Software Engineering and Applications 软件工程与应用, 2023, 12(6), 989-1000
Published Online December 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.126097

Hans汉斯

基于陶瓷电商产品评论的情感分析研究

杨利华*，李鑫扬，徐思雨，严　帅，陈泽民

景德镇陶瓷大学信息工程学院，江西　景德镇

收稿日期：2023年9月26日；录用日期：2023年12月21日；发布日期：2023年12月29日

摘　要

本文以天猫商城的景德镇红叶陶瓷青花瓷餐具盘碗碟套装为研究对象。通过爬取该研究对象的评论并对其进行文本挖掘预处理和词云绘制，采用基于BosonNLP和SnowNLP情感词典的两种情感倾向性分析方法进行情感分析和可视化展示，通过对比BosonNLP和SnowNLP两组方法得出的实验结果，从而为买家提供选购的参考标准，为卖家提供建设性的改进方向。此外，再进行相应的预处理后采用LDA主题模型进行分析。最后，利用交互可视化的方式把语句情感分数的分布情况再次直观展示，有效地验证了该产品情感偏向积极，得出了消费者可放心购买的实验结论。

关键词

情感分析，情感词典，词云，LDA主题模型，交互可视化

Research on Sentiment Analysis Based on Ceramic E-Commerce Product Reviews

Lihua Yang*, Xinyang Li, Siyu Xu, Shuai Yan, Zemin Chen

School of Information Engineering, Jingdezhen Ceramics University, Jingdezhen Jiangxi

Received: Sep. 26th, 2023; accepted: Dec. 21st, 2023; published: Dec. 29th, 2023

Abstract

This article takes the Jingdezhen Red Leaf Ceramic Blue and White Porcelain Tableware, Plate, Dish Set of Tmall as the research object. By crawling through the comments on the Red Leaf Ceramic Jingdezhen Blue and White Porcelain Tableware, Dish and Dish Set on Tmall, and performing text mining preprocessing and wordle rendering on it, two sentiment analysis methods based on BosonNLP and SnowNLP sentiment dictionaries were used for sentiment analysis and visuali-

*通讯作者。

文章引用: 杨利华, 李鑫扬, 徐思雨, 严帅, 陈泽民. 基于陶瓷电商产品评论的情感分析研究[J]. 软件工程与应用, 2023, 12(6): 989-1000. DOI: 10.12677/sea.2023.126097

zation display. By comparing the experimental results obtained by BosonNLP and SnowNLP methods, reference standards for buyers' selection were provided, Provide constructive improvement directions for sellers. In addition, after corresponding preprocessing, the LDA topic model is adopted Conduct analysis. Finally, the distribution of sentence sentiment scores was visually displayed through interactive visualization, effectively verifying the positive emotional bias of the product and drawing experimental conclusions that consumers can purchase with confidence.

Keywords

Sentiment Analysis, Sentiment Lexicon, Word Cloud, LDA Topic Model, Interactive Visualization

Copyright © 2023 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

数字时代，云计算等技术为电子商务的发展提供了全方位支持，直播电商等新业态和新模式，为用户提供了多元化的消费体验[1]。随之而来的是各大电商平台的评论数量不断增加，这些在线评论对消费者的购买决策产生了显著影响，因为它们是消费者获取产品质量信息的重要参考渠道[2]。

现如今，电商评论情感分析有很多研究。如 HU 等[3]通过情感分析挖掘用户对产品的情感倾向；李琴等[4]基于情感词典对在线景区评论进行情感分析得到情感类别倾向性与门票波动之间客观存在的联系；凌万云[5]等使用情感分析工具设计基于景区满意度的排序方法，为游客提供个性化的景点游览序列；Gunawan Leonard [6]使用 SVM (支持向量机)对雅加达的餐厅客户满意度进行分类，来获取用户对餐厅食物和服务质量的反馈。

本研究则通过网络爬虫对天猫商城的红叶陶瓷景德镇青花瓷餐具盘碗碟套装的评论进行采集，通过一系列文本挖掘预处理，绘制 stylecloud 高频词云图，后运用基于 BosonNLP 和 SnowNLP 情感词典的情感分析方法进行分析，然后对比两组实验结果，得出结论。之后再通过 LDA 主题模型对其评论进行分析，得到最优结果，再次证实结论。最后利用交互可视化工具展示评论数据，直观感受消费者对商品的满意度。

2. 陶瓷电商评论研究

2.1. 评论数据采集

本研究基于 Jupyter Notebook 云环境从天猫商城电商平台上采集“红叶居家日用旗舰店”的国瓷红叶陶瓷景德镇青花瓷餐具盘碗碟套装的在线评论(如下表 1)。

Table 1. Review data of Hongye ceramic tableware set
表 1. 红叶陶瓷餐具套装评论数据

序号	Comment
1	瓷器很好，没有瑕疵。
2	做为江西人，还是喜欢国瓷
3	很好，有个别破损，店家很爽快的补寄了
4	总体不错，就是有几个盘子底边有点粗糙。待改进。
...	...
258	这套餐具釉非常好，手感温润，的确是国瓷精品

2.2. 评论数据预处理

由于爬取到的景德镇青花瓷餐具盘碗碟套装评论信息有些可能是“刷单”、“刷的好评”，因此需要对这类无意义且重复的数据进行去重，同时进行数据清洗以过滤掉无效数据。另外，为了提升情感分析的准确性，我们还需要对剩下的景德镇青花瓷餐具盘碗碟套装有效评论进行中文分词、词性标注以及去除停用词等操作。

2.2.1. 文本去重

文本去重是指去除景德镇青花瓷餐具盘碗碟套装评论数据中的重复部分。这种情况可以分为两类：一类是京东平台为避免客户长时间未进行评论而设置的机制，如果用户超过规定时间仍未进行评论，系统会自动替客户做出好评；另一类是部分购买了景德镇青花瓷餐具盘碗碟套装的用户为了省事或刷好评(或差评)，直接复制粘贴其他人的评论。这两类重复的陶瓷商品评论数据，如果保留下来，不仅没有任何分析价值，且会干扰情感分析结果的准确性，所以都得进行去重处理(如表 2)。

Table 2. Partial text deduplication results
表 2. 部分文本去重结果

Comment
做为江西人，还是喜欢国瓷
很好，有个别破损，店家很爽快的补寄了
总体不错，就是有几个盘子底边有点粗糙。待改进。
很喜欢的一款餐具，很喜欢的一个牌子，放心，安心
有的有点瑕疵，不想换了
这个价格这样的 32 件套真的不便宜哦，有一些细小的瑕疵，包装就很到位，祝店家生意兴隆！
第一次寄套装包装有个薄泡沫，品锅破了。补发一只又破，包装连泡沫也省了。再次补发又要 3 到 7 天内发货，坑人的节奏，真怀疑是不是补发一个破锅来搪塞的，大家评评
质量还行，但烧的比较薄

2.2.2. 评论数据清洗

评论数据清洗是对数据进行重新校验和核对的过程，包括检查数据一致性，无效值和缺失值的处理等。本项目数据清洗的任务是过滤那些对我们进行商品评论情感分析无作用的数据，如各种数字、符号、表情，以及针对京东商城红叶陶瓷餐具盘碗碟套装评论中“京东”、“红叶”“陶瓷”、“餐具”等出现频数较大且无分析价值的词。确认过滤掉那些无效数据后再进行后续操作。

2.2.3. 中文分词

不同于英文句子，中文句子根据断句方式的不同，往往可能表达出不同甚至完全相反的含义，因此中文分词方法的选择尤为重要[7]。

jieba 分词的语料库来源主要有两类，一类是网上能下载到的笼统的语料库。另一类是自己收集或建立的针对陶瓷产品领域术语的语料库。本项目采用的是 jieba 中文分词，有时候按照 jieba 正常分词，会把我们不希望分开的词语分开，影响后续情感分析结果，这时我们可以调节单个词语的词频，使其能(或不能)被分出来(如表 3)。

Table 3. Chinese word segmentation results of some texts
表 3. 部分文本中文分词结果

Comment
瓷器 很 好，没有 瑕疵。
做 为 江西人，还是 喜欢 国瓷
很 好，有 个别 破损，店家 很 爽快 的 补 寄 了
总体 不错，就是 有 几个 盘子 底边 有点 粗糙。待 改进 。
很 喜欢 的 一款 餐具，很 喜欢 的 一个 牌子，放心，安心
包装 挺 好 的，碗 摸 起来 手感 不错
包装 完整 牢靠，面碗 质量 很 好。很 喜欢 这 款 面碗，性价比 很 高。服务 热情，耐心 周到。
白瓷，高亮，光洁，不 烫手，快递 很快，价格 也好。

2.2.4. 词性标注

分词成功以后，我们通常需要提取关键词，而关键词通常是名词、动词、动名词等其他词性的词组，所以在提取关键词之前，我们可以先对提取出来的词语做一下词性标注，以便于后续分类(如表 4)。

Table 4. Part-of-speech tagging results
表 4. 词性标注结果

词语	词性	权值
质感	n	7
喜欢	v	25
超赞	v	12
满意	v	16
...	...	...
损坏	v	2
瑕疵	n	12
掉色	v	1

2.2.5. 去停用词

停用词(Stop Words)是一些完全没有用或者没有意义的词，例如助词、语气词等。本项目采用的是哈工大停用词表，并结合项目本身“评论数据情感分析”需求制作出的新停用词表，进行二次过滤(如表 5)。

Table 5. Partial stop word removal results
表 5. 部分去停用词结果

Comment
瓷器 很 好 没有 瑕疵
做 喜欢 国瓷
很 好 破损 店家 很 爽快 补 寄
总体 不错 几个 盘子 底边 有点 粗糙 改进
很 喜欢 一款 餐具 很 喜欢 牌子 放心 安心
有点 瑕疵 不想 换
很 好 餐具 质量 很 好 品牌 值得 信赖
正品 景德镇 陶瓷 质量 非常 不错 家里 亮 不少 细腻

3. 基于词典的情感分析

3.1. 分词与生成词云图

通过 Jieba 分词和 StyleCloud 制作的红叶陶瓷青花瓷餐具盘碗碟套装的词云图(如图 1，图 2)，我们可以直观地看出该商品消费者最关注的问题，关注点主要集中在以下几个方面：

1) 对陶瓷商品质量、品质的关注；

2) 对陶瓷商品的包装的关注；

3) 对陶瓷店铺客服的关注。

同时用户对陶瓷产品的总体情感值是积极的，回看评论数据，用户对陶瓷商品的物流、客服的不满意，表现消极情绪。

Figure 1. Word cloud visualization result 1
图 1. 词云图可视化结果一

Figure 2. Word cloud visualization result 2
图 2. 词云图可视化结果二

3.2. 基于 SnowNLP 进行情感分析

SnowNLP 情感分析基于内置的情感系统所实现的，它将文本简单地分为积极与消极两类，输入文本评论信息即可获得返回值：这条评论是正面评论的概率(越接近于 1 越积极，接近 0 为消级)。[8]其底层思想为朴素贝叶斯模型，在已知样本信息的情况下推测评论信息积极的概率，具体公式如下：

$$P\left(c_1 \middle| w_1,\ w_2,\ldots,\ w_n\right)=\left(P\left(w_1,\ w_2,\ldots,\ w_n \middle| c_1\right)\times P\left(c_1\right)\right)\Bigg/\begin{pmatrix}P\left(w_1,\ w_2,\ldots,\ w_n \middle| c_1\right)\times P\left(c_1\right)\\ +P\left(w_1,\ w_2,\ldots,\ w_n \middle| c_2\right)\times P\left(c_2\right)\end{pmatrix}$$

其中 c_1 代表某条评论分类为积极，而 w_1，w_2,...，w_n 代表每条评论的样本信息(每条评论被划分为 n 个词向量，每个词向量都有一定的积极性权重，相加即可得到评论的积极性得分信息) [8]。

基于此，利用 SnowNLP 先对已经预处理完的青花瓷餐具盘碗碟套装的评论数据逐行计算其情感极性

值，并将其限定在[0, 1]范围内作为情感极性返回值，该值接近 1 表示情感情绪积极，接近 0 表示情感情绪消极。再分别生成与每段情感分数出现的频率对应的图。通过观察图 3 可知，靠近 1 的积极评论整体而言更密集且数量多。

在上述基础上再进行评论数据的情感波动情况分析，将情感区间设定在[−0.5, 0.5]范围内，其中大于 0 的评论被视为积极评论，小于 0 的评论被视为消极评论。这样做可以更直观地展现消费者对收到的盘碗碟套装实物的情感波动情况。通过观察图 2.3 可知，0 以上的好评要远远多于差评(对比 0 以上及以下曲线的疏密程度)。生成的情感柱状图和曲线图如下图 3，图 4 所示。

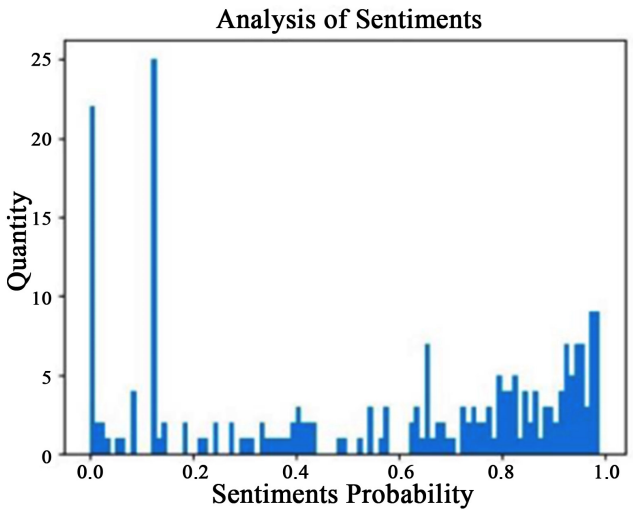

Figure 3. SnowNLP review data sentiment histogram
图 3. SnowNLP 评论数据情感柱状图

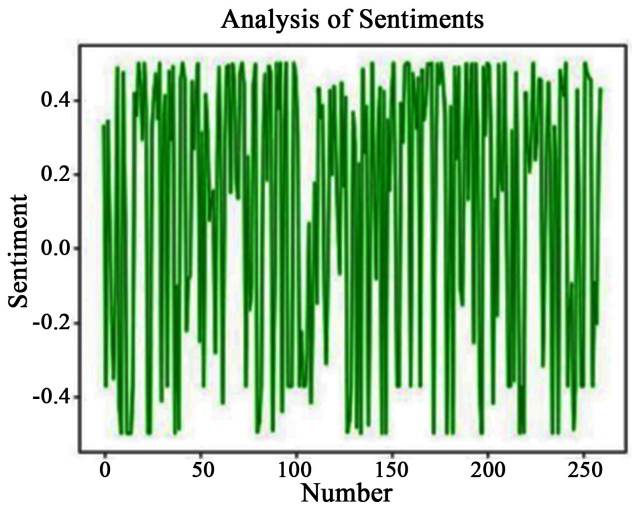

Figure 4. SnowNLP review data sentiment curve chart
图 4. SnowNLP 评论数据情感曲线图

3.3. 基于 BosonNLP 进行情感分析

0 为基值，大于 0 为乐观情感，且数越大乐观度越明显；小于 0 为低落情感，且数越小低落度越明显。对于爬取得到的陶瓷商品评论数据进行整体评论分值整合，得到最终的评论陶瓷商品的情感分值。

其做法是在数据预处理基础上对已分词完毕的评论与已有的人工标注的领域情感词典、否定词和程度副词词典进行模式匹配，查找对应的正面情感词汇和负向情感词汇，将相应的分词进行赋值，然后将整条评论的分词对应的分值进行数值相加，依据程序内部计算结果得出该条评论的最终情感的分值，最后对于爬取得到的青花瓷餐具盘碗碟套装的评论数据进行整体评论分值整合，得到最终的评论陶瓷商品的情感分值。

3.3.1. 领域情感词典的构建

因为 BosonNLP 的情感词典是专门针对微博、新闻、论坛等特定数据来源构建的，如果直接将其应用于本项目对青花瓷餐具盘碗碟套装的评论进行情感分析，准确率可能会大幅下降。因此，本项目的基础情感词典构建在 BsonNLP 情感词典的基础上参考陶瓷产业评论数据进行人工标注完善(如表 6)。

Table 6. Sentiment dictionary in some fields
表 6. 部分领域情感词典

词语	权值
可以	0.511318461412
不错	2.65135097006
喜欢	2.58769933437
惊艳	3.29873462578
...	...
粗糙	−1.307097667979
破损	−2.279476913059
垃圾	−2.58769933437
辣鸡	−2.58769933437

3.3.2. 否定词库构建

否定词表达负面情感，为负向含义。在我们爬取到的景德镇青花瓷餐具盘碗碟套装评论数据中，不乏有很多条评论包含否定词(如下表 7)，如果对这些否定词不加以处理，会对情感分析结果产生影响，甚至可能导致与原本表达的意思完全相反的结果。例如“发货很快，包装很到位，没有破损。碗的釉面很光滑、很厚重。”这句评论中的“破损”是一个负面情感词，而前面的“没有”表达否定含义，使评论情感倾向发生改变，变为正面情感。

同时在建立否定词库时还要考虑诸如“这些瑕疵不是不会影响日常使用”这种双重否定表肯定的情况。在这种情况下，我们可以根据否定词的个数来判断情感词的强度。如果否定词的个数是奇数，那么情感词的强度乘以−1；如果否定词的个数是偶数，那么情感词的强度乘以 1。

Table 7. Partial Negation Vocabulary
表 7. 部分否定词库

否定词库
不大、不丁点儿、不甚、不怎么聊、没怎么、不可以、怎么不、几乎不、从来不、从不、不用、不曾、不该、不必、不会、不好、不能、很少、极少、没有、不是...

3.3.3. 程度副词词典构建

本研究对景德镇青花瓷餐具盘碗碟套装评论数据所采用的程度副词词典程度值标注基于 CSDN 博客上 Petrichoryi 博主提供的方法。该方法将程度副词词典中的词语赋予不同的程度值，以表示情感加强或弱化。如极其(1.8)、超(1.6)、很(1.5)、较(1)、稍(0.7)和欠(0.5)。

3.4. 情感分析最终结果

Table 8. Partial results of SnowNLP sentiment analysis
表 8. SnowNLP 情感分析部分结果

评论数据	情感值
非常好，品牌好质量就有保障，下次还有再买。	0.946639683
色差太大，又薄	0.015436505
总体不错，就是有几个盘子底边有点粗糙。待改进。	0.146839914
……	……
感觉一般，价格较高，平均下来一把筷子 100 元	0.244507998
质量不错，包装完整，款式干净漂亮，是正品，满意好评。	0.999562519

Table 9. Partial results of BosonNLP sentiment analysis
表 9. BosonNLP 情感分析部分结果

评论数据	情感值	机器判断情感倾向
瓷器很好，没有瑕疵	4.86657	积极
做为江西人，还是喜欢国瓷	2.30638	积极
买了三套餐具，破了一大一小两个碗。一直拖着不给补发。	−0.26355	消极
...	...	...
碎一个	−1.75701	消极
质量很好，客服态度很好	4.49984	积极

通过上述两种情感分析方法的对比(表 8，表 9)，可以发现 SnowNLP 情感分析方法优点在于打分情况可以明显区分，步骤简单，耗时短；缺点在于没有正负之分，没那么直观。BosonNLP 情感分析方法的优点之一是其结果展示的分值范围较大，并且能够区分正面和负面情感，但耗时长。虽然 BosonNLP 在结果展示方面更直观，但是根据准确率来看，SnowNLP 的打分准确率优于 BosonNLP。因此，我们决定采用 SnowNLP 的情感分析结果，并综合分析考虑，得出本项目研究的天猫商城的红叶陶瓷景德镇青花瓷餐具盘碗碟套装总体评价情感是积极的，少部分消极情感评论主要是物流暴力分拣运输，发给买家的陶瓷因此发生不同程度损坏所致。

4. LDA 主题模型分析

Latent Dirichlet Allocation 是非监督机器学习技术，由 Blei [9]等提出。其被称为三层叶贝斯概率模型，包含文档层、主题层和词层 3 层结构，其利用文档中词的共现关系对词进行主题类聚，得到“文档—主题”“主题—词”的分布矩阵[10]。其文本中的词由“文本以一定概率挑选的主题是确定的，再从被挑选主题中以一定概率挑选词语”过程得到，是文本主题生成模型之一。在多、杂文本数据中运用 LDA 模型，能降低文本的维度，避免灾难[11]。

4.1. 确定 LDA 主题模型的最优主题数

一个概率模型或概率分布预测样本的好坏程度由 perplexity (困惑度)度量，最优模型主题数由 LDA 用 perplexity 来评估主题数的模型性能选择。低困惑度的概率分布模型能更好地预测样本[12]。运用计算困惑度的公式计算出并得到本次研究红叶陶瓷景德镇青花瓷餐具盘碗碟套装产品评论的最优的主题数量：

$$perplexity\left(D_{test}\right)=\exp\left\{\frac{-\sum_{d=1}^{M}\log\left(p\left(w_d\right)\right)}{\sum_{d=1}^{M}N_d}\right\}$$

该式用来测试语料库的大小为 M，Nd 表示第 d 篇的单词个数。

$$\sum_{z} p(z)p(w|z,gamma)$$

这个式子中主题是用 z 字母表示，文档用 w 字母表示，训练集学出的文本用 Gamma 单词表示，即代表主题分布。其概率取值范围为[0, 1]，按照对数函数的定义，分子值是一个大数，而分母是整个测试集的单词数目。当每次生成不同模型时，其能力越强，困惑度值越小，表明此次模型对所要研究的数据样本的预测能力更好。由此观之本次研究可以采取具有最低困惑度值的主题数(称为“折肘”)用来确定本次红叶陶瓷景德镇青花瓷餐具盘碗碟套装产品评论中的最优方案。利用困惑度生成的图如下图 5 所示。

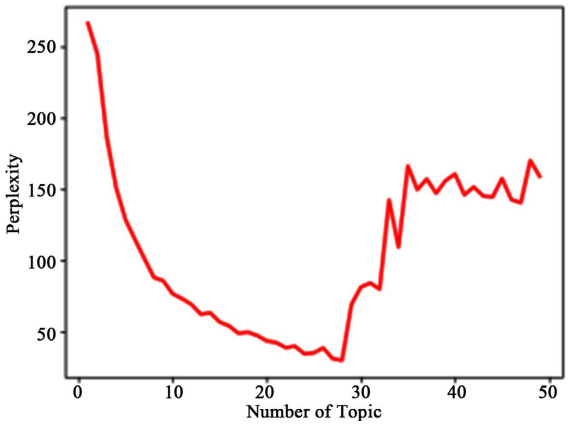

Figure 5. Perplexity—Hongye ceramics product review data topic distribution chart
图 5. Perplexity——红叶陶瓷产品评论数据主题分布图

据图观察得知横轴的值越大，困惑度的值越来越小。当横轴的值约为 28 时，存在一个显著拐点；大于 28 时，困惑度以小幅上升为正常。据其特点而言，最佳主题数值在拐点处取得。由此观之此研究选定 28 为该红叶陶瓷景德镇青花瓷餐具盘碗碟套装产品评论下 LDA 主题模型最优主题数。

4.2. 进行主题分析

根据主题的数量，分析评论的 lda 主题。评论后的数据为该评论在该主题下的概率，主题下所有评论概率和为 1，主题的中文需要人工分析。由于前面已确定最优主题数为 28，设置每个主题下展现前 20 条评论。此时 LDA 主题模型运行结果部分展示如下表 10。

Table 10. LDA analysis probability of some Hongye ceramic product reviews under the optimal number of topics
表 10. 最优主题数下部分红叶陶瓷产品评论 LDA 分析概率

主题(28 个)	评论数据(每个主题生成前 20 条)	该主题下此评论概率
Topic 0th: (品质)	我说实话：这套碗比较厚，品质一般，质量不怎么样，特别是汤碗很小气，饭碗拿到手里没有品像。根本不值这个价钱！望其他要买的亲看清才下手哟！	0.0742101396032329
	...	...
	很喜欢的一款餐具，很喜欢的一个牌子，放心，安心	0.0007347538574577517
	外观材质：有瑕疵	0.06913073237508556
	...	...
Topic 1th: (质量)	第一次寄套装包装有个薄泡沫，品锅破了。补发一只又破，包装连泡沫也省了。再次补发又要 3 到 7 天内发货，坑人的节奏，真怀疑是不是补发一个破锅来搪塞的，大家评评	0.0006844626967830253

Continued

...	...	...
	非常好，中国景德镇瓷厂，国营企业，值得信赖!	0.07421013960323292
Topic 27th: (品牌)	...	...
	做为江西人，还是喜欢国瓷	0.0007347538574577517

结果中只会返回人为规定的主题数下的固定条评论，每个评论后面的小数可是这条评论在这个主题中出现的概率，概率越大，则该评论在此主题下出现的次数越多，主题下所有评论的概率和为 1，而这个主题表示什么需要人工分析。

对天猫商城的景德镇红叶陶瓷青花瓷餐具盘碗碟套装商品评论进行情感分析，可以了解到消费者的情感倾向。根据上文实验得出的结果，SnowNLP 分析结果准确，可借助 Python 中的 SnowNLP 情感分析器对评论进行情感值计算，输出结果范围为[0, 1]，该结果分数越高代表情感倾向越积极，反之则越消极。天猫商城的景德镇红叶陶瓷青花瓷餐具盘碗碟套装商品评论情感分析的结果如上表 10 所示，整体情感均值偏高，表明消费者对于天猫商城的景德镇红叶陶瓷青花瓷餐具盘碗碟套装商品情感态度整体较为积极，多数消费者对天猫商城的景德镇红叶陶瓷青花瓷餐具盘碗碟套装商品的满意度较高，认为该商品可以达到自己的预期。由上表 10 可知，其中情感分析大于 0 接近 0.07 的极端积极评论较多，得分小于 0.03 的极端消极评论次多，而处于中立和两个极端情绪之间的相对评论较少，说明消费者对于天猫商城的景德镇红叶陶瓷青花瓷餐具盘碗碟套装商品的情感有明显倾向。

4.3. 主题可视化交互分析

人工的直接分析效率不高，不全面。使用 LDA 可视化交互分析工具——pyLDAvis 分析各个主题之间的联系。在最优主题数下，本次研究情况结果如下表 11。

Table 11. Partial word frequency
表 11. 部分词频

0：		3：		7：		13：	
“包装”	(0.043)	“超赞”	(0.115)	“功能”	(0.027)	“不错”	(0.205)
“红叶”	(0.018)	“完美”	(0.029)	“材质”	(0.027)	“品质”	(0.036)
“不好”	(0.012)	“不错”	(0.012)	“外观”	(0.023)	“补发”	(0.016)

其中括号里面的数值仍可认为是每个单词属于这个主题的概率，主题下所有评论的概率和为 1。分词的可视化分析结果如下图 6。

图 6 有许多圆圈，表示相异的主题，数字为主题编号，多少个即表明主题个数；右边的柱状图表示每个主题下的内容；我们如果将鼠标缓慢或者快速移入，亦或是点击每个圆圈，右边都会对应显示出标有红色标识的主题内容；移入或点击则可查看每个单词或者评论在左边圆圈所代表的的主题中哪里会出现；右边的柱状图则显示的是本次研究所爬取红叶陶瓷景德镇青花瓷餐具盘碗碟套装产品的评论数据中出现次数的最多的词。根据本研究表明，“不错”“好评”及“好”等表述积极情感的词语占比较多；除此之外，还有消费者关心的“质量”“包装”以及“外观”方面也占比较高，移入相应的主题下我们就可以看到对应的某方面下的情感，也是偏向积极的；但也由少量消极情感，比如该产品偶尔会出现工艺瑕疵，包装破损等现象。

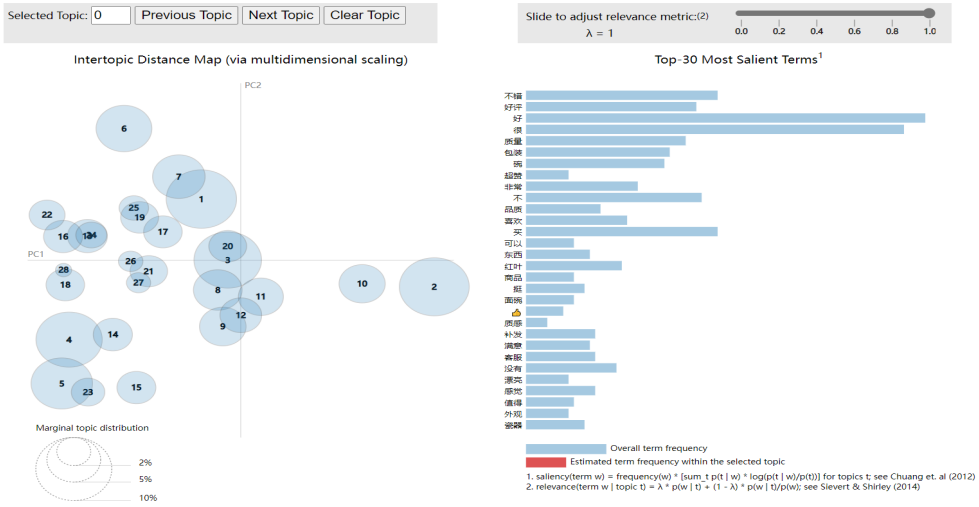

Figure 6. Visual analysis chart under the LDA theme
图 6. LDA 主题下可视化分析图

5. 结语

在当下而言，世界越来越技术化了，生活中许多地方所运用的技术已经离不开情感分析，其也在人类的使用下不断完善。本次研究在现阶段通过对比实验结果，发现 SnowNLP 打分准确率优于 BosonNLP，为了改进 BosonNLP 情感分析准确度，我们提出以下改进思路：

1) 需要提高情感词典的准确度。添加并训练有关陶瓷产品领域词汇。

2) 算法还有待优化，语气词的权重也得增大，以此来提高准确度。

3) 仅用构建的词典库来判断情感极性是不够的，还需要留意上下文的承接和联系。

同时，用于情感分析的 LDA 模型也不是非常精准，但其确是一个让用户直观明了的了解商品评论情感倾向的模型工具，所以本研究后期对于 LDA 模型还要进行更深入的使用，比如对 LDA 模型的主题数量的确定还需要进行完善等，让商品评论信息更加准确。

以上改进思路也将是本文后续工作中的重点。

基金项目

景德镇市科学技术项目资助(编号：20212GYZD009-05)；景德镇陶瓷大学大学生创新训练项目资助(编号：202110408024)。

参考文献

[1] 武聪. 基于多模态学习的消费者情感与购买意愿分析方法研究[D]: [博士学位论文]. 大连: 东北财经大学, 2022. https://doi.org/10.27006/d.cnki.gdbcu.2022.001213

[2] Zhu, F. and Zhang, X.Q. (2010) Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing*, **74**, 133-148. https://doi.org/10.1509/jm.74.2.133

[3] Hu, M.Q. and Liu, B. (2004) Mining and Summarizing Customer Reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, 22-25 August 2004, 168-177.

[4] 李琴, 李少波, 王安虹, 等.基于在线评论数据的景区门票浮动制测评分析方法[J]. 科学技术与工程, 2018, 18(1): 273-279.

[5] 凌万云, 方升, 张晓如. 基于用户在线评论的情感分析及景点优选排序[J]. 计算机与数字工程, 2022, 50(6): 1312-1316.

[6] Leonard, G., Anggreainy, M.S., Wihan, L., Santy, Lesmana, G.Y. and Yusuf, S. (2023) Support Vector Machine Based

Emotional Analysis of Restaurant Reviews. *Procedia Computer Science*, **216**, 479-484. https://doi.org/10.1016/j.procs.2022.12.160

[7] 白杨. 大数据环境下的文本挖掘教学内容探讨[J]. 无线互联科技, 2018, 15(9): 86-87.

[8] 魏丽. 基于 LDA 主题模型的电商评论数据分析[J]. 信阳农林学院学报, 2023, 33(3): 112-116. https://doi.org/10.16593/j.cnki.41-1433/s.2023.03.023

[9] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[10] 谭春辉, 陈晓琪, 梁远亮, 等. 隐私泄露事件中社交媒体围观者情感分析[J]. 情报科学, 2023, 41(3): 8-18.

[11] Dhillon, I.S. and Modha, D.S. (2001) Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning*, **42**, 143-175. https://doi.org/10.1023/A:1007612920971

[12] 关鹏, 王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术, 2016(9): 42-50.