

医药电商行业消费者满意度关键影响因素识别

——基于抗感冒药品在线评论的研究

胡晓铮,董景峰,陶新民

(东北林业大学 土木与交通学院,哈尔滨 150040)

摘要:本文以消费者满意度为切入点,聚焦医药电商平台服务质量优化路径,为精准提升服务质量提供数据支撑.基于60396条抗感冒药品有效评论数据,构建文本挖掘、主题特征提取、影响因素排序三维分析框架.运用LDA主题模型确定最佳主题数量,并挖掘评论文本潜在主题,识别出物流配送服务、药品价格等7个核心影响因素.进而以评论的文档-主题概率分布向量为输入特征,构建XGBoost分类模型,通过与逻辑回归、随机森林、支持向量机、CatBoost等模型在准确率等5类性能指标上的对比,证实该模型对数据具有显著适配性,经参数优化后,模型性能达到最优水平,并输出特征的重要性得分.研究表明不同剂型药品的消费者满意度影响因素存在显著差异,从而为医药电商消费者行为研究提供有效的理论支撑.

关键词:消费者满意度;在线评论;LDA主题模型;XGBoost;影响因素

中图分类号:TP181

文献标志码:A

文章编号:1000-2367(2026)04-0083-08

随着互联网技术迭代与网络购物渗透率提升,医药电商已成为医疗健康领域的重要组成部分.与普通电商不同,医药电商因涉及药品安全性、合规性等特殊要求,消费者在线评价不仅关乎购物体验,更直接关联公众健康权益.在此背景下,消费者满意度作为企业竞争力的核心指标,其影响机制因医药电商的特殊性而趋于复杂,单一评论分析方法已难以揭示多因素交叉作用下的内在驱动逻辑.

消费者满意度是指消费者消费后的感知绩效与购买前形成的期望进行比较后形成的满意或失望的心理感受^[1-2].当前关于消费者满意度的基础研究主要集中在以下方面:一是与其他理论间的联系,如期望不一致理论、绩效模型、公平模型;二是评估与测量,通过多种方法衡量消费者满意度^[3];三是与消费者行为的关系^[4-5];四是驱动机制及影响因素分析,即哪些因素会影响消费者满意度^[6-7].

在消费者满意度影响因素研究中,国内外学者已围绕生鲜、汽车、酒店、服装等行业展开广泛探索,但针对医药电商领域的研究仍相对有限.在数字经济深度渗透的驱动下,医药消费场景加速向线上迁移,近年来医药领域消费者满意度相关影响因素的研究热度持续上升.熊百妹^[8]利用逐步回归法建立药品消费者满意度测评的回归模型,指出药品种类、药店位置和药品质量是影响药品消费者满意度的主要因素.Zhao等^[9]则聚焦线上场景,利用多属性决策方法分析抗感冒药的在线评论,进而识别影响消费者满意度的关键因素.总体

收稿日期:2025-08-04;**修回日期:**2025-08-25.

基金项目:国家自然科学基金(62176050).

作者简介:胡晓铮(2000—),女,河北邢台人,东北林业大学硕士研究生,研究方向为基于机器学习的物流与消费市场交叉研究,E-mail:xiaozheng@nefu.edu.cn.

通信作者:董景峰(1974—),男,黑龙江虎林人,东北林业大学讲师,博士,研究方向为可持续供应链优化和竞争策略以及风险评估研究,基于机器学习的智慧交通优化与物流管理研究,E-mail:dongjingfeng74@126.com.

引用本文:胡晓铮,董景峰,陶新民.医药电商行业消费者满意度关键影响因素识别:基于抗感冒药品在线评论的研究[J].河南师范大学学报(自然科学版),2026,54(4):83-90.(Hu Xiaozheng,Dong Jingfeng,Tao Xinmin. Identifying factors influencing consumer satisfaction in the pharmaceutical e-commerce sector: A study based on online reviews of Anti-Cold Drugs[J]. Journal of Henan Normal University(Natural Science Edition), 2026, 54(4): 83-90. DOI:10.16366/j.cnki.1000-2367.2025.08.04.0003.)

而言,现阶段医药电商消费者满意度的研究成果相对较少,尤其在利用在线评论与机器学习技术探索其影响因素方面存在明显缺口。

当前,学者们对在线评论的探索已深入方法论层面:倪昭鑫等^[10]构建 CNN-BiLSTM-Attention 模型对在线评论进行情感分析,并针对分类后的正负面评价进行 LDA 主题建模,挖掘影响生鲜物流服务质量评价的关键因素.郭先利等^[11]基于 LDA 主题模型挖掘多平台在线评论中的用户偏好演化规律.何勇等^[12]以生鲜电商为背景,构建融合评论影响效用的 Mixed Logit 模型,挖掘在线评论中异质用户的隐性需求及偏好.Krishna 等^[13]通过 LDA 主题模型识别影响食物共享平台用户满意度的八大挑战.上述研究证实,在线评论文本是连接用户主观评价与满意度客观因素的关键桥梁,为从评论中提取影响因素提供了方法论基础。

已有的消费者满意度影响因素研究已形成多方法体系,结构方程模型^[14]依托问卷调查数据侧重理论假设验证,情感分析^[10]聚焦极性判断,偏最小二乘法与聚合案例分析^[15]则适用于小样本因果推断.其中,LDA 主题模型凭借无监督学习优势,高效客观地从大量文本评论中提取潜在主题分布,将用户评论转化为可量化的影响因素^[9,16],这与在线评论研究的方法论需求高度契合.然而,LDA 主题模型仅能完成主题识别,缺乏对因素重要性的相对排序.尽管 PROMETHEE II^[9]等方法可通过构建评估矩阵实现排序,但其偏好函数依赖人工设定且更适用于中小规模数据.相比之下,XGBoost 模型通过梯度提升机制客观计算特征重要性得分,精准量化各主题在预测满意度时的相对贡献度,弥补 PROMETHEE II 处理大规模数据的不足,为消费者满意度影响因素的深度挖掘提供新思路。

为突破现有方法的局限性,本文提出一种融合无监督学习与监督学习的混合框架:利用 LDA 主题模型从海量评论中挖掘潜在影响因素主题,确定影响消费者满意度的因素;进而通过 XGBoost 算法量化各主题(因素)对消费者满意度的相对重要性,从而兼顾了文本语义丰富性与特征排序客观性。

1 基于在线评论的 LDA-XGBoost 框架

本文提出的 LDA-XGBoost 框架流程如图 1 所示.应用 Python 从京东在线平台爬取 60 396 条有效评论,构建抗感冒药品文本评论数据集.首先,采用 NLP 技术对评论数据进行清洗预处理;其次,运用 LDA 主题模型提取潜在主题,实现对消费者满意度的关键影响因素的识别和结构化表征,并将主题转化为可量化的概率特征;最后,遍历每条评论并提取其文档-主题概率分布向量构成特征集,以满意度标签为目标变量,利用 XGBoost 算法进行满意度标签分类,通过特征重要性分析确定各影响因素的相对重要性排序.同时与逻辑回归(LR)、随机森林(RF)、支持向量机(SVM)、CatBoost 等模型在评论分类任务上的性能进行比较,以验证模型的准确性和可靠性。

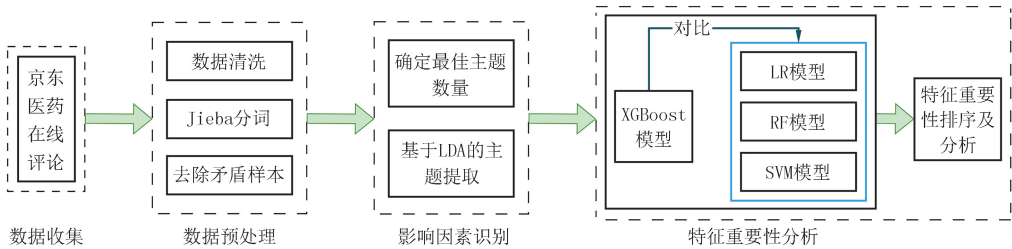


图1 框架流程图

Fig.1 Framework flowchart

1.1 基于 TF-IDF 方法的 LDA 主题提取

隐含狄利克雷分布(LDA)主题模型是一种基于贝叶斯框架的生成式概率图模型,作为无监督主题挖掘方法,它通过“词-主题-文档”三层贝叶斯结构,从非结构化文本数据中提取潜在主题,揭示隐含主题空间与观测文本集合的映射关系,其生成机制如图 2 所示。

其中 M' 表示文档数量, K' 表示主题数量, N 表示文档中的单词数量, z 表示文档中单词的主题, w 表示文档中的单词, θ 表示文档-主题的多项式分布, α 表示文档-主题分布的 Dirichlet 参数, ψ' 表示主题-单词分布, β 表示主题-单词分布的 Dirichlet 参数值. 该模型核心表达式为:

$$P(\text{单词} \mid \text{文档}) = P(\text{单词} \mid \text{主题})P(\text{主题} \mid \text{文档}), \tag{1}$$

式(1)中 $P(\text{单词} \mid \text{文档})$ 表示给定文档条件下某单词出现的概率; $P(\text{单词} \mid \text{主题})$ 表示给定主题条件下某单词出现的概率; $P(\text{主题} \mid \text{文档})$ 表示给定文档条件下某主题出现的概率.在医药电商消费者满意度影响因素的分析中,LDA 主题模型可解构海量评论的语义特征,识别影响满意度的潜在影响因素.

在进行主题提取前,需要对非结构化的文本数据进行向量化处理,文本向量化的方法有很多,其中 TF-IDF 算法^[17]可以抑制普通高频词(如的、是等)对主题提取的影响,聚焦于有区分度的词汇,提升主题可解释性.因此,本文在采用 TF-IDF 算法提取关键词语,实现文本向量化.

与此同时,合适的主题数量是保证主题提取质量的前提,为此,LDA 主题模型常用一致性分数(coherence score, C_s)与困惑度(perplexity, P_y)两个指标进行性能评估^[18],从而确定最佳主题数量.其计算公式为:

$$C_s(D) = \lg\left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}\right), \tag{2}$$

$$P_y(D) = \ln\left(\frac{\sum_{d=1}^M \ln(p(w_d))}{\sum_{d=0}^M N_d}\right), \tag{3}$$

其中, D 代表分词词汇集合, M 表示文本数量, N_d 和 $p(w_d)$ 分别为文本 d 的词汇数目和生成概率, $P(w_i)$ 和 $P(w_j)$ 表示单个词汇 w_i 与 w_j 出现的概率, $P(w_i, w_j)$ 则表示两者同时出现的概率.困惑度用于评估模型对新文档的预测能力,困惑度值越低意味着模型对新文档的预测能力越强,但其未考虑模型的可解释性和语义一致性,单独使用易造成过拟合和领域依赖性.为此,本文采用一致性分数以评估模型性能.一致性分数值越高说明模型生成主题的一致性和稳定性越好,即主题内部词语相关性更强.综合考虑困惑度与一致性分数两个指标以确定最佳主题数量,并对其进行可视化.

1.2 极端梯度提升算法(XGBoost)

XGBoost(eXtreme gradient boosting)由 Chen 等^[19]首次提出,是一种基于梯度提升框架的集成学习算法,其通过迭代训练多个弱学习器(通常为决策树),并将训练结果加权组合,最终形成强学习器.相较于传统梯度提升方法,XGBoost 目标函数由损失函数(用来衡量真实值与预测值之间的偏差)和正则化项(用以降低模型复杂度,防止模型过拟合)构成,具体公式如下:

$$Obj^t = \sum_{i=1}^N l(y_i^{t-1}, \hat{y}_i^{t-1} + f_t(x_i)) + \sum_{k=1}^t \Omega(f_k), \tag{4}$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2, \tag{5}$$

$$Obj^t = \sum_{i=1}^N l(f_t(x_i)g_i + \frac{1}{2}(f(x_i))^2 h_i) + \Omega(f_t), \tag{6}$$

该算法中,给定数据集 $D = \{(x_i, y_i)\}$,其中 $i \in 1, 2, 3, \dots, N$, x_i 为第 i 个样本的特征向量, y_i 为第 i 个样本对应的标签.假定模型使用棵树,并预测每个样本的得分为 $f_k(x_i)$, t 表示迭代次数.式(5) 为正则化项的具体展开, T 代表树中的节点数, w_j 代表叶子节点权重, γ, λ 表示树结构、叶节点权重大小的参数.式(6) 将目标函数的损失函数进行二阶泰勒展开,其中, g_i 为第 i 个样本的一阶导数, h_i 为第 i 个样本的二阶导数.这使得该算法能够进一步计算分裂节点的增益,从而选择最优分裂点.

2 京东医药电商消费者满意度实例分析

2.1 数据来源

使用 Python 抓取 2023 年 3 月至 2025 年 3 月京东平台 6 种主流抗感冒药品的在线评论信息 87 461 条,形成数据集,其包括药品名称、购买时间、评分、评论文本等信息,选定药物是非处方药(OTC),抓取的在线

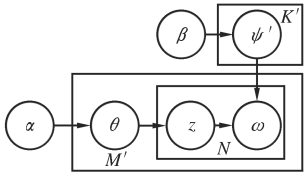


图2 LDA主题模型生成机制
Fig.2 LDA topic model generation mechanism

评论示例见附录表 S1.

抓取完成后,对评论文本进行以下处理工作:

1)对数据中用户信息进行脱敏化处理,删除评论中涉及京东平台、个人 ID、药品名称等字符,模糊购买时间等.

2)筛选并删除无效、重复、营销类文本.

3)爬虫获取的评论具有随机性,为验证消费者评分与评论内容的语义一致性、提升模型的抗干扰能力与主题提取的准确性,应用 BERT 模型对数据集进行情感分析,在大规模文本数据上进行预训练,从而学习到丰富的语义特征.按 8 : 1 : 1 的比例划分为训练集、验证集与测试集,并对训练集进行人工标注,剔除部分矛盾样本.在模型架构上,将评分信息以[评分 X]的形式拼接至评论文本前端,利用改进的 BERT-www-ext 预训练模型(基于全词遮蔽与扩展训练的多层 Transformer 编码器)提取联合语义表征,通过 CORAL 损失函数进行有序分类.

为提升抗干扰性,引入 FGSM 对抗训练与 Focal Loss 缓解类别不平衡.实验表明,模型在测试集上加权 F1 达 83.7%,MAE 为 0.51,一致性准确率为 88.1%.最终累计得到 60 396 条有效评论数据.其中,颗粒类药品评论共计 20 424 条(999 感冒灵颗粒 11 039 条,白云山板蓝根颗粒 9 385 条)、胶囊类药品评论共计 18 341 条(连花清瘟胶囊 8 851 条、芬必得布洛芬缓释胶囊 9 490 条),口服液类药品评论共计 21 631 条(抗病毒口服液 12 057 条、三精双黄连口服液 9 574 条).

4)添加自定义词典(如“快递小哥”等词语),使用 Jieba 分词进行中文分词保留有效词汇,使用百度停用词去除标点、数字及特殊符号等无关词汇,消除噪声,手动去除表情等非文本内容,并把分词结果存储在新文档以支持后续分析.

5)为简化 XGBoost 模型训练过程,根据原始用户评分构建离散目标变量,定义满意度标签,其中 1~2 分表征不满意,预设负面情感类别,3 分表征中立,预设中性情感类别,4~5 分表征满意,预设正面情感类别.

2.2 确定最佳主题数量

为精准量化评论中各因素的影响强度,应用 LDA 主题模型挖掘评论文本潜在语义结构,并提取出核心主题及其概率分布.在此之前,使用 TF-IDF 算法对语料库进行预处理,对词汇进行加权,突出重要词汇并减少噪声干扰,以提升模型训练效果和主题分析准确性.采用一致性分数与困惑度两个指标来寻找最佳主题数量,如图 3 所示.

在图 3 中,蓝色折线代表一致性分数伴随主题数量的动态变化曲线,红色则是困惑度随着主题数量的变化曲线.鉴于一致性分数与困惑度对模型的特定影响,K 的值取 6、7、8、10 时,模型性能相对更佳.这是由于 K<6 时,一致性分数较低,主题下的词语相关性较弱,而 K>10 时,困惑度相对较高且主题数量过多易造成主题重叠,因此确定最佳主题数 K 为 7.主题数量确定后,使用 LDA 主题模型对先前已经预处理好的抗感冒药文本评论数据集进行主题建模和抽取,此次建模采用 Python 中的 gensim 模块,提取潜在主题后再次更新停用词表,去除品牌、疾病症状、时间等无意义的高频词汇,得到最终主题提取结果.

考虑每个主题下概率最高的前 10 个词语及其逻辑关系,为每个主题分配了如下名称:物流配送服务(A1)、药品价格(A2)、用药感官体验(A3)、药品物流状态(A4)、在线及售后服务(A5)、药品疗效(A6)、包装质量(A7).将上述 7 个主题视为影响消费者对抗感冒药满意度的因素,如附录表 S2 所示,该表显示了每个主题下包含的前 10 个加权高频关键词及其相应的概率值.表 S2 中提到的词汇概率是指通过 LDA 模型学习的主题-词语概率分布量化计算,本质是词汇在特定主题内的相对重要性概率,具体计算公式如下:

$$P(\text{单词} \mid \text{主题}) = \frac{\text{主题 } z \text{ 中词汇 } w \text{ 出现的次数}}{\text{主题 } z \text{ 的总词汇数}}.$$
 (7)

2.3 消费者满意度影响因素排名

系统遍历每条评论,提取其文档-主题概率分布向量,形成特征变量集,以基于预设规则的满意度标签作为目标变量.考虑到用户评论数据具有随机性及不确定性,3 类满意度标签呈现明显的类别分布不平衡,为尽可能保留原始数据完整性,在各模型训练前采用 SMOTE 过采样技术进行数据平衡处理,提升模型分类准确性.

随后,按 8 : 1 : 1 的比例划分为训练集、验证集和测试集,并设置随机种子确保结果可复现,构建 XGBoost 三分类模型。为量化各特征对预测结果的贡献度,模型训练完成后,采用信息增益(Gain)准则计算特征重要性,即该特征在所有决策树的分裂节点上所带来的不纯度减少总量。特征的信息增益越大,说明其通过分裂节点对降低预测不确定性的贡献越大,特征重要性分数越高。

为进一步验证模型有效性,本研究还将 LR、RF、SVM、CatBoost 模型从准确率、精确度、F1 分数、召回率、AUC-ROC 指标与 XGBoost 模型调优前后进行模型性能对比,结果如图 4 所示。

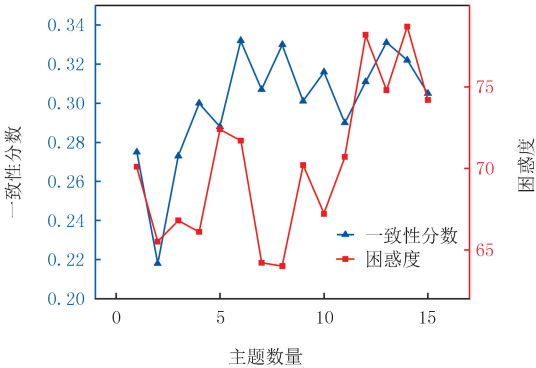


图3 不同主题数量下困惑度与一致性分数对比

Fig.3 Comparison of perplexity and coherence scores across different numbers of topics

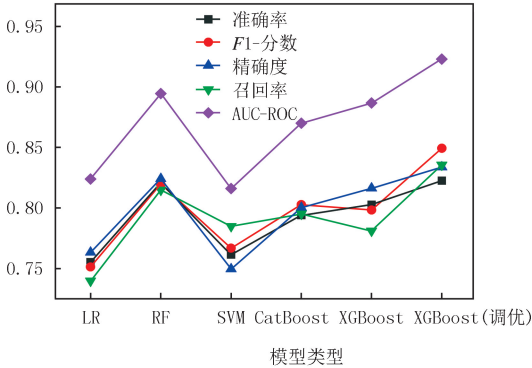


图4 不同模型性能指标比较

Fig.4 Comparative performance metrics across models

在参数调优前,默认参数设置下,XGBoost 模型与其他模型相比已展现良好性能,相比之下,XGBoost 在不同场景下的性能波动更小,整体表现更为稳定可靠。为进一步提高 XGBoost 模型性能,在参数寻优阶段,采用网格搜索与 5 折交叉验证相结合的方法^[20],定义超参数空间,寻找最优超参数(见表 1),结果显示,参数调优后的 XGBoost 模型各评价指标均有提升,模型性能最佳。

2.3.1 不同剂型药品消费者满意度影响因素排名

通过 XGBoost 模型输出的特征重要性分析结果以量化各因素对不同剂型药品的消费者满意度预测的贡献度,生成各影响因素相对重要性排序,如图 5 所示。对于不同剂型的药品来说,影响因素特征重要性排序存在一定差异。整体来看,消费者对用药感官体验因素最为敏感,用药感官体验作为消费者使用药品时的直观体验,在网购场景下,消费者常常会将此次用药感官体验与过往经历对比,这种对比直接影响消费者对药品的评价。包装质量的好坏往往直接呈现给消费者,直接而深刻影响消费者的第一印象,因此对于满意度同样有着较为显著的作用。

表 1 超参数含义及其搜索范围

Tab. 1 Hyperparameter descriptions and search ranges

超参数	含义	超参数范围
n_estimators	决策树个数	{200,300,500,800,1000,1200}
max_depth	每个决策树最大深度	{5,7,9}
learning_rate	学习率	{0.01,0.05,0.10}
subsample	子采样率	{0.7,0.8,0.9}
colsample_bytree	列采样率	{0.8,0.9,1.0}
gamma	树分裂最小损失减少量	{0,0.1,0.2}
scale_pos_weight	类别不平衡调整参数	{1,5,10,20,50}

2.3.2 各影响因素重要性比较

为更清晰直观地呈现 XGBoost 模型特征重要性分析结果,综合各药品影响因素的特征重要性得分生成如图 6 所示的各因素重要性雷达图。药品价格、用药感官体验及包装质量构成消费者满意度的核心影响维度,其重要性因药品剂型与市场环境呈现显著分化。雷达图的多维度对比特性,为企业、平台营销策略制定提供了可视化决策依据。

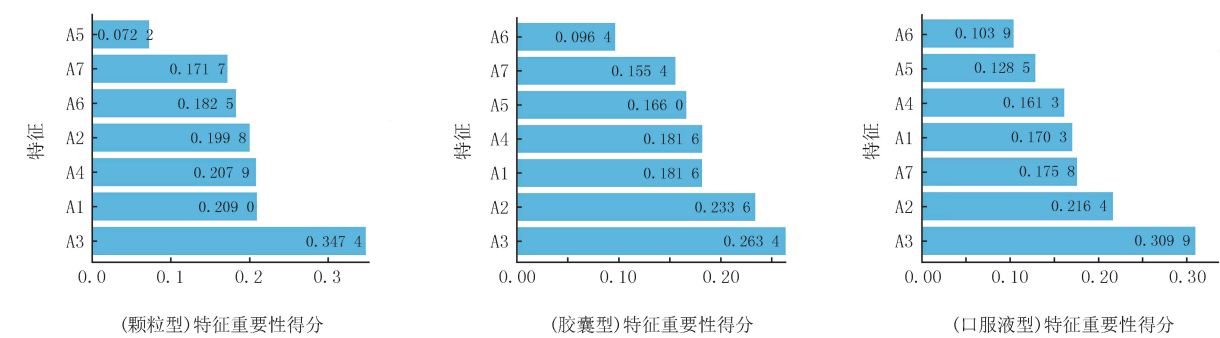


图5 不同剂型药品特征重要性得分

Fig.5 Feature importance scores for different dosage forms of medications

依据感知价值理论,价格作为感知利失的核心指标之一,在消费者的购买决策过程中扮演重要角色.价格直接左右着消费者对产品性价比的整体判断,进而决定消费者对产品价值的感知以及支付意愿.对于抗感冒药这类高频次使用、低风险的常规药物来说,消费者在进行购买决策时,往往更倾向于依赖价格作为价值判断的启发式线索.这也就解释了在6款药品中价格因素5次位列前三的现象.

跨药品维度的对比分析显示,用药感官体验与包装质量的重要性呈现剂型特异性分化:口服液因液态特性对口感与包装密封性的要求更高,而片剂则更依赖吞咽便利性设计.这种差异化特征共同构成了6款目标药品满意度影响因素的共性规律——二者在图6中均呈现高权重分布,且与价格、物流等因素形成互补影响网络,共同主导消费者的决策逻辑与评价体系.

用户感觉体验直接关联消费者即时感知、情感连接、使用决策,既是消费者与产品的第一接触点,也是用户需求的直接反馈.以白云山板蓝根颗粒的甜味设计为例,通过味觉记忆影响消费者满意度及其复购意愿,特别是儿童和长期服药患者通常会形成体验阈值,当不适度超过该阈值时,可能会增加差评比率.这一结果也与前文不同剂型药品消费者满意度影响因素排名结果相符.在信息不对称的医药电商场景,包装的密封性、说明书专业度等作为可观测信号,有效降低消费者对药品真伪与储存安全的疑虑.这也就揭示了用药感官体验、包装质量在6种药品的满意度影响因素排名中较为关键的缘由.

京东以其强大的物流配送服务在众多在线购物平台中脱颖而出,其物流配送服务多由京东自营物流提供.自营物流能够送货上门且相对于其他购物平台(如天猫、阿里健康药房)物流耗时更短、速度更快,这也是许多消费者偏好从京东购物平台购物的原因.

感冒作为一种自限性疾病,患者通常会随着时间推移自然康复.同时,服用药品后需等待一定时间才能发挥作用,药品疗效因人而异,因此药品疗效在影响消费者满意度的因素中处于相对次要地位.信任和经验也会影响消费者的药品购买决策,这与先前的研究结果一致.鉴于此,消费者在选择药物时,可能更多依赖自身经验或已知的品牌声誉,而非在线咨询服务.故而,在线服务与售后服务在各项影响因素的特征重要性排序中相对靠后.

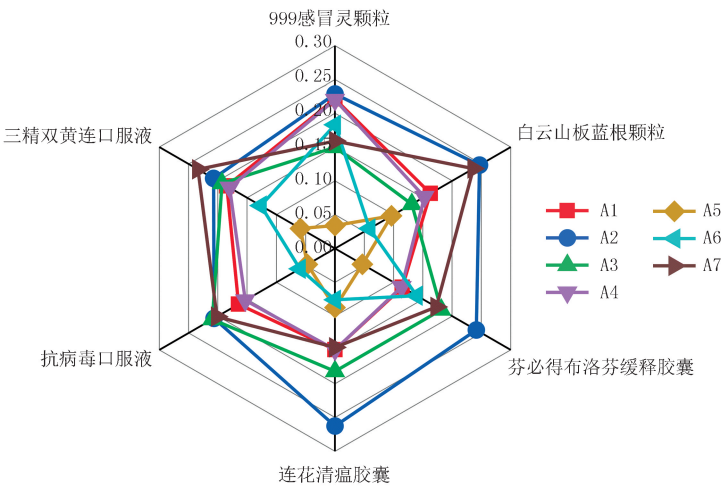


图6 各药品因素重要性雷达图

Fig.6 Feature importance Radar Chart for multiple drugs

3 结论

本文根据京东在线评论数据,借助 NLP 对文本进行预处理,将无监督学习的 LDA 主题模型与监督学习的 XGBoost 算法深度融合,构建了消费者行为研究领域的全新分析框架,这极大丰富了 NLP 与机器学习技术在医药电商行业的应用研究维度,更为突破解决传统研究方法的固有局限、开辟新研究路径提供了极具价值的思路。

通过上述框架,深入挖掘影响医药电商领域 3 种不同剂型抗感冒药消费者满意度的关键影响因素,主要集中在物流配送服务、药品价格、用药感官体验、药品物流状态、在线及售后服务、药品疗效、包装质量等 7 个维度,结合各主题高频词及词汇概率可知,在在线及售后服务、药品物流状态两个主题下,有多个负面词语,如“泡水”“破损”等,是未来平台改进的重点方向。通过对比不同剂型药品消费者满意度影响因素排名,发现消费者满意度对用药感官体验因素最为敏感,而包装质量则对口服液型药品影响权重较高。而且,从 6 种药品可视化结果整体来看,药品价格、用药感官体验、包装质量等 3 个因素对于消费者满意度的贡献较大,物流配送服务和药品物流状态紧随其后,药品疗效和在线及售后服务贡献相对较小。未来研究可延伸至多个电商平台,扩充在线评论数量,从技术层面入手进一步挖掘评论信息,对消费者行为进行更深入的研究。

附录见电子版(DOI:10.16366/j.cnki.1000-2367.2025.08.04.0003)。

参 考 文 献

[1] Oliver R L.A cognitive model of the antecedents and consequences of satisfaction decisions[J].Journal of Marketing Research,1980,17(4):460-469.

[2] Churchill G A Jr,Suprenant C.An investigation into the determinants of customer satisfaction[J].Journal of Marketing Research,1982,19(4):491-504.

[3] 王宁.数字化背景下乡村旅游消费者满意度测评模型构建[J].商业经济研究,2021(8):57-60.
Wang N.Construction of rural tourism consumer satisfaction evaluation model under the background of digitalization[J].Journal of Commercial Economics,2021(8):57-60.

[4] 廖名岩,王璜,张世兵,等.消费者购买可数字化追溯猪肉的影响因素[J].系统工程,2025,43(4):27-38.
Liao M Y,Wang H,Zhang S B,et al.The factors influencing consumer purchase of digitally traceable pork[J].Systems Engineering,2025,43(4):27-38.

[5] 朱郭奇,钱冬,李文文,等.社交电商平台消费者忠诚建立:对信任与满意度的再认识[J].管理工程学报,2025,39(5):74-91.
Zhu G Q,Qian D,Li W W,et al.Building consumer loyalty to social e-commerce platforms:a re-understanding of trust and satisfaction[J].Journal of Industrial Engineering and Engineering Management,2025,39(5):74-91.

[6] 陈丁丁,崔荣荣,牛犁.基于 Kano 模型的裘皮服饰设计需求研究[J].丝绸,2025,62(12):16-25.
Chen D D,Cui R R,Niu L.Research on fur clothing design requirements based on the Kano model[J].Silk,2025,62(12):16-25.

[7] 李敏,姜夏楠,陈彩霞.服装品牌智能客服服务满意度的影响因素研究[J].丝绸,2025,62(9):33-42.
Li M,Jiang X N,Chen C X.Research on the influencing factors of artificial intelligence customer service satisfaction of fashion brands[J].Journal of Silk,2025,62(9):33-42.

[8] 熊百妹.基于逐步回归法的药品消费者满意度影响因素实证研究[J].中国药房,2013,24(17):1548-1550.
Xiong B M.Empirical study on influential factors to drug customer satisfaction based on stepwise regression method[J].China Pharmacy,2013,24(17):1548-1550.

[9] Zhao X Q,Huang Z.A method for exploring consumer satisfaction factors using online reviews:a study on anti-cold drugs[J].Journal of Retailing and Consumer Services,2024,81:103895.

[10] 倪昭鑫,舒帆.基于自注意力机制的 CNN-BiLSTM 生鲜物流服务质量影响因素[J].科学技术与工程,2025,25(16):6821-6830.
Ni Z X,Shu F.Factors affecting fresh logistics service quality based on self-attention mechanism of CNN-BiLSTM[J].Science Technology and Engineering,2025,25(16):6821-6830.

[11] 郭先利,廖虎昌.多平台在线评论主题挖掘与消费者偏好演化趋势分析[J/OL].[2025-07-10].https://link.cnki.net/urlid/31.1977.N.20250606.1437.002.
Wu X L,Liao H C.Topic mining and consumer preference evolution trend analysis based on online reviews from multiple platforms[EB/OL].[2025-07-10].http://en.cnki.com.cn/Article_en/CJFDTOTAL-XTGL20250605001.htm.

[12] 何勇,陈静,李姗姗,等.生鲜电商用户在线选择行为研究:考虑评论的影响效用[J].管理科学学报,2025,28(5):70-83.
He Y,Chen J,Li S S,et al.Online choice behavior for fresh e-commerce customers:Considering the social influence of reviews[J].Journal of Management Sciences in China,2025,28(5):70-83.

[13] Krishna B,Puram P.The impact of trust-based challenges on user satisfaction in food sharing platforms;a text mining approach[J].Technological Forecasting and Social Change,2025,216:124159.

[14] 李伟,徐华,李媛媛.节假日时段国内自驾游客满意度研究:基于结构方程模型的实证分析[J].河南师范大学学报(自然科学版),2020,48(6):1-9.
Li W,Xu H,Li Y Y.Satisfaction of domestic self-driving tourists during holidays:Empirical analysis based on SEM[J].Journal of Henan Normal University(Natural Science Edition),2020,48(6):1-9.

[15] Wang G Q,Zhu B W,Wang K Y,et al.Links between cue combinations of physical environments and consumer satisfaction in themed restaurants from a systematic approach-avoidance perspective[J].International Journal of Hospitality Management,2025,126:104055.

[16] 吴东雪,沈桂兰.一种基于 LDA 模型的新兴主题识别与探测方法[J].河南师范大学学报(自然科学版),2024,52(2):72-80.
Wu D X,Shen G L.An emerging topic identification and detection method based on LDA model[J].Journal of Henan Normal University (Natural Science Edition),2024,52(2):72-80.

[17] 蒋昊达,赵春蕾,陈瀚,等.基于改进 TF-IDF 与 BERT 的领域情感词典构建方法[J].计算机科学,2024,51(S1):150-158.
Jiang H D,Zhao C L,Chen H,et al.Construction method of domain sentiment lexicon based on improved TF-IDF and BERT[J].Computer Science,2024,51(S1):150-158.

[18] 韩坤,刘忠轶.基于主题模型与情感特征增强的食品安全网络舆情演化分析方法[J].食品科学,2025,46(23):367-375.
Han K,Liu Z Y.A method for analyzing the evolution of online public opinion on food safety based on topic modeling and sentiment feature enhancement[J].Food Science,2025,46(23):367-375.

[19] Chen T Q,Guestrin C.XGBoost:A Scalable Tree Boosting System[C].New York:Association for Computing Machinery,2016:785-794.

[20] 温廷新,白云鹤.融合 RF-GA-XGBoost 和 SHAP 的虚假新闻群体互动质量可解释模型[J].数据分析与知识发现,2024,8(12):86-100.
Wen T X,Bai Y H.An interpretable model for fake news Groups' Interaction quality based on RF-GA-XGBoost and SHAP[J].Data Analysis and Knowledge Discovery,2024,8(12):86-100.

Identifying factors influencing consumer satisfaction in the pharmaceutical e-commerce sector

——A study based on online reviews of Anti-Cold Drugs

Hu Xiaozheng, Dong Jingfeng, Tao Xinmin

(School of Civil Engineering and Transportation, Northeast Forestry University, Harbin 150040, China)

Abstract: This paper takes consumer satisfaction as the entry point and focuses on the optimization path of service quality in pharmaceutical e-commerce platforms, aiming to provide data-driven support for the precise improvement of service quality. Based on 60 396 valid review entries for anti-cold medications, a three-dimensional analytical framework encompassing text mining, topic feature extraction, and influencing factor ranking was constructed. The LDA topic model was employed to determine the optimal number of topics and to uncover latent themes within the review texts, identifying seven core influencing factors, including logistics delivery service and drug pricing. Subsequently, using the document - topic probability distribution vectors from the reviews as input features, an XGBoost classification model was developed. By comparing its performance against models such as logistic regression, random forest, support vector machine, and CatBoost across five performance metrics(e.g., accuracy), the proposed model was demonstrated to exhibit superior adaptability to the dataset. After parameter optimization, the model achieved optimal performance, and the importance scores of the features were output. The findings reveal significant differences in the factors influencing consumer satisfaction across different dosage forms of medications, thereby providing robust theoretical support for research on consumer behavior in pharmaceutical e-commerce.

Keywords: consumer satisfaction; online reviews; LDA topic model; XGBoost; factors

附录

表 S1 在线评论数据示例

Tab. S1 Sample online review data

序号	评论内容	评分
1	居家必备良药！国民大品牌值得信赖！趁活动入手性价比超高，网上购物真是太方便了，足不出户就可以货比三家，全五分好评！非常满意的一次购物体验，以后有需要会再次购买，* * * * 杠杠滴！	5
2	还不错吧，治感冒还不错.	3
3	性价比很低，* * 上七八块一盒，凑单买的，认栽了！	2
4	快递收到了，好慢，等了好几天.	1
5	OK 收到，包装无损，* * 胶囊还可以还搭配消炎最近老感冒，吃了几次好多了 * * * 药房买的放心，一次愉快的购物体验！	4

表 S2 各主题高频词及词汇概率

Tab. S2 Topic-specific high-frequency terms and rerm probability

主题	主题名称	高频词	词汇概率
A1	物流配送服务	物流配送, 快递, 物流, 令人满意, 服务态度, 破损, 小哥, 专业, 超时, 配送	0.059 7, 0.044 7, 0.042 6, 0.030 5, 0.028 5, 0.024 8, 0.020 8, 0.018 9, 0.018 4, 0.018 4
A2	药品价格	价格, 略高于, 市场, 活动, 优惠, 稍显, 超值, 改进, 接受, 波动	0.098 0, 0.041 5, 0.041 5, 0.039 9, 0.039 9, 0.038 4, 0.032 3, 0.028 3, 0.027 5, 0.025 9
A3	用药感官体验	口感, 宜人, 清凉, 舒服, 正宗, 微苦, 中药味, 吞咽, 甜甜, 参半	0.950 6, 0.063 7, 0.062 9, 0.060 3, 0.055 1, 0.048 3, 0.030 2, 0.027 0, 0.023 7, 0.023 7
A4	药品物流状态	物流, 质量, 完好, 破损, 盒装, 送货, 追踪, 玻璃瓶, 防潮, 泡水	0.420 9, 0.301 0, 0.089 0, 0.086 9, 0.067 3, 0.037 4, 0.037 1, 0.024 7, 0.019 6, 0.017 2
A5	在线及售后服务	客户服务, 专业, 体验, 可靠, 高效, 售后服务, 客服, 不换, 不退, 恶劣	0.988 4, 0.069 9, 0.066 6, 0.054 1, 0.052 4, 0.033 9, 0.028 5, 0.020 7, 0.020 7, 0.009 7
A6	药品疗效	因素, 疗效, 药品, 物超所值, 偏贵, 波动, 有限, 常规, 尚可, 质量	0.277 6, 0.248 1, 0.245 9, 0.091 8, 0.069 0, 0.065 8, 0.020 7, 0.018 5, 0.015 4, 0.100 3
A7	包装质量	包装, 质量, 完好, 盒装, 痕迹, 拆封, 常规, 精致, 脆弱, 泡沫	0.103 2, 0.053 5, 0.022 9, 0.021 5, 0.020 5, 0.020 5, 0.018 6, 0.018 3, 0.018 0, 0.017 5