基于机器学习的电商店铺分类研究

——以服装业为例

# 选题背景（国内、国外）

中国现在已经是世界第二大经济体，其互联网科技水平也正处于世界领先地位。随着我国互联网普及率和网络购物用户数量的增加，其网络购物交易的规模也在不断扩大。根据CNNIC发布的第45次《中国互联网络发展状况统计报告》显示，截止到2020年3月，网络零售已成为我国消费增长的重要动力。

进入互联网时代，人们的衣食住行有了翻天覆地的变化。人们越来越注重生活品质，在网络购物时除了会货比三家以外，还会查看购买者的评论信息。商品信息和评价信息已被消费者认为是网络购物的重要信息来源。然而，根据中国消费者协会发布的调查报告【1】显示，大约有72.4%的受访者表示曾遭遇平台或商家“默认好评”，还有一部分的人遭遇过商家主动联系要求修改点评结论的评分。因此消费者与商家之间就存在商品信息不对称的问题，导致消费者网络购物满意度不高。虽然消费者偏向将更好的评分给予点评结论，但文本评价却更能表达消费者的真情实感【2】。如何让消费者在海量信息中找到合适自己的店铺，从而提升消费者购物满意度，这是不少平台和商家急需解决的问题。

文本处理是自然语言处理在应用领域中最基础的技术，文本评论中包含很多消费者的真实感受，比如说，货品是否货不对板，质量如何，是否会再次购买等。徐頔基于分词技术为用户提高在微信订餐的检索速度和准确性【3】。牟向前,王庆军基于词频分析技术挖掘健身APP的用户体验信息并为APP提供发展建议【4】。那么在网络购物的场景中，可以通过分词算法对文本评论进行整理，用于评价商家的商品使用情况，为后续的店铺分类建模起关键作用。分类算法的目的是生成学习器，再利用学习器把数据映射到给定类别中的某一个类别。分类算法常应用于多种场景中，比如岩性分类、文本分类、电子音乐分类等【5,6】。本课题将结合商品的定价、材质、风格和用户使用评价等因素对商家进行有效的分类。

# 选题意义

目前，随着互联网的发展，互联网经济不断衍生出新的商业模式，如直播购物、视频服务等。根据中国消费者协会发布的《2020年上半年全国消协组织受理投诉情况分析》显示【7】2020年上半年经常出现网络购物投诉问题，主要投诉的前四个方面有：一是网购商品质量低劣，存在安全隐患。二是合同违约。比如缺斤少两，产品与宣传不符等。三是平台在线客服不能及时跟进处理。四是配送不规范。这意味着消费者的权益还未得到完善体系的保障。因此，完善网络购物的消费体系，能切实帮助消费者解决实际问题，不仅可以让消费者结合自己的需求和大数据提供的参考来做决策，还有利于互联网经济的健康、可持续发展。

在互联网蓬勃发展的背景下，信息也日益增多，本论文结合文本处理和分类算法，提高店铺分类的精度和效率，为提高消费者的满意度提供了理论基础，并为进一步深入研究消费者需求和商家完善自身产品提供一定的参考价值。

# 国内外研究现状

目前，我国有不少关于文本处理与分类算法的研究 。2018年，张琰、朱燕翔、郑桂玲【8】基于jieba分词工具和TF-IDFs算法计算中文文本的关键词，再通过关键词词频分析得到消费者对产品属性的评价。2019年，鲍继彬【9】通过使用隐马尔可夫、Word2ve方法和常用的分类算法对中文文本进行分类，并在此基础上通过集成学习提升分类器的精度，提升后的精度可达90%以上。相对于国外研究，2019年ChunminLang【10】等人使用LDA模型对消费者的评价进行了深入分析。2020年FengXu【11】等人基于朴素贝叶斯理论在电商产品评论文本的情感分析做了两个方面的优化：一是提高了处理大规模评论文本的计算效率，二是强化学习器的学习能力。OLADAPO OYEBODE【12】等人基于分词技术对Google Play和App Store 的用户评论进行情感分析和主题分析，并为优化应用程序提出有效建议。

# 研究目的与方法

论文的研究目的是基于商品的定价、材质、风格和消费者在网络购物的文本评价等因素对商家进行有效的数据挖掘并分类，从而提高消费者的消费满意度和商家的商品质量。

本论文主要用到了四种研究方法，分别是文献综合法、跨学科研究法、比较分析法和描述性研究法。首先，文献综合法为论文研究的前期提供资料，能让我充分了解前人的研究方向、方法和结论，并且在此基础上思考创新点。其次，之所以运用跨学科研究法，是因为该论文涉及多学科的理论知识，比如说，数学、计算机和数据科学等学科知识。通过跨学科研究法，论文能够取其精华，去其糟粕，提高论文的科学性和有效性，从而形成一篇优秀的论文。然后，比较分析法是论文中期最重要的研究方法，会基于同一个数据集来比较不同的分类算法，从而得到相对较优的结果，增强课题的说服力。最后，描述性研究法是本论文最后涉及的研究方法，会通过数据可视化的方法来呈现结果，从而提高论文的可读性。

# 文章结构

1. 绪论
   1. 选题背景及意义
      1. 研究背景
      2. 研究意义
   2. 国内外研究现状
      1. 国内研究现状
      2. 国外研究现状
   3. 研究方法
2. 基础知识或相关算法
   1. 文本处理技术
   2. 分类算法的原理
3. 核心内容
   1. 中文分词
   2. 分类算法的比较
   3. 性能指标
4. 实例研究和分析
   1. 展示数据
   2. 数据预处理
   3. 算法的实现
   4. 实验结果和分析
5. 总结与未来工作
6. 参考文献
7. 附录

[1]张贵峰. 网购被“默认好评”有损信用消费环境[N]. 中国商报,2019-03-20(P02).

[2]武新丽,李柔,令狐大智.电子商务用户评价信息有效性研究——基于淘宝服装类商品[J].沿海企业与科技,2019(06):8-13.

[3]徐頔. 中文分词技术在微信订餐导航中的应用[D].沈阳理工大学,2018.

[4]牟向前,王庆军.基于词频分析的健身APP用户体验现状及发展建议[J].体育学刊,2020,27(02):64-68.

[5]康乾坤,路来君.随机森林算法在测井岩性分类中的应用[J].世界地质,2020,39(02):398-405.

[6]易伶.多特征融合和机器学习算法的电子音乐分类模型[J].微型电脑应用,2020,36(09):117-119.

[7][http://www.cca.org.cn/zxsd/detail/29728.html ，2020年10月26](http://www.cca.org.cn/zxsd/detail/29728.html%20，2020年10月26)日浏览。

[8]张琰,朱燕翔,郑桂玲.基于网购评论文本挖掘的手机类产品属性评价研究[J].现代商贸工业,2018,39(22):49-51.

[9]鲍继彬. 基于机器学习的网购平台产品评论情感分析[D].哈尔滨工业大学,2019.

[10]Chunmin Lang, Muzhen Li, Li Zhao,Understanding consumers’ online fashion renting experiences: A text-mining approach,Sustainable Production and Consumption,Volume 21,2020, Pages 132-144,ISSN 2352-5509,https://doi.org/10.1016/j.spc.2019.12.003.

[11]Feng Xu, Zhenchun Pan, Rui Xia,E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework,Information Processing & Management,Volume 57, Issue 5,2020,102221,ISSN 0306-4573,https://doi.org/10.1016/j.ipm.2020.102221.

[12]O. Oyebode, F. Alqahtani and R. Orji, "Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews," in IEEE Access, vol. 8, pp. 111141-111158, 2020, doi: 10.1109/ACCESS.2020.3002176.