

学士学位论文

基于机器学习的东方树叶茶饮评论文本消费者满意度感知研究

姓 名

学 号

院 系

年 级

专 业

指导教师

二〇二三年四月十日

学位论文原创性声明

本人所提交的学位论文《 》，是在导师张利民的指导下，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。

论文作者（签名）： 指导教师（签名）：

年 月 日 年 月 日

# 摘 要

在线评价是指网络用户对商品、服务等给予实时反馈和讨论，能够对消费者的购物决策以及企业的品牌形象产生重要的影响，相较于商品公开的信息，具有更高的可信度以及参考性。当消费者对产品或服务体验感到满意或者糟糕时，可以在平台中留下对应情绪的评论以及星级等等以此来表达自己的真实感受，因此，商家可以通过在线评论了解消费者的诉求和满意度倾向，这为分析消费者满意度提供了一种新的视角。本研究选择茶饮行业中的东方树叶品牌作为代表，运用LDA主题模型和基于情感词典的情感分析方法，来确定产品满意度受到哪些因素的影响。研究结果表明，消费者对茶饮产品的满意程度主要体现在产品质量与价格、物流服务等方面。本研究利用文本挖掘技术揭示在线评论文本与消费者满意度之间的相关性，为相关企业的市场营销和品牌定位提供决策支持和重要的参考价值。这些研究成果对于加强市场竞争和提升产品消费者体验都有着重要的意义

关键词：消费者满意度；情感分类；LDA

ABSTRACT

Online evaluation refers to the real-time feedback and discussion given by network users to products and services, which can have an important impact on consumers' shopping decisions and the brand image of enterprises. Compared with public information on products, it has higher credibility and reference. When consumers are satisfied or bad with the product or service experience, they can express their true feelings by leaving comments and star ratings corresponding to their emotions on the platform. Therefore, merchants can understand consumers' demands through online reviews and satisfaction tendency, which provides a new perspective for analyzing consumer satisfaction. This study chooses the Oriental Leaf brand in the tea industry as a representative, and uses the LDA topic model and the sentiment analysis method based on the sentiment dictionary to determine which factors affect product satisfaction. The research results show that consumers' satisfaction with tea products is mainly reflected in product quality, price, and logistics services. This study uses text mining technology to reveal the correlation between online review text and consumer satisfaction, and provides decision support and important reference value for relevant companies' marketing and brand positioning. These research results are of great significance for strengthening market competition and improving product consumer experience

**Key words:** consumer satisfaction; sentiment classification; LDA

目 录

[第1章 绪论 1](#_Toc132901101)

[1.1 研究背景 1](#_Toc132901102)

[1.2 研究现状 1](#_Toc132901103)

[1.3 研究内容、研究方法及思路 2](#_Toc132901104)

[1.3.1 研究内容 2](#_Toc132901105)

[1.3.2 研究思路 2](#_Toc132901106)

[第2章 评论文本对消费者满意度的影响理论基础 4](#_Toc132901107)

[2.1 文本挖掘相关理论 4](#_Toc132901108)

[2.1.1 文本特征提取相关理论 4](#_Toc132901109)

[2.1.2 文本情感分析相关理论 4](#_Toc132901110)

[2.2 LDA主题模型 5](#_Toc132901111)

[2.2.1 LDA主题模型简介及原理 5](#_Toc132901112)

[2.2.2 LDA先验知识 5](#_Toc132901113)

[2.2.3 LDA生成文档过程 6](#_Toc132901114)

[2.3 木桶原理 7](#_Toc132901115)

[第3章 提取评论文本中的消费者满意度影响因素 7](#_Toc132901116)

[3.1 评论采集及预处理 7](#_Toc132901117)

[3.1.1 数据采集及清洗 8](#_Toc132901118)

[3.1.2 数据分词 9](#_Toc132901119)

[3.2 情感分类 12](#_Toc132901120)

[3.2.1 情感词匹配 12](#_Toc132901121)

[3.2.2 修正情感倾向 12](#_Toc132901122)

[第4章 LDA模型进行主题分析 1](#_Toc132901123)

[4.1 建立词典及语料库 1](#_Toc132901124)

[4.2 寻找最优主题数 1](#_Toc132901125)

[4.3 评价主题分析结果 2](#_Toc132901126)

[结 论 6](#_Toc132901127)

[参考文献 7](#_Toc132901128)

# 第1章 绪论

## 1.1 研究背景

消费者满意度是企业竞争力和市场份额的重要指标。随着电商平台的兴起，消费者对商品和服务的需求变得更加多元化。然而，由于消费者难以直接了解商品质量和售后服务的真实情况，因此消费者满意度成为了电商平台和企业关注的重点指标。传统的调查方式存在局限性，因为消费者难以准确表达自己的感受和需求。使用评论文本数据来研究消费者对商品和服务满意度的程度，可以增强对消费者观点和看法的全面和客观了解。

运用评论文本数据研究，可以发掘消费者对产品、服务和价值的评价，进而为企业提供更具针对性的产品建议和意见。同时，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，可以为实现精准营销、提高客户满意度提供新思路和方法。

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客观、全面的购物指南，促进相关产业链的健康可持续发展。同时具有重要的实践意义和研究价值，能够为相关企业提供准确的市场调研数据和产品设计改善思路，有效提高消费者的购物和服务体验，同时也有助于电商平台，企业，消费者之间的互利共赢高效发展。

## 1.2 研究现状

当今互联网时代飞速发展，一些包括电商在内的高新产业开始蓬勃发展，基于评论数据研究消费者对产品满意度逐渐成为国内学者研究的热点。目前我国的研究主要聚焦于以下两个领域：一是在不同领域的电商平台上对消费者满意度的研究；二是基于不同技术手段的评论文本分析方法研究。

在不同领域的电商平台上，国内的研究主要集中在淘宝、京东、天猫等大型电商平台。例如，一些学者通过对淘宝网上某个品牌的产品评论进行情感分析，探究消费者满意度与情感极性之间的关系。还有一些学者将评论文本数据与星级评价数据相结合，分析不同星级评价下消费者的满意度特征和差异。此外，还有学者通过对电商平台上的购物体验、物流配送、客服服务等方面的评论文本进行研究，从不同角度了解消费者的满意度和需求。

在评论文本分析方法方面，国内的研究主要包括传统文本挖掘方法、情感分析方法、主题模型方法等。传统文本挖掘方法主要包括关键词提取、分类、聚类等，可以用于分析消费者对商品和服务的评价特点和关注点。情感分析方法则通过分析文本中的情感词语、情感极性等来判断消费者的情感倾向，从而得出消费者对商品和服务的满意度。主题模型方法则可以从大量的评论文本中提取出隐含的主题信息，例如消费者关注的产品特性、服务质量、价格等方面。

总体来说，国内的研究基于评论文本的消费者满意度研究已经初步具备规模和成果，但仍存在一些问题和挑战。例如，评论文本数据的质量良莠不齐、标注标准不统一等问题，以及如何将研究结果与实际业务应用相结合等挑战。未来需要进一步深入研究和探索，推动基于评论文本的消费者满意度研究更加深入、全面、客观地展开。

## 1.3 研究内容、研究方法及思路

### 1.3.1 研究内容

基于评论文本的消费者满意度研究的核心是利用消费者在各种渠道上留下的评论文本，通过收集、清洗、分析和挖掘的手段，以获得有关消费者满意度的信息和见解。具体而言，该研究涉及四个主要方面：一是评论文本的收集和清洗，其中包括数据源的选择、噪声和重复评论的清除等；二是评论文本的特征提取和分析，其中包括情感分析、主题分析和关键词提取等；三是消费者满意度的建模和预测，其中包括基于机器学习和自然语言处理技术的模型；四是研究结果的解释和实际应用，其中包括消费者满意度的诊断以及针对产品和服务的改进建议。以上研究内容表明，基于评论文本的消费者满意度研究可以为企业提供客观、全面、及时的消费者反馈信息，有助于企业改善产品和服务质量，进而提高企业的竞争力。

### 1.3.2 研究思路

首先，了解市场：了解东方树叶茶饮市场的概况和消费者行为，主要可以通过文献调研和市场调研等方法，掌握茶饮品牌在市场上的地位和消费者购买行为等信息，从而确定研究目标和范围。

其次，数据采集与预处理：本文通过Python爬虫技术，爬取了东方树叶品牌下相应产品的评论文本数据，包括消费者对不同茶饮口味、价格、服务等方面的评价

再次，数据分析：使用文本挖掘技术，对文本数据进行情感分类，并探索不同因素对消费者满意度的影响，如物流、口感、价格和售后等因素。

最后，研究结果解释和总结[1]：分析研究结果，概括出像文本情感分类分析在东方树叶茶饮消费者满意度感知研究中的应用、茶饮品牌口味和服务如何影响消费者满意度等结论，并提出可操作性较强的实践建议，如茶饮品牌如何提升消费者服务体验，设计有效的营销策略等等。

大致思路流程如图1.1所示：

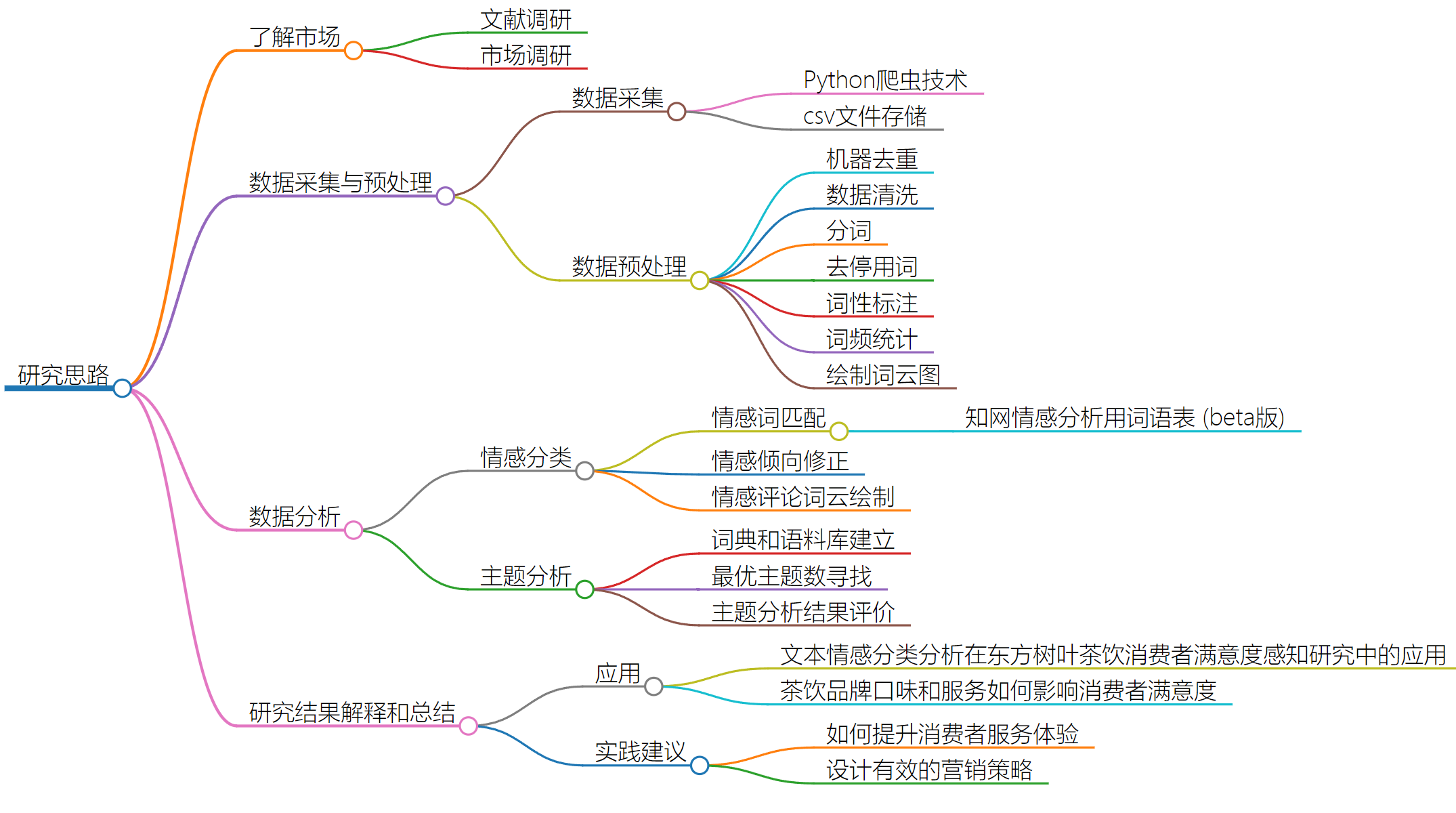


图 1.1整体流程思路

# 第2章 评论文本对消费者满意度的影响理论基础

## 2.1 文本挖掘相关理论

文本挖掘[2]是指从文本数据中自动地获取有用信息的技术，它结合了自然语言处理、信息检索、机器学习和数据挖掘等多个学科的理论和方法。其中，自然语言处理技术用于解决文本的语义理解和分析，信息检索技术用于快速地从大规模文本数据中检索出相关信息，机器学习技术用于构建文本分类、情感分析、主题分析等模型，数据挖掘技术用于挖掘隐藏在文本数据中的关联规律和模式。这些理论和方法为文本挖掘提供了强大的支持，使得文本数据的价值得到了更加充分的挖掘和利用。

### 2.1.1 文本特征提取相关理论

提取在线文本的属性特征是研究在线文本属性情况的过程，消费者的评论文本通常包含对商品、服务以及某些相关特征的看法。这些观点对于企业了解自身缺点以及为潜在消费者提供购物前的参考信息非常有帮助，因此文本属性和特征的提取至关重要。主题分析是一种常用的技术，它可以揭示隐藏在文本语料库中的潜在语义。因此，本文采用LDA主题模型对在线评论中产品或服务相关特征进行提取。

### 2.1.2 文本情感分析相关理论

文本情感分析是一种利用计算机技术识别和提取文本中所存在的情感信息的方法，涉及知识范围广，包括但不限于自然语言处理、机器学习等相关领域的技术。情感分析的方法主要有情感词典、机器学习，以及近年来越来越受到关注和研究的基于深度学习的技术，该方法主要通过神经网络来提高情感分类的准确性和效率。除此之外，情感分析在情感检索、舆情监测、产品推荐、品牌管理等方面都得到广泛应用。

1. 基于情感词典

该方法能够通过对一段文本中的情感词语进行统计分析，判断该文本的情感色彩，通常会分为正面、负面和中性情感。情感词典是一种词汇资源库，其中包括了一系列情感词汇及其情感极性（正面、负面或中性）。文本会被拆分成词语，然后被匹配到情感词典中对应的情感词，以此来计算文本的情感倾向。此外，在进行情感分析时，会针对否定词、程度副词等修饰词语的影响进行考虑，以便更精确地评估文本的情感倾向。该情感分析方法在短文本、社交媒体和产品评论等不同场景中广泛应用，具有高效和实时的特点。

1. 基于机器学习

此方法是一种利用计算机算法对经过人工训练的模型进行情感分类的技术。相比于以情感词典为基础的方法，它不依赖于事先构建的情感词典，可以处理新的或未知的情感词汇。通过对已标注的情感数据进行训练，机器学习可以学习不同情感类别的特征，常见的机器学习算法包括KNN、SVM、NB和ME等，可以通过这些特征对新的文本数据进行情感分类。该方法相较于基于情感词典的方法准确率高、鲁棒性强，并在文本挖掘、社交媒体分析、市场营销等领域得到广泛应用。

## 2.2 LDA主题模型

### 2.2.1 LDA主题模型简介及原理

消费者在撰写商品评价时，通常会确定几个主题并表达满意度或不满意度。这些主题是消费者购物过程中关注的重点或影响其满意度的因素。为了挖掘出评论中隐藏的主题，本文使用LDA主题模型进行分析。LDA主题模型是一种用于挖掘文本中主题和它们的相互关系的自然语言处理技术，具有可解释性和可扩展性，广泛应用于信息检索、文本分类、情感分析、推荐系统等领域。

LDA主题模型的原理介绍：我们研究一个包含n篇文档的文本集合，每篇文档包含N个单词。我们的目标是从这个文本集合中挖掘出潜在的主题以及它们相关的主题词。为此，假设每篇文档由多个主题共同组成，其中不同主题中包含着不同概率分布的单词。接着，我们通过对文本集合中每个单词的概率统计分析，推断出每篇文档的主题分布和每个主题的单词分布，再由此间接推断某个单词所属的主题。在这个过程中，通常采用贝叶斯推断方法，例如Gibbs采样算法。

### 2.2.2 LDA先验知识

LDA模型的推导过程中，需要用到下面概率和统计学的先验知识，其中包括但不限于以下先验知识

1. 多项分布

多项分布用于描述在多次试验中同时出现多个不同离散值的概率分布，其概率密度函数为：

1. Gamma函数

Gamma函数定义：



满足 ，

1. Beta分布

Beta分布定义

其中是正实数参数,是Beta函数，满足

1. Dirichlet分布

Dirichlet的概率密度函数为:

其中是正实数参数向量，​表示第i个类别作为先验出现的次数，, ,是多元Beta函数，满足

1. Gibbs采样

一种蒙特卡洛方法，通过从联合分布中抽取样本进行概率近似计算，其中每个变量通过给定其他变量的条件概率分布进行采样。

1. 情感分析

一种自然语言处理技术，用于判断文本中表达的情感极性，如正面、负面或中性。情感分析的任务有很多应用，如产品评论的可视化分析、舆情监测和营销策略制定等。

### 2.2.3 LDA生成文档过程

LDA 中，生成文档的过程如下：

1. 对某个文档（m=1,2,···,n）抽取其主题。通过使用参数为的狄利克雷分布来生成“文档-主题”分布, 其是一个多项式分布
2. 根据, 随机生成文档中第 n 个词语的主题
3. 由参数为的狄利克雷分布找到主题下词语的分布
4. 最后，由找到对应主题的词语

以上，我们利用LDA模型将文档转换为“文档-主题”分布和“主题-词语”分布，从而准确确定文档中隐藏的主题及其相关的词汇，并通过分析实现对文本的深入解读。

模型生成过程如下图2.1

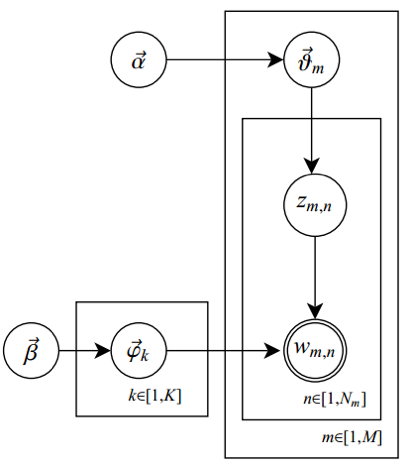


图 2.1 LDA主题模型

利用 LDA 主题模型，精准识别文本的主题和对应的词汇，为分析在线评论提供了重要的洞察。主题分析揭示了文本中呈现的主要观点和关注点，这对于揭示消费者不满意的原因至关重要。企业可以根据主题分析结果改进服务，提升用户满意度。本文采用LDA主题模型对消费者网购评论进行主题分析，将其作为探究消费者满意度的重要变量。

## 2.3 木桶原理

木桶原理是一种理论，根据该理论，水桶的容量由最短的木板决定。企业或组织也存在薄弱点，该点影响着企业的发展水平和竞争力，必须加以提升。随着网络购物和在线评论机制的广泛使用，消费者可以在评论中表达对产品或服务的满意度，这些评论可以被视为企业的“短板”，并可能影响消费者忠诚度和回购率，从而限制企业的发展。因此，电商企业须利用文本挖掘技术，发现影响消费者满意度的相关因素和企业的薄弱点，以改进产品和服务质量，促进企业的发展。

# 第3章 提取评论文本中的消费者满意度影响因素

## 3.1 评论采集及预处理

本文选择反爬措施较薄弱的京东作为数据源，并针对消费者喜好的四款茶饮产品收集和分析评论文本数据。鉴于茶饮产品种类繁多，无法对所有产品进行数据采集和分析，因此本文选取了常年销量领先的几种产品，分别爬取了好评、中评及差评信息按照顺序排列作为本文数据源，并且对其进行了相应的数据预处理操作[3]。

### 3.1.1 数据采集及清洗

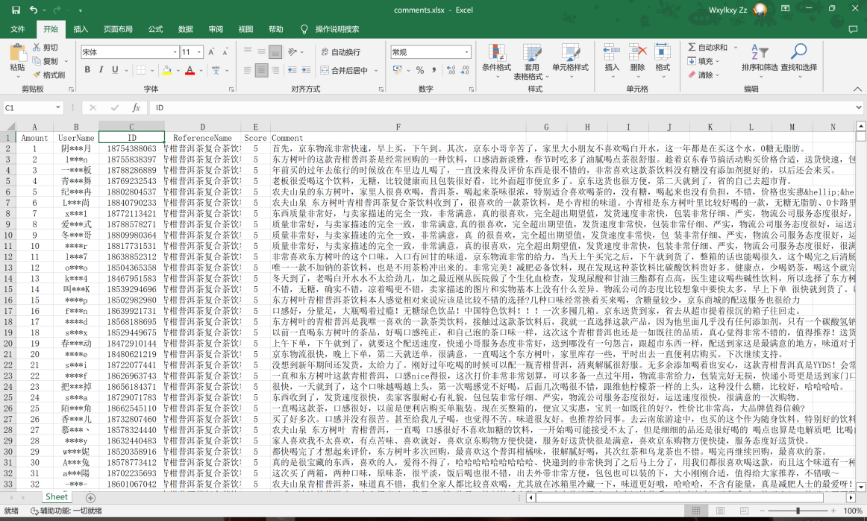
在京东平台的农夫山泉官方自营店中搜索”东方树叶”产品。随后，本研究按照评论排序，本文收集了超过6000条消费者评论文本信息。同时，我们利用Python编写了异步爬虫，数据的部分样本如图3-1所示。

图 3.1 数据采集样图

由于数据中掺杂着大量与实验不相关、重复和无效数据，为了确保分析结果的准确性，对爬取的数据进行清洗处理。

首先进行去重处理，然后定义了一个清理函数，该函数用于去除标点符号、数字和特殊字符，并进行停用词处理和词形还原，最后去除了多余的空格和换行符。为进一步提高文本分析的准确性，还剔除了纯数字或纯字母评论。清洗过程中所删除的无效数据及其分类详见表3-2。

##### 表 3.2 无效评论文本数据示例

|  |  |
| --- | --- |
| 类别 | 评论文本 |
| 1 数字类 | 1111 |
| 2 符号类 | ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ |
| 3 无效文本类 | ,lhjsdfasdfsdbgnfgngh |
| 4 未填写类 | 此用户未填写评价内容 |

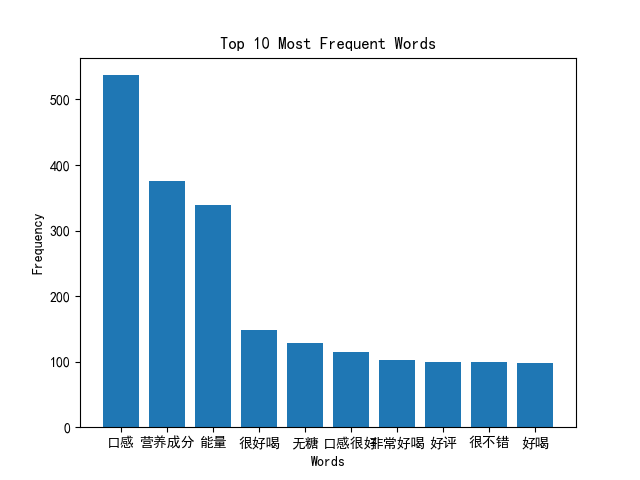
##### 表 3.2 续

其次，对清洗后的评论进行了分析。

例如统计评论的数量以及平均长度等基本信息， 其次统计最常见的词汇，取前10个最常见的词汇并且生成柱状图详见表3-3。

Total number of comments: 5760 （总评论数）

Average length of comments: 55.75（平均评论长度）



###### 图 3.3 高频词统计柱状图

### 3.1.2 数据分词

数据分词是文本处理的基础，它将连续的文本数据按照一定规则或算法切分成独立的词语或符号序列。这个过程将整个文本数据转化为一个个单独的词汇单元，是自然语言处理、文本挖掘、机器翻译等领域中必不可少的预处理步骤。准确的分词可以显著提升计算机对文本信息的理解和识别能力，从而有效提高后续处理工作的准确性和效率。中文分词是一个特别重要且具有挑战性的任务，因为中文的基本书写单位是字而不是单词。准确的中文分词能让所形成的词串反映句子的原意，而不会混淆篇章结构。因此，准确的分词是进行文本处理的必要过程。

本文中，我们采用广泛应用的中文分词工具“jieba”处理了清洗过的文本数据，并得到了详细的分词结果表格（见表格3-4）。

##### 表 3.4 分词结果示例

|  |  |  |
| --- | --- | --- |
| 序号 | 评论文本数据 | 分词结果 |
| 1 | 特别好喝的农夫山泉饮料。 | 好喝 农夫山泉 |
| 2 | 茉莉红茶真的是最爱了，有一股茶香的味道！ | 茉莉 红茶 真的 爱 一股 茶香 味道 |
| 3 | 很不错，方便饮用，省心 | 不错 方便 饮用 省心 |
| 4 | 很好，口感不错，物流也很快物有所值，下次还会再来光顾的 | 口感 不错 物流 很快 物有所值 下次 还会 光顾 |

经过文本分词后，一般会分词分出一些无意义的词语，例如语气词、程度副词等，比如“呢”、“了”、“的”、“么”、“另一方面”等。我们需要将其加入停用词表以去除这些无意义的词。为了确保研究结果的普遍性，本文使用了Github平台开源的中文停用词表，并将一些产品名称如“东方”、“树叶”、“饮料”、“喝”、“茉莉花茶”等也加入到了其中。  
 然后通过使用 Python 程序进行词频统计，我们得到了一个包含 8550 个数据分词的词频统计表。下表 3-5 展示了前 10 个词语的词频统计结果。

##### 表 3.5 排名前 10 位的特征词

|  |  |  |  |
| --- | --- | --- | --- |
| 排名 | 特征词 | 词频 | |
| 1 | 喜欢 | | 2343 |
| 2 | 口感 | | 1981 |
| 3 | 味道 | | 1886 |
| 4 | 不错 | | 1614 |
| 5 | 好喝 | | 1581 |
| 6 | 口味 | | 892 |
| 7 | 物流 | | 818 |
| 8 | 活动 | | 656 |
| 9 | 健康 | | 634 |
| 10 | 方便 | | 632 |

本研究的目标是分析该产品在茶饮行业中成为常胜将军的原因。为此，我们选定含有产品特征描述的评论进行分析。但是存在一些无法明确特征的评论，例如“无限回购”、“继续支持”等，无法明确提取出哪些产品特征受到了用户的满意。因此，我们只挑选出含有明确名词，如“味道好”、“口感好”或其他专有名词的评论进行分析。我们采用jieba和ICTCLAS标记方法对分词后的词语进行了词性标注[4]，具体结果见图3-6。

###### 

###### 图 3.6 词性标注样图

在进行了数据预处理之后，我们可以通过词云图来进行分词效果的展示，以获得对分词效果的直观感受。



###### 图 3.7 高频词词云分布

图3-7为茶饮产品评价中高频词云图，图中的字号大小反映了每个词语的出现频率。我们可以发现，“健康”、“口感”、“味道”、“物流”、“活动”等词汇出现的频率极高，这也表明消费者在购买茶饮产品时关注这些方面，并对其有很高的期望和评价标准。因此使用LDA主题模型对评论进行分析，以便深入研究影响消费者满意度的因素。

## 3.2 情感分类

### 3.2.1 情感词匹配

本文重点分析产品的优势劣势，并使用情感词典匹配方法识别评论文本的情感倾向，而无需深入分析每条评论的情感程度。本文采用了知网的开源情感分析词语表。然而，基于词表的情感分析方法的效果受限于词表中词语之间的关联性。为了提高情感分析的效果，需要根据具体场景进行词表优化。例如，在本文中，可以根据网购评论的情景增加一些只会在特定场景中出现的词语，如“好评”、“超值”、“差评”等。因此，可以根据日常生活网购中经常使用的网络评价用语例如“种草”、“ 宝藏”、 “无糖”等词语添加到正面情感词表内，在负面情感

词表内中加入“避雷”、“踩雷”、“拔草”和“惯例三星”等词语。详情可见图3-8。

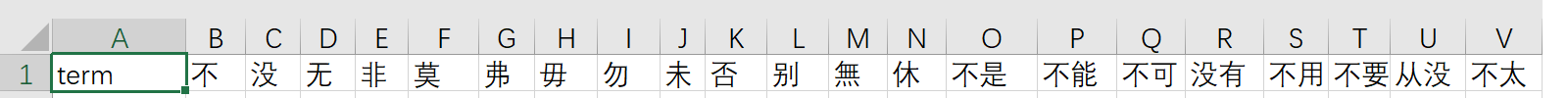
###### 

###### 图 3.8 情感词表

### 3.2.2 修正情感倾向

在情感分析中，对于包含否定词的文本，我们往往需要对情感极性进行修正，以准确反映文本的情感倾向。使用否定词表是一种常见的修正情感极性的方法。对于跟在否定词后面的情感词，修正是必要的：如果情感词和否定词之间只隔了一个词，则情感词的极性需要反转；如果这两个词之间有多个词，则需要进一步对语义进行分析，判断修正程度和情感倾向。

否定字表如下图：



###### 图 3.9 修正词表

# 第4章 LDA模型进行主题分析

## 4.1 建立词典及语料库

LDA主题模型是解决指代问题的有效工具。在分词后的评论中，可能出现“贵”一词被独立分离，我们难以确定它是指运费贵还是产品贵等其他情况。想要解决指代不明的问题，简单的词频统计和情感分析并不足够。因此，我们需要采用LDA主题模型，该模型可计算出词汇在主题中的概率分布，从而确定“费用”所属的主题，并计算出同一主题下其他特征词的概率，以解决这种问题。通过这种方法，我们可以有效地解决多种指代问题。

## 4.2 寻找最优主题数

本文采用采用自适应最优LDA模型选择方法，其基于相似度进行实现。该方法具有自适应性和灵活性，可以根据相似度选择最优的主题数从而提高LDA模型的效率和稳定性。同时，该方法还可以减少过拟合的可能性，提高模型的准确性和可靠性。

对所有主题数进行LDA模型训练，得到对应的主题分布矩阵。

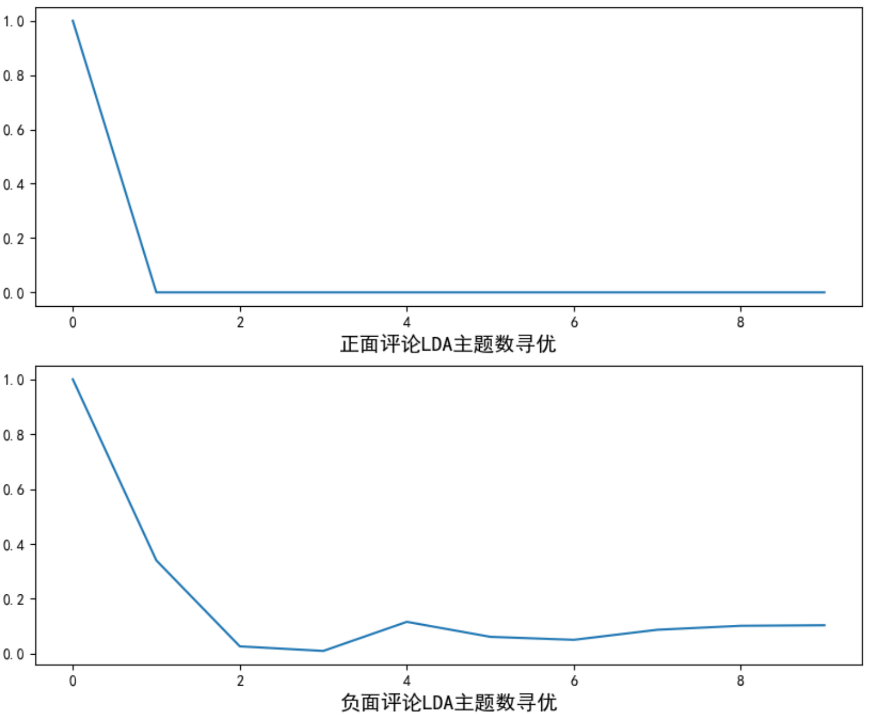
计算每个主题数的主题分布矩阵与其他主题数对应的主题分布矩阵之间的相似度。

根据相似度选取相似度最大的一组主题数作为最优的主题数。

使用所选的最优主题数重新进行LDA模型训练，得到最终的主题分布和话题结构

在相似度度量方法中，余弦相似度是非常常用的一种，它适用于计算两个矩阵或向量之间的相似程度。对于LDA模型，我们可以将不同主题之间的主题分布矩阵视为向量，同时把余弦相似度作为衡量各个主题间相似性的度量标准。具体而言，从词频入手，采用词向量模型，计算主题所包含词项的相似度，通过相似度值的大小来判断两个主题的内容是否相近。这种方法能够更加全面地反映主题的语义信息，减少信息冗余，提高主题模型的稳定性和准确性。

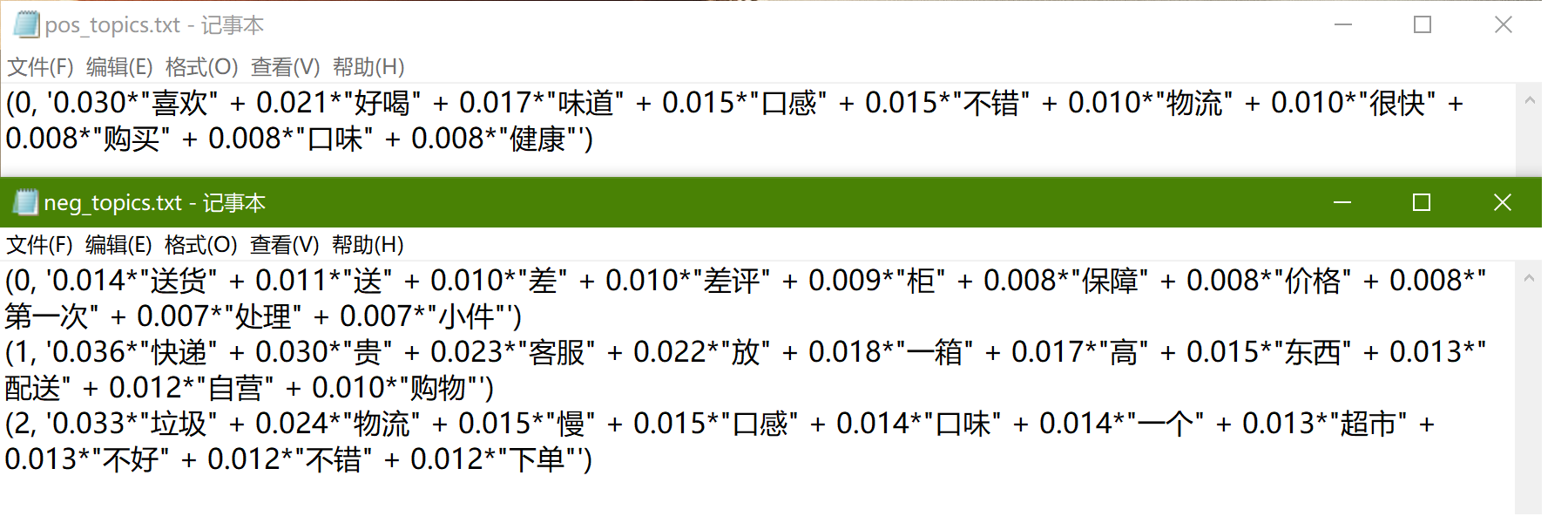
假设与是两个n维向量，记为=（）, =（）,则和的夹角的余弦值可以通过以下公式计算

借助LDA主题模型可找出不同主题数下的主题词，选取每个模型中前n个主题词生成两两主题间的词频向量，这里n取100并且计算它们之间的余弦相似度。相似度越大，说明两个主题之间的词汇更为相似。我们反复在不同的主题数下执行此过程，最终得到各主题数下平均余弦相似度的值。通过分析这些值，我们可以找到最佳的主题数。此外，我们还可以制作主题间平均余弦相似度图来展示主题之间的相关性和差异性，从而有效地优化主题模型，提升其效果和应用价值。

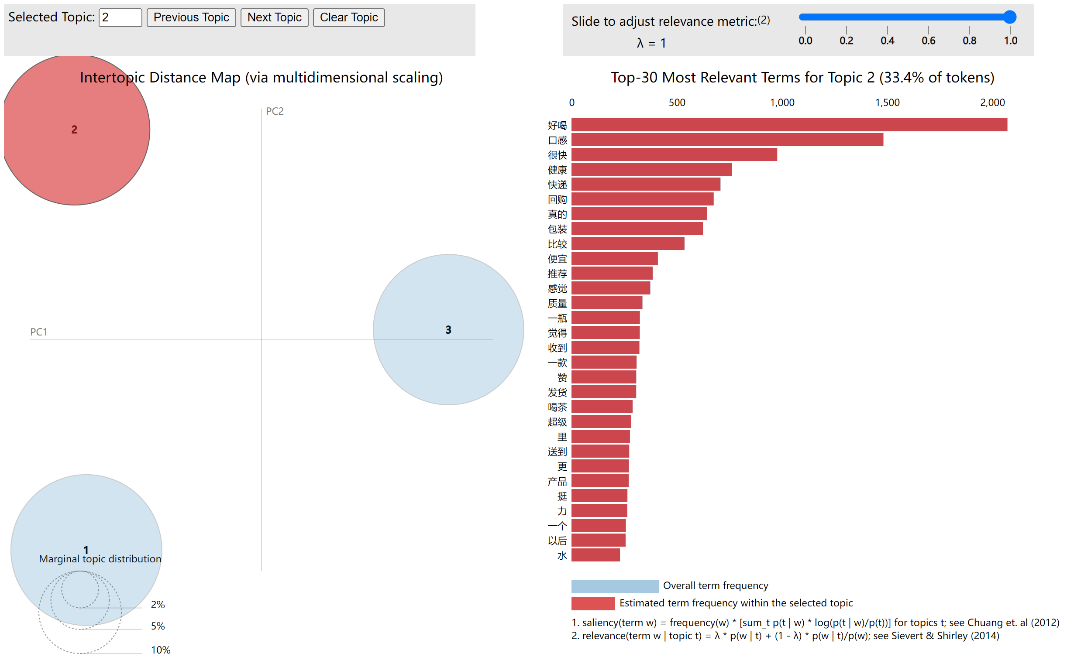
###### 图 4.1 主题间平均余弦相似度图

从图4-1可知，当正面评论主题数为1时，平均余弦相似度在不同主题数之间达到最小值。这说明在LDA模型中，正面评论具有较高的相似性和一致性，大部分正面评论文档都可以划分至同一个主题下。因此，主题数为1时的分类效果较好。负面评论当主题数为3时，该值达到最小。这说明文本在主题数为3时，具有相对较好的区分度和分类效果，可以对各个主题进行较好的划分和分类。

## 4.3 评价主题分析结果

根据上一节得出的最佳主题数，我们分别针对正面和负面评论数据建立了LDA主题模型，接下来，进行LDA主题分析并提取了每个主题中概率最高的10个单词，详见下图4.2。

###### 图 4.2 抽取的关键词

为更清晰地展现这些主题词信息，我们使用pyLDA.vis进行了可视化处理，结果分别在图4.3和图4.4中展示。通过这些分析处理，我们深入研究了消费者对东方树叶茶饮的评论情况[5],为后续分析提供了有力支持和基础。

###### 图 4.3正面评论LDA主题可视化

###### 图 4.4 负面评论LDA主题可视化

为更准确地描述消费者对东方树叶茶饮的评论情况，我们从每个主题中抽取了前10个关键主题词，并将正面和负面主题词分别列在表4.5和表4.6中。

通过表4.5，我们可以清晰地了解到正面评价文本中的潜在主题。

##### 表 4.5 正面潜在主题

|  |
| --- |
| Topic 1 |
| 喜欢 |
| 好喝 |
| 味道 |
| 口感 |
| 不错 |
| 物流 |
| 很快 |
| 购买 |
| 口味 |
| 健康 |

在该主题下，由高频特征词能看出东方树叶

1. 味道好：说明东方树叶在饮品配方和加工工艺方面都投入了大量付出，同时注重饮品口感的均衡，让消费者体会到真正的茶香四溢
2. 口感不错：在严格的质量控制上，东方树叶茶饮品牌不断注重改进口感，在推出新产品时，进行多轮商品调试，针对消费者反馈做出调整，确保口感味道达到最佳。
3. 物流快速：东方树叶茶饮品牌对物流的管理非常严格，配送渠道快速，让消费者更快收到心仪的产品。
4. 健康：东方树叶茶饮品牌选用了天然的茶叶作为原材料，不含任何防腐剂和人工添加剂，保证了产品的健康和营养价值。
5. 购买方便：东方树叶茶饮品牌旗下渠道较多，包括实体店和网店，方便消费者购买。

综上，东方树叶茶饮品牌的优点主要在于味道口感好、物流速度快、口味丰富和健康营养等方面。这些优点有助于提升品牌形象和口碑，品牌方面需要在这些方面持续发力，以提升产品的市场竞争力。

表4.6展现了在负面评价文本中可能存在的主题。

##### 表 4.6 负面潜在主题

|  |  |  |
| --- | --- | --- |
| Topic 1 | Topic 2 | Topic 3 |
| 送货 | 快递 | 垃圾 |
| 送 | 贵 | 物流 |
| 差 | 客服 | 慢 |
| 差评 | 放 | 口感 |
| 柜 | 一箱 | 口味 |
| 保障 | 高 | 一个 |
| 价格 | 东西 | 超市 |
| 第一次 | 配送 | 不好 |
| 处理 | 自营 | 不错 |
| 小件 | 购物 | 下单 |

针对主题1中的高频特征词汇, 我们可以看到它们主要关注物流配送方面。具体而言，送货速度、服务态度等方面存在差评。在展示柜上的销售方面，产品可能会被摆放得不整齐且价格相对较高，这可能会导致初次体验不佳以及小件处理不及时等一系列问题。这些问题需要制造商和经销商进一步优化和改进，例如优化库存布局以减少物流运输距离，通过多种营销活动适度降低销售价格，线下柜台注重整理等等，以提高产品的市场竞争力。

就主题2中的高频特征词而言，它们主要与价格，售后和服务等内容有关。这表明，部分消费者对于东方树叶茶饮的售后服务可能存在不满意。具体来说，与市场其他同类品牌相比，东方树叶茶饮品牌的价格可能较高。在消费者咨询、投诉等问题方面，东方树叶茶饮品牌的客服服务可能存在缺陷，这可能会影响到消费者的购物体验。此外，在物流和销售等方面，东方树叶茶饮品牌可能存在物品放置不当、保护不足等问题。最后，可能会存在自营模式限制：东方树叶茶饮品牌的自营模式可能会导致产品种类、库存等方面有所限制，这限制了消费者的购物选择。针对这些问题，生产商需要考虑产品价格并优化售后服务体系，以提高产品可选性，以满足消费者的需求，不断改进，提高产品竞争力，增加市场份额。

主题3中的高频特征词汇主要与口感口味等方面有关，表明存在消费者对该品牌产品的口感和口味不够满意或产品种类单一的问题。为了解决这些问题，制造商可以增强产品研发力度，推出不同品类、口味和口感的饮品，同时通过市场调查分析消费者的口味需求和偏好，以便在市场反馈及时调整产品口感和口味。此外，可以提高生产加工工艺，确保饮品口感、香味和口味的平衡，使消费者享受到更加舒适、自然和美味的口感。同时，要不断留意消费者的反馈和需求，不断优化产品和品牌形象，推动品牌的持续发展。

# 结 论

本研究采用文本挖掘与情感分类方法分析东方树叶茶饮产品评论文本，对消费者满意度的相关指标进行挖掘。相比传统的问卷调查更加客观和有效，能够提供详细的消费者评论信息，准确地表达消费者的情感信息。研究表明，消费者对产品质量、价格、信任度以及物流服务的体验和感知是影响其满意度的主要因素之一。

首先，将重点加强消费者对产品质量的感知具有重要意义。消费者对产品的口感、健康程度等方面的评价对产品满意度具有很大的影响。因此，企业应该加强对产品研发和提高生产加工工艺，采用高品质原材料，精细化加工工艺等方式提高产品质量

其次，为确保消费者的价格敏感度适中和提高信任度，企业还需要建立消费者对产品的信任体系,减少消费者的满意度受到价格敏感性和信任感影响的概率。在提高消费者信任度上，企业需要采取具体措施，例如提供良好的售后服务，建立合理高效的物流体系和完善的客户服务等。旨在提升消费者对产品的忠诚度和满意度，进而提高产品的市场竞争力。

最后，改进物流服务设施对于提升消费者满意度具有至关重要的作用。作为电商交易的重要环节，商家应该优化库存布局，减少物流运输距离；引进先进的物流技术和系统，提高物流效益；增加物流网建设资金，提高配送能力和服务水平；建立多元化的物流配送体系，响应不同用户需求；加强物流流程管控和监管，及时解决各种物流问题。这些举措有助于电商平台提升物流服务品质和用户体验，提升市场竞争力。在大力改进物流服务的同时，商家还应着眼于用户需求动态变化，制定相应的物流服务策略和服务标准，不断提高服务品质和水平，以进一步提高消费者满意度。

# 参考文献

[1] 王涛, 刘星亮,王泓淇,于志军.基于机器学习的富硒茶评论文本消费者满意度感知研究[J]. 湖北农业科学, 2022, 61(1): 146-152.

[2] 张心悦.生鲜农产品在线评论文本内容对消费者满意度的影响研究[D].哈尔滨工业大学, 2020.

[3] 李金海.基于在线评论挖掘的网络购物混合推荐模型及策略研究[D]. 镇江: 江苏大学, 2016.

[4] 杜学美,丁璟妤,谢志鸿,雷丽芳. 在线评论对消费者购买意愿的影响研究[J]. , 2016, 28(3): 173-183.

[5] 朱丽叶, 袁登华, 张静宜. 在线用户评论质量与评论者等级对消费者购买意愿的影响——产品卷入度的调节作用[J]. 管理评论, 2017, 29(2): 87-96.

[6] 魏华, 黄金红. 在线评论对消费者购买决策的影响——产品卷入度和专业能力的调节作用[J]. 中国流通经济, 2017, 31(11): 78-84.

[7] 李蕾,林家宝,黄士. 农产品电子商务顾客满意度和忠诚度的形成机制研究——基于感知价值和服务质量的视角[J]. 世界农业, 2017 (11): 97-103.

[8] 郝媛媛, 叶强, 李一军. 基于影评数据的在线评论有用性影响因素研究[J]. 管理科学学报, 2010, 13(8): 78-88.

[9] 李道和,杨俊,陈江华.茶叶电子商务消费者满意度影响因素分析——基于在线评论的探索分析[J].林业经济,2019(10).

[10] 林春桃,苏宝财,叶乃兴,郑德勇.福建乌龙茶的品牌忠诚影响因素实证研究——以福州消费者为例[J].林业经济问题,2009,29(6).

[11] 王家琦, 张耀荔, 陈静. B2C 网络购物模式下物流服务质量与顾客满意的关系研究[J]. 中国商贸, 2013.

[12] 袁伟,范进, 赵定涛,等. 第三方物流服务质量对顾客忠诚影响的实证研究[J]. 中国科学技术大学学报, 2017, 47(6): 508-513.

[13] 林家宝,万俊毅,鲁耀斌.生鲜农产品电子商务消费者信任影响因素分析:以 水果为例[J].商业经济与管理,2015(05):5-15.

[14] 吴键, 李秀军. 基于文本挖掘的特色农产品网络零售满意度影响因素与提升策略研究——以随州香菇为例[J]. 湖北农业科学, 2022, 61(24): 236-242.

[15] 王悦.基于在线评论的京东平台大米消费者满意度研究[D].黑龙江八一农垦大学,2019.

[16] Xu X, Li Y. The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach[J]. International journal of hospitality management, 2016, 55: 57-69.