评论的顾客满意度研究:以健康监测穿戴产品为例[J].数据分析与知识发现,2023,7(5):145-154.

[2] 李宗敏,张琪,杜鑫雨.基于辟谣微博的互动及热门评论情感倾向的辟谣效果研究:以新冠疫情相关辟谣微博为例[J].情报杂志,2020,39(11):90-95.

[3] AL-SHARUEE M T,LIU F,PRATAMA M. Sentiment analysis: an automatic contextual analysis and ensemble clustering approach and comparison [J]. Data & Knowledge Engineering,2018,115:194-213.

[4] 赵康.面向主题的网络爬虫系统的设计与实现[D].北京:北京邮电大学,2019(5):37-49.

[5] 裘惠麟,邵波.多源数据环境下科研热点识别方法研究[J].图书情报工作,2020,64(5):78-88.

[6] 赵蓉英,戴祎璠,王旭.基于LDA模型与ATM模型的学者影响力评价研究:以我国核物理学科为例[J].情报科学,2019,37(6):3-9.

[7] 朱韦光.基于在线评论的智能手机需求偏好判别及客户细分模型构建研究[J].计算机时代,2023(9):132-135.

[8] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2,2013:3111-3119.

[9] 王科,夏睿.情感词典自动构建方法综述[J].自动化学报,2016,42(4):495-511.

[10] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003(3):993-1002.

[11] 董丽丽,赵繁荣,张翔.基于领域本体、情感词典的商品评论倾向性分析[J].计算机应用与软件,2014,31(12):104-108.

[12] 韦婷婷,陈伟生,胡勇军,等.基于句法规则和HowNet的商品评论细粒度观点分析[J].中文信息学报,2020,34(3):88-98.

[13] 朱连江,马炳先,赵学泉.基于轮廓系数的聚类有效性分析[J].计算机应用,2010,30(S2):139-141.

[14] 王晓东,张姣,薛红.基于蝙蝠算法的K均值聚类算法[J].吉林大学学报(信息科学版),2016,34(6):805-810.

Emotion analysis and clustering analysis of Jingdong mobile phone user comments

Su Shufei, Lin Cong*

(School of Statistics and Mathematics, Guangdong University of Finance & Economic, Guangzhou 510320, China)

Abstract: In order to help merchants understand consumers' demand preferences for commodity and the composition of consumer bases, product demand preferences in online comments and customer group models are constructed based on dictionary-divided sentiment analysis and K-means clustering. The online comments of Huawei Mate60 series mobile phones from the Jingdong platform are crawled and then processed. The LDA topic model is used to determine consumers' topic of interest, and sentiment analysis using HowNet dictionary combined with custom dictionary is used to calculate sentiment scores. Finally, consumer segmentation is obtained based on word vectors and K-means clustering algorithm. This can help merchants develop clear product positioning and characteristics according to the characteristics of different clustering groups to meet consumer demands.

Keywords: online commentary; sentiment analysis; K-means clustering; topic mining; text preprocessing