

基于BERT-LDA模型的消费者在线评论研究

李智, 陈郁

上海工程技术大学纺织服装学院, 上海

收稿日期: 2024年5月22日; 录用日期: 2024年6月12日; 发布日期: 2024年8月12日

摘要

本研究旨在通过文本挖掘方法研究消费者的需求和偏好。通过收集和预处理天猫商城的服装商品的在线评论数据, 应用BERT-LDA模型进行分析, 发现消费者在购物体验、服装特性和服装品质方面呈现出多样化的关注度和情感积极率。研究结果表明, 虚拟试穿等新型产品体验方式将深刻影响消费者的购买决策。消费者提高了对服装的可持续性的关注程度, 倾向于选择实用性强、易于回收利用, 且能“一衣多穿”的服装。基于该研究结果, 本文为服装电商行业的市场营销提供了有益的参考和指导。

关键词

服装, 在线评论, 文本挖掘, 消费需求, LDA模型, BERT模型

Research on Consumer Online Reviews Based on BERT-LDA Model

Zhi Li, Yu Chen

School of Textiles and Fashion, Shanghai University of Engineering Science, Shanghai

Received: May 22nd, 2024; accepted: Jun. 12th, 2024; published: Aug. 12th, 2024

Abstract

The purpose of this study is to study consumers' needs and preferences through text mining methods. By collecting and preprocessing online review data of clothing products on Tmall and applying BERT-LDA model for analysis, it is found that consumers show diversified attention and positive emotional rate in terms of shopping experience, clothing characteristics and clothing quality. The results show that new product experience methods such as virtual trying on will profoundly affect consumers' purchasing decisions. Consumers are paying more attention to the sustainability of clothing, and tend to choose clothes that are practical, easy to recycle, and can be worn more than once. Based on the research results, this paper provides useful reference and guidance for

the marketing of apparel e-commerce industry.

Keywords

Clothes, Online Reviews, Text Mining, Consumption Demand, LDA Model, BERT Model

Copyright © 2024 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

功能性服装在运动实践中作为必要的装备, 随着消费者需求的多样化而备受瞩目。然而, 由于普通消费者在选择功能性齐全的服装时缺乏相对专业知识, 尤其是在进行网络购物时, 通常优先根据先前消费者发布的在线评论来做出决策。在数字化零售时代, 这种基于在线口碑的消费方式已经广泛存在。在线评论不仅为消费者的购物决策提供了重要的参考依据。服装品牌将根据该在线评论调整电商销售计划, 以更好地满足市场需求。

目前的研究大多基于问卷调查和结构方程模型分析, 该方法使消费者易受到社会期望的影响, 导致回答结果与实际行为或态度存在偏差。此外, 针对大量在线评论数据和复杂的变量结构, 传统文本挖掘方法难以准确触达消费者的潜在倾向和真实需求。为解决上述问题, 本研究结合 BERT 模型和 LDA 模型的优势, 根据新增的服装评价维度词向量, 以有效地处理在线评论数据并进行情感分析和评价维度建模, 进而揭示消费者的真实电商消费需求和偏好。研究结果表明, 虚拟试穿等新型产品体验方式将深刻影响消费者的购买决策。消费者越来越注重服装的可持续性, 更倾向于选择实用性强、易于回收利用的多功能服装。本文还为服装品牌提出建议, 以更好地满足消费者需求并提高产品竞争力。

2. 研究现状和方法

2.1. 服装消费需求现状研究

针对服装电商的消费需求和穿着偏好研究非常丰富。Lauren 等[1]采用多元回归方法进行假设检验, 整合计划行为理论(TPB)提出服装消费者购买意愿模型, 发现消费者态度和感知行为控制对购买服装的意愿有显著的正向影响。Hong 等[2]采用结构方程模型分析方法, 研究关于服装复购意向数据, 发现影响顾客复购意向的关键因素包括产品的穿着动机和功能性。其中男性和女性顾客对产品的满意程度也有所不同, 男性更注重服装的功能性, 而女性则更加关注服装如何融入日常生活以及其设计美学。Park 等[3]基于定性方法采用先例分析和个体采访方式, 探讨婴儿潮一代对服装的需求和偏好。研究结果发现婴儿潮一代期望在运动时避免暴露身体, 且更青睐透气有弹性的面料, 对服装制造过程有道德关注, 强调环保制造; 不偏好本土制造, 不喜欢豪华品牌或价格过高的服装。

上述研究大多基于问卷法或结构方程模型调查分析结果, 消费者容易受到社会期望的影响, 存在回答与其实际购买行为或态度不一致的问题。由于需要预先定义研究变量之间的不同关系, 对于未知的大量在线文本数据和复杂的变量结构, 难以挖掘潜在的消费者倾向和真实偏好需求。

2.2. 文本挖掘现状研究

相比于通过假设检验等方法来研究消费者需求, 消费者的主观在线评论更能反映出真实情况。网购

平台产生了涵盖不同购买情境、产品特性和消费需求的大量在线评论, 其语言表达即时地反馈了消费者体验。在互联网背景下, 文本挖掘的优势在于处理冗杂的在线评论数据更高效便捷。先前的研究采用 K-means 聚类[4]和基于自然语言处理的文本挖掘[5]等技术, 来提取消费者的特征并进行情感分析。Mangi [6]等人采用隐马尔可夫模型的情感分析方法, 通过语义聚类信息定义隐藏变量, 从而拟合计算句子的情感取向。但对长距离依赖关系的文本建模能力有限, 且情感在文本分布中的复杂动态性导致对序列标注任务中内部结构理解不足。Arvind 等人[7]应用双向编码器表示的合成注意力, 并结合梯度提升分类器, 用于对电子产品评论数据集中的情感极性进行分类和预测, 解决了准确捕捉相应的细粒度情感问题。但在效率和正确率二者中不能兼顾, 单一模型的运行上存在局限性。Baris 等人[8]对主题模型提取常用的潜在狄利克雷分配算法(Latent Dirichlet Allocation, LDA) [9]进行改进, 提出了用于提取电子产品评论方面的无监督句子分段(SS-LDA)算法。由于无需任何注释作为训练数据, 仅用情感词典作为输入的测试效果, 在精确率上有较好的提升。

随着深度学习的发展和更多研究领域的涉足, 对于在线评论的价值挖掘方法在不断精益求精。传统的主题建模方法产生的评价维度重叠和解释模棱两可的局限性越发明。最近研究者[10] [11]提出的双向编码表征算法 BERT 和 LDA 模型的结合, 很好地通过增强评价维度连贯性和情感分析来解决这类问题。BERT-LDA 模型在电子产品、在线评论、社交网络等领域充分应用, 该方法较少出现在产业链复杂且亟待可持续发展的纺织服装行业。由此, 服装在线评论的价值挖掘具有广阔的发展前景, 其研究结果将有助于推动服装品牌向零售数字化方向迈进。

2.3. 研究方法

因此本文按照以下 4 个步骤进行研究: 首先, 对消费者的在线评论进行收集和预处理。然后, 应用 BERT-LDA 模型的文本挖掘和评价维度建模, 对聚集到不同维度的数据进行统计分析。第三步, 进行情感分析并构建情感字典, 训练情感分析模型。识别消费者对每个不同类别关注维度积极或消极的情绪, 使满意度结果可视化。最后, 分析评价维度的消费者情感积极率结果, 总结消费者在购买时的真实需求, 为服装电商的未来生产过程提供指导性建议。

3. 构建服装评价维度

本文使用 python 网络爬虫采集了天猫商城从 2020 年 11 月至 2023 年 11 月, 电商在线购物平台零售服装链接下公开的消费者评论信息。筛选出 18 款有效评论数量大于 5000 条的服装套装产品, 其中包含迪卡侬、李宁、安踏等 6 个品牌, 经初步筛选各品牌提取 5000 条左右数据, 最终保留在线评论数据 30,395 条。由于通过网页抓取收集的数据包含不相关的信息, 使用 python 中的正则表达式删除非文本内容、特殊的非字符和标点符号, 以避免影响后续步骤。采用 Python 中的 jieba 分词工具包, 将长句拆分成多个独立的单词。并设置大量无实际含义的中文停用词, 对服装在线评论数据进行去除停用词和分词处理。以 Tf-idf 值较高词汇作为表征评价维度的备选词, 最终得到 384 个可用评价维度词汇。

对在线评论文本数据集进行评价维度建模中, 常用困惑度作为衡量文本语言概率模型性能的指标。该指标根据各词汇来估计一句话出现的概率, 其公式为:

$$PP(S) = P(w_1, w_2, \dots, w_N)^{\frac{1}{N}}$$

通过该公式计算模型的测试数据集中每个文档的平均对数似然来进行评估, 越低表示预测能力越好, 一般随着评价维度数量的增加而降低。在对服装文本数据计算困惑度后发现, 当评价维度数设置为 9 时, 模型可解释性较高, 能够把相关服装评价维度词汇聚集在同一评价维度内。因此本文选择 $K = 9$ 作为初

始评价维度数。

困惑度以概率为基础, 评估模型整体性能。但从服装电商知识的角度分析, 评价维度和服装关联更为重要。由于困惑度仅代表基础评价维度, 对特定领域如服装电商评价常用词汇特征无法准确捕捉, 难以涵盖在线评论文本中的语义信息。为解决这个问题, 本文根据 BERT-LDA 模型对评价维度重新进行聚类, 通过 pyLDAvis 可视化工具包实现图 1 中的评价维度聚类可视化。这种交互式探索功能有效地挖掘了服装领域的主要评价维度, 并详细展示各评价维度的词汇含量比例。图 1 中左侧气泡大小表示评价维度出现频率, 气泡之间距离越远, 则相似性就越低。右侧表示评价维度中排名前 30 的特征词。浅蓝色代表整个文档中各词汇的出现频率, 而深红色代表各词汇在特定评价维度中的权重。与右侧相邻的是可调参数 λ , 用于修改词汇和评价维度之间的相关性, λ 越接近 1 则表示模型对于文本数据的理解程度越高。

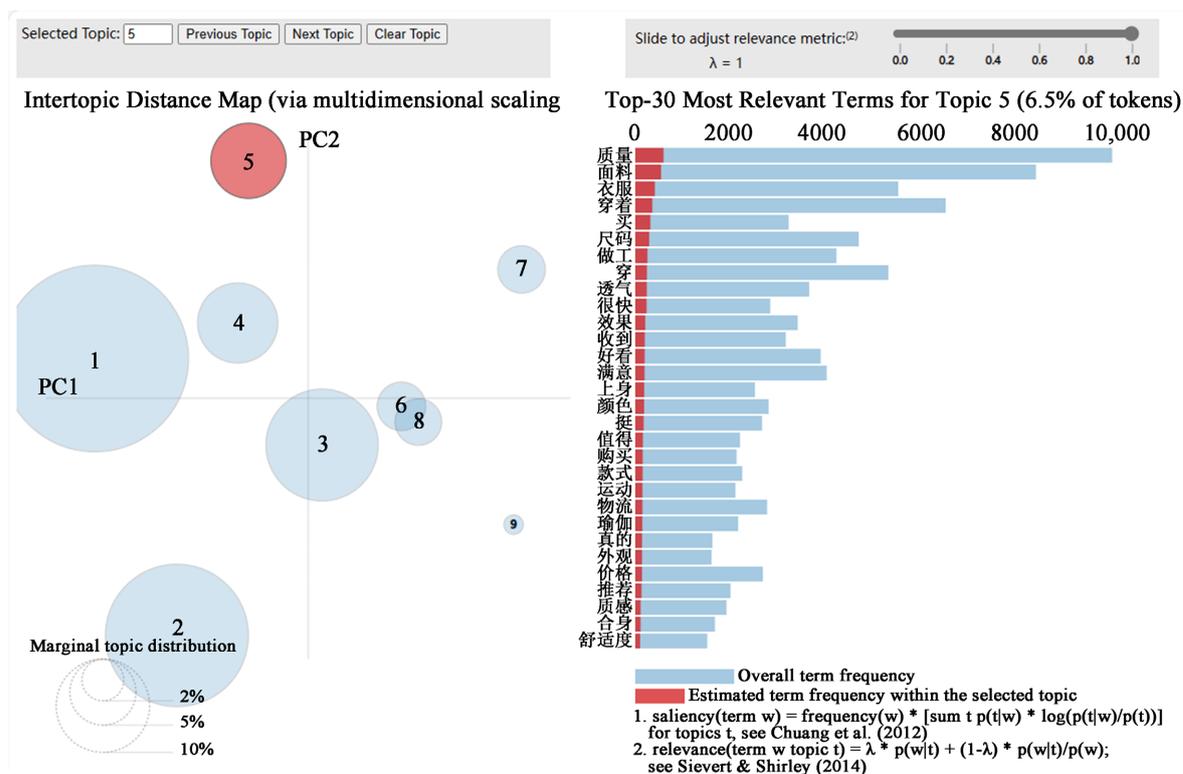


Figure 1. Evaluation dimension visualization results

图 1. 评价维度可视化结果

通过观察图 1 评价维度可视化的表征词汇, 对半自动化标注的不同评价维度的聚类结果进行分析。依据服装电商网购服务性质对“产品服务”和“物流服务”评价维度进行合并, 对应图 1 中第 6 和 8 簇。“面料质量”“热湿舒适性”“尺码样板”是服装产品共有的评价维度。据参考文献[12]中功能性服装重点关注其合身设计的研究方式, 服装的“产品性能”将和“款式设计”一同作为服装特性的二级维度。根据研究[13]将“购物体验”再次细分为“网购服务”和“产品体验”。第 9 簇聚类词汇“耐久性”融入“服装品质”类中, 可将其删除。因此, 从最初的 9 个评价维度变为表 1 中主要的 7 个, 构成了服装电商在线文本评价维度。

4. 分析消费评价维度

为研究消费者对评价维度的关注优先级和情感积极率, 根据上述构建的 7 个主要服装电商在线评价

维度, 对消费者文本评论数据进行情感分析。首先构建自定义的基础词典, 其中包括基础情感词典和服装领域词典。基础情感词典源自大连理工大学情感本体库, 共选取了 25,648 个情感词。这些情感词被归类为 7 个大类和 21 个子类, 情感强度从弱到强分为 1、3、5、7、9 五个等级。初始收集的服装在线文本评论中, 有不符本文模型研究的评价维度和未含情感词典标注的在线文本评论, 将其删除后最终包含有效的评论条数总和为 24,661 条。依据 6 类服装品牌案例的在线评论情感倾向标注结果, 计算出每类服装电商各维度的情感积极率情况, 如表 1 所示。

Table 1. The positive emotional rate of each dimension of the case product

表 1. 案例产品各维度情感积极率

产品编号	面料质量	热湿舒适性	尺码样板	款式设计	产品性能	网购服务	产品体验
A	76.56	85.58	84.01	84.40	81.50	86.42	61.82
B	82.98	87.74	79.54	80.46	84.61	92.86	69.49
C	83.74	74.56	86.32	84.82	92.31	90.08	94.53
D	86.40	80.17	96.39	87.29	92.00	86.86	81.95
E	90.09	80.31	88.19	88.12	97.84	74.53	67.33
F	71.85	92.73	77.44	92.17	88.07	76.41	95.87

分析表 1 得出以下结论: 服装能够在“面料质量”和“热湿舒适性”等方面表现出色, 但对于电商类目的“网购服务”和“产品体验”方面表现略差, 例如 A 和 E 品牌。反之, C 和 F 品牌在“产品体验”上有数值较高的情感积极率, 但在“面料质量”和“尺码样板”方面表现不突出。这一情况的出现分析认为, 服装电商需要在保证面料品质的前提下控制成本, 进而会牺牲一些电商营销的资金和人力。服装电商的稳定发展难以均衡包括服装的外观、功能性和品牌体验等多方面因素。针对此类情况, 需要及时跟踪市场动态调整, 权衡各因素的重要性和影响力, 找到适合服装电商发展的着力点, 并根据市场反馈和服装销售数据不断调整和优化营销策略。例如 D 品牌, 在各消费者评价维度上情感积极率数值都达到良好且平衡, 才能使品牌服装更有竞争力。

通过对各服装情感倾向标注的结果进行汇总计算, 得到如表 2 所示各评价维度的关注度与情感满意度。

Table 2. Attention and emotion of each evaluation dimension

表 2. 各评价维度关注度与情感情况

名称	一级维度				二级维度		
	关注度	情感积极率/%	情感中性率/%	情感消极率/%	名称	关注度(组内排序)	情感积极率/(组内排序)
服装品质	0.2470	85.48%	9.24%	5.28%	面料质量	0.2770 (3)	87.45% (1)
					热湿舒适性	0.3322 (2)	87.22% (2)
					尺码样板	0.3908 (1)	83.89% (3)
服装特性	0.4115	87.72%	6.82%	5.46%	款式设计	0.4280 (2)	90.90% (1)
					产品性能	0.5720 (1)	85.34% (2)
购物体验	0.3411	77.46%	11.48%	11.06%	网购服务	0.5829 (1)	87.27% (1)
					产品体验	0.4171 (2)	63.75% (2)

分析表 2 中服装评价维度数据, 可得以下结论:

1) 服装评价维度关注度从高到低依次是“服装特性”“购物体验”“服装品质”, 关注度值分别为 0.4115、0.3411、0.2470, 这表明消费者对于产品整体情感倾向, 正在从“服装品质”相关维度向“服装特性”转变。这一情况的出现分析认为与服装满足个人审美需求有关, 即消费者通过浏览服装电商产品详情页中的宣传视频、买家秀图片等服装展现方式, 来初步判断哪类服装更具有吸引力。由于面料详细参数和尺码表等文字信息, 难以让未触摸过产品的消费者对服装的舒适性做出准确的判断, 因此服装的款式美观性逐渐成为新的关注焦点。

2) 服装购物体验的“网购服务”维度在情感积极率和关注度中排名靠前, 分析认为与传统服务相比, 在线评论的及时性、全面性和互动性为消费者提供服装体验参考, 可提高购物决策的效率和准确性。“产品体验”维度情感积极率仅有 63.75%, 其中“虚拟试穿”一词也出现在“产品体验”维度表征词表中, 这类新产品体验方式将影响消费者的购买意愿。虚拟试穿提供新的体验营销, 吸引消费者的注意力和兴趣。通过虚拟试穿应用程序, 消费者可将试穿效果分享到社交媒体上, 这在一定程度上满足了消费者对服装的“购物体验”需求, 也扩大了品牌的曝光度和影响力。

3) “服装特性”中的二级维度“产品性能”排名靠前, 依据结果分析认为消费者更关注服装使用的可持续性。选择可再生、可回收和环保的材料是可持续服装的基础。消费者更愿意选择实用性强, 易修复回收的服装, 这样可以减少购买新服装的频率。相比之下, “款式设计”通常和个人时尚搭配有关, 对服装的实用性、功能性影响较小, 消费者会优先考虑选择“产品性能”作为评价服装的重要指标。此外, “款式设计”情感积极率高达 90.9%, 分析认为服装设计需要考虑到其在运动过程中对身体的保护功能, 避免瑜伽运动过程中产生不适或受伤。优秀的设计细节例如口袋设计、内置文胸、腰部剪裁等更好地满足了消费者对“服装特性”的需求。

4) “服装品质”中, 消费者对于“尺码样板”的关注度在排名中最高, 但情感积极率最低为 83.89%, 而“面料质量”的关注度在排名中最后, 但情感积极率最高, “热湿舒适性”适中。综合分析该数据结果得出, 高质量的面料可能会让消费者感到满意和愉悦, 但对面料质量的认知程度较低, 可能并不了解面料的具体特性。尺码直接关系到穿着者的身体感受, 因此更容易引起消费者的重视。在“尺码样板”合适的前提下, 消费者通常对具有良好舒适性和质感的面料产生积极的情感体验。

综上所述, 服装在具备功能性的同时, 还需要保证款式设计的创新性, 以及维护好消费者的购物体验。具体建议如下:

在购物体验方面, 为提升网购消费者的整体情感积极率, 本文建议采用增强现实 AR 技术或虚拟试穿技术, 让服装“穿着”在消费者身体上, 增强试穿的真实感和互动性。品牌在虚拟试穿平台中增加社交分享功能, 让消费者分享到社交媒体上获得交流意见与反馈, 从而提升品牌服装的购物体验。

在服装特性方面, 消费者更关注服装的“产品性能”维度中的“可持续”表征。建议品牌优化生产流程, 采用节能、低碳的生产工艺, 减少废水排放对环境的影响。设计服装时采用“一衣多穿”的理念, 可适用于不同场合活动, 使得服装具有时尚感的同时, 满足运动、休闲和日常生活的需要。通过践行可持续发展理念来提升服装品牌的社会责任形象, 吸引更多消费者对服装的认同和支持。

从服装品质方面, 消费者对“面料质量”有 87.45%的情感积极率。本文建议服装电商选用有机棉、竹纤维、再生聚酯等环保材料。在服装上标识清晰的材料成分和回收标识, 方便消费者识别和分类处理废弃服装。设计服装时考虑到可回收性和再利用性, 例如可拆卸装饰物、拉链、扣子等, 方便在服装报废后将不同部件分开处理, 提高回收利用率, 推动循环经济发展。可持续发展的服装设计不仅满足了消费者对“服装品质”的要求, 还有助于实践环保意识和可持续生活方式。

5. 结语

通过对服装在线评论的文本挖掘, 采用 BERT-LDA 模型揭示了消费者对服装电商的消费偏好。研究发现虚拟试穿这类新产品体验方式将影响消费者的购买意愿, 消费者更注重服装的可持续性, 即倾向于选择实用性强、易于回收利用和多场景搭配的服装。基于此, 本文提出了一系列针对服装电商的建议。在“购物体验”方面, 建议采用虚拟试穿技术提升网购消费者的整体购物体验。在“服装特性”方面, 设计多功能的“一衣多穿”服装。在“服装品质”方面, 建议使用环保服装材料, 考虑可持续发展的设计理念。

此外, 基于目前收集的部分数据和产品案例, 还不能囊括消费者对服装需求的全部特征。因此在未来的研究中, 通过增加产品文本数据来更好地了解消费者的习惯和需求, 从用户角度出发进行更精细的分析。结合大数据和机器学习模型, 准确地预测市场动态, 从而帮助服装电商更合理地安排生产和管理。

基金项目

上海市教育委员会东方学者项目(TP2017074)。

参考文献

- [1] Watts, L. and Chi, T. (2018) Key Factors Influencing the Purchase Intention of Activewear: An Empirical Study of US Consumers. *International Journal of Fashion Design, Technology and Education*, **12**, 46-55. <https://doi.org/10.1080/17543266.2018.1477995>
- [2] Lee, H.J. (2023) A Study on Korean Customers' Intentions to Repurchase for the Sustainable Growth of the Athleisure Market. *Sustainability*, **16**, Article 69. <https://doi.org/10.3390/su16010069>
- [3] Park, S. (2021) Exploring Yoga Wear Needs for the Baby Boomer Generation. Ryerson University Library and Archives.
- [4] Vincent, O.R., Makinde, A.S., Salako, O.S. and Oluwafemi, O.D. (2018) A Self-Adaptive K-Means Classifier for Business Incentive in a Fashion Design Environment. *Applied Computing and Informatics*, **14**, 88-97. <https://doi.org/10.1016/j.aci.2017.05.002>
- [5] Nawaz, Z., Zhao, C., Nawaz, F., Safeer, A.A. and Irshad, W. (2021) Role of Artificial Neural Networks Techniques in Development of Market Intelligence: A Study of Sentiment Analysis of eWOM of a Women's Clothing Company. *Journal of Theoretical and Applied Electronic Commerce Research*, **16**, 1862-1876. <https://doi.org/10.3390/jtaer16050104>
- [6] Kang, M., Ahn, J. and Lee, K. (2018) Opinion Mining Using Ensemble Text Hidden Markov Models for Text Classification. *Expert Systems with Applications*, **94**, 218-227. <https://doi.org/10.1016/j.eswa.2017.07.019>
- [7] Mewada, A. and Dewang, R.K. (2022) SA-ASBA: A Hybrid Model for Aspect-Based Sentiment Analysis Using Synthetic Attention in Pre-Trained Language BERT Model with Extreme Gradient Boosting. *The Journal of Supercomputing*, **79**, 5516-5551. <https://doi.org/10.1007/s11227-022-04881-x>
- [8] Ozyurt, B. and Akcayol, M.A. (2021) A New Topic Modeling Based Approach for Aspect Extraction in Aspect Based Sentiment Analysis: SS-LDA. *Expert Systems with Applications*, **168**, Article 114231. <https://doi.org/10.1016/j.eswa.2020.114231>
- [9] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **27**, 680-691.
- [10] Tan, X., Zhuang, M., Lu, X. and Mao, T. (2021) An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model. *IEEE Access*, **9**, 15860-15871. <https://doi.org/10.1109/access.2021.3052566>
- [11] Venugopalan, M. and Gupta, D. (2022) An Enhanced Guided LDA Model Augmented with BERT Based Semantic Strength for Aspect Term Extraction in Sentiment Analysis. *Knowledge-Based Systems*, **246**, Article 108668. <https://doi.org/10.1016/j.knosys.2022.108668>
- [12] Liu, R., Little, T., Williams, and Eugene, M. (2012) Evaluation of Elite Athletes Psycho-Physiological Responses to Compression Form-Fitted Athletic Wear in Intensive Exercise Based on 5Ps Model. *Fibers and Polymers*, **13**, 380-389. <https://doi.org/10.1007/s12221-012-0380-9>

- [13] Rose, S., Hair, N. and Clark, M. (2011) Online Customer Experience: A Review of the Business-to-Consumer Online Purchase Context. *International Journal of Management Reviews*, **13**, 24-39.
<https://doi.org/10.1111/j.1468-2370.2010.00280.x>