

学士学位论文

论文题目：严格控制在25个汉字（符）以内，如题目语意未尽，可用副题名补充说明报告论文中的特定内容，论文题目字体采用一号黑体，居中书写，副题名字体采用小一号黑体，副题名字体另起一行，并采用1.25倍行距。

写论文作者姓名。姓名之间没有空格，少数民族或外籍作者姓名，样例：哈里森•沃尔德伦三号（16 pt）仿宋。居中。

填写论文作者学号。Times New Roman 三号（16 pt）。

姓 名

学 号

分别填写论文作者所在院系和专业。三号（16 pt）仿宋。居中。

专业示例：生物科学

院 系

年 级 20XX级/20XX级（专升本）

专 业 生物科学

指导教师

填写论文作者的第一指导教师。姓名之间没有空格，三号（16 pt）仿宋。居中。

年 月 日

此部分填写论文成文打印的日期，用三号（16 pt）宋体汉字，不用阿拉伯数字。年份示例：

二〇二二年，年月日之间不留空格

学位论文原创性声明

本人所提交的学位论文《××××××××××××》，是在导师×××的指导下，独立进行研究工作所取得的原创性成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中标明。

本声明的法律后果由本人承担。

论文作者（签名）： 指导教师确认（签名）：

年 月 日 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解衡水学院有权保留并向国家有关部门或机构送交学位论文的复印件和磁盘，允许论文被查阅和借阅。本人授权衡水学院可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其它复制手段保存、汇编学位论文。

**如果不需要此行，请删除。**

保密的学位论文在\_\_\_\_\_\_\_年解密后适用本授权书。

论文作者（签名）： 指导教师（签名）：

年 月 日 年 月 日

# 摘 要

本文旨在通过基于评论文本的研究方法，探索消费者对干某一产品或服务的满意度，并为个业提供改进建议，以提高客户满意度和忠诚度，实现商业价值的最大化。文章首先介绍了评论文本的研究背景和意义，接着，我们对国内外相关研究现状进行了简要分析发现消费者对于产品或服务的满意度有着各种各样的维度和影响因素。最后，我们设计了一个基于评论文本的研究框架，包括数据收集、文本预处理、情感分类和TF-IDF,LDA主题模型等步骤，并以某电商平台的用户评论数据为例进行了实证分析，识别消费者对产品满意度影响因素，研究发现,消费满意度主要体现在对产品的信任感知、营销感知、质量感知、物流服务和价格感知5大因素,质量感知与信任感知是影响消费者满意度差评的主要因素。通过机器学习与数据挖掘识别出不同维度的消费者满意度的指标,并提出了针对性的改进建议。文章的研究成果对于企业改善产品或服务质量，提高客户满意度和忠诚度具有重要的实践意义。

关键词：机器学习；情感分类；LDA；TF-IDF；满意度感知

ABSTRACT

This paper aims to explore consumer satisfaction with a certain product or service through a research method based on review text, and provide improvement suggestions for the industry to improve customer satisfaction and loyalty and maximize business value. The article first introduces the research background and significance of review texts, and then, we briefly analyze the current research status at home and abroad and find that consumers' satisfaction with products or services has various dimensions and influencing factors. Finally, we designed a research framework based on comment text, including data collection, text preprocessing, sentiment classification and TF-IDF, LDA topic model and other steps, and conducted an empirical analysis using the user comment data of an e-commerce platform as an example , to identify the influencing factors of consumers' product satisfaction. The research found that consumer satisfaction is mainly reflected in the five major factors of product trust perception, marketing perception, quality perception, logistics service and price perception. Quality perception and trust perception are the two factors that affect consumers. The main factor of poor satisfaction rating. Through machine learning and data mining, the indicators of consumer satisfaction in different dimensions are identified, and targeted improvement suggestions are put forward. The research results of the article have important practical significance for enterprises to improve product or service quality, improve customer satisfaction and loyalty.

**Key words：**Machine Learning；Sentiment Classification；LDA；TF-IDF；Perception of satisfaction

# 目 录

[摘 要 I](#_Toc128841907)

[目 录 III](#_Toc128841908)

[第1章 绪论 1](#_Toc128841909)

[1.1 研究背景和意义 1](#_Toc128841910)

[1.2 国内外研究现状及其分析 1](#_Toc128841911)

[1.3 研究内容、研究方法及思路 2](#_Toc128841912)

[1.3.1 研究内容 2](#_Toc128841913)

[1.3.2 研究方法及思路 2](#_Toc128841914)

[第2章 在线评论文本内容对消费者满意度影响的理论基础 4](#_Toc128841915)

[2.1消费者满意度相关理论 4](#_Toc128841916)

[2.1.1 消费者满意度概念 4](#_Toc128841917)

[2.1.2 消费者满意度指数模型 4](#_Toc128841918)

[2.1.3 期望不一致理论 4](#_Toc128841919)

[2.1.4 双因素理论 4](#_Toc128841920)

[2.2 文本挖掘相关理论 5](#_Toc128841921)

[2.2.1 文本属性特征提取相关理论 5](#_Toc128841922)

[2.2.2 文本情感分析相关理论 6](#_Toc128841923)

[第 3 章 基于评论文本的消费者满意度影响因素提取 8](#_Toc128841924)

[3.1 产品在线评论数据抓取及预处理 8](#_Toc128841925)

[3.1.1 数据采集 8](#_Toc128841926)

[3.1.2 数据清洗 8](#_Toc128841927)

[3.1.3 数据分词 9](#_Toc128841928)

[3.2 生鲜农产品消费者满意度影响因素提取 10](#_Toc128841929)

[3.2.1 LDA 主题模型简介 10](#_Toc128841930)

[3.2.2 主题结果分析 11](#_Toc128841931)

[3.3 消费者满意度文本数据量化 11](#_Toc128841932)

[3.3.1 情感词典的构建 12](#_Toc128841933)

[3.3.2 情感值计算 13](#_Toc128841934)

[3.3.3 情感值量化结果 13](#_Toc128841935)

[3.4 本章小结 13](#_Toc128841936)

[第 4 章 在线评论文本内容对消费者满意度影响的实证分析 13](#_Toc128841937)

[4.1 研究假设的提出 13](#_Toc128841938)

[4.1.1 产品质量因素对消费者满意度的影响及假设提出 14](#_Toc128841939)

[4.1.2 商家服务质量因素对消费者满意度影响及假设提出 14](#_Toc128841940)

[4.1.3 物流服务质量对消费者满意度的影响及假设提出 14](#_Toc128841941)

[4.1.4 感知价值对消费者满意度的影响及假设提出 14](#_Toc128841942)

[4.1.5 感知价值的中介作用及假设提出 14](#_Toc128841943)

[4.1.6 消费者满意和不满意影响因素的比较分析及假设提出 14](#_Toc128841944)

[4.2 模型建立 14](#_Toc128841945)

[注释 16](#_Toc128841946)

[参考文献 18](#_Toc128841947)

# 第1章 绪论

## 1.1 研究背景和意义

消费者满意度一直是企业关注的重要指标，消费者的满意度直接关系到企业的竞争力和市场份额。随着互联网和电商的快速发展，消费者对于商品和服务的需求越来越多元化，而电商平台已经成为消费者获取商品和服务信息、进行购物的重要途径。然而，电商平台上的商品数量庞大、价格差异大、同类商品质量也参差不齐，消费者难以直接了解商品质量和售后服务的真实情况，因此，消费者的满意度成为电商平台和企业关注的重要指标。

传统的消费者满意度研究多采用问卷调查、电话访谈等方式，但是这些方法存在着一定的局限性，因为消费者往往难以准确表达自己的感受和需求。而基于评论文本数据的消费者满意度研究，可以通过挖掘和分析消费者在电商平台上的评价文本，更全面、客观地了解消费者对于商品和服务的评价，为电商平台和企业提供更具参考价值的数据支持，进而优化产品和服务，提高消费者满意度。

在电商平台上，消费者的评价文本涵盖了大量的信息，包括对商品的评价、对服务的评价、对价格的评价等。基于评论文本数据的消费者满意度研究可以挖掘出这些信息，从而更加全面地了解消费者的需求和偏好，为企业提供更具有针对性的产品和服务。

同时，基于评论文本研究消费者满意度，具有重要的实践意义和研究价值。随着自然语言处理、文本挖掘和机器学习等技术的不断发展，越来越多的研究者开始探索基于评论文本的消费者满意度研究方法，为实现精准营销、提高客户满意度提供了新的思路和方法。此外，基于评论文本的消费者满意度研究还可以为消费者提供更加客观、全面的购物参考，促进电商行业的健康发展。

综上所述，基于评论文本研究消费者满意度，可以更加全面、客观地了解消费者的需求和偏好，为企业提供更具有针对性的产品和服务，为消费者提供更加客观、全面的购物参考，促进电商行业的健康发展。因此，基于评论文本的消费者满意度研究具有重要的实践意义和研究价值，可以为企业提供更加准确的市场调研数据和产品设计思路，为消费者提供更加优质的商品和服务体验，同时也可以促进电商平台和企业之间的互利共赢关系的建立和发展。

## 1.2 国内外研究现状及其分析

近年来，随着电子商务行业的快速发展，越来越多的消费者倾向于通过网络平台进行购物消费。与此同时，网络上的评论文本也逐渐成为衡量消费者满意度的重要指标。因此，基于评论文本的消费者满意度研究逐渐受到国内外学者的重视。

在国外，基于评论文本的消费者满意度研究已经逐渐成熟。例如，美国密歇根大学的Xiong等人提出了一种基于情感分析的评论文本挖掘方法，用于研究消费者对于不同电商平台的满意度。此外，英国格拉斯哥大学的Karami等人也对于基于评论文本的消费者满意度进行了探讨。

在国内，基于评论文本的消费者满意度研究也逐渐兴起。例如，南开大学的何迪等人采用主题模型方法对于微博上的旅游评论进行主题分析，研究消费者对于不同旅游景点的满意度。此外，中南财经政法大学的蒲钢等人也采用基于情感分析的方法对于淘宝网上的商品评论进行分析，研究消费者对于不同商品的满意度。

综上所述，基于评论文本的消费者满意度研究已经成为国内外学者关注的热点问题，目前的研究主要集中在情感分析、主题模型、机器学习等方面，这些方法已经被广泛应用于不同的电商平台和商品领域。

## 1.3 研究内容、研究方法及思路

### 1.3.1 研究内容

主要研究内容是探讨利用消费者在网络平台上发表的评论文本，通过对其进行语义分析和情感计算，来评估消费者对商品或服务的满意度。该研究首先采用数据爬虫技术，从网络平台上获取大量的消费者评论文本，并对其进行预处理，包括去除标点符号、停用词和数字等，同时进行词性标注、分词和词形还原等操作，以便进行下一步的关键词提取和情感分析。

在关键词提取方面，该研究采用了TF-IDF算法和LDA模型，分别从文本的高频词汇和主题出发，提取关键词，并通过情感分析技术计算每个关键词的情感值。在情感计算方面，该研究采用了情感词典和情感分类器，对每个关键词进行情感极性和强度的计算，并通过加权平均的方式，对每条评论文本的情感值进行综合计算。最后，该研究通过可视化的方式展示了商品或服务的满意度水平，为企业提供了指导意见，帮助其提供更加优质的商品和服务体验。

总之，基于评论文本研究消费者满意度是一个非常有意义的研究方向，它可以通过挖掘消费者在网络平台上的评论信息，了解消费者的需求和偏好，优化产品设计、改进营销策略，提高企业的竞争力和用户满意度。

### 1.3.2 研究方法及思路

1. 数据收集：收集与研究主题相关的评论文本数据。这些评论文本数据可以来自于各种电商平台、社交媒体平台等，例如淘宝、京东、微博、微信等。

2. 数据预处理：对于收集到的评论文本数据进行预处理，包括去除无关信息、分词、去除停用词、词性标注等。

3．特征提取：将预处理后的文本数据转换成计算机可处理的特征形式。常用的特征提取方法包括词袋模型、TF-IDF模型等。

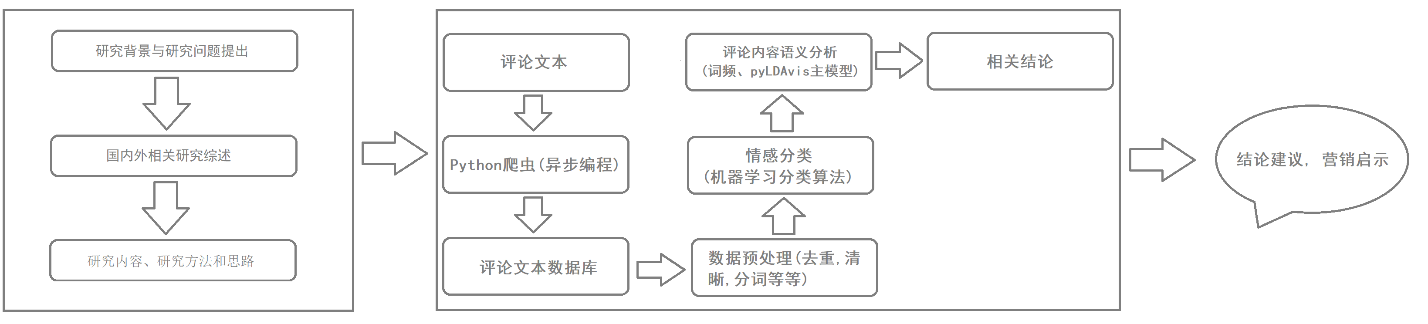
4. 情感分析：采用自然语言处理、机器学习等技术，对于提取出的文本特征进行情感分析，得到文本的情感极性。

5. 主题分析：采用主题模型等方法，对于提取出的文本特征进行主题分析，得到文本的主题分布。

6. 消费者满意度分析：将文本的情感极性和主题分布与消费者满意度进行关联分析，得出消费者对于不同维度不同方面的满意度评价。

通过以上的研究内容，可以对于消费者的满意度进行深入的研究和分析，得出有效的结论和建议，从而为电商企业提供更加准确的市场调研数据和产品设计思路，为消费者提供更加优质的商品和服务体验。

本研究的技术路线如下图 1-1 所示



图表 1-1 论文流程图

# 第2章 在线评论文本内容对消费者满意度影响的理论基础

## 2.1消费者满意度相关理论

### 2.1.1 消费者满意度概念

早在 1965 年，美国学者 Cardozo 将消费者满意度引入营销领域以来，学者们对消费者满意度做了大量研究，消费者满意度早已概念化。Cardozo[62]通过实验研究了顾客满意，他指出顾客满意度受购买产品所花费的精力以及对产品的期望的影响，它可以对消费者重复购买产生促进作用；Oliver[63]认为消费者满意度是期望一致—期望不一致的函数，它的产生基于消费者购买前的预期与购买后的情感体验的比较的评估。因此本文对消费者满意度的定义采用 Oliver 的观点，即消费者满意度是消费着对产品和服务的体验与其期望是否一致的情感状态。

### 2.1.2 消费者满意度指数模型

早在 1965 年，美国学者 Cardozo 将消费者满意度引入营销领域以来，学者们对消费者满意度做了大量研究，消费者满意度早已概念化。Cardozo[62]通过实验研究了顾客满意，他指出顾客满意度受购买产品所花费的精力以及对产品的期望的影响，它可以对消费者重复购买产生促进作用；Oliver[63]认为消费者满意度是期望一致—期望不一致的函数，它的产生基于消费者购买前的预期与购买后的情感体验的比较的评估。因此本文对消费者满意度的定义采用 Oliver 的观点，即消费者满意度是消费着对产品和服务的体验与其期望是否一致的情感状态。

### 2.1.3 期望不一致理论

期望不一致理论是学者 Oliver[63]提出的有关消费者满意度的理论，该理论认为消费者购买产品和服务时会将其购买后的体验与购前期望相比较，如果购买后的体验超出购前期望，则消费者就会感到满意，如果购买后的体验低于购前期望，则消费者会感到不满意。期望不一致理论的重点在于将消费者期望与实际感知绩效的差异作为判断满意不满意的依据。

### 2.1.4 双因素理论

双因素理论是赫兹伯格[69]提出的有关工作满意度的著名理论，也成为“激励—保健”因素。赫兹伯格认为满意和不满是两个相互独立且不同的连续事物，他将导致工作满意的因素称为激励因素，激励因素通常与工作自身内容性质相连，如赞赏、晋升、责任感等，它能够对员工起到激励作用；赫兹伯格将引起员工不满的因素成为保健因素，保健因素通常与工作以外的因素相连，如工作环境、人际关系等，它会导致员工产生不满。赫兹伯格认为激励因素与满意相连，若激励因素得到满足，员工会产生巨大的满足感，但是其缺失可能并不会导致消费者不满；相反保健因素与不满意相连，它的满足可能不会导致员工产生巨大的满足感，

但是一旦它缺失，将会导致员工强烈的不满。因此，满意与“没有满意”而不是“不满意”相互对立，“不满意”与“没有不满意”而不是“满意”相互对立，因此员工对工作满意的提升只能通过激励因素来实现，员工对工作不满意的改善只能通过保健因素来实现。在双因素理论里，保健因素是先决性和基础性条件，它的存在在于防止员工积极性的下降和不满的产生；激励因素十分重要，它将起到激励员工获得满足感的作用。双因素理论在满意度理论已经得到广泛应用.

随着研究的深入，学者们不仅依据该理论来研究员工满意度，还将其扩展到其他领域，如研究消费者、读者、游客满意度等：比如陈平[70]

将双因素理论运用到高校图书馆的读者满意度上，发现读者满意和读者忠诚分别是保健因素和激励因素，并针对性地提出了建议；苗志娟[71]

基于双因素理论为零售企业促进顾客满意度提出了建议；Kim[72]基于双因素理论来识别比较影响酒店消费者满意和不满意的影响因素，发现影响消费者满意的因素和影响消费者不满意的因素不同；李志勇[74]

以双因素理论的视角来研究博物馆游客的满意度，发现只有激励因素对

游客整体满意度具有显著正向影响等，因此本文将该理论应用到生鲜农产品消费者上，来研究影响消费者满意和不满的激励因素和保健因素。

## 2.2 文本挖掘相关理论

在线评论文本数据是一种半结构化或非结构化的数据类型，其中蕴含着大量有价值的信息和内容。这些评论文本包含了消费者对产品或服务的态度和看法信息，可为消费者购买决策和企业提升产品服务质量提供参考。因此，利用文本挖掘技术帮助企业改进服务已成为研究的热点。在线评论文本挖掘主要应用于提取在线文本属性特征和进行情感分析等方面。

### 2.2.1 文本属性特征提取相关理论

在线文本属性特征提取研究的是消费者评论文本中涵盖的商品、服务和其他关键属性的观点。这些观点对于企业了解其不足之处非常有帮助，并且可以为潜在的消费者提供购物前的参考信息。因此，文本特征和属性的提取至关重要。主题分析是一种常见的方法，可用于提取文本语料库中隐含的语义信息。因此本文利用 LDA 主题分析 对在线评论中产品或服务特征进行提取。

### 2.2.2 文本情感分析相关理论

情感分析是利用计算机技术、文本分析技术以及自然语言处理技术对文本中的情绪、态度和情感倾向进行提取、挖掘、归纳推理的方法。它可以在短时间内对大量的文本数据进行情感态度的提取，是文本挖掘领域一个重要的研究领域。它可以从文档、句子、词语和属性的层次角度来情感分析所提取的文本。从文档角度进行情感分析就是对整篇文档进行情感分析，将文档按照积极情感和消极情感分类；句子层次的情感分析则是分析一句话的情感倾向；词语角度则是将词语短语表达的情感进行分类；属性角度的情感分析则是抽取并量化文本中评价对象的属性情感值的过程。情感分析的方法有很多种，这些方法可以总结为以情感词典为基础和以机器学习为基础的情感分析两类。文本情感分析具有较大的实用价值，在学术界得到广泛利用。

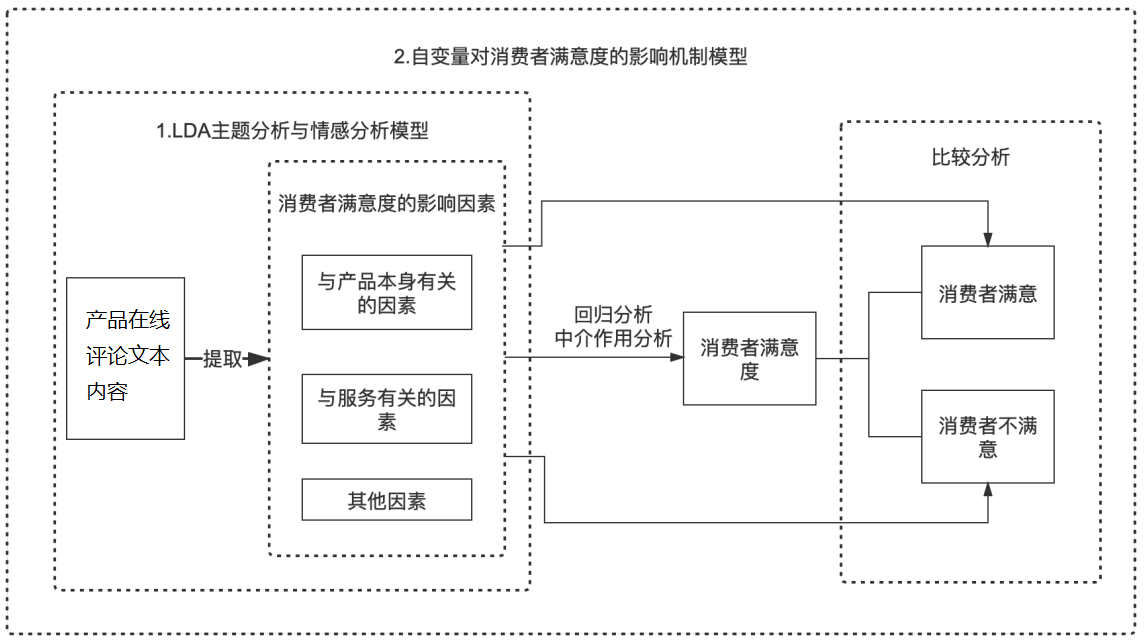
（1）以情感词典为基础的情感分析由于利用情感词典的情感分析是利用事先构建情感词典来对文本、句子词语进行情感分析，因此该方法需要首先对情感词典进行构建。该方法是通过比较句中积极情感词和消极情感词的个数差异来计算句子的情感倾向，如果句中积极情感词的数量超过了消极情感词的数量，此时该句的情感值为正，说明句子的情感是正向的；反之，如果句中消极情感词的数量比积极情感词的数量更多，说明该句的情感值为负，说明句子的情感偏消极；如果积极情感词的数量和消极情感词的数量相等，此时情感值为 0，说明该句子是中性的。因此通过情感词典我们可以快速识别出句子的情感倾向，且操作容易、算法简便，因此受到了学者们的青睐，不少学者利用了情感词典对文本进行情感分析，李鸣[75]等通过构建酒店领域情感词典来对酒店在线评论进行细粒度情感分析，来研究酒店特征与消费者满意度之间的关系；杨春晓[76]则利用构建好的情感词典情感分析了卷烟的在线评论。

（2）以机器学习为基础的情感分析以机器学习为基础的情感分析是利用计算机的某些算法对人工训练好的模型进行情感分类的方法，其使机器像人一样“学习”，这种方法的优势在于其结果与情感词的存在与否没有关系，因此其具有较高的准确率，在机器学习领域，K-最近邻法（KNN）、支持向量机(SVM)、朴素贝叶斯（NB）、最大熵(EM)等是常用的方法。同样基于机器学习的情感分析方法同样在研究中得到了广泛的使用。Baltas[77]等通过比较不同方法对 Twitter 文本情感分析的效果，发现朴素贝叶斯的效果比决策树和逻辑回归效果要好。Xiang[78]等利用朴素贝叶斯来分析著名的三个购物网站的在线评论文本的情感倾向，发现评论有用性与文本的情感倾向等通过将 PMI 和支持向量机相结合的方法，情感分析了商品评论;利用支持向量机对维吾尔文进行了情感分类。

2.3 木桶原理

木桶原理是一个管理学概念，它是由美国管理学家彼得提出的，它的核心观点是一只水桶盛水的多少由筒壁上最短的那块木板决定，而不是由筒壁上最高的木板决定，要想使木桶多盛水，就需要补齐那块短板，因此木桶原理又可称为短板理论。对于一个企业或组织来说，木桶上最短的木板就代表企业或组织最薄弱的地方，它代表了整个企业或组织的发展水平，因此企业要想做大做强，必须想办法补上短板。而随着网络购物的蓬勃发展和在线评论机制的建立，消费者可以畅所欲言，任意将自己的评价展示在网购平台上，表达自己的满意不满意，消费者通过在线评论表达的有关企业产品或服务的不满和意见可以被看做是该企业的“短板”，该“短板”使消费者产生不满意，进而影响到消费者忠诚度对企业的忠诚度和回购意愿，从而制约企业的发展。因此对于电商企业来说，通过在线评论找出影响消费者满意度的因素，进而找出企业的“短板”十分重要。因此文本将通过文本挖掘技术基于在线评论文本对消费者满意度的影响因素进行探索，并发掘生鲜电商的“短板”所在，帮助生鲜电商来改进产品质量和服务质量，从而促进生鲜电商的发展。

2.4 理论模型的建立

通过对国内外现状及相关理论整理，结合前人的研究成果确立了产品消费者满意度分析的研究模型，如下图 2-5 所示，图 2-5 展示了本研究的研究思路和各个变量之间的关系，依据以下框架，我们进行了第 3 章和第 4 章的研究，研究过程如下所述：首先根据文本挖掘相关理论，基于 LDA 主题模型和基于知网情感词典的情感分析来对产品在线评论文本进行挖掘，并将挖掘出的主题作为影响消费者满意度的可能性因素，根据以往研究经验，提取的因素往往与产品和服务相关；然后依据在线评论和消费者满意度研究成果，将提取的影响因素作为自变量，将消费者满意度作为因变量，根据提出的假设进行实证分析，包括回归分析、中介作用分析等，来分析提取的影响因素对消费者满意度的影响；最后再根据现有文献研究，将消费者满意度分为消费者满意和消费者不满，通过比较分析来探究消费者满意和不满意因素的差异。

图表 2-5 在线评论文本内容对消费者满意度的影响研究理论模型

2.5 本章小结

通过综合国内外的现状和相关理论，结合前人的研究成果，确定了产品消费者满意度分析的研究模型，该模型如图2-5所示，展示了本研究的研究思路和各个变量之间的关系。依据该模型，我们进行了第3章和第4章的研究。具体研究过程如下：首先，根据文本挖掘相关理论，采用LDA主题模型和基于知网情感词典的情感分析来挖掘产品在线评论文本，并将挖掘出的主题视为影响消费者满意度的潜在因素。根据以往的研究经验，提取的因素通常与产品和服务有关。其次，根据在线评论和消费者满意度研究成果，将提取的影响因素作为自变量，将消费者满意度作为因变量，依据提出的假设进行实证分析，包括回归分析、中介作用分析等，以分析提取的影响因素对消费者满意度的影响。最后，根据现有的文献研究，将消费者满意度分为满意和不满意两种情况，通过比较分析来探究影响消费者满意度和不满意度的因素的差异。

# 第 3 章 基于评论文本的消费者满意度影响因素提取

## 3.1 产品在线评论数据抓取及预处理

目前有很多平台销售该产品，在对多个平台进行比较之后，本文决定选择京东平台销售数据来作为研究对象，由于产品种类繁多，没办法对所有产品的数据进行收集分析，因此本文选取了京东农夫山泉东方树叶整箱装500ml\*15瓶产品的评论文本数据。 京东商场在线评论页面每页显示 10 条评论，一共显示 100 页，消费者可以自由选择阅读好评、中评和差评。因此为了保证研究的准确性，本研究收集了上述产品共4种产品按好评、中评和差评划分并按时间顺序的所有评论信息作为研究对象。

### 3.1.1 数据采集

本文选用 Python 爬虫程序作为数据抓取的工具，该程序可以在短时间内获取大量格式化数据，它爬取迅速，容错性强，自定义丰富，并且能够对爬取的数据进行处理，非常方便，本文利用 Python 爬虫数据共抓取了 13000 条评论数据， 每种类别产品的评论数量汇总如下表 3-1 所示。

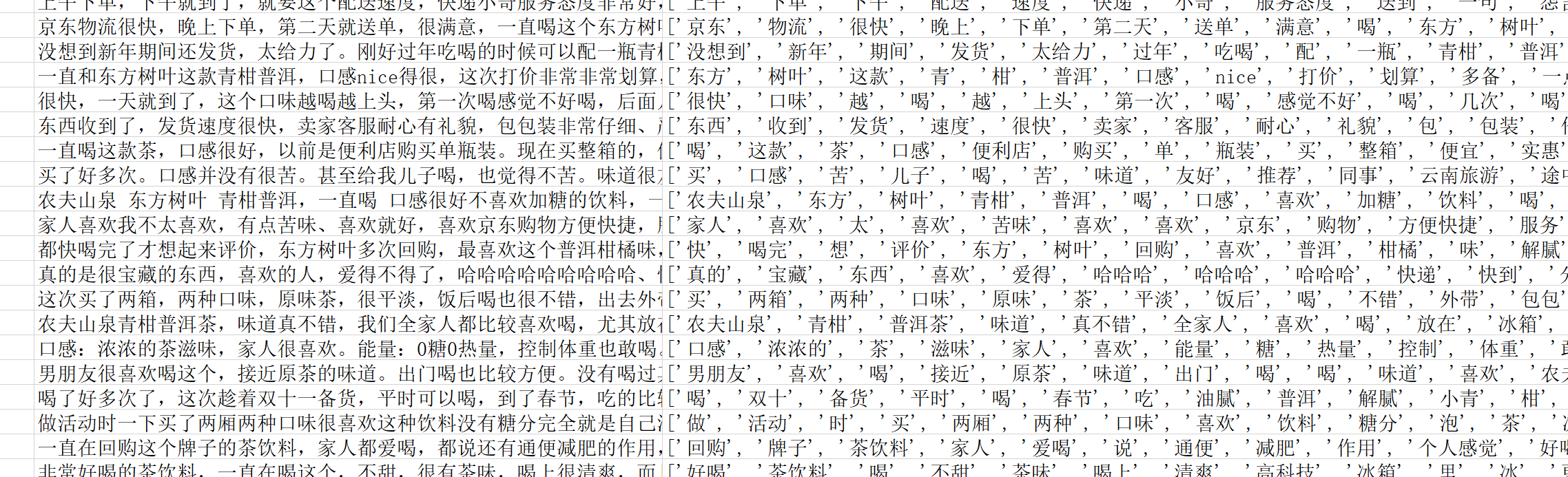
### 3.1.2 数据清洗

为了确保研究结果的准确性，并更好地对在线评论文本进行分析，需要对搜集到的文本数据进行清洗，以去除大量的无关数据、重复数据和无效数据等，这些数据可能会对研究结果产生影响。本研究主要从以下两个方面对文本进行分析：首先，对搜集到的评论文本进行清洗，删除没有实际意义的无关数据、重复数据和无效数据，然后对文本数据进行分词，再对分词数据进行清洗，去除一些无效和无意义的分词。经过文本清洗，本研究得到了5926条评论数据。表3-2展示了清洗掉的部分无效数据及其分类。

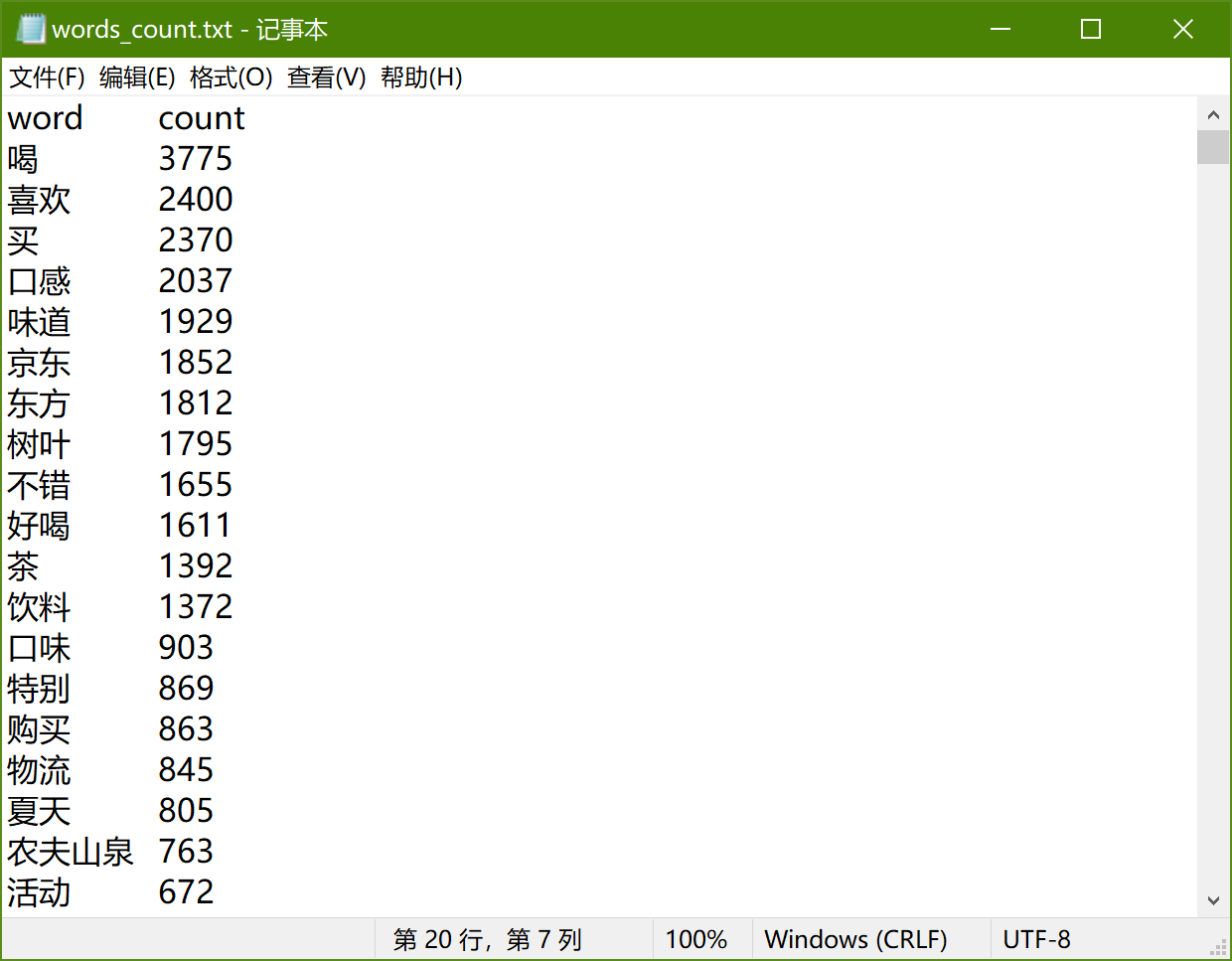
|  |  |
| --- | --- |
| 类别 | 评论文本 |
| 数字类 | 1113322 |
| 符号类 | @###！！ |
| 无效文本类 | 无效文本类 |
| 空值类 |  |

### 3.1.3 数据分词

由于分词是对文本分析的基础，因此首先我们对清洗好的文本数据进行分词。 目前普遍利用“jieba”分词工具来对文本进行分词，因此本文选用 jieba 对数据 进行分词操作，处理结果如下表 3-3 所示。



在对文本进行分词之后，得到的分词中同样也有一些无意义的词语，这些词语主要是一些没有实际意义的语气词、程度副词等，比如“的”、“了”、“么”、 “具体地说”、“另一方面”等，需要将这些词加入停用词表。运行 Python 程序，我们得到 8033个分词数据的词频统计表，下表展示了 排名前 20 的分词结果，如表 3-4 所示



图表 3-4

运用全部的词频结果，制作了词云图如图 3-1 所示。



词云图 3-1 显示了商品评价中高频词语的词云分布，词语的频数越大，该词云图上该词语的字号越大，因此从图 3-1 中我们可以看出，“好喝”、“口感”、 “物流”、“口味”等词语的频数较大，说明这些词语在消费者撰写评论时经常提及到，是消费者购买生鲜农产品时关注的重点。为了进一步了解消 费者满意度的影响因素，本文利用 LDA 主题模型来对在线评论文本进行分析， 提取出影响消费者满意度的关键因素。

## 3.2 生鲜农产品消费者满意度影响因素提取

### 3.2.1 LDA 主题模型简介

由于消费者在撰写评论时，往往是先确定几个主题，然后围绕确定的主题来 发表自己的意见和看法，来表达自己的满意和不满意，其主题就是消费者在购物 过程中关注的重点或者导致消费者满意不满意的因素，因此我们利用主题模型来 对在线评论文本内容进行主题分析，挖掘出在线评论中隐藏的主题，LDA 就是 其中最受关注的主题模型之一。因此本文利用 LDA 主题模型来对在线文本内容 进行主题提取。

LDA（Latent Dirichlet Allocation）是一种基于概率图模型的主题模型，可用于对文本数据进行主题建模。LDA模型的基本假设是文本中的每个单词都来自于某个主题，而每个文档则是由多个主题构成的概率分布。该模型使用无监督学习的方法，自动从文本中学习主题并确定每个文档的主题分布和每个主题的词分布。

借助 LDA 模型，一篇文档可以由以下几个步骤生成：

首先要抽取某个文档（m=1,2,····,n）的主题，根据参数为α的狄利克雷分布来生成“文档-主题”分布，它是一个多项式分布；接着根据,随机生成主题作为文档中第 n 个词语的主题；然后对于再根据参数为 β 的狄利克雷分布来找到主题下词语的分布，其同样是一个多项式分布；最后根据来找到对应于该主题的词语。因此通过 LDA 模型，将一篇文档变成了“文 档—主题”分布和“主题—词语”分布，这样我们就可以清楚地找出文档的隐含主题和主题对应的词语，从而达到了对文本分析的目的。

下图 3-2 所示为 LDA 生成文档的过程：

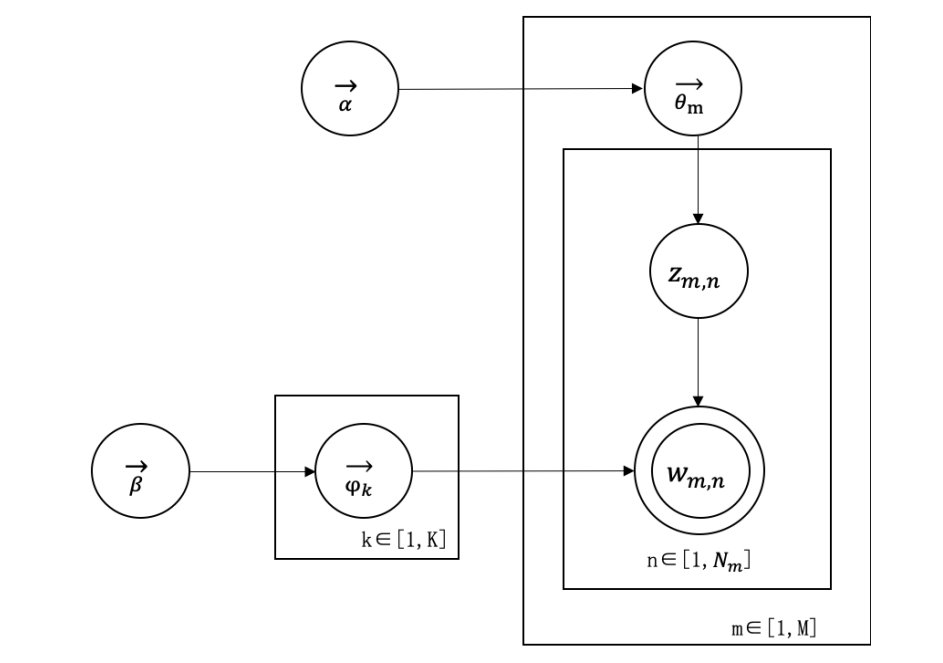


图 3-2 基本 LDA 文档生成原理

因此，通过 LDA 主题模型，我们可以清楚地知道文本的主题和主题对应的词语，对于在线评论来说，我们需要了解消费者在评论中阐述的重点，通过 主题分析，我们可以比较直观地了解评论中的主题，这往往代表着消费者的观 点和关注的重点，这些主题可能就是导致消费者对产品或服务不满的原因，可 以帮助企业改进服务，提高消费者满意度，因此本文利用 LDA 主题模型来对 在线评论文本内容进行主题分析，并将提取的主题作为可能影响消费者满意度 的因素，来对消费者满意度进行探索。

3.1 中处理好的数据导入到 LDA 模型中进行模型训练，对于 LDA 主题 数的选取，虽然有计算困惑度、HDP 等方法，但是目前总常用的还是基于人工观察和主观判断方法，它具有操作性强的优点。因此本研究通过人工观察的方法来确定本文的主题个数，对在线评论文本内容进行主题分析，找出在线评论文本内容中蕴含的主题，并以此作为影响消费者满意度的因素，为下文的研究做铺垫。

### 3.2.2 主题结果分析

为了得到比较理想的结果，本研究将主题数设定为 5-10，并通过观察这 5 个 LDA 的结果，发现当主题数为 9 时 LDA 模型的结果较好，因此本研究设定 LDA 主题模型的主题数为 9，各个主题的命名以及主题下词语以及概率的大小如下 3 个表所示。

## 3.3 消费者满意度文本数据量化

接下来，我们通过情感词典对提取的 9 大主题进行情感分析，将定性数据转 化成定量数据。我们主要利用知网情感词典对评论文本内容及提取的主题进行情 感值计算，通过对在线评论进行观察，可以将在线评论内容的结构总结为“特征 词+情感词”、“特征词+否定词+情感词“、“特征词+情感副词+情感词”三类。 比如一条评论的内容为“玉米口感很软糯，价格划算，物流不快”，按其结构可 以分解为“口感+很+软糯”、“价格+划算”和“物流+不+快”，正好符合上述 三种结构，因此情感分析实质上就是通过计算修饰特征词的情感词和副词的值来 算出该特征词的情感倾向。 由于 LDA 模型只是大体上对文本进行分类，可能会存在各个主题下的特征 词对应不准的情况，因此本文根据 LDA 主题结果对 LDA 每个主题词下的特征 词进行人工调整，使得各个主题下的主题词更加清晰贴合主题，方便接下来进行 的情感分析，整理的结果如下表 3-8 所示。

在对商品进行评价时，评价所用的情感词和程度词的不同都会影响特征词和 评论文本的情感倾向，因此本研究利用知网情感词典，结合生鲜农产品创建了本 研究专有情感词典，对抽取的特征词进行赋值。

### 3.3.1 情感词典的构建

正负面情感词典、否定词词典以及程度副词词典构成了本研究的情感词典。 （1）基础情感词典 本文首先合并修改了词典中的正面情感词和评价词，将两者统一为正面情感 词典，接着按同样的方法形成负面情感词典。在合并的过程中，根据实际情况加 入一些词典中没有的词语，最终获得 10274 个正面情感词，13759 个负面情感词， 以此作为情感词的情感倾向性分析的基础参考词典。 （2）否定词词典 由于否定词具有否定的含义，因此否定词可以使文本和特征词的情感倾向发 生转变。比如“牛肉不新鲜”中就因为加入了否定词不导致对牛肉新鲜的评价由 正面变成了负面。本文归纳整理了 71 个常见的否定词，词典中的部分否定词如 下表 3-9 所示。

表格 3-9 部分否定词

|  |  |
| --- | --- |
| 否定词 | 分值 |
| 不、勿、否、无、非、不能、不是、从不、不好、几乎不、不要、禁止、没有、很少、绝不、木有、永不、从未 | -1 |

（3）程度副词词典 在程度词的修饰下，文本的情感倾向会增强或者减弱，例如“羊肉非常新鲜” 和“羊肉比较新鲜”，虽然都是对羊肉新鲜特征的评价，但是由于“非常”和“比 较”这两个程度词的不同导致这两句话对新鲜的评分完全不同。因此本文采用知 网情感词典中的程度级别词语作为程度副词词典，并将各个程度副词按照情感强 度进行赋值，范围在 0.5-3 之间，共 219 个程度副词，表 3-10 中列出了部分程度 副词及其分值。

|  |  |
| --- | --- |
| 程度副词举例 | 强度 |
| 有些、略微、偏、出头、不为过… | 0.5 |
| 半点、不大、轻度、稍微、少许… | 1 |
| 还、一些、比较、相当、有些… | 1.5 |
| 挺、多、更、还、愈加、更加… | 2 |
| 非常、特别、尤其、太、很… | 2.5 |
| 最、极度、完全、最为、十足 | 3 |

### 3.3.2 情感值计算

(1)影响因素特征词情感值计算 构建了情感词典之后，我们对每个影响因素的特征词进行情感值计算，将其 转化成定量数据。假设某个特征词为，该特征词的情感值基于以下公式得出： 对于正负情感词来说，若修饰特征词的情感词为正面情感词则+1，若为负面 情感词则-1，计算公式如下（3-1）所示：

### 3.3.3 情感值量化结果

## 3.4 本章小结

本章使用文本挖掘技术对文章所需的文本数据进行处理和分析。首先，利用网络爬虫搜集京东东方树叶茶饮的四种产品在线评论文本数据。在对搜集到的文本数据进行清洗之后，使用Python运行jieba分词库对文本数据进行分词，以得到词频和词云，以了解消费者在网购茶饮类产品中关注的内容。然后，建立LDA主题模型，提取了感知价值、新鲜、味道、物流包装保护性、外观、商品描述可靠性、售后服务质量、物流配送速度、口感这9个主题，并将其作为自变量来影响消费者满意度。接下来，基于知网词典对提取出的满意度影响因素进行情感分析，并计算其情感得分，以将其转化为定量数据。通过对评论文本进行情感分析，作为衡量消费者满意度的因变量的值。通过以上操作，我们得到了消费者满意度矩阵，为接下来分析自变量对消费者满意度的影响提供了数据基础。

# 第 4 章 在线评论文本内容对消费者满意度影响的实证分析

上一章通过对收集到的在线评论文本数据的内容进行分析，对在线评论文 本内容进行主题分析和情感分析，我们从在线评论文本内容中得到了 9 个文本 主题作为影响消费者满意度的因素，并将其转化成了定量数据，因此本章将对 各个影响因素对满意度的影响路径进行进一步的探讨，实证分析从评论内容中 提取的各个影响因素是如何影响消费者满意度，主要由以下三个研究：研究一 是检验第 3 章提取到的 9 个影响因素对满意度影响因素的分析；研究二是检验 在感知价值这一中介作用下，消费者满意度是如何收到自变量的影响；研究三 是进一步细分总体消费者满意度为消费者满意和消费者不满意，来比较分析消 费者满意和不满意影响因素差别，得到的结论能够为商家提出建议。

## 4.1 研究假设的提出

对上述影响因素进行归类我们将这九个主题进一步概括，发现感知价值、 生鲜农产品的新鲜、味道、外观和口感属于产品质量因素；物流包装保护性和 物流配送速度属于物流服务质量因素；商品描述可靠性和售后服务质量属于商 家服务质量因素。接下来我们就对各个影响因素提出假设。

### 4.1.1 产品质量因素对消费者满意度的影响及假设提出

### 4.1.2 商家服务质量因素对消费者满意度影响及假设提出

### 4.1.3 物流服务质量对消费者满意度的影响及假设提出

### 4.1.4 感知价值对消费者满意度的影响及假设提出

### 4.1.5 感知价值的中介作用及假设提出

### 4.1.6 消费者满意和不满意影响因素的比较分析及假设提出

## 4.2 模型建立

通过文章的第三章，我们已经提取了本文的自变量和因变量，并且对其进行 了赋值，接下来我们要通过线性回归、中介分析和比较分析来具体研究提取的影 响因素对消费者满意度的影响。

# 注释

黑体小三号字，居中书写，段前40 pt，段后20 pt，行距20 pt。“参考文献”四个字之间不需要空格。

**如不需要注释，请删除此页。**

①

②

③

正文部分用五号字，汉字用宋体，英文用Times New Roman体，行距采用固定值16 pt，段前空3 pt，段后空0 pt。

……

# 参考文献

黑体小三号字，居中书写，段前40 pt，段后20 pt，行距20 pt。“参考文献”四个字之间不需要空格。

[1]刘国钧,陈绍业.图书馆目录[M].北京:高等教育出版社,1997:15-18.

[2]钟文发.非线性规划在可燃毒物配置中的应用[C]//赵玮.中国运筹学会第五届大会论文集.西安:西安电子科技大学出版社,1996:468-471.

[3]张筑生.微分半动力系统研究[D].北京:北京大学数学院数学研究所.1983.

[4]冯西桥.核反应堆压力管道与压力容器的LBB分析[R].北京:清华大学核能技术设计研究院,1997.

[5]袁庆龙,候文义.Ni-P合金镀层组织形貌及显微硬度研究[J].太原理工大学学报,2001,32(1):51-53.

[6]谢希德.创造学习的新思路[N].人民日报,1998-12-25(10).

[7]汉语拼音正词法基本规则:GB/T 16159—1996[S].北京:中国标准出版社,1996.

[8]姜锡洲.一种温热外敷药制备方案:881056073[P].1989-07-26.

[9]王明亮.关于中国学术期刊标准化数据库系统工程的进展[EB/OL].(1998-10-04)[2019-4-5].http://www. cajcd.edu.cn/pub/wml.txt/980810-2.html.

[10]万锦坤.中国大学学报论文文摘(1983-1993)英文版[M/CD].北京:中国大百科全书出版社,1996.

正文部分用五号字，汉字用宋体，英文用Times New Roman体，行距采用固定值16 pt，段前空3 pt，段后空0 pt。