**Statistics and Application** 统计学与应用, 2024, 13(3), 758-765
Published Online June 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.133077

Hans汉斯

# 基于文本分析的预制菜市场消费倾向研究

**章文豪**

浙江财经大学数据科学学院，浙江 杭州

收稿日期：2024年5月30日；录用日期：2024年6月19日；发布日期：2024年6月28日

## 摘　要

**本文旨在研究预制菜市场中消费者的消费现状，并聚焦预制菜消费者的行为，总结消费者关注重点。基于此，本文通过Python爬虫获取重点预制菜企业商品的评论数据，运用语义网络分析与LDA-coherence主题评价模型进行了文本挖掘，提取出多个关键词作为调查维度，以期了解对影响消费者购买预制菜意愿的因素，并作出如下总结：消费者评论关键词可以分为5个维度：产品质量、产品价格、口感味道、简单快捷、物流运输。73.62%的评论是非消极的情绪，表明大多数消费者对预制菜的现状和发展持乐观和期待态度。**

## 关键词

**预制菜，语义网络，LDA-Coherence主题模型，情感分析**

# Research on Consumption Tendency of Prefabricated Vegetable Market Based on Text Analysis

**Wenhao Zhang**

School of Data Science, Zhejiang University of Finance & Economics, Hangzhou Zhejiang

Received: May 30th, 2024; accepted: Jun. 19th, 2024; published: Jun. 28th, 2024

## Abstract

**The purpose of this paper is to study the current consumption status of consumers in the prefabricated food market, focus on the behavior of prefabricated food consumers, and summarize the key points of consumer concern. Based on this, this paper obtains the review data of key prefabricated food enterprises through a Python crawler, uses semantic network analysis and LDA-coherence topic evaluation model to carry out text mining, and extracts multiple keywords as**

文章引用: 章文豪. 基于文本分析的预制菜市场消费倾向研究[J]. 统计学与应用, 2024, 13(3): 758-765.
DOI: 10.12677/sa.2024.133077

**survey dimensions, in order to understand the factors affecting consumers' willingness to buy prefabricated dishes, and makes the following summary: consumer review keywords can be divided into five dimensions: product quality, product price, taste and taste, simple and fast, and logistics and transportation. 73.62% of the reviews were non-negative, indicating that most consumers are optimistic and expectant about the current situation and development of pre-made dishes.**

## Keywords

**Pre-Made Dishes, Semantic Web, LDA-Coherence Topic Model, Sentiment Analysis**



## 1. 引言

### 1.1. 研究背景

预制菜自 20 世纪 90 年代起在中国市场逐渐普及，最先传入上海，后发展到与上海经济水平趋同的江浙地区，近年来得到越来越多消费者认可。2022 年，中央一号文件将培育发展预制菜产业列入规划，各地也在积极推动预制菜产业高质量发展[1]。如今，小小的预制菜串联起田间地头和居民餐桌，既关联着经济增长，也关联着人民群众对美好生活的向往。

国家层面：预制菜产业可以促进产业链和结构升级，振兴乡村经济。该行业链接农业和餐饮消费，提升了农产品的附加值和销售渠道，为农业和经济发展带来了机遇。中央一号文件首次提及预制菜的发展，这标志着发展预制菜被提升为乡村振兴的重要战略举措，预制菜行业将迎来全新的发展机遇。

企业层面：使用预制菜能提高出餐速度，实现口味可控。餐饮成本主要包括原材料、人工、租金和能源成本。人力成本是最大的支出之一。提高人效和坪效是餐饮企业主要追求的目标。

消费者层面：预制菜可以为消费者提供方便和快捷的食品选择，特别是在忙碌的现代生活中。预制菜可以减少烹饪的时间和复杂度，同时也提供了各种营养均衡的选择。此外，预制菜还能够保证食品质量和安全，并减少食品浪费。[2]

### 1.2. 研究方法

文本分析的研究包括自然语言处理、语言学、本体学、数据挖掘、机器学习、概率论和统计分析在内的多个学科，是一格综合性的研究领域。刘娜娜、张强[3]通过利用 ROSTCM6 软件进行情感分析后用 SAS 变量聚集法建立消费者需求模型。赵杨等[4]基于特定 APP 用户数据进行情感分析并加入了 CNN-SVM 深度学习模型，得到多维度的用户信息。崔连超[5]改进了情感词的分类算法并提高对中文语料库的查询准确率。

本文的研究工作包括：对京东 APP 的预制菜商品评论进行爬取和预处理，运用 Python 的 jieba 包与常用的停用词表进行分词，进行词频统计，利用 ROSTCM6 软件的 NetDraw 工具绘制语义网络图，进行对象属性的可视化分析，在通过 SnowNLP 模型进行情感分析，最后通过 LDA 主题模型总结归纳。

## 2. 文本挖掘

### 2.1. 数据采集

通过 Python 的 requests 包，抓取京东平台部分热销预制菜品牌商品评论。共抓取数据好评 15088 条，中差评合计 3526 条，合并后共计 18614 条。由于电商平台存在商家雇佣水军采取作弊行为刷高销量和用户因网络问题重复评论等现象，进行原始数据的清洗尤为重要。报告采用文本比较方法，删除插入的图片和 HTML 超链接；删除用户 ID、发表时间等无用文本；评论内容完全为英文字母、数字和标点符号的，视为随意发表的评论，予以删除；大量重复出现，视为无意义的评论，予以删除。经过预处理后，统计得到有效评论共 17300 条，其中好评 13964 条，差评 3336 条，文本有效率为 92.94%。

在数据抓取过程中，获取的数据是可检索的结构化数据 json 数据，使用正则化匹配出需要的数据，用表格形式储存，如表 1 所示：

**Table 1.** Part of the information crawled by a web crawler
**表 1.** 网络爬虫爬取的部分信息

| 用户 ID | 商品类别 | 时间 | 评论内容 |
|---|---|---|---|
| 188****3650 | GUO LIAN 小霸龙 蒜香烤鱼 | 2023-02-22 | 吃了一盒，感觉不错，京东送货很快，价格便宜，还会继续买吃起来很方便，还可以当火锅吃。 |
| 186****6318 | 美好 农家小酥肉 | 2023-01-08 | 小酥肉以前在我们家都是用烤箱烤一下在吃今年冬天把它放在大白菜里面炖着吃，味道很不错。 |
| 186****3233 | 珍味小梅园 红烧狮子头 | 2023-01-14 | 过年就靠这些菜做一桌年夜饭啦，家里老人比较多，还是不要去外面比较保险，珍味小梅园的品质还是很相信的，京东物流送货也很快。 |

### 2.2. 词频统计

通过 Python 的 jieba 包对其进行分词，分词后，在通过 CSDN 下载停用词表将无意义的词载入并过滤，完成最后的文本数据清洗，之后将高频词进行统计，得到分析结果如图 1 所示

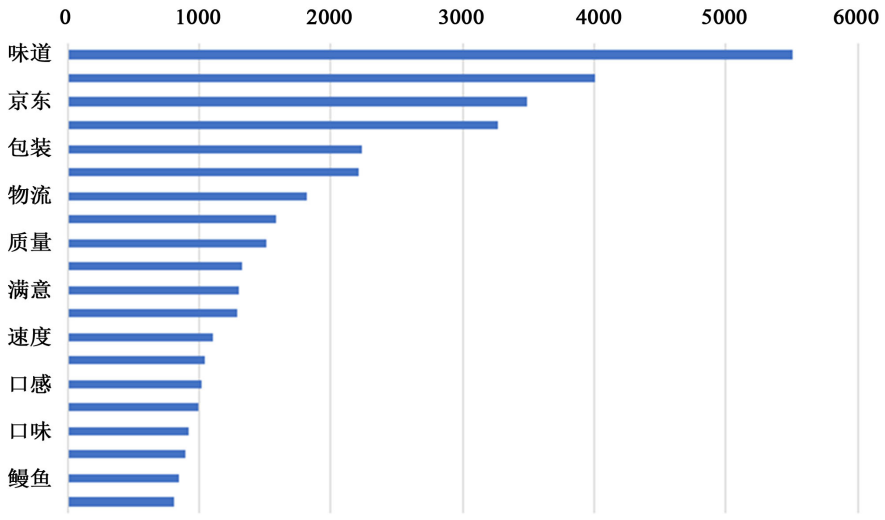


**Figure 1.** Statistical chart of word frequency of pre-made dish evaluation
**图 1.** 预制菜评价词频统计图

## 3. 语义网络分析

语义网络是一种社会网络分析的方法，由大量的常识与概念构成。由网络节点和有向线段组成完整的语义网络关系图，在图中，概念之间的从属关系用箭头的方向来表示。本文使用 ROSTCM 软件，该软件常用于文本分析领域，借助 ROSTCM 软件中的 NetDraw 工具可以绘制语义网络图，利用语义网络图实现评价对象间属性的可视化分析。

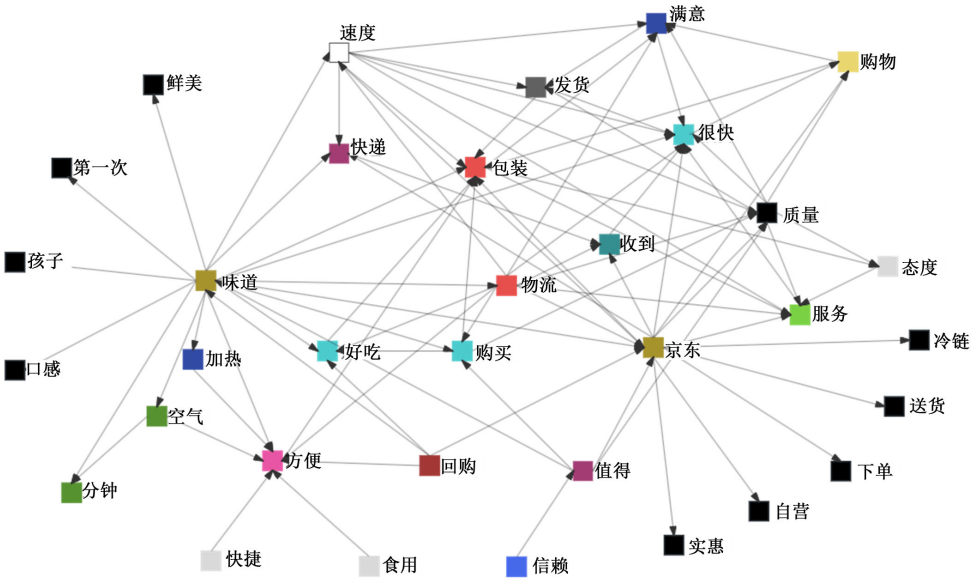


**Figure 2.** Semantic network diagrams
**图 2.** 语义网络图

通过图 2 的网络评论的语义网络分析结果，我们发现：质量、包装、物流、方便、满意是受多方面影响的节点，也是评论中的高频词，是消费者评论时总会指向的几个方面，所以预制菜企业要着重改进产品质量，强化物流网络建设，配套更方便的产品使用设备，通过最核心的节点提升顾客满意度。

## 4. 情感分析

### 4.1. 情感强度计算

SnowNLP 中的情感分类基本模型是贝叶斯(Bayes)模型，对于有两个类别的分类问题来说，其特征为，特征之间是相互独立的，属于类别的贝叶斯模型的基本过程为：

$$P\left(c_1 \middle| w_1,\cdots,w_n\right)=\frac{P\left(w_1,\cdots,w_n \middle| c_1\right)\cdot P\left(c_1\right)}{P\left(w_1,\cdots,w_n\right)} \tag{1}$$

其中：

$$P\left(w_1,\cdots,w_n\right)=P\left(w_1,\cdots,w_n \middle| c_1\right)\cdot P\left(c_1\right)+P\left(w_1,\cdots,w_n \middle| c_2\right)P\left(c_2\right) \tag{2}$$

贝叶斯模型预测的过程则用到了上述公式(1)，对其简化有：

$$P\left(c_1 \middle| w_1,\cdots,w_n\right)=\frac{1}{1+\exp\left[\log\left(P\left(w_1,\ldots,w_n \middle| c_2\right)P\left(c_2\right)\right)-\log\left(P\left(w_1,\ldots,w_n \middle| c_1\right)\cdot P\left(c_1\right)\right)\right]} \tag{3}$$

其中，分母中的 1 可以改写为：

$$1=\exp\left[\log\left(P\left(w_1,\cdots,w_n\left|c_1\right.\right)\cdot P\left(c_1\right)\right)-\log\left(P\left(w_1,\cdots,w_n\left|c_1\right.\right)P\left(c_1\right)\right)\right] \tag{4}$$

通过标记好的评论正负样本训练模型并对其余评论预测得分，结果如下：

**Table 2.** Sentiment score for some product reviews
**表 2.** 部分商品评论情感得分

| 评论 | 情感得分 |
|---|---|
| 这个佛跳墙春节就买过，真的感觉很不错，吃起来口感超赞，最主要的是里面的料太丰盛了，这个价位买的真的很值了，以前也吃过其他品牌的真心觉得这个更胜一筹，还会回购。 | 0.999977317 |
| 京东快递一如既往送货神速，晚上下单隔天上午送进家门。家人一直想尝试下佛跳墙，网上平台铺天盖地不知真假，最终还是相信京东自营的，年夜饭吃掉了，味道还不错，就是有点腻。 | 0.782334396 |
| 已经做了一餐，味道很不错料也很实在，应该多备点了，就是冰箱已经被塞满，这一袋很大，能吃两餐了，一次半袋，家里三四口人正好。 | 0.704372012 |
| 第一次买正大的羊蝎子太让人失望了根本不是羊蝎子都是些边角料一块羊蝎子都没有而且每块都特别肥，加了蔬菜结果搞的菜都特腻了，不知道是不是因为一批质量不行所以搞活动清仓抢了三份本来还很开心现在光发愁怎么消耗完，砸牌子。 | 0.472968147 |
| 这个猪肚鸡，用的全是琵琶腿小的那一头，皮包骨，根本没图片中拍的这种大块肉。第一包打开吃出来一堆骨头端也是无语了。还有买了包 200 多，隔天评价时再看，才 49 一包吐血。 | 0.455101312 |
| 产品难吃，吃完后还拉肚子。根本就不是猪肚鸡的味道，份量小用材差。 | 0.312619665 |

## 4.2. 情感强度分布

通过使用 Python 的 snownlp 程序包进行情感分析，最终会获得对一个评价的得分 P，情感分数置于[0, 1]之间，得分在[0, 0.3]的归为消极情绪，[0.3, 0.7]的归为中性情绪，[0.7, 1]的归为积极情绪，得到结果如图 2~4 所示。

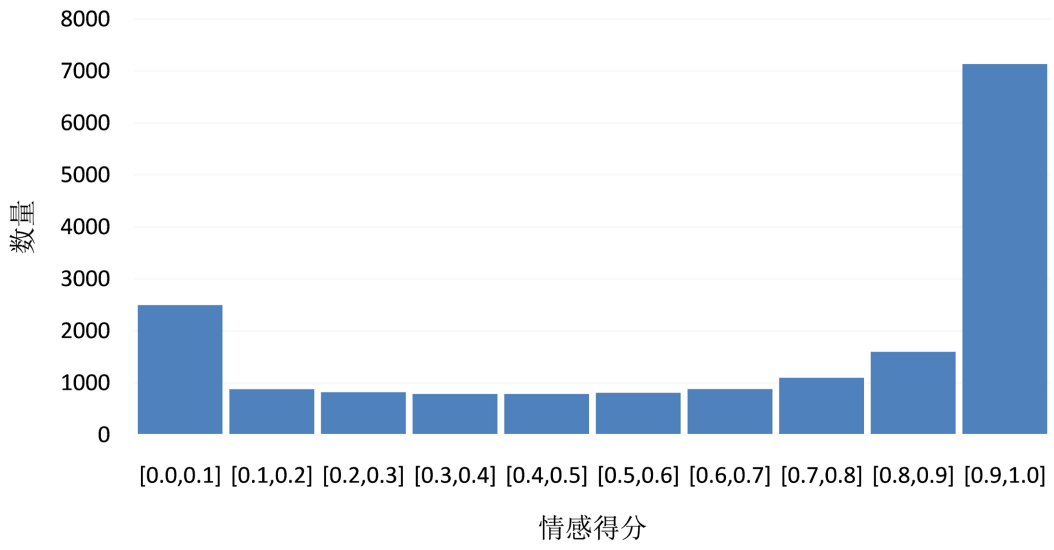


**Figure 3.** Sentiment score graph
**图 3.** 情感得分分布图

从图 3 中可以看出，情感得分大多分布在 0.8 到 1.0 之间，统计得到消极情绪评论共 4613 条，积极

情绪评论 12874 条，具体的积极、中性与消极情绪各部分占比如图 2~5 所示：

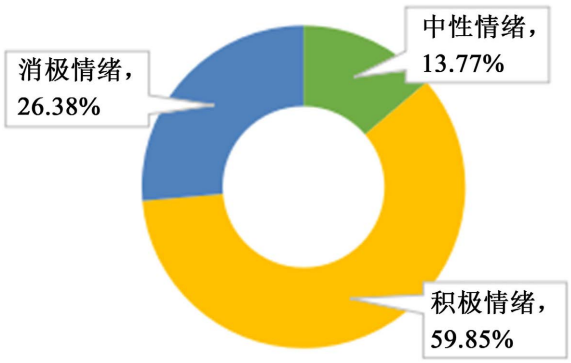


**Figure 4.** Proportion of sentiment comments
**图 4.** 情绪评论占比图

从图 4 中可以看出，有 73.62%的评论是非消极的情绪，表明大多数消费者对预制菜的现状和发展持乐观和期待态度，同时消极评论所反应出的问题也能为预制菜产品的改进提供有效参考。

## 5. LDA 模型分析

为了了解预制菜消费者主要倾向于关注产品相关的哪些领域，使用 LDA 主题模型对评论文本进行主题的聚类。LDA 主题模型不关心文档中单词的顺序，通常使用词袋特征(bag-of-word feature)来代表文档。对分完词并去除停用词的评论进行主题分类，一般用来评价 LDA 主题模型的指标有困惑度(perplexity)和主题一致性(coherence)，困惑度越低或者一致性越高说明模型越好。一些研究表明 perplexity 并不是一个好的指标，所以我们选用 coherence 来评价模型并选择最优主题，得到的主题-coherence 变化情况如图所示：

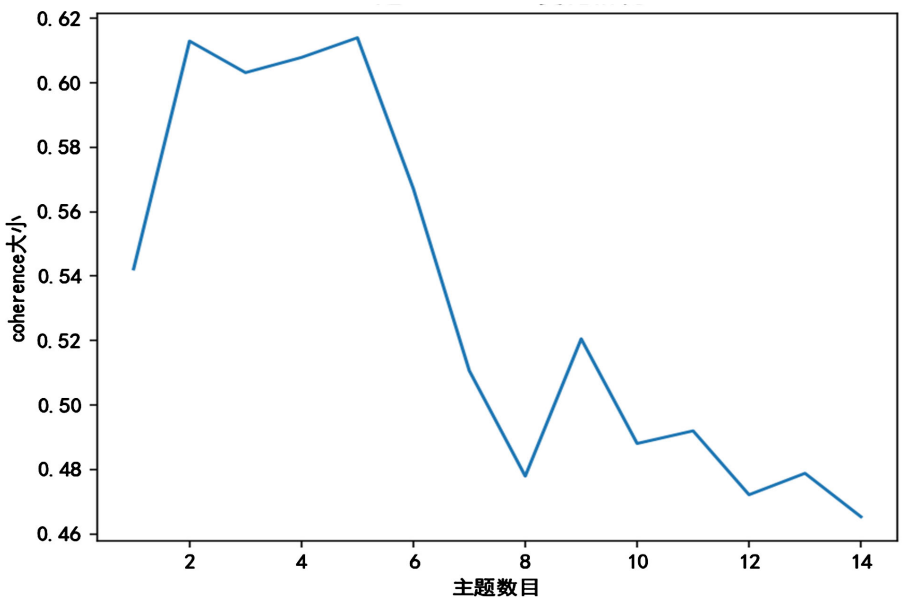


**Figure 5.** Topic-coherence plot with number of topics
**图 5.** 主题-coherence 随主题数目变化图

由图 5 可得主题数为 2 或 5 时分类效果最好，主题数取 2 时并不能很好的进行分类，所以接下来设定主题数为 5，并输出每个文档最有可能对应的主题，同时用 pyLDAvis 对 LDA 模型结果进行可视化，得到如下结果。

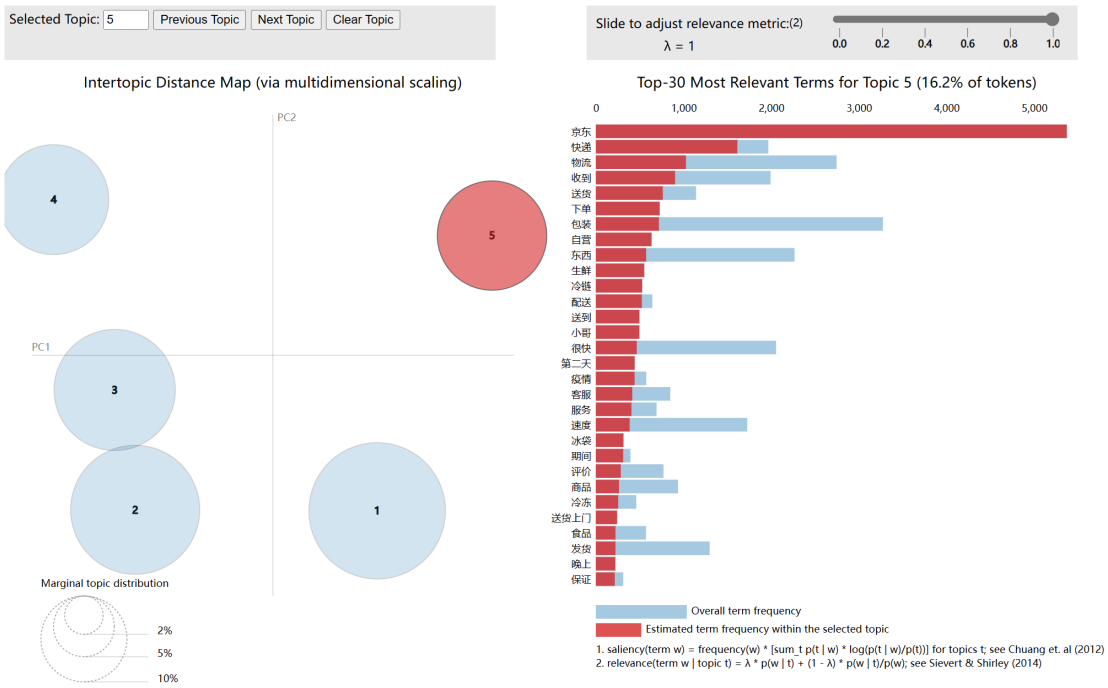


**Figure 6.** LDA Model Topic Classification Bubble Chart
**图 6.** LDA 模型主题分类气泡图

图 6 中左侧每个不同的气泡表示一个不同的主题，右侧是主题内前 30 个特征词表示各个词语对主题的贡献度，浅蓝色的表示这个词在整个文档中出现的频率(权重)，深红色的表示这个词在这个主题中所占的权重。五个气泡分布均匀，重叠区域极少且只存在于 2、3 主题之间，模型较好。将五大主题的高频率关键词进行整理如表 2 所示：

**Table 3.** Topic keyword distribution
**表 3.** 主题关键词分布

| 产品质量 | 产品价格 | 口感味道 | 简单快捷 | 物流运输 |
|---|---|---|---|---|
| 不错 | 活动 | 味道 | 炸锅 | 京东 |
| 质量 | 价格 | 喜欢 | 空气 | 快递 |
| 包装 | 便宜 | 好吃 | 分钟 | 物流 |
| 满意 | 性价比 | 新鲜 | 油炸 | 收到 |
| 品牌 | 划算 | 正宗 | 火锅 | 送货 |

通过表 3 可得，主题 1 的关键词出现频率较多的有包装、质量、品牌等，所以本文将 topic1 归为产品质量。主题 2 的高频率关键词为活动、价格、性价比等，将 topic2 归为产品价格。主题 3 的高频率关键词为味道、新鲜、好吃和正宗等，将 topic3 归为口感味道。主题 4 的高频率关键词为炸锅、分钟、油炸等，价格，topic4 归为简单快捷。主题 5 的高频率关键词为快递、物流、送货等，将 topic5 归为物流运输。

## 参考文献

[1] 张晔, 张运, 周杰. 让预制菜产业健康可持续发展[N]. 科技日报, 2024-04-19(005).

[2] 陈嘉. 预制菜搅动大众餐桌[N]. 张家口日报, 2024-04-10(004).

[3] 刘娜娜, 张强. 基于电商平台的消费者需求及产品数据挖掘技术分析[J]. 内蒙古统计, 2019(1): 38-41.

[4] 赵杨, 李齐齐, 陈雨涵, 等. 基于在线评论情感分析的海淘 APP 用户满意度研究[J]. 数据分析与知识发现, 2018, 2(11): 19-27.

[5] 崔连超. 互联网评论文本情感分析研究[D]: [硕士学位论文]. 济南: 山东大学, 2016.