E-Commerce Letters 电子商务评论, 2024, 13(1), 248-259
Published Online February 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.131031

Hans汉斯

基于电商平台在线评论的运动相机消费者偏好趋势挖掘

陈　思

江南大学商学院，江苏　无锡

收稿日期：2024年1月3日；录用日期：2024年1月12日；发布日期：2024年2月29日

摘　要

当前电商对于消费者在线评论关注度越来越高。其包含了消费者的使用体验、产品偏好等信息，能够帮助商家了解消费者满意度和未来偏好等，针对性地进行产品升级与营销调整，以更加迎合消费者购买倾向。本文结合文本挖掘、情感分析和Lasso-SVM筛选预测模型，挖掘消费者偏好趋势，为商家提供从在线评论文本提取隐含信息，有助于商家进一步了解未来产品优化方法。本文选取京东的运动相机进行实例分析，探索消费者偏好趋势挖掘，为在线评论的文本分析与偏好趋势挖掘提供了参考依据。

关键词

在线评论，偏好预测，信息增益，Lasso-SVM筛选预测模型

Mining Consumer Preference Trends of Action Cameras Based on Online Reviews on E-Commerce Platforms

Si Chen

School of Business, Jiangnan University, Wuxi Jiangsu

Received: Jan. 3rd, 2024; accepted: Jan. 12th, 2024; published: Feb. 29th, 2024

Abstract

At present, e-commerce companies pay more and more attention to consumers' online reviews. It contains consumers' use experience, product preferences, and other information. For merchants, this implicit information can help them understand consumers' satisfaction and future prefe-

文章引用: 陈思. 基于电商平台在线评论的运动相机消费者偏好趋势挖掘[J]. 电子商务评论, 2024, 13(1): 248-259.
DOI: 10.12677/ecl.2024.131031

rences, so as to carry out targeted product upgrades and marketing adjustments, so as to better cater to consumers' purchase tendencies. In this paper, text mining, sentiment analysis and Lasso-SVM screening prediction model are combined to study and analyze online review text to predict consumer preference and provide a method for merchants to extract implied information from online review text and predict consumers' future preference, which is helpful for merchants to further understand the method of future product optimization. This paper selected JD's sports camera for example analysis to explore consumer preference trend mining and prediction, providing a reference for text analysis and preference trend mining of online reviews.

Keywords

Online Reviews, Prediction of Preference, Gain of Information, Lasso-SVM Prediction Model

1. 引言

近年随着互联网社交平台的推广和发展，越来越多人倾向于发布 Vlog、Plog 等方式记录和分享日常生活、户外运动、旅行等，相应所需电子设备的市场需求也在不断增加。与传统视频录制设备相比，运动相机具有体积小、重量轻、防抖、防震、便于携带等优点，可提供多场景环境下的高清视频拍摄和录制。在此背景下，运动相机市场的竞争日趋白热化，促使运动相机厂商加速产品的优化升级和调整产品的营销策略。为在激烈的市场竞争中胜出，提供更符合消费者及时需求的产品，短时间内把握与预测消费者需求偏好是商家竞争胜出的关键所在。伴随着互联网时代的到来与物流水平的提高，许多消费者倾向于通过在线电子商务平台购买商品。消费者可以通过电商平台购买产品、分享体验。消费者购买产品后，会在电商平台上对购买的产品的不同方面进行在线评论，如价格、性能、用户体验等。

在线评论包含文字、图像、表情和视频等多种内容。近年来，很多学者基于在线评论在很多领域开展了诸多研究，如消费者满意度、情绪分析、购买意向预测等等。Zhou 等[1]根据词频统计计算指标权重，利用 DIPCA 算法动态评估客户满意度。Barta 等[2]针对消费者的负面在线评论使用结构方程模型来研究差评的影响以及评论的感知说服力、有用性和可信度。Sim 等[3]使用具有空间概率模型和卷积神经网络(CNN)来衡量在线评论对住宿预订意图的影响。Kim 等[4]使用 CONCOR 方法聚类分析以及因子分析和线性回归分析进行定量分析，研究在线评论对顾客满意度影响。Zhang 和 Niu [5]提出了一种新的基于长短期记忆交互的卷积神经网络(LICNN)模型利用在线评论来预测酒店需求。当前研究基于文本聚类等手段，针对在线评论文本的关键词来探索定性数据，或基于在线评论进行定量化数据处理。Pineda-Jaramillo J 和 Pineda-Jaramillo D [6]使用来自在线评论的非结构化数据来预测城市地铁系统的出行满意度。Yadav 和 Sagar [7]结合文本分析和社交网络方法，利用 LDA 模型进行主题分类，然后使用 NetworkX 进行网络分析。Zhu 等[8]提出了一种基于论点质量和来源可信度的扩展信息采用模型，以研究与在线评论的感知有用性相关的因素。该模型以读者信任倾向为个体因素、产品类型为调节变量，构建了在线评论感知有用性相关因素的权变模型。

在线评论的情感分析是通过提取在线评论的文本中的情感倾向和情感倾向强度，绝大多数研究针对情绪分析进行分析挖掘。Zhang 等[9]基于细粒度情感分析与卡诺(Kano)模型，完成了情感分析和需求识别的结合，最终得出消费者对产品属性的需求偏好。Zhang 和 Guo 等[10]提出了一种将直观模糊

TODIM (IF-TODIM)方法与情感分析相结合的产品选择模型。Wang 等[11]使用多注意力双向 LSTM 来识别视频中消费者评论的情感极性，能更好地挖掘消费者在各个主题上的情感之间的关系。Obiedat 等[12]提出了一种将支持向量机(SVM)算法、粒子群优化(PSO)算法和过采样技术相结合，对在线评论中存在的不平衡数据分布进行情感分析。许多研究对在线评论情感分析后没有对得出结果再进一步分析挖掘。Bhuvaneshwari 等[13]基于双向长短期记忆自注意力卷积神经网络模型对用户评论进行情感分析。Lucini 等[14]探索了基于客户满意度维度的航空公司推荐预测，基于贝叶斯分类器的情绪分析后使用逻辑回归模型预测航空公司的推荐。多模态情感分析领域的注意力机制法，能通过提取模态的关键特征提取情感分析的权重信息，进而增加模型准确度。Shi 等[15]提出一种基于多尺度卷积神经网络和交叉注意融合机制的情感分析模型 MSCNN-CPLCAFF，有效分析了抖音平台的视听和文本数据的跨模态特征融合问题。

目前研究在线评论中使用的模型，大多数研究人员倾向于使用机器模型和深度学习模型。沈超等[16]首先使用决策树分类模型来识别消费者偏好的关键和非关键属性。然后利用时间序列来分析非关键属性的未来重要性。Yakubu 等[17]使用 Shapley 值和 Choquet 积分进行产品属性重要性识别研究，再使用模糊时间序列进行产品属性的变化趋势预测。挖掘与分析在线评论的机器学习模型，以支持向量机(SVM)、梯度提升、随机森林、逻辑回归和决策树等为主。Nilashi [18]使用分类回归树(CART)模型进行客户偏好预测。Lee 等[19]使用了四种机器学习算法对大型餐厅的在线评论进行对比研究和预测。Zibarzani 等[20]使用学习向量量化(LVQ)算法对在线评论进行聚类分析，并使用分类回归树(CART)算法预测客户满意度和偏好。Hussain 等[21]基于 IOWA 算子的自适应神经模糊系统，将预测模型与模糊 C 均值、减法聚类和网格划分相结合，对在线服务质量执行复杂的预测。Khorsand 等[22]使用多种机器学习模型根据酒店和乘客的在线评论信息来预测入住率，横向比较了每种算法的结果。深度学习模型可以利用非线性激活函数提取在线评论中的特征向量和其相关联系，从而提高研究精度。研究模型以长短期记忆网络神经网络(LSTM)、递归神经网络(RNN)和卷积神经网络(CNN)等为基础。Wang 等[23]通过深度神经网络预测消费者评论感知效用，选取 BiLSTM 模型和传统 LSTM 神经网络模型进行预测并比较 2 种模型的精确率。Jain 等[24]选择四种机器学习模型预测乘客推荐，并用分层 K 折交叉验证来测试和验证模型，比较四种预测模型结果。Su 等[25]提出卷积注意力－长短期记忆网络(CA-LSTM)模型，通过注意力机制和卷积运算从在线评论中获得真实的情感特征指标具有更好的分类性能。Chang 等[26]用长短期记忆网络对在线评论的评分与文本进行情感分析，预测每月酒店入住率。Oh 等[27]基于 RNN 处理酒店信息的分类变量，后使用 LSTM 模型处理在线评论文本信息和卷积神经网络提取酒店图像特征。Olmedilla 等[28]使用一维卷积神经网络模型对在线评论有用性进行预测，此模型展示了对评论有用性的预测和分类的高性能。

纵观现有研究，在线评论研究多片面集中于消费者满意度、情绪分析、购买意向预测等其中一个方向，或只专注于模型优化，缺少针对在线评论整体深入挖掘研究。针对客户需求偏好的变化，多数研究止步于通过在线评论文本对客户的偏好需求进行满意度评价与总结，较少关注消费者偏好倾向动态发展趋势。偏好预测模型研究多集中于预测模型算法改进与优化方面，缺少对属性偏好维度动态变化的考虑，没有充分考虑消费者对于产品属性的情绪动态走向。研究大多数集中在消费者需求的获取、识别、映射与转化等方面，对于消费者多样化需求偏好的研究相对较少。面对消费者多样化和差异化需求时，厂商和电商在产品的生产和销售无法及时根据消费者需求调整策略方向。缺少以优化与升级产品性能为目的消费者的偏好倾向挖掘研究。在此基础上，本文在深入研究了产品在线评论文本挖掘、情感倾向分析、预测模型相关理论和研究现状后，提出消费者偏好预测模型构建，采用实证分析对用户评论进行了深入挖掘，得出消费者偏好未来趋势。

2. 模型构建

本文构建的模型主要包括三个部分，即偏好特征提取、偏好特征重要性计算和关键偏好特征筛选与预测。具体而言，首先进行情感倾向分析得出消费者情感倾向值，之后利用基于困惑度的 LDA 模型提取在线评论中包含的偏好特征，然后进一步使用信息增益法结合在线评分与情感倾向得出偏好特征的重要性。最后使用 Lasso-SVM 模型进行偏好特征的筛选与预测，如图 1 所示。

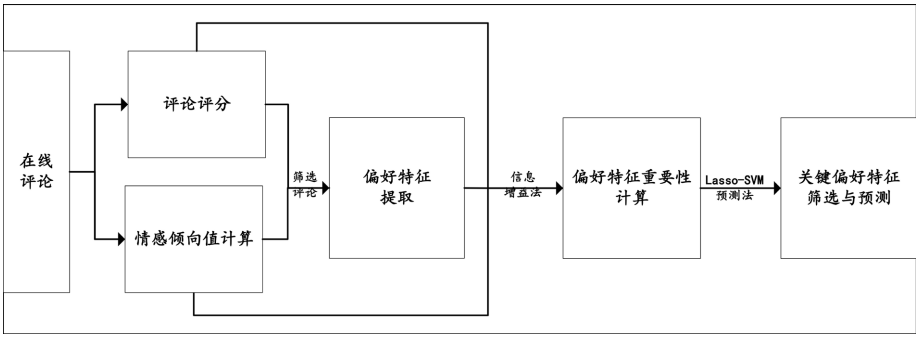


Figure 1. Research model
图 1. 研究模型

2.1. 偏好特征提取与情感分析

在线评论数据通常包含评论文本和客户对产品的打分。根据收集的大规模在线评论，本文首先采用 Blei 等[29]提出的生成式主题模型 LDA 法提取潜在产品属性词语，然后通过人工剔除其中常见的非属性词语，并对剩余潜在属性词语进行同义词合并，生成产品属性词典。LDA 选取主题数量需要同时考虑模型构建的复杂性和数据信息的覆盖范围，本文引入困惑度来选取最佳主题数 K，防止数据过拟合现象或信息缺失，具体计算公式如下[30]。

$$perplexity(D)=\exp\left(-\frac{\sum \log\log p(w)}{\sum_{d=1}^{M} N_d}\right) \tag{1}$$

其中分母为数据集中所有单词之和。$p(w)$为数据集中每个词语出现的概率，其计算公式为：

$$p(w)=p(z/d)*p(w/z) \tag{2}$$

其中 $p(z/d)$为每个文档中各个主题出现的概率，$p(w/z)$为是词库中的各个词语在其中一个主题中出现的概率。

因为消费者的在线评论中有大量词语，一些出现频率低或者偏僻词语对于本研究属于噪声，设定上限可以剔除数据的噪声，之后把文本数据集中的词汇向量化，即将其转化为词频矩阵，再计算各个词频矩阵中每个词汇的 TF-IDF 权重，转化向量以及提取主题关键词。了解用户的看法需要通过评论文本得知消费者偏好倾向，因此需要进行情感分析。本文使用 Python 中的 Snownlp 模块，识别各个特征在每个出现评论中的情感倾向。

2.2. 基于信息增益法的偏好特征重要性计算

在产品设计术语中，信息熵可以表示为在数据集 S 中，区分一个类别和另一个类别的不确定性。在数据集 U 中，区分一个类别和另一个类别的不确定性。假设初始数据集 U 包含 n 条评论数据 $(d_1,d_2,\cdots,d_n)$，每条评论数据都有一个对应的类别 $(c_1,c_2,\cdots,c_n)$，则初始数据集(包含所有特征)的信息

熵为：

$$Entropy(U) = -\sum_{r=1}^{k} p(c_r)\log_2 p(c_r) \tag{3}$$

其中，$p(c_r)$表示数据集 U 中的类变量 c_r 的概率，k 表示类变量的个数。

本文的类变量包含三个：c_1 到 c_3 分别表示评分的高、中和低。为了确定最大能力的偏好特征，减少选择集的不确定性，根据偏好特征变量的取值划分为 n 个子数据集。给定一个特定的偏好特征 a，信息熵是该偏好特征的每个唯一值的信息熵的总和为：

$$Entropy_a(U) = \sum_{j=1}^{n} \frac{|U_j|}{|U|} Entropy(U_j) \tag{4}$$

其中 U_j 表示训练数据 U 的子集，包含属性的互斥结果值，$|\cdot|$表示集合所包含的元素个数。本文的偏好特征有三个互斥结果值(高、中和低)，则训练集 U 将被划分为三个数据子集，U_1 将包含偏好特征值为高的所有数据实例。根据前文，本文将消费者对商品的评分(高、中、低)作为类变量，视为消费者满意程度，结合偏好特征情感(正面，负面和中性)，运用信息增益方法计算每个偏好特征对于客户满意度的影响大小。因此，偏好特征 a 的信息熵为：

$$Entropy_a(U) = \frac{|U^+|}{|U|} Entropy(U^+) + \frac{|U^-|}{|U|} Entropy(U^-) + \frac{|U^0|}{|U|} Entropy(U^0) \tag{5}$$

其中 U^+、U^-和 U^0 分别表示产品属性 a 为正面、负面以及中性的评论。

之后，计算信息增益值。信息增益越大，说明该词语特征含有的信息量就越大，也就是该偏好特征越重要，即属性 a 的 $Entropy_a(U)$越低，其增益 IGF 越高，这两者的关系可表示为词语特征的信息增益等于初始信息熵减去已知词语特征的信息熵，计算公式如下：

$$IGF = Entropy(U) - Entropy_a(U) \tag{1}$$

根据公式(6)可以计算每个产品属性的信息增益。根据信息增益 IGF 的值，将所有偏好特征的信息增益值按照从大到小排序作为偏好特征重要性排序，以及下一步偏好趋势分析的输入值。

2.3. 基于 Lasso-SVM 模型的关键偏好特征趋势分析

为了获得更加客观和细致化的用户偏好趋势特征，本文在偏好趋势分析的分析中使用 Lasso 与 SVM 模型相结合预测特征趋势。SVM 模型进行长时间序列预测上具有较强的精度。增加 Lasso 算法进行偏好特征的筛选，剔除了对于消费者未来购买行为不产生影响的偏好特征，减少不相干偏好特征对模型精度的干扰，从而改善建模品质。

Lasso 算法[31]可以将不相关的自变量收缩为零，以此达到优化回归模型中的多重共线性问题。Lasso 通过对最小二乘估计加入 L_1 范数作为罚约束，使某些系数估计为 0，减少参数数量。使用 Lasso 回归，剔除相关性较小的因素，对消费者偏好进行筛选，是基于惩罚方法对数据进行变量选择，通过对原本的系数进行压缩，将不显著的变量系数压缩至 0，再将此类特征直接舍弃，Lasso 回归预测模型目标函数表示为：

$$\sum_{j=1}^{n}\left(y_i - \lambda\sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\lambda\sum_{j=1}^{p}|\beta_j| \tag{7}$$

式中：RSS 是实际值减去估计值的差的平方和；β 是参数向量；λ 是调优参数；p 为参数个数。根据式(7)可知，由于 Lasso 回归模型的目标函数包含惩罚项系数 λ，故在计算模型回归系数前，需要得到最理想的 λ 值，λ 值的确定可以通过定性的可视化方法和定量的交叉验证方法。在本文中，将偏好特征的信息增益

值作为向量 X_1 到 X_n，将消费者每条评论类型的信息熵作为 Y 值输入 Lasso 模型中，将结果中向量 X_1 到 X_n 的向量系数为 0 的偏好特征剔除，将筛选后的特征向量输入 SVM 时间序列预测模型中。

支持向量机(SVM)是一种监督学习的分类方法。它在解决小样本、非线性和高维模式识别方面具有许多独特的优势，可广泛应用于函数拟合等机器学习问题。其原理和解决过程如下：

给定一个包含两类数据的样本数据集：

$$V=\left\{\left(x_i,y_i\right)\middle|x_i\in R^m,y_i\in\left\{-1,+1\right\},i=1,\cdots,n\right\} \tag{8}$$

则支持向量机判别函数为：

$$f\left(x\right)=sign\left(\sum_{i=1}^{n}A_iy_iK\left(x,x_i\right)+B\right) \tag{9}$$

其中：

$$B=y_i-\sum_{i=1}^{n}A_iy_iK\left(x,x_i\right) \tag{10}$$

i 为支持向量个数，本文中 i 作为偏好特征向量个数，$K\left(x,x_i\right)$ 为核函数。将样本数据集 V 作为训练集输入后，根据实际数据集选取核函数与惩罚因子求解优化问题的最优解向量A_i。本文将 Lasso 筛选过的偏好特征向量分别作为特征向量输入 SVM 中，通过已有的 1 至 N 阶段作为时间序列值，对第 $N+1$ 阶段进行预测，之后将得到的 $N+1$ 阶段作为已知数据，将第 2 至 $N+1$ 阶段再次作为时间序列值，对第 $N+2$ 阶段进行预测；以此类推，预测未来 3 期的趋势走向。因此，本研究所使用的 Lasso-SVM 方法，首先使用 Lasso 方法进行变量选择，得到特征变量；然后使用支持向量机作为分类器，利用式(9)对特征变量进行训练。并采用 k 倍交叉验证方法对预测模型的性能进行检验。

3. 实证分析

3.1. 数据收集及预处理

本文选取运动相机为研究对象，从京东收集了销量排名前 20 的运动相机的所有评论内容，共计收集到 15,301 条评论运动相机的在线评论，数据的时间跨度从 2020 年 1 月到 2022 年 12 月。每条评论主要由评论者名称、评价星级、评论内容、评论时间、商品属性、评论类型等组成。一些例子如表 1 所示。

Table 1. Examples of online reviews of action cameras
表 1. 运动相机在线评论示例

用户名称	评价星级	评价内容	时间	商品属性	评论类型
t****	star5	非常小巧很实用，就是容易发烫，很棒的。	2022-07-16	Action 2 续航套装	好评
-****	star5	运动拍照效果不错，拍照效果好。	2022-07-16	Action 2 续航套装	好评
p****	star2	只能拍广角，买回来基本上没用。	2022-10-04	【重磅新品】S3 运动相机	中评

消费者在电商平台上发布的在线评论一般含有很多噪声，比如一些重复评论、无意义符号、外语单词等，这些噪声对于研究过程和最终结果都有干扰，需要除去数据的噪声，再进行后续研究。故在使用 LDA 模型前需要将所获取的数据进行预处理。本文通过 Python 的 pandas 模块，利用 drop_duplicates 方法进行数据去重，最终得到 15,296 条评论文本。再将 jieba 文包导入 Python，进行文本清洗以及文本分词。处理后的示例如表 2 所示。

Table 2. Data preprocessing example
表 2. 数据预处理示例

在线评价文本	预处理后
非常小巧很实用，就是容易发烫，很棒的。	非常_小巧_很_实用_就是_容易_发烫_很_棒
运动拍照效果不错，拍照效果好。	运动_拍照_效果_不错_拍照_效果_好
只能拍广角，买回来基本上没用。	只能_拍_广角_买_回来_基本_没用。

3.2. 在线评论偏好特征识别

3.2.1. 偏好特征提取

本文使用 jieba 文包以及 sklearn 工具包通过 LDA 模型进行偏好维度的提取。

其中主题数目和困惑度折线图如图 2 所示，由图得，当主题数目为 4 时，达到折线最低点，因此最佳主题数 K = 4。将 K = 4 代入 LDA 模型，求出每个主题前 20 个关键词如表 3。

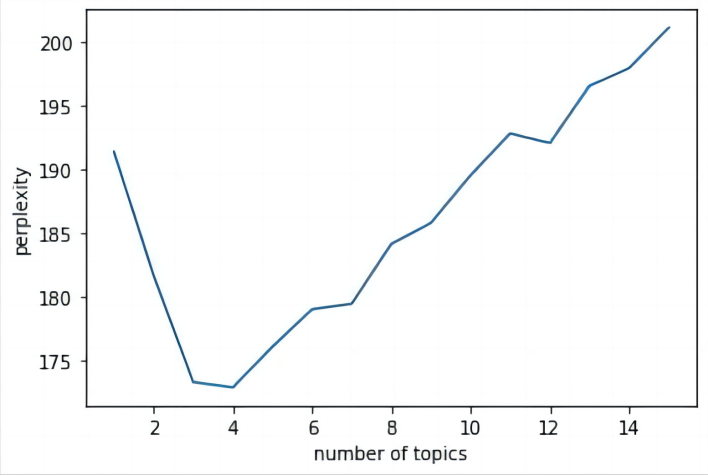


Figure 2. Line chart of topic number and perplexity
图 2. 主题数与困惑度折线图

Table 3. Keywords of online reviews
表 3. 在线评论关键词

主题	关键词
1	效果、画质、防抖、性能、特色、小巧、画面、测试、口袋、视频、光线、想象、稳定性、夜景、全景、杠杠、地方、手机、色彩、跑步
2	产品、质量、客服、物流、价格、速度、发货、活动、购物、品牌、朋友、服务、体验、性价比、下单、商品、信赖、售后、颜值、速度快
3	手机、感觉、视频、评价、机器、功能、连接、用户、内容、研究、开机、客服、电池、死机、设备、屏幕、试用、升级、软件、剪辑
4	视频、运动、小巧、续航、记录仪、电池、行车、配件、记录、头盔、小时、镜头、功能、时间、摩托车、模式、清晰度、感觉、设计、像素

结合表 3 和图 2，主题 1 主要反映运动相机的使用功能；主题 2 主要反映平台的服务体验；主题 3 主要表现消费者对相机的使用体验感；主题 4 表现运动相机的使用场景。综合来看，4 个主题无重叠，拟合较好。

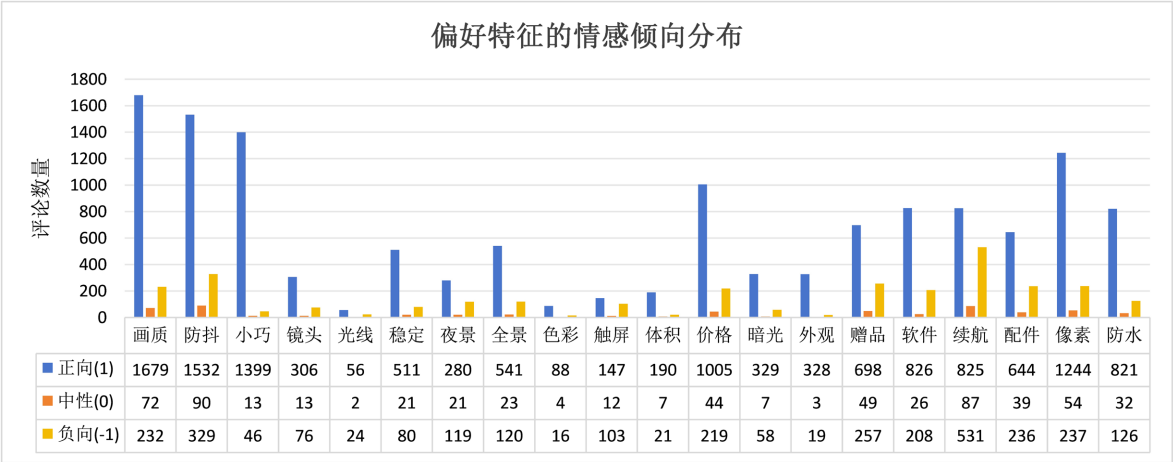
3.2.2. 情感分析

本研究使用 snowNLP 库自带字典进行分析，并设情感得分大于 0.6 为正向情感倾向评论，0.6 到 0.4 之间为中性情感倾向评论，小于 0.4 为负向情感倾向评论。限于篇幅，展示部分文本情感得分如下表 4 所示。

Table 4. Emotion score of online review text by emotion orientation
表 4. 各情感倾向在线评论文本情感得分

评论文本	情感得分	情感倾向值
超级不好用，三脚架转接头按上就拔不下来。千万别买	0.187095115	−1
唯一优点就是小了，续航真的很一般，每次拍几分钟，就要放到充电盒，吃灰的可能性大	0.50224667	0
相对于一代来说提升巨大，真正成为了一款非常好的，谢谢产品	0.99307648	1

经过情感分析算法分类，共得 11,307 条正面情感评论、3053 条中性评论和 936 条负面情感评论。由于京东商城在线评分采用 5 星制原则，为了便于对比，本文将 5 星作为正向倾向，4 星与 3 星作为中性情感倾向，2 星与 1 星作为负向倾向。将得出的情感倾向性估值与在线评分对比，将情感倾向与评分不一致的数据剔除，剔除 458 条数据，最终留下 14,838 条有效评论。



	画质	防抖	小巧	镜头	光线	稳定	夜景	全景	色彩	触屏	体积	价格	暗光	外观	赠品	软件	续航	配件	像素	防水
正向(1)	1679	1532	1399	306	56	511	280	541	88	147	190	1005	329	328	698	826	825	644	1244	821
中性(0)	72	90	13	13	2	21	21	23	4	12	7	44	7	3	49	26	87	39	54	32
负向(-1)	232	329	46	76	24	80	119	120	16	103	21	219	58	19	257	208	531	236	237	126

Figure 3. Distribution map of emotional tendency of preference characteristics
图 3. 偏好特征情感倾向分布图

基于前文的偏好特征识别以及情感分析法，本案例研究识别出 20 个消费者偏好特征的情感分布，偏好特征的情感分布情况如图 3 所示。可以看出消费者对大多数偏好特征持有正向情感倾向，“画质”、“防抖”和“体积”等偏好特征是消费者关注较多的偏好特征。而消费者对于“价格”、“电池”和“配件”的负面情感较多。以“体积”为例，在所有评论中共筛选出包含“体积”的评论 1611 条，其中正面情感评论数 1005 条。

3.2.3. 偏好特征筛选与趋势分析

本文采用 3 年的在线评论数据，共计有效评论 16,388 条，从 2020 年 1 月到 2022 年 12 月，分为 12 个阶段，每个阶段代表 1 个季度。将处理好的数据集划分为 12 节时间段，使用信息增益法分别计算每节时间段偏好特征重要性，下面表 5 展示了部分偏好特征信息增益值。

使用 Lasso-SVM 模型进行消费者偏好特征的筛选。本文采用 sklearn 子模块 linear-model 中的 Lasso 类中目标函数所包含的惩罚项系数进行计算，采用 5 重交叉验证的方法得到 Lasso 的最佳的 λ 值，Lasso$_\lambda$ = 6.421185709064758e − 07。最后基于最佳的 λ 值分别得到 Lasso 模型变量系数，经过 Lasso 筛选后，相关系数输出结果如表所示，可看出可知中，发现特征 *X*4、*X*7、*X*7、*X*9、*X*11、*X*12 其变量系数值为零，将这 6 个偏好特征剔除，最终筛选得到 10 个关键偏好特征，如表 6 所示。

Table 5. Information gain value of preference feature
表 5. 偏好特征信息增益值

季度	像素	防水	配件	赠品
2020 年 1 季度	0.00072	0.00861	0.00000	0.00425
2020 年 2 季度	0.00317	0.00407	0.00032	0.00113
2020 年 3 季度	0.00000	0.01342	0.00037	0.01022
2020 年 4 季度	0.00186	0.00993	0.00459	0.00495
2021 年 1 季度	0.00187	0.00455	0.00268	0.00501
2021 年 2 季度	0.00164	0.00236	0.00177	0.00361
2021 年 3 季度	0.00177	0.00168	0.00139	0.00301
2021 年 4 季度	0.00113	0.00179	0.00139	0.00308
2022 年 1 季度	0.00464	0.00153	0.00022	0.00160
2022 年 2 季度	0.00060	0.00186	0.00080	0.00153
2022 年 3 季度	0.00052	0.00157	0.00069	0.00170
2022 年 4 季度	0.00065	0.00094	0.00085	0.00139

Table 6. Lasso regression output values
表 6. Lasso 回归输出值

偏好特征	*X*i	输出值
像素	*X*1	2.255119
防水	*X*2	6.793466
配件	*X*3	0.808191
电池	*X*4	0.000000
功能	*X*5	12.715979
赠品	*X*6	−1.028311
外观	*X*7	−0.000000
夜景	*X*8	−1.250289
价格	*X*9	−0.000000
体积	*X*10	45.518130
镜头	*X*11	−0.000000
色彩	*X*12	0.000000
稳定性	*X*13	8.336630
全景	*X*14	10.947069
防抖	*X*15	−3.977190
画质	*X*16	−0.000000

将 Lasso 筛选过的偏好特征向量分别作为特征向量输入 SVM 中，通过已有的 1 至 12 阶段作为时间序列值，对第 13 阶段进行预测，之后将得到的 13 阶段作为已知数据，用 2 至 13 阶段预测第 14 阶段；以此类推，预测未来 3 期，如下图 4 所示。

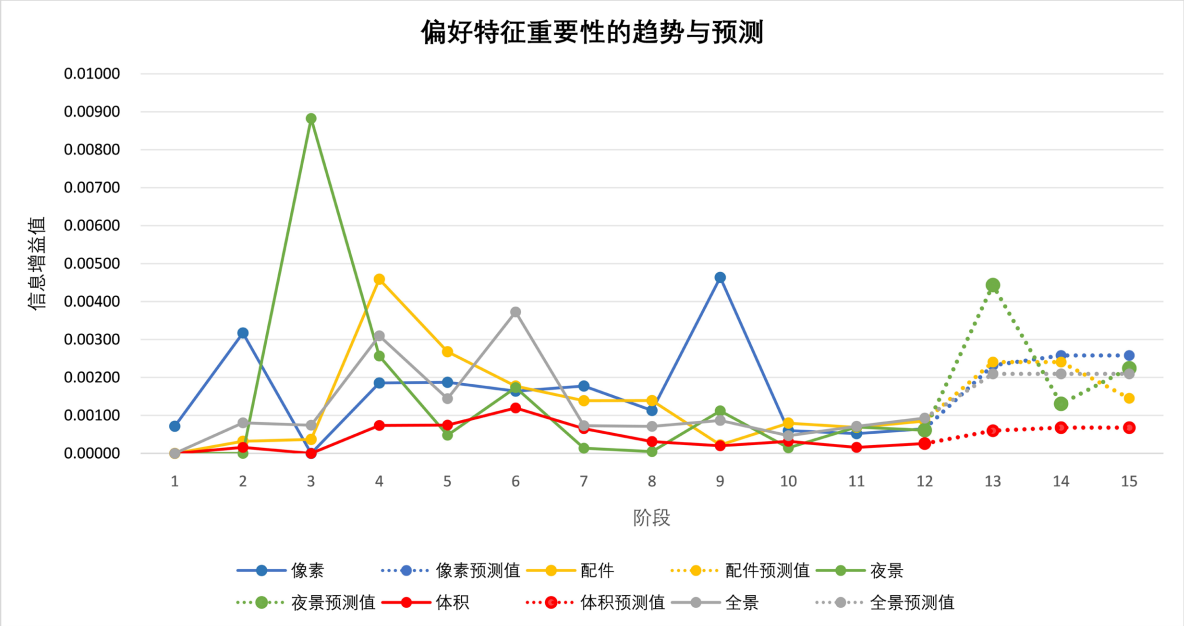


Figure 4. Trends and predictions of the importance of preference features
图 4. 偏好特征重要性的趋势与预测

图 4 展示了“像素”、“夜景”、“体积”、“配件”和“全景”属性在 12 个阶段的变化趋势以及未来 3 阶段预测值。可以观察到“夜景”和“像素”整体信息增益值较高；“全景”在 4、6 阶段信息增益值增长明显；“配件”在第 4 期达到峰值随后逐渐下降；“体积”的信息增益值较低，但呈现缓慢上升趋势。以未来 3 阶段预测值为未来消费者偏好走向，则未来消费者对于“配件”方面的偏好在短暂上升之后继续下降趋势，“像素”和“体积”的偏好在上升。针对“配件”，消费者对于运动相机配件的偏好度在 4 期到 8 期关注度较高，说明在这几个阶段，电商平台上的运动相机的配件参差不齐，而随着时间的推移，大部分的运动相机配件在逐渐改良，这是厂商对相机配件不断加强重视的结果，在 13 到 15 阶段消费者偏好度短暂升高后继续降低，因此未来一阶段厂商不需要将大量的资源集中于相机配件的提升。

4. 讨论

本文将文本聚类、情绪分析与偏好预测结合起来，针对在线评论文本进行消费者偏好预测挖掘。构建的模型适用于电商进行消费者产品偏好趋势分析与预测，从而更好的针对产品属性进行调整优化，使其更符合消费者未来购买倾向。同时以实际在线评论为例验证其有效性，进而帮助商家更加了解用户的不同偏好和变化，了解影响消费者购买行为的影响因素、关系和程度，更加合理精准地制定营销策略，提高营销效果。

本文使用情感分析以及 LDA 模型挖掘消费者情感倾向与偏好特征，在此基础上使用信息增益法与 Lasso 算法等改进 SVM 预测法来进行消费者偏好趋势预测，补充了当前文献中将文本分析、情感分析与偏好特征趋势分析结合起来的空白。本文也扩展了偏好特征提取相关研究，绝大多数研究使用 LDA 模型

倾向于进行客户市场细分、主题提取等方向，本文使用 LDA 进行偏好特征的提取，增加了 LDA 模型的使用方向。此外，大部分研究的预测倾向于使用机器模型或者深度学习模型，本文在 SVM 模型基础上使用 Lasso 模型进行偏好特征筛选也为未来研究提供了新思路。

消费者的在线评论不仅包括评分和文字，还包括一些表情符号和一些图片和视频。因此，将来在提取消费者偏好特征的过程中需要扩大自变量维度，更全面地分析消费者偏好倾向。同时在进行情感分析时，权重评分分配和分层也比较简单，未来可与情感细粒度分析相结合使偏好维度的情感评分更加真实。此外，本文在研究过程中舍弃了一些偏好维度。这些偏好维度在情感倾向上没有明显的波动，但可能对消费者的偏好倾向产生隐性影响。在未来的研究中，我们可以挖掘和预测这样的偏好维度，以探究它是否对消费者偏好产生影响。

参考文献

[1] Zhou, G. and Liao, C.L. (2021) Dynamic Measurement and Evaluation of Hotel Customer Satisfaction through Sentiment Analysis on Online Reviews. *Journal of Organizational and End User Computing*, **33**, 1-27. https://doi.org/10.4018/JOEUC.20211101.oa8

[2] Barta, S., Gurrea, R. and Flavián, C. (2023) Consequences of Consumer Regret with Online Shopping. *Journal of Retailing and Consumer Services*, **73**, Article ID: 103332. https://doi.org/10.1016/j.jretconser.2023.103332

[3] Sim, Y., Lee, S.K. and Sutherland, I. (2021) The Impact of Latent Topic Valence of Online Reviews on Purchase Intention for the Accommodation Industry. *Tourism Management Perspectives*, **40**, Article ID: 100903. https://doi.org/10.1016/j.tmp.2021.100903

[4] Kim, Y.J. and Kim, H.S. (2022) The Impact of Hotel Customer Experience on Customer Satisfaction through Online Reviews. *Sustainability*, **14**, Article 848. https://doi.org/10.3390/su14020848

[5] Zhang, D. and Niu, B. (2024) Leveraging Online Reviews for Hotel Demand Forecasting: A Deep Learning Approach. *Information Processing & Management*, **61**, Article ID: 103527. https://doi.org/10.1016/j.ipm.2023.103527

[6] Pineda-Jaramillo, J. and Pineda-Jaramillo, D. (2022) Analysing Travel Satisfaction of Tourists towards a Metro System from Unstructured Data. *Research in Transportation Business & Management*, **43**, Article ID: 100746. https://doi.org/10.1016/j.rtbm.2021.100746

[7] Yadav, H. and Sagar, M. (2023) Exploring COVID-19 Vaccine Hesitancy and Behavioral Themes Using Social Media Big-Data: A Text Mining Approach. *Kybernetes*, **52**, 2616-2648. https://doi.org/10.1108/K-06-2022-0810

[8] Zhu, Z., Liu, J. and Dong, W. (2022) Factors Correlated with the Perceived Usefulness of Online Reviews for Consumers: A Meta-Analysis of the Moderating Effects of Product Type. *Aslib Journal of Information Management*, **74**, 265-288. https://doi.org/10.1108/AJIM-02-2021-0054

[9] Zhang, J., Lu, X. and Liu, D. (2021) Deriving Customer Preferences for Hotels Based on Aspect-Level Sentiment Analysis of Online Reviews. *Electronic Commerce Research and Applications*, **49**, Article ID: 101094. https://doi.org/10.1016/j.elerap.2021.101094

[10] Zhang, Z., Guo, J., Zhang, H., Zhou, L. and Wang, M. (2022) Product Selection Based on Sentiment Analysis of Online Reviews: An Intuitionistic Fuzzy TODIM Method. *Complex & Intelligent Systems*, **8**, 3349-3362. https://doi.org/10.1007/s40747-022-00678-w

[11] Wang, Z., Gao, P. and Chu, X. (2022) Sentiment Analysis from Customer-Generated Online Videos on Product Review Using Topic Modeling and Multi-Attention BLSTM. *Advanced Engineering Informatics*, **52**, Article ID: 101588. https://doi.org/10.1016/j.aei.2022.101588

[12] Obiedat, R., Qaddoura, R., Ala'M, A.Z., Al-Qaisi, L., Harfoushi, O., Alrefai, M.A. and Faris, H. (2022) Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution. *IEEE Access*, **10**, 22260-22273. https://doi.org/10.1109/ACCESS.2022.3149482

[13] Bhuvaneshwari, P., Rao, A.N., Robinson, Y.H. and Thippeswamy, M.N. (2022) Sentiment Analysis for User Reviews Using Bi-LSTM Self-Attention Based CNN Model. *Multimedia Tools and Applications*, **81**, 12405-12419. https://doi.org/10.1007/s11042-022-12410-4

[14] Lucini, F.R., Tonetto, L.M., Fogliatto, F.S. and Anzanello, M.J. (2020) Text Mining Approach to Explore Dimensions of Airline Customer Satisfaction Using Online Customer Reviews. *Journal of Air Transport Management*, **83**, Article ID: 101760. https://doi.org/10.1016/j.jairtraman.2019.101760

[15] Shi, W., Zhang, J. and He, S. (2023) Understanding Public Opinions on Chinese Short Video Platform by Multimodal

Sentiment Analysis Using Deep Learning-Based Techniques. *Kybernetes*. https://doi.org/10.1108/K-04-2023-0723

[16] 沈超, 王安宁, 陆效农, 彭张林, 张强. 基于在线评论的客户偏好趋势挖掘[J]. 系统工程学报, 2021, 36(3): 289-301.

[17] Yakubu, H. and Kwong, C.K. (2021) Forecasting the Importance of Product Attributes Using Online Customer Reviews and Google Trends. *Technological Forecasting and Social Change*, **171**, Article ID: 120983. https://doi.org/10.1016/j.techfore.2021.120983

[18] Nilashi, M., Ahmadi, H., Arji, G., Alsalem, K.O., Samad, S., Ghabban, F., Alarood, A.A., *et al*. (2021) Big Social Data and Customer Decision Making in Vegetarian Restaurants: A Combined Machine Learning Method. *Journal of Retailing and Consumer Services*, **62**, Article ID: 102630. https://doi.org/10.1016/j.jretconser.2021.102630

[19] Lee, M., Kwon, W. and Back, K.J. (2021) Artificial Intelligence for Hospitality Big Data Analytics: Developing a Prediction Model of Restaurant Review Helpfulness for Customer Decision-Making. *International Journal of Contemporary Hospitality Management*, **33**, 2117-2136. https://doi.org/10.1108/IJCHM-06-2020-0587

[20] Zibarzani, M., Abumalloh, R.A., Nilashi, M., Samad, S., Alghamdi, O.A., Nayer, F.K., Akib, N.A.M., *et al*. (2022) Customer Satisfaction with Restaurants Service Quality during COVID-19 Outbreak: A Two-Stage Methodology. *Technology in Society*, **70**, Article ID: 101977. https://doi.org/10.1016/j.techsoc.2022.101977

[21] Hussain, W., Merigó, J.M., Raza, M.R. and Gao, H. (2022) A New QoS Prediction Model Using Hybrid IOWA-ANFIS with Fuzzy C-Means, Subtractive Clustering and Grid Partitioning. *Information Sciences*, **584**, 280-300. https://doi.org/10.1016/j.ins.2021.10.054

[22] Khorsand, R., Rafiee, M. and Kayvanfar, V. (2020) Insights into TripAdvisor's Online Reviews: The Case of Tehran's Hotels. *Tourism Management Perspectives*, **34**, Article ID: 100673. https://doi.org/10.1016/j.tmp.2020.100673

[23] Wang, H., Liang, T. and Cheng, Y. (2021) Prediction of Perceived Utility of Consumer Online Reviews Based on LSTM Neural Network. *Mobile Information Systems*, **2021**, Article ID: 7054016. https://doi.org/10.1155/2021/7054016

[24] Jain, P.K., Patel, A., Kumari, S. and Pamula, R. (2022) Predicting Airline Customers' Recommendations Using Qualitative and Quantitative Contents of Online Reviews. *Multimedia Tools and Applications*, **81**, 6979-6994. https://doi.org/10.1007/s11042-022-11972-7

[25] Su, Y. and Shen, Y. (2022) A Deep Learning-Based Sentiment Classification Model for Real Online Consumption. *Frontiers in Psychology*, **13**, Article ID: 886982. https://doi.org/10.3389/fpsyg.2022.886982

[26] Chang, Y.M., Chen, C.H., Lai, J.P., Lin, Y.L. and Pai, P.F. (2021) Forecasting Hotel Room Occupancy Using Long Short-Term Memory Networks with Sentiment Analysis and Scores of Customer Online Reviews. *Applied Sciences*, **11**, Article 10291. https://doi.org/10.3390/app112110291

[27] Oh, S., Ji, H., Kim, J., Park, E. and delPobil, A.P. (2022) Deep Learning Model Based on Expectation-Confirmation Theory to Predict Customer Satisfaction in Hospitality Service. *Information Technology & Tourism*, **24**, 109-126. https://doi.org/10.1007/s40558-022-00222-z

[28] Olmedilla, M., Martínez-Torres, M.R. and Toral, S. (2022) Prediction and Modelling Online Reviews Helpfulness Using 1D Convolutional Neural Networks. *Expert Systems with Applications*, **198**, Article ID: 116787. https://doi.org/10.1016/j.eswa.2022.116787

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

[30] 王婷婷, 韩满, 王宇. LDA 模型的优化及其主题数量选择研究——以科技文献为例[J]. 数据分析与知识发现, 2018, 2(1): 29-40.

[31] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x