学号：201975088　　密级：_____

长江大学

硕士研究生学位论文

基于语义分析的电商用户评论多维度文本挖掘研究

专业领域：　计算机技术

研究方向：　机器学习与人工智能

研 究 生：　王　宁

指导教师：　周云才 教授

　　李盟禹 工程师

论文起止日期：2021 年 4 月至 2022 年 5 月

学号：201975088　　密级：_____

长江大学

硕士研究生学位论文

基于语义分析的电商用户评论多维度文本挖掘研究

专业领域：　计算机技术

研究方向：　机器学习与人工智能

研 究 生：　王　宁

指导教师：　周云才　教授

　　李盟禹　工程师

论文起止日期：2021 年 4 月至 2022 年 5 月

Research on multi-dimensional text mining of e-commerce user comments based on semantic analysis

Field：Computer Technology

Direction of Study：Machine Learning and Artificial Intelligence

Graduate Student： Wang Ning

Supervisor：Prof. Yuncai Zhou、RD. Mengyu Li

School of Computer Science
Yangtze University
April,2021 to May,2022

摘 要

随着互联网技术的发展，新冠疫情的肆虐，导致电商平台火热，越来越多的人选择网购的方式。用户在购买商品后，在对应电商平台进行商品评论是重要的购物体验反馈渠道，是对整体购买流程的综合评价。有购买倾向的用户、商家、平台、厂商、监管部门都会通过用户的评论与评分情况，对商家的经营状态与商品的质量进行评估，从而做出决策，商品评论的重要程度可见一斑，但是面对数以万计的评论，各方都无法获得直观有效的信息，对商品评论进行文本挖掘就显得尤为重要。

本文基于手机对生活的重要作用以及不同人群购买手机时看重手机功能不同的特点，制定了以手机商品评论数据为数据来源，从多维度对手机商品评论进行文本挖掘，并依据挖掘结果，从多维度为用户、厂商提供建议的实验目标。在准备好研究所需的相关理论知识与运行环境后，本文详细论述了实验过程。首先利用 python 编写京东平台商品评论的爬虫，分别对占据国内手机市场份额较高的 oppo、三星、小米、华为、苹果五大手机厂商 2021 年主打机型的手机商品评论进行爬取，共爬取 8010 条评论。在对爬取到的评论中的重复评论、无意义评论、停用词、无意义字符和后续会干扰文本挖掘的词汇进行去除后，使用结巴分词对评论进行分词和标注词性工作。准备好文本数据之后，采用有监督的机器学习决策树进行分析，以爬取评论中的商品评分作为训练数据，进行模型训练，参数调优，并画出决策树。对每条评论中各个词汇表达的正负向情感进行求和计算，分析整个评论的情感倾向，并通过词云图展示正负向评论中的高频词。由于决策树和情感倾向分析虽然可以分析出评论的正负倾向，以及高频词汇，但是不能展示出正负向情感词汇是在描述手机的哪一个主题属性，为了解决这个问题，本文进行了基于 LDA 模型的主题挖掘，利用平均余弦相似度寻找最优主题数，并对实验过程及结果进行分析，进行实验改进，通过导入手机类别的搜狗细胞词库与手动添加方式，补充手机领域的分词词库，通过词汇同义替换与手动添加删除的方式，补充手机领域情感词词库后，再次进行 LDA 模型主题挖掘。

本文实验最终以测试集 0.92 的准确率画出决策树，以 0.86 的准确率分析出评论情感倾向，在补充完善手机领域情感词库与分词词库后，基于 LDA 模型进行了较为准确的主题挖掘，并以此结果为基础，从手机拍照、音效、系统、外观、充电与电量、客服、物流、保价等多个维度对用户购买与各个厂商新产品改进方向提出建议。文章结尾总结了在实验中可以改进的地方，对构建领域情感词库与领域分词词库的可行性及效果进行展望。

关键词：文本挖掘，LDA 模型，决策树，情感分析

Abstract

With the development of Internet technology and the raging of the new crown epidemic, e-commerce platforms have become popular, and more and more people choose the way of online shopping. After a user purchases a product, the product review on the corresponding e-commerce platform is an important feedback channel for shopping experience and a comprehensive evaluation of the overall purchase process. Users, merchants, platforms, manufacturers, and regulatory authorities who are inclined to purchase will evaluate the business status of merchants and the quality of products through user reviews and ratings, so as to make decisions. The importance of product reviews is evident. For tens of thousands of reviews, all parties cannot obtain intuitive and effective information, so it is particularly important to perform text mining on product reviews.

Based on the important role of mobile phones in life and the characteristics that different people value different functions of mobile phones when purchasing mobile phones, this paper formulates the data of mobile phone product reviews as the data source to conduct text mining on mobile phone product reviews from multiple dimensions. Dimension provides suggested experimental goals for users and manufacturers. After preparing the relevant theoretical knowledge and operating environment required for the research, the experimental process of this paper begins. First, use python to write a crawler for product reviews on the Jingdong platform, and crawl the product reviews of the five major mobile phone manufacturers in 2021, namely oppo, Samsung, Xiaomi, Huawei, and Apple, which occupy a high share of the domestic mobile phone market. 8010 comments. Remove duplicate comments, meaningless comments, stop words, meaningless characters and subsequent words that would interfere with text mining in the crawled comments. Afterwards, we use stuttering word segmentation to tokenize comments and tag part-of-speech. After preparing the text data, the supervised machine learning decision tree is used for analysis, and the scores in the crawling comments are used as training data to train the model, adjust the parameters, and draw the decision tree. After that, the positive and negative sentiments expressed by each word in each comment are summed and calculated, the sentiment tendency of the entire comment is analyzed, and the high-frequency words in the positive and negative comments are displayed through the word cloud graph. Although decision tree and sentiment analysis can analyze the positive and negative tendencies of comments and high-frequency words, they cannot show which theme attributes of mobile phones are described by positive and negative sentiment words. In order to solve this problem, this paper based on The topic mining of the LDA model uses the average cosine similarity to find the optimal number of topics, analyzes the experimental process and results, and makes experimental improvements. By

importing the Sogou cell thesaurus of the mobile phone category and adding manually, the word segmentation in the mobile phone field is supplemented. Thesaurus, through the synonymous replacement of words and manual addition and deletion, after supplementing the emotional word thesaurus in the mobile phone domain, the topic mining of the LDA model is carried out again.

In the experiment in this paper, the decision tree was drawn with the accuracy rate of 0.92 in the test set, and the sentiment tendency of comments was analyzed with the accuracy rate of 0.86. After supplementing and improving the emotional thesaurus and word segmentation thesaurus in the mobile phone field, a relatively accurate topic mining was carried out based on the LDA model. , and based on the results, from various dimensions of mobile phone photography, sound effects, system, appearance, charging and power, customer service, logistics, price protection and other dimensions, put forward suggestions for user purchases and the improvement direction of new products of various manufacturers. At the end of the article, the areas that can be improved in the experiment are summarized, and the feasibility and effect of constructing domain sentiment dictionary and domain word segmentation thesaurus are prospected.

Key words: Text Mining, LDA Models, Decision Trees, Sentiment Analysis

目录

第 1 章 绪论

1.1 课题的研究背景及意义

1.1.1 研究背景

根据中国互联网络信息中心（CNNIC）发布的第 49 次《中国互联网络发展状况统计报告》显示，截至到 2021 年底，我国网民的规模达惊人的 10.32 亿，截至 2021 年 6 月，我国有 8.2 亿人网购[1]。事实上，自 2019 年末新冠疫情爆发，线下购物受多种条件制约，导致线上购物愈发火热。这也促使了抖音电商、拼多多等电商平台后起之秀的发展。再加上我国的老牌电商平台，如淘宝、京东、苏宁等，目前我国大型电商平台的竞争十分激烈[2]。多种电商平台的发展也给用户在购物平台选择上增加了更多样的选择，比如主营服饰的唯品会、兴趣为王的抖音电商以及以数码 3C 闻名的京东平台。选择购物平台只是消费者购物的第一步，查看商品评论往往是消费者进行购物过程的第二步，因为通过商品评论，消费者可以得到参考依据，迅速了解产品，获得其他购物者的使用体验。除去商品的专有属性，消费者在进行网购时考虑的往往是图片是否真实、细节描述是否详细、评价是否有中差评、卖家服务态度好不好、售后的保障等等方面[3]。这些方面的评价往往决定着能否进行一次成功的购物。

随着科技的发展，手机也越来越成为生活中每个人的必需品，有数据分析单位公布数据分析汇报，2018 年 12 月我国网民平均每人每日玩手机的时间 5.69 个小时[4]。一部手机的平均使用寿命在二到三年，基于全国网民人口总量[5]，手机市场依旧十分火热，根据 CINNO Research 报告，去年上半年，国内共销售智能机 1.9 亿部，量级巨大。用户在购买手机时，往往从多方面进行考虑[6]，不同的人群有不同的需求，如女性群体会更加注重外观与拍摄功能，男性群体会更注重运行速度与参数，老年群体更注重字体大小等等。手机厂商受限于成本及技术原因，在手机产品功能的侧重点上会有所不同，但为了使自己产品成为更多群体的购买选择，在产品宣传时往往会对手机的全属性都赋予完美的宣传，这样不免会有些夸张的成分，就造成了用户在购买前手机之前，往往通过商品评论，借鉴其他用户购买后的使用体验，来辅助自己进行判断，最终决定是否购买。其他用户在进行评论时，不仅仅会给出购买流程的综合评价，也会从购买流程中的多个维度进行点评，如产品本身上的质量、性能、外观、使用体验等方向；服务方向上的客服、物流、售后等方向；平台方向上的活动、界面等方向。然而一款商品往往众口难调，本文实验希望通过文本挖掘从大量的评论中提取最具代表大众的观点，对用户、手机厂商、平台从多维度给出合理的改进建议。

1.1.2 研究意义

电商平台的商品评论具有大量性、快速迭代性、多样性、表达自由性等众多属性，购买者、商家、电商平台方都无法直接从平台中的商品评论获得当前商品的一个全面的信息。所以需要对商品评论进行文本挖掘，得到更直观，更具象的信息。王伟等[7]人证明了特征观点对用户消费存在作用。进行电商平台评论数据挖掘技术的整个流程会涉及到许多相关领域的技术研究，比如：爬虫、数据清洗、自然语言处理等方面。虽然流程复杂，但是还是十分具有研究意义的：对用户来说，可以帮助用户更加详细的了解手机商品的信息，了解其他用户到手后的使用情况，从而辅助用户做出决策；对商家来说，可以帮助商家了解当前产品的优势与不足，判断商品是否成功，了解当下用户的需求，帮助商家定位市场，发现商机，为以后的产品设计制订方向，同时，去研究竞对商品的商品评论，也可以了解对手的现状，做出更有利于商家的决策，还可以通过挖掘商品评论，对第三方服务进行评估考核；对平台来说，可以通过商品评论挖掘，发现违规商品、劣质商品、甚至违法商品，对此类商品进行管控，也可发现受用户喜爱的商品、用户回购率高等特征的商品，对此类商品进行流量扶持，提供精准推荐的参考依据等平台生态操作，在控制平台调性的同时提高服务质量，增加平台影响力与收入；对政府监管部门来说，可以通过商品评论挖掘，得出用户行为、商家行为、平台行为的基本情况，为监管提供有力依据。

1.2 用户评论分析挖掘国内外研究现状

电商平台的商品评论的文本挖掘研究是近些年的热点，每年国内外都有相当数量的科研人员对电商平台的商品评论进行挖掘与研究，并且取得了相当不错的成果。评论分析挖掘涉及到的内容有很多，主要包括用户评论关键词抽取、情感倾向分析和主题分析，本章将依次介绍这三部分的国内外发展现状。

1.2.1 用户评论关键词抽取国内外研究现状

用户评论关键词主要为产品特征词、平台服务词、厂商服务词，顾名思义，产品特征指的是描述一款产品本身的属性、性能、外观特点等词语或者短语。平台服务词则是基于电商平台的一些列服务相关的词汇，包括但不限于平台本身的体验、活动、第三方配送服务等词语或短语。厂商服务词就是基于厂商对此产品的一些服务保障，包括但不限于售后服务、客服服务、优惠政策、宣传是否相符等词语或者短语。用户关键词的提取目的主要是为了服务后续的情感分析，它要从大量的非结构化的商品评论中提取出用户所表达观点的特征。通常用户对商品的各个属性、各个功能评价指标不同，比如“大”这个词，“屏幕大”为正向情

感，而“噪声大”却是负向情感，这也是评论关键词抽取需要攻克的难点之一。

在目前的国内外研究中，用户评论的关键词抽取主要有人工提取与自动提取两种方式。其中人工提取的准确率非常高，但是费时费力，通用性差，每当有新的评论就需要人工再去提取，显然是不可取的。所以省时省力的自动提取成为用户评论关键词抽取的主要手段。根据是否需要给算法提供人工标注的数据去训练模型，可以将自动抽取分为有监督和无监督两类不同的抽取方法。其中有监督的用户评论关键词自动抽取方法准确率与成本都要比无监督的用户评论关键词自动抽取方法高，成本高是因为它需要提供人工标注的数据去进行模型训练[8]，准确率高则是经过了研究人员大量的实验证明。国外研究中， Luhn[9]早在上世纪五十年代就提出以文章中词汇出现的频率与文章中词汇出现的位置来确定关键词的开创性关键词抽取方法。在 Luhn 提出关键词抽取方法之后，国内外学者基于这个方法，通过不断地科研探索实验，得到了许多准确率更高、效果更好、速度更快的关键词抽取方法，其中 TF-IDF 算法是这些关键词抽取办法中的佼佼者，它的主要思路是通过统计一个文档中的所有单个词汇在本文档中出现的频次与这些单个词汇在其它文档中出现的文档个数之间的比值来确定文档中这个词汇对文档的关键程度。在实际应用过程中，研究人员往往会在 TF-IDF 算法原有的基础上进行优化调整，使得算法更加符合要进行分析的文本数据的特征。钱爱兵等[10]认为词汇本身的词性与长度、在文档中出现的位置、在文档中出现的次数都是在提取文档关键词需要考虑的因素，而且对应的权重也不同，于是提出了多因素综合加权的方式抽取文章关键词的方法，并且通过实验验证了该方法在实际应用上要比传统的 TF-IDF 算法效果优秀。胡学钢[11]将词汇出现在文本的词汇、位置联接为词汇链，然后把前后文的内容联接起来，使得备选关键词的特征信息得到相对的补充，并通过实验证明这种方法下的关键词抽取效果有所改善。张建娥[12]注意到汉语中词语并不是独立的，而是与其他词语之间有一定联系的特点，在传统的 TF-IDF 算法上以注重词语之间的关系进行改进对评论文本进行关键词的提取方法，实验数据表示，这种经过改进抽取关键词方式的实验结果要优于传统 TF-IDF 的统计方式，有效提高了平均召回率。王立霞[13]等利用词汇之间彼此的映射关系建立了网络结构，然后通过词汇语义与统计特征进行得分计算的方式获取关键词，解决传统 TF-IDF 这种关键词抽取方法缺乏对文本语义的理解缺陷，并用实验证明了该方法优于传统 TF-IDF 算法。由于传统 TF-IDF 方法在提取关键词时，没有认识到词汇彼此之间的联系，Madhu Kumari[14]等人为了解决这一缺陷 ，提出了术语加权方案(SBT)，主要思路是通过对词汇在文章中的含义相近词汇归类统计，以此来替换词汇在本文本与其他文本中出现的次数，减轻因未考虑词语之间的联系而影响关键词提取结果准确性的情况，并通过实验数据验证了该方法比传统 TF-IDF 算法抽取结果更好。

随着机器学习的不断发展，研究者又先后提出了利用 K-mean 聚类算法分析、朴素贝叶斯模型[15]和支持向量机算法[16]去进行关键词抽取的方法。为了提高抽取文章关键词的准确率与召回率，朱德泽[17]等提出了利用 LDA 模型对文章关键词进行抽取，并给出实验数据，证明了 LDA 抽取关键词的方法，效果符合预期。基于图的排序算法 TextRank[18]将词语与词语之间的关系度量当作图的边，单个词语当作图的结点，并以排序结果对关键词进行抽取。梁伟明[19]提出有向加权图比无向加权图效果更好的观点，并通过把 TextRank 模型关键词的长度、位置等信息进行融合的实验，证明了此方法是有效的。Yan Chen 等人[20]提出了一种比 TF-DF 和 PageRank 的方法性能更好的关键字提取方法，该方法基于图通过字共现图的高阶结构特征。郎冬冬[21]等将朱德泽和梁伟明的算法进行结合，利用 LDA 模型进行主题挖掘，并加入主题相关影响因素，构造无向加权图时利用特征信息，实验表明抽取的关键词覆盖率更高、表达意向更强。 在深度学习的影响下，李跃鹏[22]等构造了一种在长文本中表现优异的算法，该算法利用 Word2vec 进行文本分词并完成词嵌入，并以词汇之间的相似度进行聚类，获取文本关键词，并通过实验证明该算法的长处。Qinjun Qiu[23]等人则是把 Word2vec 算法和朴素贝叶斯模型整合在一起，并利用实验证明了该方法的效果，对比 TF-DF 在正确率与召回率的表现，都有不小的进步。Xiangling Fu[24]为了更准确的获取日常生活语言的表述特征，使用基于双向长短时记忆(BiLSTM)的模型去获取目标文本中的关键字，通过实验数据，可以发现这种方式有较好的效果。聂青阳与姚天昉[25]等人发布了一款分析论坛评论的系统，该系统会分析出用户对汽车这一商品的多种属性的满意程度与给出的改进建议。

1.2.2 情感倾向分析国内外研究现状

情感倾向性分析是文本挖掘的重点内容，通常进行数据所表达的感情极性，把文本数据表达的情感区分为积极情感评价、消极情感评价和中性情感评价三个部分。从数据的细粒度属性上进行区分，可以分为词汇、语句和篇章三个类型。我们对这三个类型分别探讨。

（1）词汇的情感极性判断

国内外现有研究中，词汇情感极性的判断主要建立语义词典的方法和基于语料库的方法。首先是用语义词典，这种方法也叫做基于语义倾向的方法，核心思想是对文本进行分词后，利用现有或自己新构造的情感词典对评论中的词汇进行匹配，根据评论中正向词汇和负向词汇的个数来判断整个评论表达的情感倾向。基于语料库的方法本质上是机器学习的方法，核心思想是通过事先标注好的训练集，构造高效的分类器，经过模型学习训练之后，寻找词汇之间的特征，完成情感倾向的判断。

国外英文类型判断大部分选用 Wordnet 和 General Inquirer，其中 WordNet 本体库是普林斯顿大学研发的在线词汇参照系统，General inquirer 词典含有 4207 个正负情感词汇，词典其中所有词汇都进行了词性、情感倾向、情感倾向强度的打标。极大的方便了使用者在进行文本情感倾向分析时的使用过程。目前，Kamps[26]等认为 WordNet 有效地限制了它们对名词和动词句法范畴的适用性。这是因为 WordNet 的相似性几乎完全集中在 WordNet 的分类关系中。我们通过研究 WordNet 最重要的图论模型，确定同义形容词语义取向的主观因素意义。Esuli[27]等提出了一种对这些术语定量分析的方法来确定主观条件，这些术语在联机字典中表示半监督分类。Andreevskaia[28]等提出了一种利用情感标签提取 WordNet 程序提取情绪形容词，该方法做了 58 步并且在唯一的非相交的种子列表中运行，对于每一个单词，计算出净重叠得分，使用网络重叠得分作为一个词汇隶属度的度量，获得最高网重叠分数的就是核心形容词。Liu[29]等认为 WordNet 是根据同义词组成的集合来判断词汇表达的情绪，但是同义的词汇可能存在不是或者没有相似情绪的情况，这样会对我们词汇极性判断时造成一些困扰。

在进行中文词汇情感判断的时候，我们往往会选择中国知网发布的情感词典 HowNet 和台湾大学中文情感极性词典。其中中国知网情感词典 Hownet，是 2007 年知网将《情感分析用词语集》发布于其官网，共发布英文和中文文件各 6 个。知网依据汉语词汇表达的含义将词汇进行了划分，各类别的具体情况为:程度级别词语，如:百分之百、绝对等共计 219 个；负面评价词语，如:不好使、不好用等共计 3116 个；正面评价词语，如:带劲、工整等共计 3730 个；负面情感词语，如:悲哀、可怜等共计 1254 个；正面情感词语，如:拜服、感恩等共计 836 个；还有一些主张词语。台湾大学中文情感极性词典共包含 11090 个褒义贬义情感倾向词。在现有研究中，各种情绪分析方法很少有人对中文词汇之间的关系做出判断， Dang[30]等提出了一个解决词汇分析中结构歧义的好方法：利用语义相关度来解决，并且 Dang 提出一个构建：认为文档中的每个汉语词汇与相邻词汇是无关的，在基于人工标注的词汇情感倾向，通过知网情感词典中各个词性之间的关系，从而使不同词性之间的关联度得到提升。语义相似度这个概念是有对称性的[31]，可以去算影响的因素，这个方法的核心思想在于对我们要探知其情绪状况的词汇，选择一些已知其正负情感倾向的词汇，再把这些词汇汇集在一起形成种子。Qu[32]等人为使用 bag-of-opinions 模型对文章的情感倾向进行计算，将汉语中的词汇分为三种词性：情感词、修饰词和否定词。为了获得更多未知情绪属性词汇的极性，Xu Y[33]等用类似于知网的 HowNet 这种情绪词典，以句子内部语法关系为基础，寻找与要进行分析正负向情感的词汇语义相似的一些词汇，并将这些语义相似的词汇参考已知词汇的极性也放到上面提到的集合，以此来判断未知情绪属性词汇的极性，这种方法取得了不错的效果，目前还在继续研究阶段，

是因为其过于依靠于已知的种子词汇[34]的情感极性情况，是这种方法最大的弊端。在利用情感词典分析词汇情感之外的方法主要有用无监督的方法来分析词汇的情感与用人工标注的语料库来分析未知情绪的文本情感的方法。

情感分析大多是基于情感词典对文本数据进行分析，所以情感词典中的正负向情感词完备充足，是否准确是文本分析的关键。在如何构建情感词典与使用什么样的情感词典，目前主要分为两类，一类是使用如知网 Hownet 和台湾大学中文情感极性词典这种已经很成熟的通用词典，但是显然通用的情感词典在特定领域上的准确度就会下降，进而影响整体的文本分析结果，显然不能满足我们日益精细、日益严格的要求；另一类是建立领域情感词典，在要分析的专业领域下，总结收集该领域下特有的情感词，接着与通用的情感词典进行结合去重，这样就打造了专业领域的情感词典。但是这种方法也有一个问题，不用专业领域之间的特有情感词也不尽相同，甚至不同领域之间的情感词倾向会存在冲突的情况，面对如此繁多的专业领域，短时间内，科研人员在使用时只能临时构建，没有统一标注，这也造成了同一领域内，两个不同的科研人员的情感词典可能会相差较大的情况。但是从长远来看，通过在通用词典上添加基准词建立专业领域的情感词典才是未来的发展方向。例如，Chen[35]等人添加微博领域专属情感词汇建立领域情感词典，进行微博内容分析。

（2）语句的情感极性判断

对颗粒度为语句的情感极性判断需要我们认识语句的本质：语句是多个词汇以一种顺序结构排列在一起的词汇的集合。同时语句的情感极性也是接下来判断篇章类文本极性的一个前置步骤部分，所以判断一个语句的正负倾向情感很是关键[36]。通常研究一个语句的正负情感倾向主要有两种方法，一种是机器学习，另一种是根据情感词汇来综合判断的方法。

对于语句的正负情感倾向分析，分为主客观与语句整体情感倾向两种类型。在主客观分析中，Wiebe[37]使用放弃标注语句和篇章，转而去标注颗粒度较细的词汇与短语的方法，利用分布相似度对词汇情感倾向聚类，实现分类。Wiebe[38]还介绍了一个利用手动注释意见，情绪，猜测，评价和其他私人状态语言的方式，产生对方案的描述和使用的例子——注释项目研究语料库。

Hatzivassiloglou[39]等通过对动态形容词作用的研究，面向语义形容词建立主观性强的预测。通过使用新的训练方法 SIMFINDER，计算语句之间的相似度，结合统计性的介绍和评价这两个指标，完善已有的自动分配定位标签技术，达到进行主客观分类的目的。但是，对于句子正负情感极性的判断，Hu[40]等通过构建积极词典和消极词典，对语句中的每个情感词确定语义方向，利用 WordNet 中词汇之间的联系，考虑句子中每个词汇的情感倾向，最终句子中词汇情感倾向比重占比大的词汇的情感倾向为这整句话情感倾向。Zha[41]通过分析研究语句之间的

依存关系来对文本进行无监督分析。Turney[42]等在研究语句正负情感倾向时，判断的方法选择了 PMI 判断方法，该方法的思路是首先抓取主观语句，通过形容词种子集来给所需要判断的情感倾向的句子中的词汇进行评分，再根据得分综合去判断整个句子的情感倾向。李钝[43]等人基于语言学的相关知识，将权重优先的计算方法引入到句子文本情感倾向分析，通过计算句子中各情感词的情感倾向及对应的权重，得到句子最终的情感倾向。

（3）篇章类型的情感极性判断

同样的，颗粒度为篇章级的文本本质是很多语句的集合，分析多个语句集合在一起后所表达的正负情感倾向，除了需要对集合中的每一条语句进行正负情感分析外，还要考虑到语句之间的相互关系、篇章整体的表达以及一些与篇章内容相关的其他知识，比如一些词汇通过不同的排列组合就会产生不一样的情感倾向，同样的，句子之间的排列组合顺序也会造成不同的情感倾向。所以我们进行篇章级颗粒度文本的研究时，通常选择机器学习的研究方式，比如 LDA 主题模型、SVM、决策树、K 近邻等算法。2002 年，Pang[44]等人创新般的使用机器学习技术对电影的评价文本基于主题进行做情感分析，主要用到了 LDA 主题模型、SVM 等算法，并取得了不错的效果。刘志明[45]等人分别使用朴素贝叶斯、支持向量机和汉语语言模型三种分类算法对微博文本数据进行分析，得出支持向量机算法对微博文本情感分析的效果优于其他两种算法的结论。

总的来说，中文的情感倾向性分析要比英文困难很多，主要原因是中文语句的句子成分要比英文复杂很多，句子中几乎每一个分词都会影响整体语句的情感倾向，这与英文语法环境是完全不同的，所以国外的研究应用在中文文本上效果并不好。名词搭配上形容词可以表达两种完全相反的情感情绪，程度副词、否定词和连词也会对句子含义产生较大的改变，而国内研究现状大多数是研究情感倾向性只涉及到情感词，只有少数的研究考虑到了程度副词的作用，采取将情感词和程度副词相结合的方式，进而提高了分析语句情感的正确率。结合中文语言特点，国内对情感倾向性分析主要有：谭松波、程学旗[46]等在训练分类器时，选择了多种方法；黄永文[47]等使用 SVM 处理消费者意见；张艳辉、李宗伟[48]等对用户发出的评论内容是否为真，和作用大小进行了摸索，确定了评论中真实内容较多时，产生的作用较大，施乾坤[49]等将 LDA 模型与可视化结合，取得了较为直观的展示效果，阮光册[50]利用一种复合模型对篇幅较短的内容进行主题发现。

1.2.3 主题分析国内外研究现状

主题模型是在给定的语料中，统计分析出这批语料所描述的主题，主题个数可以根据实际情况由研究人员设定[51]。文档中出现一个主题，就会有许多代表这个主题的词汇反复出现，文档中有多个主题，就会有多个可以代表不同主题的词

汇会反复出现，此时主题模型，可以从多个文档中寻找出那些代表文档主题的词汇使用的规律以及对应的主题，这样我们就从多个非结构化的文档中，提取出了这些文档讨论的主题。起初,人们通常认为文本的表达方式是 VSM(向量空间模型,vector spacemodel)[52]。上世纪六十年代，VSM 被 Gerald Salton[53]等人认为是主要用于信息检索，VSM 模型认为文档在词典空间内被表示，是词汇与文档之间关系的映射，是多对一的映射。随着文本技术的发展，科研人员逐渐认识到向量空间模型（VSM 模型）不重视词汇与词汇之间的关系，不注重判断是否是表达语义相同的关系。针对这种情况， Landauer 提出隐含语义分析。LSA 的核心思路是统计，由此发现了词汇之间存在关联，通过大量的文本数据进行实验，可以得到在同话题下文本的词汇，之间关系密切。于是研发人员借助线性代数的方法提取与词汇之间有联系的其他词汇，并研究词汇之间的语义联系，这个过程就是提取潜在语义。Hofmann[54]试着对隐含语义模型优化，在潜在语义模型的基础上加入了概率的因素，构造了概率隐含语义模型（PLSA）。PLSA 是新的改进，自动文档搜索潜在模型的基础上计算数据的因素，用概率去评估词汇、文档、隐含语义三者之间的关系。这个方法能很好的解决隐含语义模型没有注重词汇之间的关系的情况。但是在实际使用过程中，PLSA 经常会过拟合。因此， Blei 在 PLSA 的基础上进行改进优化[55]，提出了潜在狄里克雷分配模型（即 LDA，latent dirichlet allocation）来解决这个问题。LDA 模型灵活，具有良好的基础，使用范围广泛，是目前常用的主题挖掘模型，在本文第二章将会简单讲解它的原理。

1.3 本文研究内容与组织结构

1.3.1 论文的主要工作

本文主要针对京东平台上的苹果、小米、三星、华为、oppo 手机官方旗舰店在 2021 年售卖的旗舰机机型苹果 13、小米 12、三星 Galaxy S21、华为 P50、oppo Find X3 进行评论文本挖掘。

通过 python 爬虫技术，总计爬取了 8010 条各型号下的用户评论信息。然后对这些爬取下来的数据进行去重、清洗、分词、词性标注、商品评分映射为积极或消极情绪等工作，再之后分别利用决策树、情感分析、LDA 主题模型技术进行文本挖掘，通过词云图、京东用户评分对比、模型准确率等方式展示分析结果，并适当调整参数使结果更加准确。接下来对本次实验进行分析总结，找出实验的不足，提出建立手机领域专属语料库等改进建议与解决方法，并简单进行验证，并依据改进后的实验结果，给各个手机厂商、京东平台提供建议。

1.3.2 论文的组织结构

本文一共分为六个章节：

第一章是绪论，共有三个部分，第一部分是研究背景和意义；第二部分是介绍用户评论分析挖掘的国内外研究动态；第三部分是本文的研究内容与结构安排。

第二章是对本文用到的实验理念、实验工具、实验环境进行介绍。

第三章是数据预处理部分，第一部分介绍了文本挖掘的流程；第二部分是基于 python 爬虫技术对京东电商平台的手机评论进行采集；第三部分是数据预处理阶段，主要是去重、清洗、分词、词性标注等工作。

第四章是模型的构建与分析，第一部分是决策树分析；第二部分情感倾向分析的实现；第三部分是基于 LDA 模型的主题挖掘。

第五章是基于第四章的实现结果，对 LDA 主题模型中的不合理的步骤进行改进，通过添加领域情感词与领域分词的动作，成功改进实验。

第六章是总结与展望，对本文的主要研究工作进行总结，发现研究上的不足与待改进的地方，为未来的工作提出建议。

本实验的流程图如图 1-1 所示：

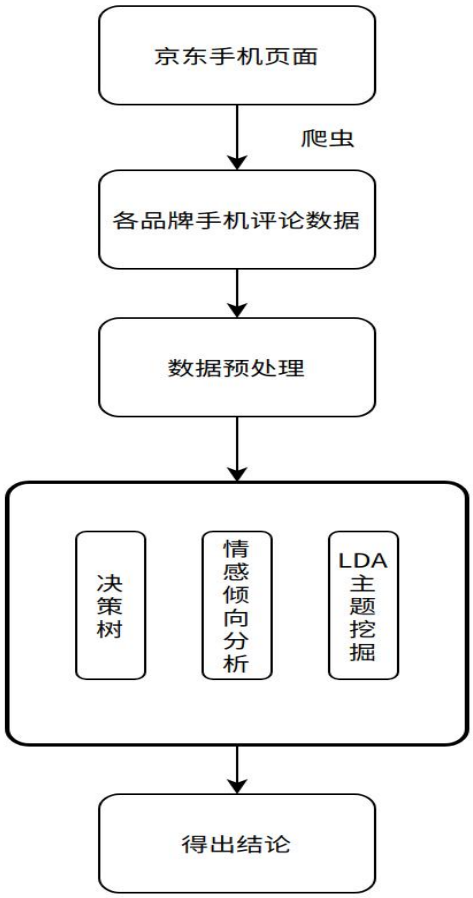

图 1-1 实验流程图

Figure 1-1 Experimental flow chart

第 2 章 情感分析相关理论概述

本章主要针对于本文文本挖掘实验中所用到的相关技术、理论知识及实验环境进行介绍，在文本挖掘过程中会涉及到网页爬虫、分词、停用词、Word2vec 词云展示、决策树、LDA 主题模型等相关技术。

2.1 数据的获取与处理

2.1.1 网络爬虫介绍

文本挖掘最基础的是文本数据的获取，现在互联网上公开的数据集种类较少，很难满足研究人员的研究需求。这时就需要研究人员根据自己的需求，在法律允许的范围内来获取文本数据。网络爬虫就是研究工作人员获取数据的首选，它是一个按照研发人程序设定，模仿浏览器，从目标网页自动高效准确的获取网

页内容的程序，并不需要用户自己去浏览网页获取信息，节省用户获取信息的时间与精力，也避免了因为人工注意力问题导致采集出错。通用的网络爬虫通常以给定初始页面的 URL 开始，在模拟浏览器获得响应后，将网页内容保存下来，进行初始网页布局分析，在网页源代码中找到所需要的内容，并通过正则等方式，将数据保存下来。当然，我们也可以根据自己的需求，通过循环等方式，设置爬虫自动爬取多个目标网站。上文只是简单介绍了爬虫的思路，python 中的 urllib、urllib2、re 等一些爬虫基本库可以很好的实现这些基础爬虫的功能，但更完善的爬虫能力，则需要爬虫框架、爬行策略等技术与思想进行支持，本文不展开介绍。随着爬虫技术的发展，各网站的反爬措施也日益完善，用户在进行数据采集时，还须遵纪守法。

2.1.2 中文分词及词性标注

词是最小表达有价值意义的单位，它可以由一个或多个连续的字组成，是 NLP 的基础。可以说分词的好坏直接影响后续所有的分析结果。由于中英文语言文字的差别，英文文本分词可借助词汇直接的空格进行很好的分词，而中文文本就比较困难，要结合字与字之间的关系。目前中文分词方法主要可以分为三大类：基于词典方法、基于词频方法和基于机器学习方法。各种方法的介绍如下：

1、基于词典分词方法。该方法是将文本和已有的词典进行比较和匹配。匹配的方式也可分为正向最大匹配法、逆向最大匹配法和双向最大匹配法，假如此时需要进行分词的中文文本为：

中文文本：“游戏开发从此变得很简单了”

词典：dict[]={“游戏开发”，“游戏”，“简单”}

正向最大匹配法会从文本中的“游”开始扫描，在扫描到“游戏”时，因为此时“游戏”既是词典里的词，又是词典中“游戏开发”的前缀，所以此时不能直接分词仍需继续扫描，若接下来的文字不为“开”的话，则分词为“游戏”，若为“开”，则需继续扫描，最终得出结论“游戏开发”是可以匹配的最长词，扫描结束，由此可见，最大匹配出的词必须保证下一个扫描不是词表中的词或词的前缀才可以结束[56]。逆向最大匹配法的匹配过程则与正向相反，首先定义最大分割长度为 4，然后从文本末尾取出 4 个汉字进行比对，即“很简单了”，由于“很简单了”没在词库表，去掉最左边接着匹配，“简单了”依旧没在词典中，去掉最左边接着匹配，最后剩下单字“了”，定为一个分词，然后根据最大分割长度 4，取出“得很简单”，进行上述匹配过程，以此类推，最终得到的最大匹配词分词也为“游戏开发”。而双向最大匹配则是综合了正向和逆向的优势，比较两者的结果，以词越长越好、单字越少越好的原则进行对比，并选择相对较优的结论输出。

基于词典分词这个方法的优点是简单、效率较高，缺点是它本身的匹配机制需要在汉语语言的复杂多样性的情况下，有一个完备覆盖的词典，显然中文语言博大精深，词典很难覆盖完全，这也造成了这种方法不适合大规模的文本分词处理。

2、基于词频分词方法。这个方法实际上就是通过文本中的“词频”进行统计排序，但是这里的词，是将文本中的任意相邻的两个字组合成词汇，然后统计这个组成的词在文本中出现的频率，以此类推，最后对所有语句中通过这种方式构成的词，进行词频的排序，当达到事先设定的阈值时，得到分词结论。但是这种方法的缺点也是显而易见的，这种相邻的字组合成的词，很大概率不是词汇，在计算时会浪费大量的运行成本，而且这种方法在忽视了长词汇的情况下，计算时的空间和时间复杂度依旧较高，不推荐使用。

3、基于机器学习分词方法。这种分词方法目前在国内外相对而言是比较前沿的，许多大型互联网公司都是采用这个方法。这个方法的主要思路是：通过已有的、成熟的语料集，结合合适的算法进行不断的训练优化，最终完成分词。由于基于机器学习分词方法过度依赖于语料集，语料的好坏直接影响最后的分词结果，所以这个方法需要大量的人工去标注语料集，剔除噪音数据，以保障语料的高质量性，供机器去学习训练模型，该方法在实现上需要巨大的成本。一般科研人员想通过此方法得到精准的分词，是有一定实现难度的。

基于上文的三种分词方法，现在研究人员已经开发出比较成熟的分词工作包。常用的中文分词包主要有结巴分词包、清华大学研究室发布的 THULAC、中科院计算所的张华平、刘群推出的 ICTCLAS 分词系统以及近两年北京大学推出的 pkuseg 分词包。结巴分词目前是在 nlp 领域使用最多，广受研发人员喜爱的分词包，它有全模式、精准模式、搜索引擎共三种模式。其中全模式是把句子按每个字进行切分，然后试图把每两个相邻的字进行组合，是词包中的词语就保存下来，这么做可能造成结果会出现重叠，造成一个字分别和前后两个字组成词汇；精准模式则是试图做到最精准的分词，结果不会出现全模式下的重复词汇，适合在文本情感分析时使用；搜索引擎模式是在精准模式结果的基础上，再对长词进行分词，不适合用于文本情感分析。THULAC 的特点是能力强、准确率高、速度快，THULAC 是利用了世界上规模最大的人工分词和词性标注中文语料库，大约有 5800 万字训练而成，模型标注能力强大。THULAC 标准数据集 Chinese Treebank（CTB5）上分词表现优异，是在该数据集上最好的几个分词效果之一。THULAC 在分词时可以做到 15 万字 s 秒，只进行分词时，运行速度可达到 1.3MB/s，同时进行分词与词性标注时，运行速度为 300KB/s。ICTCLAS 分词系统的 Free 版开放了源代码，为研究者提供了非常珍贵的学习参考资料，且支持多种操作系统，多种主流开发语言。ICTCLAS 拥有中文分词、词性标注、支持用户词典、新词识别等

多种国能，广受好评。ICTCLAS 的缺点在于每隔大概一个月左右的事件，需要更新证书，不然在运行 ICTCLAS 进行中文分词时会报错，更新证书才可以正常使用。pkuseg 与上文分词包相比，除了保持分词的高准确率外，相比于其他的分词工具包还新增了自己的特色：多领域分词与支持用户自训练模型，pkuseg 根据不同领域的领域名词特点，预训练了许多不同领域的分词模型供用户调用，在预训练模型未覆盖用户的领域时，也支持用户用自己领域的全新标注数据，训练所需的领域模型。实验证明，在专有领域上，pkuseg 的表现要优于结巴分词、THULAC 等一些通用的中文分词包。

词性标注(Part-of-Speech tagging, POS tagging),又称词类标注,是指对分词结果中的每个分词进行词性打标的程序，根据系统词包，确定每个分词是名词、否定词、动词、形容词或其他词性，对于标点符号也有统一的词性标注。中文中的词性标注相对英文词性标注较为简单，因为中文词汇中有多个词性的情况非常少见，大多词汇仅仅拥有一个词性，少部分词汇拥有两个属性，但是第一词性出现的频次要远远高于第二位的词性。通常国内的中文分词包在分词时，也会同步完成词性标注。

2.1.3 停用词与 TF-IDF

停用词(Stop Words)是文本挖掘过程中要用到的一个很重要的工具，它主要是指文本数据中高频次、低价值的词汇，去除掉文章中的停用词，会降低文本特征的维度或者提升文本数据特征的质量。在进行文本数据挖掘任务中，一些词汇是基于中文语法规范，用来充当其他词汇之间的联接，使得语句通畅，不能提供任何有价值的文本数据信息，即使把这些词去掉，也不会影响文本数据的大意，只是会造成语句的不通畅，但是保留这些词汇时，则会因为这些词出现的频次高，而对文本数据分析产生干扰噪音。那么停用词表又是如何产生的呢？

其实停用词表并没有标准的定义，通用的停用词表它只是相对较好的一个版本，目前一些开源的停用词表，如哈工大停用词表、百度停用词表、四川大学停用词表等完全可以满足一般文本数据分析的需求，研究者也可以自己制作停用词表，这就需要一个很重要的统计方法做支撑——TF-IDF。TF-IDF 是一种常用于信息检索与数据挖掘的加权技术，算法简单高效，常被用于文本数据的清洗。TF-IDF 有两层意思，一层是"词频"（Term Frequency，缩写为 TF），另一层是"逆文档频率"（Inverse Document Frequency，缩写为 IDF）。TF 的含义为词汇在单个文档中出现的频率，而 IDF 则为该词汇在所有文档中出现的频率，TF 计算方法为某个词汇在文章中出现的总次数/文章的总词数，IDF 的计算方法为 log（语料库的文档总数/（包含该词汇的文档数+1））。当计算出 TF(词频)和 IDF(逆文档频率)后，将这两个值相乘，就能得到一个词的 TF-IDF 的值。词汇在某个文本

中的 TF-IDF 越大，那么这个词在这个文档中越重要，相反词汇在某个文本中的 TF-IDF 越小，说明这个词对文档越不重要。这时我们将 TF-IDF 值很低的 k 个词汇总结出来，就是一个停用词表的雏形，在经过对这些词汇的筛选，删除有价值的词汇，就得到了可用的停用词表。在进行文本挖掘时，就可以在分词之后，对分词后的词汇与停用词表进行对比，若在停用词表中，就剔除掉，以此来提高文本挖掘的准确性。

2.1.4 词云图

词云图，又被称为文字云、标签云、关键词云，是一种将文本数据可视化的方式，它的核心价值在于可视化表达出文本中的高频词汇，使得文本传达的重要信息一目了然。它通过提取文本数据中的高频词汇，将这些词汇按照不同颜色、大小组成某一图案。由此可见，词云图的要素是由高频词汇、字体颜色、字体大小和图案行政这四个方面构成的，通过这种过滤掉了大量的文本信息，结合词汇词频越高，字体越大、颜色越引人注目、在图案中的位置越核心的展示方式，对用户有一定的视觉冲击效果，使得用户看一眼词云图，就可以大概的获得文章描述的主要内容。

词云图这种展示形式有一定的闪光点，也有一定的不足之处。它的优势在于：（1）词云图制作难度低，上手简单。（2）词云图比传统的饼图、折线图、条形图更有视觉冲击力，更有吸引力，迎合了人们的速读的心理。（3）词云图内容展示更直接，经过对文本内容的浓缩与精简，词云图展示的都是文本数据的核心关键点，使用户粗略的做到望图而知意，节省用户的时间。（4）词云图的应用场景多，它可以运用在舆情分析、主题抓取、用户分析等多项工作内，也可展示在数据分析报表、工作总结、PPT 等多个文档中。它的不足之处在于：（1）词云图展示时，造成信息的缺失。词云图的核心优势是高亮展示了高频词汇，让高频词通过字体、颜色等因素吸引目光，但是这也相应的造成了许多低频词汇不能很好的表达，这些低频词汇通常颜色暗淡、字体偏小，或许直接未被展示，这对某些有特定需求或者关注某些细节的读者需求来说，是不能满足的。（2）词云图通过字体大小来区分高频与低频词汇，这样对相差较大的词汇有很好的区分度，但是对于频率相近的词汇，并不能很好做出的区分。（3）词云图输出没有统一标准，受限于前置流程中分词选择、算法技术、词库的质量、图案的选择等原因，同一篇文本数据可能会被不同的人生成相差巨大的词云图。（4）词云图的表达内容缺少逻辑性，词云图的高频词都是从各个有逻辑结构的文本数据中提出出来的，彼此之间本身就没有很强的逻辑性，按照用户的图案展示后，几乎完全丧失了内在的逻辑结构，需要读者通过自身的知识和想象才能将高频词汇串联起来，猜测关键信息，这种逻辑结构的缺乏，也使得用户不能从词云图获得文本

数据描述的主题个数。

2.2 决策树

2.2.1 sklearn 简介

sklearn 是 python 中的机器学习库。sklearn 的全称 scikit-learn，它的功能非常强大，是基于 numpy、scipy、matplotlib 等数据科学包的基础，涵盖了机器学习中的样例数据、数据预处理、模型验证、特征选择、分类、回归、聚类、降维等几乎所有环节。sklearn 是 python 中传统机器学习的首选库，本文主要对其分类算法进行介绍。

- 线性模型，由回归任务和分类任务组成，其中回归任务中对应线性回归，分类任务对应逻辑回归，通过线性回归拟合对数几率的方式来实现二分类。
- K 近邻算法，是一个懒惰模型，无需进行训练，通过元素自身的位置，判断与周边的样本的距离，进而达到分类的目的。
- 支持向量机，源于线性分类，通过最大化间隔实现最可靠的分类边界。支持向量机的核心在于三个点：间隔、对偶、核函数。其中间隔是指由硬间隔升级为软间隔，解决了带异常值的线性不可分场景，对偶是在优化过程中求解拉格朗日问题的技巧，核函数才是支持向量机最关键的核心，它可以实现由线性可分到线性不可分的升级，同时避免维度灾难。
- 朴素贝叶斯，源于概率论中贝叶斯全概率公式，模型训练的过程就是拟合各特征分布概率的过程，是 LDA 模型的基础，本文将在 2.4 中进行详细介绍。
- 决策树，这是一个直观而又强大的机器学习模型，训练过程主要包括特征选择-切分-剪枝，本文将在 2.2.2 中进行详细介绍。

2.2.2 决策树模型概述

决策树是一个通过训练数据搭建起来的树结构预测模型，是一种监督学习模型，需要提供样本数据进行训练学习。决策树由结点和有向边组成。结点有三种类型：根节点、内部结点和叶结点，一般的，一棵决策树包含一个根结点、若干个内部结点和若干个叶结点。通过决策树模型，我们可以根据训练好的模型，高效的对新的数据进行归纳分类。决策树在使用时，是从文本数据的根节点开始，通过选择特征分支，依次向下，直至到达叶子节点，其中内部结点表示一个特征或属性，叶结点表示一个类，叶结点对应于最后的决策结果，其他每个结点则对应于一个属性测试。在数据挖掘中决策树是经常用到的很重要的技术，不仅可以用来做数据分析，还可以用来作预测。在进行预测时，每个结点包含的样本集合根据属性测试的结果被划分到子结点中，根结点包含样本全集，从根结点到每个

叶结点的路径对应了一个判定测试序列。决策树学习的目的强化泛化能力，加强对新文本数据处理的能力。

决策树学习的算法通常是一个递归地选择最优特征，并根据该特征对训练数据进行分割，使得对各个子数据集有一个最好的分类的过程。这一过程对应着对特征空间的划分，也对应着决策树的构建。实验开始时，进行根节点的构建，此时需要将所有训练数据都放在根结点，通过下文介绍的 Gini 系数选取最优特征，按照这个特征属性，对数据集进行分割，使得被分割的两个子集时在当前所有特征属性中最好的分类。如果子集未被完全正确分类，则对子集重复上述操作，若此时基本被正确分类，则构建叶节点，并将这些子集的数据划分到对应的叶节点中去。如此递归地进行下去，直至所有训练数据子集被基本正确分类，或者没有合适的特征为止。最后每个子集都被分到叶结点上，即都有了明确的类。这就生成了一棵决策树。

在决策树学习中，为了尽可能正确分类训练样本，结点划分过程将不断重复，有时会造成决策树分支过多，为了避免过拟合，可通过主动去掉一些分支来降低过拟合的风险。若当前结点的划分不能带来决策树泛化性能提升，则停止划分并将当前结点标记为叶结点，剪枝是决策树停止分支的方法之一。剪枝有分预先剪枝和后剪枝两种。预先剪枝是通过预先设定决策树生长指标，如高度、阈值，当决策树达到该高度指标时就停止决策树的生长；当决策树达到某个节点的实例个数小于阈值时，就可以停止决策树的生长。后剪枝则是先从训练集生成一棵完整的决策树，然后自底向上地对非叶结点进行考察，若将该结点对应的子树替换为叶结点能带来决策树泛化性能提升，则将该子树替换为叶结点。相比于预剪枝，后剪枝更常用，因为在预剪枝中精确地估计何时停止树增长很困难，而且后剪枝可以充分利用全部训练集的信息，但后剪枝在大样本集中的计算量代价比预剪枝方法要大得多。

2.2.3 Gini 系数

决策树是通过属性来构建节点，那么在属性的选择上就需要用到熵与 Gini 系数，它们都是决定决策树节点属性选择的指标，本文以 Gini 系数为例，Gini 系数公式如公式 2-1 所示：

$$G(\mathrm{p})=\sum_{\mathrm{k}=1}^{K}\mathrm{p_k}(1-\mathrm{p_k})=1-\sum_{\mathrm{k}=1}^{K}\mathrm{p_k^2} \tag{2-1}$$

Gini 系数代表了分支下样本种类的丰富性，样本种类越多，越混乱，Gini 系数就越大，相反，Gini 系数越小，说明分类越纯，结果越是理想，我们越能确

定它属于哪一类。当它最小为 0 的时候，是完全分类的状态。因为 Gini 系数等于零是比较理想的状态，一般实际情况下，Gini 系数介于 0 和 0.5 之间。当文本数据有若干属性值时，通过某个属性，得到的 Gini 系数都是非常小的值，可以说这个属性是个很好的区分类别的属性。例如当原始数据比较混沌，Gini 系数为 0.5 时，我们通过每种属性进行分类，使得新得到的两组数据 Gini 系数都比较小，对两组数据的 Gini 系数做加权平均就得到了此时的 Gini 系数，选择使此时 Gini 系数最小的属性作为结点，就完成了一次分类。以此类推，对新的数据继续使用此方法，直到分类结束。Gini 系数的不断最小化，实际上就是提高分类正确率的过程。构造决策树的基本思路是随着树的深度，也就是层数的增加，让 Gini 系数迅速降低，Gini 系数下降的越快，代表决策树分类效率越高。

2.2.4 决策树的优缺点

决策树的优点在于：决策树可以清晰的展示出文本数据中，哪些字段比较重要；决策树的计算量相对其他算法来说并不是很大；决策树可以生成可以理解的规则。

决策树的缺点在于：决策树作为监督学习，需要样本训练集，样本集的质量决定着决策树的准确性；在决策树中，当因类别过多产生错误时，错误增加的速度，影响的范围会很快；决策树不能进行主题挖掘。

2.3 情感倾向分析

文本情感分析又称意见挖掘、倾向性分析等。是对给定文本，通过一定的算法进行处理分析，进而推理归纳出文本想表达的感情倾向的过程。随着互联网各种社交平台、电商平台的发展，用户会发出大量的评论、点评、攻略，来表达自己对人物、舆情事件、物品、行为、平台的看法或者建议，这也促使了文本情感分析的发展。许多平台和商家从这些评论中获取有价值的信息，反馈给本身的经营策略，从而提供更好的服务。

2.3.1 电商评论的特点

电商平台商品的用户评论不同于其他文本，它有独有的特点：（1）整体评论篇幅较短，可能由几个词汇或者几个句子组成。电商用户评论通常是用户在购买使用商品之后，对整个购买流程及商品使用感受的表达，描述方式自由、随意、带有用户的主观情感色彩。（2）文本语言不规范，电商的用户评论本质还是属于网络产物，因此评论往往会带有浓浓的网络语言的特点，不遵守传统的行文规范，评论内容相对口语化，充斥大量网络用语，网络用词，比如“永远滴神”、“yyds”、“666”，这些不是传统的表达词汇，但是都表达了积极肯定

的正向情感。（3）评论内容的主题较为集中，电商用户评论主要目的还是为了分享自己的购物体验，所以话题会集中在商品性能、购买体验、售后服务等几方面进行评价。（4）情感词汇变化少，通常电商平台用户的评论是针对与某一商品的属性进行评价，评价的指标往往是固定，比如“屏幕大/小”，“像素高/低”等商品属性的描述。（5）短句成文且多特征，电商的评论文本通常都较短，可能由词汇或多个短句组成，且可能存在一句话多个短句构成，每个短句表述的属性有都不一致，比如“屏幕大，充电快，运行快，配送快，价格低”，短短一句话的五个短句描述了这次购物体验的五个维度。这些都是电商平台商品评论的特点，也是我们评论文本挖掘可以重点努力的方向。

2.3.2 构建情感词典

判断词汇的情感极性的方式主要是通过情感词典。构建情感词典的方法主要有：1、将公开的情感词词典进行整合，如知网的 HowNet 情感词典、台湾大学的 NTSUSD 情感词典，词典越多越好，经过增加、删除、合并而整合后的词典就会更精准更全面；2、对现有的情感词典进行拓展，主要是根据词典内容，补充同义词，但实际上并没有补充的新的其他情感含义的词汇；3、构建领域词典，使用通用的情感词典来识别中文中所有场景下的情感词是远远不够的。在某些特定领域中，一些非常规的情感词也可以表达出情感倾向，比如“这个键盘是防水的”，其中防水就是键盘领域中的积极倾向。这是就需要通过 PMI 点互信息与左右熵结合的方式去识别这些词汇，PMI 点互信息的核心思想是概率统计，如果这个词汇总是与某一类情感一起出现，那么这个词汇是这个情感倾向的可能性就会越大。具体方法为，选中核心情感词作为种子词汇，这个种子词必须情感明确，再找到与情感种子词关联度最高的 top-n 个词语，就可以添加到对应的情感词典中了。

在追求更精细的情感倾向分析时，仅靠情感词典去判断一个句子的情感是远远不够的。中文博大精深，语义丰富，有许多情感都是含蓄表达的，并没有通过情感词表达情感，比如：“我用来你这个面膜，脸部就开始痒痒”，这句话达的是消极的情绪，但是没有使用情感词。还有中文中反讽的场景，将正话反说，比如：“你家卖的手机很好用，我只能呵呵了“。如果用通用的情感词典去匹配，有”好用“，”呵呵“两个词汇都是表达的积极情感，但这句话加进了网络语言，表达的却是消极情绪。要处理这种复杂的文本表达方式，需要结合中文语法进行更深入的研究。

2.3.3 带权情感词

判断词汇的情感极性主要是需要定量来衡量每个单词的词汇极性，通常选择

[-1,1]中的数来表示一个词汇的情感极性，当词汇的情感极性数值大于 0 时，表示该词汇是积极的，当词汇的情感极性数值小于 0 时，表示该词汇是消极的，并且词汇的情感极性数值绝对值越趋近于 1 时，代表词汇正负向情感程度越强烈。但这种把所有情感词都归类到 0、1、-1 是不严谨的，因为同都是积极情感或消极情感时，不同情感词汇表达出的程度也是不相同的，如知网情感词库中正向情感词关于“好”的词汇就有：好、刚好、还好、超好、最好等等一些词汇，这些词汇表达出的情感是有层次的，同样的消极情感词也会有这也对应的层次表达，当一个文本句子中，既有正向情感倾向，又有负向情感倾向时，将整个句子归为中性情感是不合理的，这时就需要依据正、负情感词表达出情感所带的权重高低，去综合判断才是严谨合理的。情感词加权的方式与标准目前还没有统一的标准，需要根据研发人员自行判断添加。

2.4 LDA 主题模型

2.4.1 朴素贝叶斯

朴素贝叶斯（Navie Bayes,NB）是统计学中最为常用的分类算法，它是基于概率论的知识进行统计，从而分类。本质是对于需要分类的目标事项，经过发生概率的计算，得到属于哪一分类的概率最大，就将这个目标事项分到哪一类别。

朴素贝叶斯分类算法的基本原理是来自贝叶斯理论，贝叶斯公式如公式 2-2 所示：

$$\mathrm{P}\left(B_{\mathrm{i}} \mid A\right)=\frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)} \tag{2-2}$$

朴素贝叶斯（NB）是所有贝叶斯模型中最成功的一种。它在贝叶斯理论的基础上，增加了假设所有的特征之间都是彼此相互独立的前提，彼此发生的概率互不影响，且彼此之间的权重是相同的。在这种前提下，通过特征之间联合分布概率来计算后验概率。朴素贝叶斯算法因算法简单，实现容易，易于理解，所以常常会用来分类，而且其本身在分类方面效果不错。但是在实际应用过程中会发现往往各个特征项之间都会具有一定的依赖关系，很少有相互之间都为独立的情况。

2.4.2 LDA 主题模型

为了更好的消费者评论进行情感分析，更加直观的看出其的优势跟劣势，一般研发人员会采用 LDA 主题模型对整体商品评论分别从好评与差评出发，进行主

题分析。通过主题模型的运用，可以发现评论中各个词汇之间的联系，从而得到文本中相对有用的信息。可以通过主题模型，将某一主题下的相关词汇归类到一起，并展示出来，从而更加深入的看出消费者对商品的情感倾向。LDA 模型基于无监督学习的一种典型的主题模型。

LDA 主题模型是由 Blei 在 2003 年提出，全称为潜在狄利克雷分配（LDA,Latent Dirichlet Allocation）生成式主题模型。LDA 主题模型的本质是包含词、文档、主题三层结构的贝叶斯概率模型，即每一篇文档表示了文档中的主题所构成的概率分布，而每一个主题又代表主题中词汇所构成的概率分布，以此实现了对文本的建模。LDA 主题模型采用的是词袋模型。所谓词袋模型，就是在文档中，考虑词汇是否会出现，而不去考虑这个词汇出现的顺序。在贝叶斯概率理论中，如果后验概率和先验概率满足同样的分布律，那么，先验分布和后验分布被叫做共轭分布。而 LDA 主题模型中的狄利克雷分布就是多项式分布的共轭分布。

在 LDA 主题模型生成的流程中，我们首先要输入的是 m 篇文档，此外我们还需要输入主题个数 K，此时 K 不是最佳的，所以我们可能按照不同主题数量的值，进行多次主题分析。LDA 假设文档的主题与主题中的词先验分布都为 Dirichlet 分布，此时就会有对应主题的多项分布，由于主题产生的词是不依赖于某一个具体文档，所以文档主题分布和主题词分布是独立的，这样就有 K 个主题与词的 Dirichlet 分布和 K 个主题编号的多项分布，这样就组成了 Dirichlet-multi 共轭，基于 Dirichlet 分布的主题词后验分布能通过贝叶斯推断的方法得到，这基本就是 LDA 模型的思路。基于 LDA 主题模型可以使用 Gibbs 采样算法求出每篇文档的主题分布和每个主题下的词汇分布。

如上文所述，LDA 模型分析文本时，文本的主题数量 K 是无法事先确定的，但主题数的多少对聚类的结果有很大影响，当主题数量过多时，则会造成语义过于分散，无法抓取文本的主要主题，当设置的文本主题数量过少时，就会出现一个主题下包含着多层语义信息的情况，这两种结果都不是好的主题挖掘。所以一个合适的主题数的确定对主题挖掘的结果是非常重要的。在确定主题数的时候，我们通常的做法是计算一定范围内的主题数下，在各个主题数量时的平均余弦相似度，选择平均余弦相似度较低时，所对应的主题数作为文本分析的主题数。

2.4.3 词向量与余弦相似度

所谓词向量，本质就是将文本词语进行数字化，并通过向量的形式表示。词向量技术也是自然语言处理过程中一个很重要的技术。它可以通过映射把词汇或者短语以一个相对低维度的实数向量在空间中表示。文本分类分析中，词汇之间在语义上的相似度也是通过向量空间的相似度进行判断，将空间向量相似的词汇

划分为同一类，达到词语归类的目的。

词向量表示方式有很多种，其中比较简单的一种词向量的表示方式是 onc-hot-representation，简称独热向量，是一种离散表示。这种表示方法创建一个文本中出现的所有词汇的一个词典，把每个词都表示为一个长向量。向量的长度等于词典大小，每一个词向量的分量只有一个 1，其它都为 0。这种方法表示简单，但也有两个非常明显的缺陷，第一就是会造成维度灾难，当词典很长时，会导致词向量也同样边长，在计算时需要巨大的资源。第二个缺陷是这种表达方式不能保留词汇之间彼此的关系信息。另外一种词向量的表示方法为分布表示，这个方法通过将词语映射为一个固定长度且相对较短的向量，分布表示认为词汇表达的语义信息是在这个词汇结合了相邻词汇以后，一起来构成的。词向量空间由全部的词向量组成，其中的每一个点都表示一个词汇，通过计算这些词汇之间的距离，来判断这些词汇之间的相似度。代表工具就是 2013 年发布的 wod2vec，wod2vec 的本质是一种浅层的神经网络模型，它有两种网络结构：CBOW 和 Skip-gram。其中 CBOW 通过获取目标位置词汇相邻两边的上下文，去预测目标位置词汇，而 Skip-gram 则是通过目标位置词汇预测上下文出现的概率。由于对词汇使用了唯一的词向量，也导致 Word2Vec 不能对多义词词汇进行很好的表示与处理。

在判断两个词向量是否相似时，我们经常用到一个概念是余弦相似度，其实这个值不仅仅被用作判断词向量的相似程度，将语句、文本或者词典等其他元素用向量表示后，都可以使用余弦相似度来判断两者之间的相似程度。余弦相似度的计算方法是计算两个向量之间余弦夹角的大小，它的取值范围在[-1,1]，余弦值越接近于 1，说明两个向量相似程度越高。但是在实际应用中，余弦相似度会遇到两个问题。第一个问题是两个向量被定义的方式要是统一的标准，第二个向量是两个向量的长度需要保持一致。当文本特征向量长度不一致时，通常通过依靠主观经验设置关键词，或者剔除不重要的词汇，但是这样处理后，对准确率都会有影响。另外，余弦相似度算法需要对文本逐个向量化并进行余弦计算，需要计算资源较大，因此余弦相似度只适合短文本，而不适合长文本。

2.4.4 gensim 库

gensim 库是一款 Python 工具包，在从初始分结构化文本数据到文本的隐含主题的向量表达上表现出色，过程是主要利用无监督学习。在处理自然语言上，这个库包含了大量的常见模型，诸如 LSI、 LDA， HDP、DTM、 DIM、TF-IDF、word2vec、paragraph2vec 和基本的语料处理工具。

gensim 是使用 python 语言开发的，安装和其他工具包一样可以使用 pip 安装，十分方便。在使用 gensim 进行文本挖掘之前，我们需要将文本数据进行预

处理，将这些初始文本数据处理成 gensim 可以处理的稀疏向量格式。使用时，在将文本数据切割成词汇后，使用 dictionary = corpora.Dictionary()生成词典， 并通过 doc2bow()方法将文本数据转化为稀疏向量形式，接着通过使用 save 函数将词典持久化，保存在语料库 corpus 中。然后就可以在 models 中调用 LDA、 DIM、TF-IDF 等模型对刚刚保存的 corpus 进行模型训练、分类等处理。最后可以在 similarities 中计算余弦相似度，帮助使用者选择更好的主题数量。在这些功能中，将文本数据处理成向量是 gensim 的关键。gensim 也支持将训练好的模型进行保存，以便下一次使用。 gensim 的功能十分强大，是值得深入学习的。

2.5 本章小结

本章介绍了本文接下来实验会用到的主要原理与工具，包括数据获取与处理、决策树、情感倾向分析、LDA 主题模型等相关内容，受篇幅限制，不能详尽，但主要思路已经明确，下章开始将进入本文的实验部分。

第 3 章 电商评论文本挖掘流程及数据预处理

本章主要介绍了本文电商评论文本挖掘实验的整体流程及电商评论文本的数据预处理。介绍了电商评论建模分析的整个流程，同时通过介绍网页爬虫获取数据，并将其进行预处理，数据清洗形成可以用于文本分析的数据源，之后对数据源通过决策树、情感分析、LDA 模型分别从有监督的机器学习、无监督机器学习、主题分析挖掘进行实验，并通过对比各个模型的实验结果进行分析挖掘，挖掘出电商评论中的价值点。

3.1 电商评论文本挖掘流程

下图 3-1 为电商评论文本挖掘整体流程：

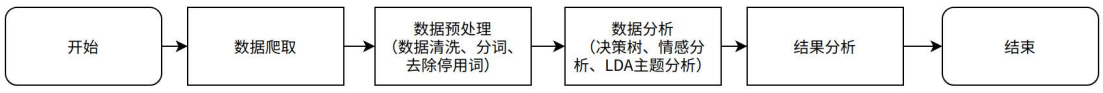

图 3-1　电商评论文本挖掘流程图

Figure 3-1 E-commerce review text mining flowchart

图 3-1 电商评论文本挖掘整体流程

- 数据爬取：选中目标平台店铺（本文以京东手机电商评论为数据来源）后，通

过爬虫技术从京东平台各大手机厂商官方店铺中采集旗舰手机评论相关的数据；

- 数据预处理：使用正则表达式对爬取到的数据进行清洗，去除重复数据以及官方自动评论等无意义数据，对获取的用户评论相关数据进行分词及去除停用词操作，为数据分析提供可直接应用的数据；
- 数据分析：通过决策树模型分析，给出评论文本中的重要词汇；通过情感分析，对整条评论进行打分，标注为积极或消息；通过 LDA 主题模型分析，分析出各个主题下的情感词，并制作词云图，给出更直观的数据感受；
- 结果分析：对各品牌的手机旗舰机商品评论分析指标化后的数据进行分类分析挖掘，从多维度得出各个品牌相应的结论，以了解用户的需求、意见、购买原因、产品的优缺点等等，以此对京东平台、各个手机厂商、用户提供建议。

3.2 数据爬取

3.2.1 爬取数据的选择

基于本文第一章研究背景及意义介绍，本文主要选取京东电商平台的手机类商品进行评论文本挖掘，根据各个手机厂商在我国占有的市场份额及用户比例，选取苹果、华为、小米、OPPO、三星五个手机厂商进行评论文本挖掘分析，选取的京东店铺为各个厂商在京东平台的官方旗舰店，选取的机型为各个厂商 2021 年发行的手机旗舰机。选取各厂商的京东官方旗舰店可以有效避免因第三方经营问题而产生的产品质量、假货、客服态度、非官方配送时效等问题，使评论更加聚焦在厂商本身的问题上，而非第三方产生的问题上。选取各厂商 2021 年旗舰机的原因是，旗舰机作为各个厂商每年的主打品牌，在手机质量、售后服务、发展方向都很好的能体现出手机厂商的状态，而其他年份或其他机型因目标年代或目标群体不同，不能准确反应手机厂商现在的状态。下表 3-2 为选取的各个品牌手机型号、发布时间、在京东的店铺名称及网址。

表 3-2 各个手机品牌型号统计

Table 3-2 Statistics of mobile phone brands and models

品牌	2021 年旗舰机	发布时间	京东店铺名称	店铺链接
苹果	苹果 13	2021 年 9 月 15 日	Apple 产品京东自营旗舰	https://item.jd.com/100026667910.htm
华为	HUAWEI P50	2021 年 7 月 29 日	华为京东自营官方旗舰店	https://item.jd.com/100014453209.htm
小米	小米 12	2021 年 12 月 28 日	小米京东自营旗舰店	https://item.jd.com/100017508685.htm
OPPO	OPPO Find X3	2021 年 3 月 11 日	OPPO 京东自营官方旗舰店	https://item.jd.com/100018535884.htm
三星	Galaxy S21	2021 年 1 月 19 日	三星京东自营官方旗舰店	https://item.jd.com/100011464877.htm

3.2.2 爬取工具及爬取内容的选择

近些年随着爬虫技术的需求与发展，市面上出现了很多典型的爬虫软件产

品，比如八爪鱼采集器、火车头采集器等。本实验最初选择八爪鱼采集器进行上述商品网址的评论采集，但是通过设置采集任务要求，在完成采集后，发现该采集器只能采取到京东评论的默认排序，也是就是只能采取到好评文本的数据，并不能采集到差评的评论文本，这显然不能满足本实验的文本要求。所以本实验采取通过 python 编程实现各个手机型号评论数据的采集，采集的数据包括京东平台展示出的该评论下的商品 id、用户昵称、评论内容、评论时间、商品颜色、商品尺寸、评论回复人数、评论的点赞数、商品评分。为后续操作方便，我们采集时将评论内容设置为 content，将商品评分设置为 content_type。受京东网址反爬策略所限，最终此次实验在本阶段共采集 8010 条京东电商平台手机评论数据，其中苹果 13 评论 2000 条、华为 P50 评论 1210 条、小米 12 评论 1540 条、OPPO Find X3 评论 1420 条、三星 Galaxy S21 评论 1840 条，包含 5 分评论 4940 条、3 分评论 60 条、1 分评论 3010 条。

3.3 数据预处理

3.31 爬取的评论情感倾向标注

有监督的机器学习需要人工标注好结果，再交给机器模型去训练分析学习，然而面对本实验中的 8000 多条评论，人工去标注不仅费时费力，而且在一条评论描述了多个论点时，主观猜测用户评论意图容易出现错误。面对这个问题，本实验通过借助京东商城的评论打分机制解决，由于每条评论都有对应发出用户对商品的打分，5 分评论为好评，1 分为差评。本实验将 5 分评论设置为正向情感（pos），1 分评论设置为负向情感（neg）。

3.32 去重及数据清理

商品的评论数据从电商平台爬取下来后，是无法直接使用的，因为其中有大量的整条评论重复、单条评论中词汇大量重复、无意义特殊符号等低质量评论，这个时候我们需要对数据进行去重及数据清理工作。

（1）评论文本去重

在爬取的评论中，我们会发现很多重复内容，这些重复的评论数据对我们后面的机器学习等分析不仅不能提供价值，还会造成一些不必要的麻烦，所以要首先去除掉，造成这些重复的主要原因为：京东平台的自动评论与用户本身的省事行为。其中自动评价主要为购买者在购买后很长一段时间没有评价的情况下，电商平台的系统会自动为购买者做出好评评价，而且评论内容也为一样；用户本身的省事行为主要体现在复制他人的评论，造成冗余的信息，这两种原因造成的评论没有分析的价值，所以在进行语义分析时，将重复数据去除。

（2）机械压缩去词

除去上文的整条评论是重复文本外，还有大量的单条评论中存在重复的词汇情况，例如：“不错不错不错不错”、“很垃圾很垃圾很垃圾”、“一般一般一般”等评论。这种类型的评论语句看起来很有内容很长，其实不然，句子中的重复词汇都是可以去除的，只需要保留整条评论中的一两个词汇就可以很好的表达出评论的中心思想，比如上述三句评论，分别保留为“不错”、“很垃圾”、“一般”就可以准确的表达出句子的情感，去重重复文本的这个过程，就是“机械压缩”的过程。

本文共爬取各品牌手机评论 8010 条，去重后还剩 7956 条。

（3）无意义符号去除

除去重复文本外，评论中的无意义符号、字母数字、一些专有名词也会对文本挖掘造成负面影响。本实验通过 python 正则表达式，去除掉了评论中的数字、字母、京东、苹果、华为、三星、oppo、小米这些专有名词。其中去除字母与数字的原因是这两种词出现频率高，但是不能表达出情感，所以在数据预处理阶段将其去除；而除去京东、苹果、华为、三星、oppo、小米是因为这些专有名词在京东平台对应的商品评论里，大概率为高频词，且对语义分析没有帮助，故也在预处理阶段将其去除。

3.33 分词

分词是文本挖掘分析的关键一步，这个步骤中我们将评论由句子转化为更容易判断情感的词汇，同时也将非结构化数据转化为结构化数据，为方便后文的使用与研究，本实验分词后构造的 DataFrame 包含：每个分词、对应的词性、分词所在原句子的 id、分词所在原句子的情感（content_type）。

本实验使用的是结巴的 psg.cut()分词，使用 psg.cut()分词对评论分词后，就会得到分词的词和它的词性，再通过特征工程降维，与该分词所在原评论所在顺序、该分词所在评论对应的 content_type 再次匹配，再次特征工程降维，就可以得到目标 DataFrame。下图为分词后构建的 DataFrame，可以发现经过数据清洗后的 7956 条评论被分割为 329232 个词汇。

3.34 去除标点符号与停用词

在 3.33 中，实验将评论分词后得到 329232 个词汇，但是这些词汇中有大量的标点符号与停用词噪音，我们仍需将这些噪音去除，并在 DataFrame 中新增分词在原评论中所在顺序的列，为后续情感分析做准备。

去除标点符号通过前文的到的分词词性，发现标点符号词性为“x”，去除词性为“x”的分词，就完成了标点符号噪音的去除。导入 5748 个停用词后，将停用词与上文分词做对比，经过运算去除掉停用词。本实验分词后的 329232 个词汇经过去除标点符号与停用词后，还剩下 146229 个分词。此时我们可以对剩下的分词进行该词在评论中位置的顺序与匹配，特征工程降维，得到含有每个分

词在语句中顺序的 DataFrame，如图 3-3 所示。

	index_content	word	nature	content_type	index_word
0	1	这款	r	pos	1
3	1	家人	n	pos	2
7	1	下单	n	pos	3
8	1	购买	v	pos	4
11	1	到货	v	pos	5

图 3-3 评论分词后的 DataFrame

Figure 3-3 DataFrame after comment segmentation

3.4 本章小结

本章通过爬虫技术对京东平台上的目标手机商品进行了爬虫爬取，获得了本次实验的基础文本数据，并通过文本去重、高频专有名词去除、无意义符号去除对文本进行了初步的清洗工作，将 8010 条数据清洗为 7956 条数据；随后又通过结巴 psg.cut()分词将初步清洗后的数据进行分词，将 7956 条评论分割为 329232 个词汇，再通过去除标点符号、去除停用词两步操作，将 329232 个分词词汇清洗为 146229 个分词，并构建了含有分词所在原句子的 id、分词、对应的词性、分词所在原句子的情感、分词所在评论的位置顺序的 DataFrame，为后续机器学习、情感分析等操作提供可直接使用的数据。下图 3-4 为分词完成后的各手机型号的词云图展示，按先从左到右再从上到下的顺序依次为 OPPO Find X3、三星 Galaxy S21、HUAWEI P50、小米 12、苹果 13 和所有型号评论汇总的词云图，可以粗略的看到拍照、屏幕、速度这几个词出现在了每个词云，通过词云图我们可以大致了解到每个手机型号的评论中出现的高频词汇是什么，但是对用户的所表达的情感与主题尚不清晰。

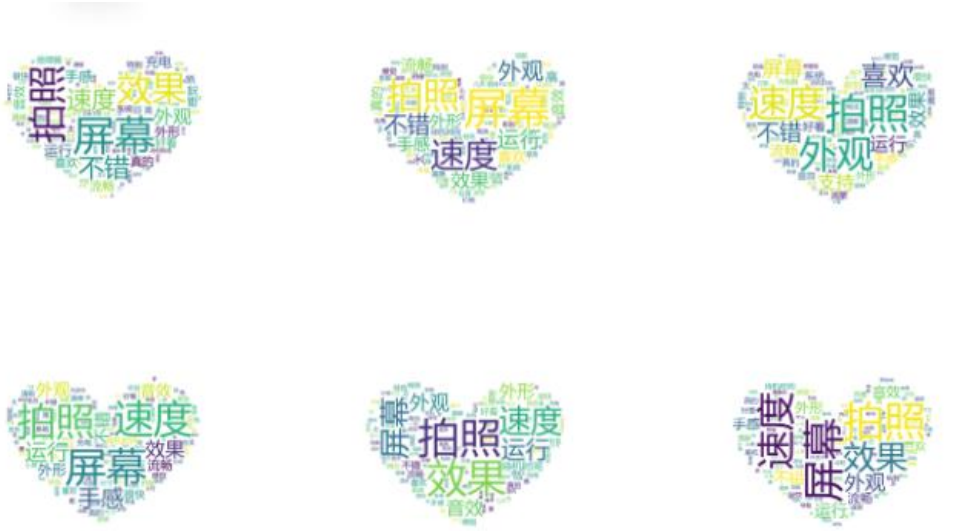

图 3-4 各型号手机评论词云图

Figure 3-4 Word cloud of mobile phone reviews of various models

第 4 章 模型构建与分析

4.1 决策树

本阶段我们用决策树做情感分类，因为决策树属于有监督的机器学习，需要提供人工标注的评论文本情感分析结果给机器去训练学习。上文提到我们已将爬取到的评论中，5 分评分标注为积极情感，1 分评分标注为消极情感。本阶段将把数据分为训练集与测试集，训练集交由机器进行学习，并用学习好的模型对测试集进行预测。

DecisionTreeClassifier(max_depth=5)

在训练集上的准确率：0.88

	precision	recall	f1-score	support
3	0.00	0.00	0.00	9
neg	0.76	0.94	0.84	584
pos	0.96	0.84	0.89	998
accuracy			0.87	1591
macro avg	0.57	0.59	0.58	1591
weighted avg	0.88	0.87	0.87	1591

在测试集上的准确率：0.87

DecisionTreeClassifier(max_depth=10)

在训练集上的准确率：0.93

	precision	recall	f1-score	support
3	0.00	0.00	0.00	9
neg	0.87	0.93	0.90	584
pos	0.96	0.92	0.94	998
accuracy			0.92	1591
macro avg	0.61	0.62	0.61	1591
weighted avg	0.92	0.92	0.92	1591

在测试集上的准确率：0.92

DecisionTreeClassifier(max_depth=15)

在训练集上的准确率：0.95

	precision	recall	f1-score	support
3	0.00	0.00	0.00	9
neg	0.88	0.92	0.90	584
pos	0.95	0.94	0.95	998
accuracy			0.93	1591
macro avg	0.61	0.62	0.62	1591
weighted avg	0.92	0.93	0.92	1591

在测试集上的准确率：0.93

DecisionTreeClassifier(max_depth=20)

在训练集上的准确率：0.97

	precision	recall	f1-score	support
3	0.00	0.00	0.00	9
neg	0.90	0.92	0.91	584
pos	0.95	0.95	0.95	998
accuracy			0.93	1591
macro avg	0.62	0.62	0.62	1591
weighted avg	0.93	0.93	0.93	1591

在测试集上的准确率：0.93

图 4-1 不同深度决策树

Figure 4-1 Decision trees of different depths

经过构造特征空间和标签，划分训练集、测试集，词转向量后，就可以构建决策树。此时一定要注意进行训练集、测试集的划分，再进行词转向量。这是因为词转向量测试集中的 transform 的实例化是由训练集训练得到的词转向量。上图 4-1 分别为决策树的深度在 5、10、15、20 的时候，决策树模型在训练集与测试集的表现，我们发现在树的深度为 10 的时候，决策树在训练集与测试集已经达到了不错的效果，训练集的准确率有 0.93，测试集的准确率有 0.92，此时的准确率为模型预测与人工标注的对比。对应的局部决策树如图 4-2 所示，共有样本 6361 个，左面的样本数量较多，有 4452 个，其中拍照，手感，打字都是比较重要的因素。

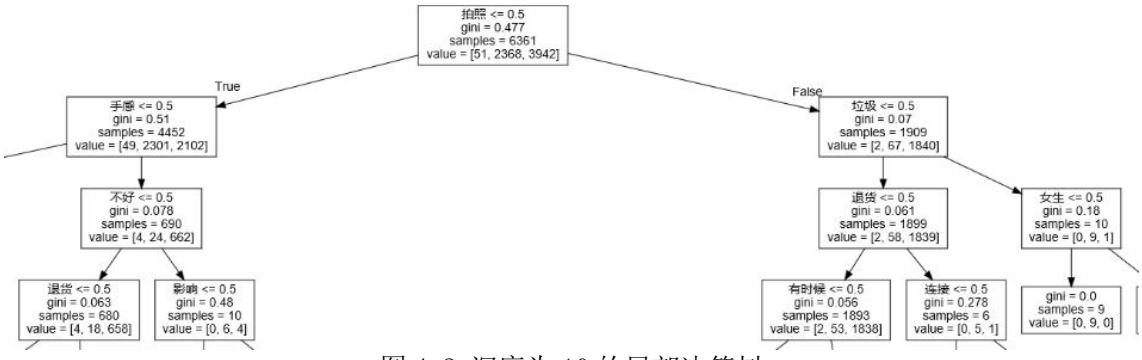

图 4-2 深度为 10 的局部决策树

Figure 4-2 Decision tree local with depth 10

4.2 情感倾向分析

4.2.1 目标与思路

本阶段情感分析的思路与目标如下，不同于有监督的决策树，情感分析是一种无监督的机器学习，并不需要数据集中人工打的情感标签，但是我们需要通过算法，让机器明白每一条评论代表的情感是正向还是负向，才能使机器开始无监督学习并对测试集进行情感标注。

4.2.2 情感修正

首先我们对每个分词进行情感判断，是正向就标记为 1，负向就标记为-1，不是情感词就标记为 0。其中在判断分词情感的时候，除去词汇本身的词性外，还要考虑词汇修正后的情感，因为在中文的语句中，否定可以改变词汇的情感，双重否定又可将否定的情感再次改变，所以在进行评论句子情感分析的时候需要考虑到被否定词作用的修正情感。

实验过程中，我们首先导入知网发布的情感分析用词语集，有 3743 个正面评价词语、3138 个负面评价词语、833 个正面情感词语、1251 个负面情感词语。同时我们还可以将近期新出现的评价词按正负向加入到情感分析用词语集中，并将每个分词对应的权重联接到前文的 DataFrame 中，下图 4-3 为联接后的 DataFrame。

	index_content	word	nature	content_type	index_word	weight
0	1	这款	r	pos	1	0.0
1	1	家人	n	pos	2	0.0
2	1	下单	n	pos	3	0.0
3	1	购买	v	pos	4	0.0
4	1	到货	v	pos	5	0.0

图 4-3 增加权重后的分词 DataFrame

Figure 4-3 Word segmentation DataFrame after adding weights

接下来我们做情感修正，这就需要用到否定词语集，本实验所用否定词语集共有 21 个否定词，将否定词导入后，即可开始对评论词情感的修正。首先我们在上文的 DataFrame 上新构造一列，用于存储修成后的情感，然后将经过数据预处理的分词进行分析，只保留情感词，本实验中的分词中 7225 个情感词。在进行情感词修正的时候，如有这个情感词被多重否定，那么当否定次数为奇数时，

情感词为否定，当否定次数为偶数时，情感词为肯定；本实验在进行修正情感时，采取观测该情感词在对应评论中前 2 个词是否为情感词，来判断是否为否定的语气。此外在判断时，有两种特殊情况，如果情感词在句首，则没有否定词，不需要判断；如果情感词是在句子的第二个位置，则只需要看前 1 个词，来判断是否是否定的语气即可。

4.2.3 计算每条评论的情感值

在获得每个情感词的情感倾向后，我们就可以计算整个评论的情感得分，通过 groupby()和 sum()函数可以将同一评论下的情感词相加，对应得到的情感分数和，就为每条评论的情感值得分，因实验目的不同，本实验此处忽略了不同情感词所占权重不同的问题。得到每条评论的情感值得分后，就可以对该评论的文本情感做出判断，即当情感值得分大于 0 时，该评论为正向评论，当情感值得分小于 0 时，该评论为负向评论。

经过计算，爬取的 7952 条评论中，只有 3933 条评论分析出了正负向情感，另外还有 4019 条评论未被分析出正负向情感，被定义为中性评论，这与我们已知的评论评分占比相差甚大，至少有一半以上的评论无法判断出它的情感。

4.2.4 情感词得分与商品评分对比

在计算每条评论的情感词得分后，可以与商品评分做对比进行比较。使用 pd.merge()表联接的方式，将评论情感得分联接到大表中，并通过 pd.crosstab()函数将评论情感得分与商品评分进行对比，下图为对比结果，可以发现预测出来情感的 3933 条评论中，有 588 条负向评论，2807 条正向评论是一致的，也还有 525 条评论是不一致的。

也可以使用 classification_report()进行更直观的展示，如图 4-4 所示，我们发现预测出情感倾向的评论情感词得分准确率为 0.86，效果尚可。

	precision	recall	f1-score	support
3	0.00	0.00	0.00	13
neg	0.63	0.76	0.69	773
pos	0.94	0.89	0.91	3147
accuracy			0.86	3933
macro avg	0.52	0.55	0.53	3933
weighted avg	0.87	0.86	0.87	3933

图 4-4 情感预测结果

Figure 4-4 Sentiment prediction results

4.2.5 词云图展示

在得到评论正负向得分后，此时我们可以通过词云图展示出正向评论与负向评论分别对应的词频，如下图 4-5 所示，左面为正向评论，右面为负向评论，我们可以更直观的发现正向评论都在关注些什么，负向评论又在关注些什么。但是这时我们又有新的疑问，评论一个手机维度有很多，例如拍照功能、运行速度、拍照效果、外观造型、充电速度等等方面，此时词云图中的喜欢、不错、满意、垃圾、不好等形容词，我们也无法判断是在形容哪一个手机属性，它的指向是不明的，为了解决这个问题，就需要引入下文的实验方法——基于 LDA 模型的主题挖掘。

图 4-5 情感分析词云图展示

Figure 4-5 Sentiment analysis word cloud display

4.3 基于 LDA 模型的主题挖掘

LDA 模型采用的是一种词袋模型。所谓词袋模型，是在一篇语料文档中，我们只去考虑这个词汇是否会出现，而不去考虑这个词的出现的顺序。在词袋模型中，每一个词都是等价的。我们通过 LDA 模型可以很好的挖掘语料中的主题个数，并分析出在不同的主题下，用户评论都有什么不同的词。LDA 的优势是：不需要人工调试，只用相对较少的迭代次数，就可以找到最优的主题结构。

4.3.1 建立词典与语料库

本阶段实验进行之前，需要准备 LDA 主题挖掘过程中需要用到 gensim 库，主要使用库中的 models()函数与 corpora()函数，所以需要提前将 gensim 库安装好。上文得到的正向评论数据与负向评论数据可以直接导入，在本阶段使用。因为 LDA 主题模型是一种词袋模型，所以我们新建立词典，并进行去重动作后，建立一种 gensim 可以识别的语料库格式。

4.3.2 主题数寻优

接下来就到了 LDA 模型的重点，主题数寻优。首先构造一个余弦相似度的函数 cos(vector1,vector2)计算两个向量之间的相似程度，用来后续判断不同主题下的关键词文本的相似度。接下来就可以构造主题寻优函数，如下文代码所示，此函数可以重复调用，解决其他项目的问题。构造函数步骤为：

- 初始化平均余弦相似度：构建平均余弦相似度的一个 list，并向列表添加加平均余弦相似度为 1 时的元素，此时主题数为 1，余弦相似度为 0，余弦数为 1，做主题分类就没有意义。
- 构建主题数为 2 到 10 的循环：遍历生成各个主题个数下的平均余弦相似度情况，本实验设定的主题个数为 2 到 10 个主题个数，当主题个数大于 10 时，主题划分过于精细，也不具有研究意义。
- LDA 模型训练：在主题数遍历条件下，分别对每个主题数使用 models.LdaModel(x_corpus,num_topics=i,id2word=x_dict)进行 LDA 模型训练。
- 提取关键词：在主题数遍历条件下，LDA 模型训练后，即可根据训练结果取出每个主题下的 TOP 关键词，进而利用分词取出该主题下的每一个关键词。
- 构造词频向量：取出每一主题下的所有关键词后，对该主题下的关键词进行去重处理。
- 构造主题词列表：行表示主题号，列表示各主题词，并统计 list 中元素的频次，返回元组。
- 主题余弦相似度：利用返回可迭代对象的所有数学全排列方式，计算出主题个数之间，两两组合的余弦相似度。
- 平均余弦相似度：最后计算出不同主题个数下的各主题两两之间的平均余弦相似度，选取平均余弦相似度最低时对应主题的个数，就为最优的主题分类。
- 主题寻优函数构造好之后，就可以把本实验中的正向评价词与负向评价词传入函数，得到在不同主题数下的平均余弦相似度，并将各个品牌正面评论主题数与负面评论主题数寻优结果展示出来。

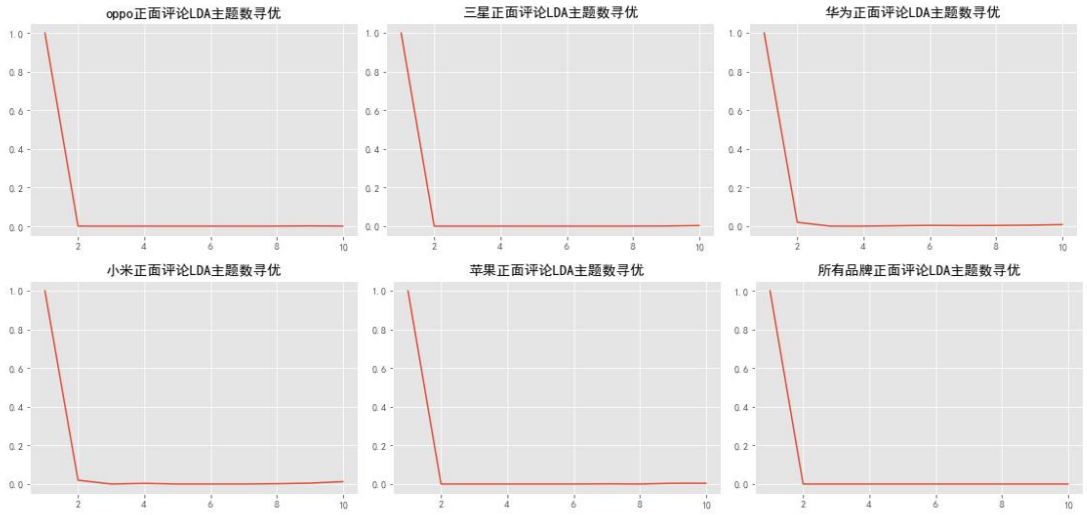

图 4-6 所有品牌正面评论主题寻优

Figure 4-6 All Brands Positive Review Topics Search

根据上图 4-6 所示，几乎所有品牌的手机正向评论在主题数为 2 的时候，就到了最优的主题分类，之后 2 到 10 个主题时，也都保持着最优的主题数，这与我们认知中的用户对手机多维度评论是不符合的，所以我们以苹果 13 手机为例，分别计算出苹果正面评论 LDA 主题数为 2 与主题数为 10 时的主题 TOP 关键词。

苹果正面评论 LDA 主题数为 2 时，运行结果如图 4-7 所示：

```
[(0,
  '0.046*"屏幕" + 0.039*"外观" + 0.033*"待机时间" + 0.032*"喜欢" + 0.031*"外形" + 0.019*"流畅" + 0.019*"清晰" + 0.017*"满意" + 0.012*"系统"
+ 0.010*"手感"'),
 (1,
  '0.056*"速度" + 0.055*"拍照" + 0.048*"运行" + 0.047*"效果" + 0.036*"音效" + 0.033*"不错" + 0.026*"很快" + 0.024*"好看" + 0.014*"特色" + 0.
013*"特别"')]
```

图 4-7 苹果正面评论 LDA 主题数为 2

Figure 4-7 The number of LDA topics in Apple's positive comments is 2

此时可以看到苹果 13 用户的正向评论分为两个主题。其中第一个主题对屏幕、外观、外形、待机时间、系统等手机属性表达了喜欢、满意的态度；在第二个主题中，用户对运行速度、拍照等手机属性表达了不错的态度。

苹果正面评论 LDA 主题数为 10 时，运行结果如图 4-8 所示：

图 4-8 苹果正面评论 LDA 主题数为 10

Figure 4-8 The number of LDA topics in Apple's positive comments is 10

此时可以看到苹果 13 用户的正向评论分为 10 个主题。这 10 个主题就可以比较全面的展示用户对苹果 13 使用体验的正向评论，第一个主题对充电、电池的属性表达了支持的态度，属于产品方面；在第二个主题中，用户对拍照时的体验质感表达了正向态度，属于产品方面；在第三个主题中，用户对看电影时音效清晰表达了正向态度，属于产品方面；在第四个主题中，用户对续航、信号能力表达了正向态度，属于产品方面；在第五个主题中，用户对玩游戏的手感方面表达了正向态度，属于产品方面；在第六个主题中，用户对快递服务表达了正向态度，属于服务方面；在第七个主题中，用户对价格合适与否表达了正向态度，属于价格方面；在第八个主题中，用户对京东自营的活动表达了正向态度，属于平台方面；在第九个主题中，用户对手机屏幕表达了正向态度，属于产品方面；在第十个主题中，用户对物流速度、双十一活动表达了正向态度，属于平台方面。

同样的，平均余弦相似度对负面评论的主题数量进行寻优。

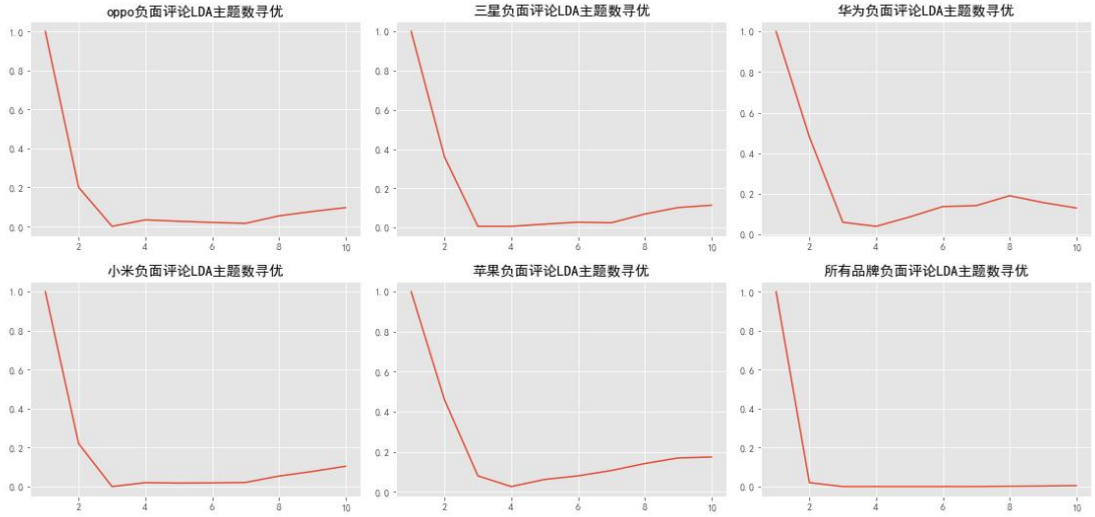

图 4-9 所有品牌负面评论主题寻优

Figure 4-9 Search for all brand negative comment topics

通过上图 4-9 所示，各个品牌的负面评价最优主题数分别为 oppo3 个、三星 3 个、华为 4 个、小米 3 个、苹果 4 个、所有品牌 3 个。我们依旧选取苹果品牌作为例子，进行分析。苹果 13 负面评论主题数为 4 时，此时各个主题下的评论词如图 4-10 所示：

```
[(0,
  '0.026*"信号" + 0.023*"好看" + 0.016*"充电" + 0.012*"保价" + 0.011*"降价" + 0.010*"几天" + 0.010*"特色" + 0.010*"头" + 0.009*"送" + 0.009
*"真"'),
 (1,
  '0.085*"屏幕" + 0.080*"拍照" + 0.057*"贵" + 0.046*"外形" + 0.035*"待机时间" + 0.023*"太" + 0.020*"真的" + 0.015*"卡顿" + 0.014*"流畅" +
0.012*"包装"'),
 (2,
  '0.078*"效果" + 0.070*"运行" + 0.068*"速度" + 0.065*"外观" + 0.054*"音效" + 0.036*"清晰" + 0.023*"卡" + 0.016*"度" + 0.016*"正品" + 0.014
*"显示"'),
 (3,
  '0.110*"不好" + 0.083*"高" + 0.040*"慢" + 0.022*"不错" + 0.021*"感觉" + 0.019*"像素" + 0.018*"手感" + 0.016*"坑" + 0.014*"充" + 0.013
*"电"')]
```

图 4-10 苹果负面评论 LDA 主题数为 4

Figure 4-10 Apple negative review LDA topic count is 4

此时不难看出，第一个主题中用户对降价、保价、送充电头等因素产生了不满，属于营销活动方面；第二个主题中用户对待机时间产生不满，属于产品方面；第三个主题中用户对运行速度和运行效果产生了不满，属于产品方面；第四个主题中用户对充电慢、充电方面产生不满，也属于产品方面。

通过对苹果 13 正向评论文本与负向评论文本的主题挖掘，我们可以大概的知道苹果 13 在京东平台售卖的过程中，哪些方面是被用户认可的，哪些方面又是被用户不认可的，可以给到京东平台及苹果手机公司一个粗略的方向建议，但是此时，在结果中我们也会发现一些问题，有一些本应该出现在正向评论中的主题词，如“好看”，却出现在了负面评论主题中，说明我们的模型还有需要改进

的地方。

第 5 章 主题挖掘实验的改进与结论

上一章中，我们已经通过 LDA 对各手机厂商的京东电商评论进行了正负向的主题挖掘，但是综合看来，仍有瑕疵。本章将总结实验中的不足，并通过调整，以期获得更准确的主题挖掘。

5.1 实验分析

分析第四章中的实验步骤，我们可以逐一可能存在问题的操作：1、在数据清洗时，我们选择了把数字字母清洗掉，其实这样会造成的一定的误伤，比如“很 nice”、“666”、“很 6”，这些词汇是近些年很火的网络词汇，也表达出很强烈的情感，而且在手机领域中，评论出现数字的频率较小；2、同样在数据清洗时，我们选择了去除标点符号，这样就可能造成了不同的词拼接在一起，产生非用户评论的词，从而导致情感的改变；3、评论大多不是由一个句子组成，通常由几个句子或短语组成，即使是有一个句子组成，也会发生在一个句子里对手机的多个属性进行点评的情况，这时候情感倾向是不一致的。比如：“手机屏幕大，运行速度快，耗电快”，这句话中包含了对手机屏幕，手机运行速度和手机电量的评价，情感倾向分别为正、正、负，但是按照现状的情感词典去判断，会将这个语句定义为正、正、正，最终表示的情感也会被定义为正向情感，这是因为缺少专业领域的情感词典所致；4、同样的，由于缺少专业领域的分词工具包，而是使用通用的结巴中文分词包，也导致了一切词汇被错误的切割，比如“息屏”就被切分为了“息”、“屏”。本章将针对这些问题点，改进实验。5、在情感词的使用上也存在着一定的问题，首先如同上文提到的一样，由于确实手机领域的专属情感词库，会导致一些情感词无法正确的识别，其次就是通用的词库在手机领域也不是全部适用，比如“手机用起来电下的很慢”和“手机运行速度很慢”两个评论中都提到了“很慢”，在现有的知网情感词库中，“很慢”是一个负面评价词语，若就此把这两条评论都归为负面评价显然是不对的。

5.2 实验改进

通过分析我们得知了上文实验中结果不是特别准确的问题所在，下面我们将逐一进行改进优化。

针对在数据清洗时，清洗掉数字字母会造成误伤，影响情感倾向的情况，我

们尝试了不清洗数字字母与将含有数字字母的网络词进行同义替换两种改进方式，其中不清洗数字字母方式，会造成少量的数字成为高频词，而且留下的网络用语并不能被分词词包或者情感词准确识别；而进行同义替换的方式相对较好，但是容易产生遗漏。

对于因数据清洗时去掉标点符号导致词汇的识别错误和评论中由多个短句组成，且短句的情感倾向不一致的情况，我们采取拉长评论数量的处理方式，利用换行符、句号等对评论进行切割，并将分割后的几条内容都视为一个完整的新评论。

在分词步骤上的改进，我们通过利用结巴分词可以增加用户自定义词库的功能解决。首先我们下载多个搜狗细胞词库，将手机相关的词汇的词库进行汇总，并转成 txt 文档。通过 jieba.load_userdict（）函数进行自定义词库的加载。这样我们就解决了大部分分词错误的问题，但是还有少部分词汇仍然不会被识别到，我们通过手工的方式继续添加手机领域的名词，如：息屏、闪退、黑屏等等词汇。然后根据分词结果，发现依旧没有识别到的词汇，重复进行上面手工添加的动作，继续完善分词词典。在改进之前，苹果 13 手机评论会被结巴分词为 72195 个词，改进后评论会被结巴分词为 72085 个词，证明改进过程中，减少了 110 个因为结巴词库未覆盖而造成的词汇未被准确分词情况。

在情感词的改进上，我们实验中使用的是知网发布的情感词，为了使得情感词更全面，我们选择在知网情感词的基础上，通过去重加入台湾大学情感词和清华大学李军的情感词，结合三大主流情感词，提高情感词的覆盖面与准确度。另外再通过人工发现的途径，手动添加手机领域的情感词。并且在之后运行的时候，发现了不合理的情感词，如负向评价词中的“不到”，进行了剔除的动作。

5.3 实验结论

在上述动作的调整后，实验的结果变得更加准确，我们根据实际情况，将正负主题数都设置为 8，下面我们将对这实验结果进行分析，并给出各个手机厂商生产经营建议与用户的购买建议。

5.3.1 对手机厂商的建议

表 5-1 oppo Find X3 主题挖掘结果

Table 5-1 oppo Find X3 theme mining results

oppo Find X3 正向评论 LDA 主题模型结果								oppo Find X3 负向评论 LDA 主题模型结果							
Topic 1	Topic2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic8	Topic 1	Topic2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic8
流畅	漂亮	特色	不错	满意	喜欢	拍照	屏幕	不好	物流	垃圾	失望	东西	发热	客服	一点
外观	细腻	完美	确实	值得	特别	清晰	运行	体验	差	显示	王者	几天	屏幕	真的	充
速度	外形	失望	手感	希望	舒服	效果	灵敏	电量	包装	失败	选择	效果	旗舰机	感觉	一加
推荐	好看	真的	精致	体验	音效	充电	太	怀疑	赠品	华为	手里	帧	一会	快递	选
显微镜	高	很棒	很快	东西	新	期待	很好	麻烦	换	拍照	一如既往	吐	玩	太	买回来
合适	耐用	解决	感觉	荣耀	购买	支持	送	特别	不到	送	无语	频繁	没想到	骗人	感受
白色	好评	足够	系统	有趣	充电器	适合	摄像头	电池	本来	电	卡	月	期待	降价	数据
苹果	颜值	电	强大	棒	华为	电池	轻薄	小时	激动	有种	系统	充电	下单	可惜	视频
做工	换	选择	发热	续航	拿到	质感	习惯	烫	打开	半个	第一次	软件	惊喜	降	惊艳
长	优惠	收到	待机	稳定	小时	真	快递	没法	价格	质量	网络	外观	货	拍	欺骗

上表 5-1 为 oppo Find X3 主题挖掘结果，通过对 oppo Find X3 的正负向评论主题挖掘可以发现，用户在购买 oppo Find X3 的过程中，满意的点有：外观漂亮、音效舒服、续航稳定、拍照清晰、屏幕灵敏等方面；不满意的点有：电池不好、物流差、玩游戏卡、容易发烫、客服骗人、买后降价等方面； 所以对 oppo 手机厂商的建议是：在产品方向上，提高产品性能，解决手机卡顿问题，在手机配件上解决产品痛点，提升电池性能，保持现有的屏幕、拍照、音效的闪光点；在服务质量上，加强对客服的管理，做好岗前培训、岗内绩效考核，提升顾客体验；在物流方面，可以根据实际情况，更换第三方服务商或提升自身出货及配送服务的效率，达到行业平均水平以上，提供保价服务让客户满意。

表 5-2 三星 Galaxy S21 主题挖掘结果

Table 5-2 Samsung Galaxy S21 theme mining results

三星 Galaxy S21 正向评论 LDA 主题模型结果								三星 Galaxy S21 负向评论 LDA 主题模型结果							
Topic 1	Topic2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic8	Topic 1	Topic2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic8
速度	流畅	漂亮	稳定	外观	完美	清晰	喜欢	差	失望	东西	卡	垃圾	保护	麻烦	讨厌
外形	屏幕	特别	舒服	精致	合适	运行	满意	退	退货	真的	特别	发热	太	怀疑	价格
音效	耐用	不错	电池	希望	音乐	效果	值得	屏幕	感觉	模糊	华为	不好	客服	无语	体验
推荐	特色	系统	感觉	高	待机时间	发热	拍照	后悔	没法	确实	随便	降价	评	激活	投诉
充电	手感	失望	惊喜	好看	很快	真的	确实	送	号	耗电	不到	购买	充电	真	一个月
轻薄	很棒	购买	细腻	价格	续航	太	习惯	建议	摄像头	问	申请	电池	不快	消耗	失败
一点	支持	体验	卡	肯定	解决	适合	服务	宝	很快	掉	不行	天	上当	没想到	到手
荣耀	信赖	不行	实在	收到	玩游戏	差	产品	厉害	耐用	续航	掉电	质量	没用	收到	理由
能力	发烫	很好	超级	颜色	强大	好评	质感	满意	运行	自营	电量	开机	信号	几天	解决
换	时间	物流	真实	放心	送	大小	赞	速度	担心	喜欢	超过	找	售后	差劲	购物

上表 5-2 为三星 Galaxy S21 主题挖掘结果，通过对三星 Galaxy S21 的正负向评论主题挖掘可以发现，用户在购买三星 Galaxy S21 的过程中，满意的点有：屏幕手感、外形轻薄、外观精致、拍照质感等方面；不满意的点有：充电不快、价格讨厌、耗电快、降价等方面；整体来说，三星 Galaxy S21 的现状是除了电池与外观外，其他方面中规中矩，既没有被用户讨厌，也没有赢得用户的喜欢，所以对三星手机厂商的建议是，首先大力解决被用户诟病的电池和电量问题，提供保价服务，然后在比较平庸的手机性能方向大力发展，毕竟用户购买时，手机性能才是核心的因素，在手机性能上没有闪光点就容易被市场淘汰。

表 5-3 华为 P50 主题挖掘结果

Table 5-3 Huawei P50 topic mining results

华为 P50 正向评论 LDA 主题模型结果								华为 P50 负向评论 LDA 主题模型结果							
Topic 1	Topic2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic8	Topic 1	Topic2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic8
值得	喜欢	特别	清晰	希望	满意	流畅	发热	不好	充电器	垃圾	失望	失望	东西	发热	信号
国产	支持	期待	拍照	速度	鸿蒙	运行	效果	麻烦	垃圾	想	支持	客服	体验	不配	垃圾
很快	漂亮	手感	屏幕	很好	感觉	推荐	很棒	视频	充电	鸿蒙	苹果	太	推荐	续航	耳机
耐用	完美	忠实	物流	信赖	音效	好看	外形	电	没想到	电池	开机	实在	送	差	拍
系统	不错	超级	速度快	收到	充电	特色	真的	感觉	时	超级	拍照	快递	真的	真	充
合适	外观	颜值	终于	东西	直屏	确实	实在	配件	怀疑	自营	新	打开	购买	担心	喜欢
开心	舒服	电	新	充电器	功能	果断	中	没用	线	舒服	两天	第二天	不到	可惜	伤心
快递	太	待机时间	性价比	国货	苹果	强大	适合	延迟	时间	开	不行	激动	屏幕	网络	系统
失望	质量	很漂亮	精致	优秀	价格	体验	足够	遗憾	流畅	抱歉	评	荣耀	内存	旧	抱歉
时间	屏	电池	游戏	舒适	颜色	高	只能	质量	一个月	气死	第一次	完美	物流	包装	年

上表 5-3 为华为 P50 主题挖掘结果，通过对华为 P50 的正负向评论主题挖掘可以发现，用户在购买华为 P50 的过程中，满意的点有：系统很快、外观漂亮、拍照清晰、物流速度快、鸿蒙系统满意等方面；不满意的点有：充电方面、客服失望、购买不到、旧包装、信号垃圾、鸿蒙系统垃圾等方面；所以对华为手机厂商的建议是：保持现有的外观、拍照优势，加强充电性能的发展，提高客户考核，用新包装对手机进行包装，提高产量以供给市场需求。这里还注意到正负向评价的主题里都提到了鸿蒙系统，说明华为自主研发的系统目前还没有获得所有人的认可，众口难调，华为作为国产之光，在自主研发的鸿蒙系统优化上任重而道远。

表 5-4 小米 12 主题挖掘结果

Table 5-4 Xiaomi 12 theme mining results

小米 12 正向评论 LDA 主题模型结果								小米 12 负向评论 LDA 主题模型结果							
Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8
手感	满意	屏幕	希望	流畅	特别	喜欢	不错	屏幕	差	欺骗	模糊	失望	发热	体验	垃圾
合适	清晰	运行	真的	完美	发热	拍照	速度	东西	麻烦	实在	流畅	不好	玩	高	打游戏
推荐	舒服	确实	颜值	外观	值得	很快	适合	缺点	真的	后悔	帧	感觉	不经	窄	视频
外形	精致	漂亮	好看	很棒	效果	电池	充电	清晰	太	荣耀	喜欢	理解	发现	王者	新
感觉	耐用	系统	太	东西	音效	差	期待	卡	开	苹果	电话	特别	两天	要死	好像
稳定	高	尺寸	提升	接受	特色	赠品	支持	退货	感	失败	信号	客服	没用	多年	电池
玩游戏	棒	好评	换	一点	体验	细腻	收到	掉	希望	手感	配置	怀疑	暂时	无语	情况
视频	送	信赖	新	轻薄	卡	苹果	大小	想	送	不到	打开	换	对不起	拍照	摄像头
屏	物流	手	小屏	小巧	米	几天	开	拍	真	第一	不灵敏	可惜	搞	玩游戏	第一次
解决	购买	点″	失望	没想到	待机时间	感	发现	米粉	耗电	曲面	指纹	软件	速度	华为	延迟

上表 5-4 为小米 12 主题挖掘结果，通过对小米 12 的正负向评论主题挖掘可以发现，用户在购买小米 12 的过程中，满意的点有：外形颜值好看、玩游戏稳定、物流满意、待机时间可以、充电很快、音效可以等方面；不满意的点有：指纹不灵敏、客服不好、发热、拍照垃圾；所以对小米手机厂商的建议是：保持现有的优势，短期针对这个型号可以大力宣传其他厂商相对薄弱的充电性能方面，获得优势。长期要解决拍照问题、指纹不灵敏等产品性能问题，同时优化客服服务。

表 5-5 苹果 13 主题挖掘结果

Table 5-5 Mining results of Apple 13 topics

苹果 13 正向评论 LDA 主题模型结果								苹果 13 负向评论 LDA 主题模型结果							
Topic 1	Topic2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic8	Topic 1	Topic2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic8
拍照	喜欢	速度	清晰	便宜	漂亮	满意	完美	难受	不好	失望	没收	瑕疵	差	降价	保价
音效	流畅	外形	运行	好看	值得	屏幕	合适	坏	麻烦	不到	东西	贵	磨损	信号	郁闷
服务	效果	很棒	特别	舒服	确实	支持	真实	天	没到	不配	着急	受不了	收到	频繁	号
习惯	特色	很好	太	不错	价格	很快	耐用	一个	充电	莫名	月	屏幕	星期	几天	莫名其妙
强大	降价	手感	星光	活动	精致	真的	高	头	感觉	慢	没用	恶心	客服	充	电
电池	外观	系统	舒适	希望	适合	续航	时间	发热	质量	卡	伤心	掉价	声音	换	配置
瑕疵	待机时间	足够	几天	细腻	购买	网络	打开	真的	黑屏	价保	外观	太	没法	活动	显示
产品	客服	发货	视频	粉色	感觉	大小	两天	真	快递	半个	玩	不好	实体店	双十	待机
商品	差异	颜色	发热	信赖	发现	长	充	服	电话	高	写	申请	退货	划痕	担心
优惠	提升	收到	屏幕显示	双十	一流	力	担心	发货	为啥	华为	速度	送	电池	情况	损坏

上表 5-5 为苹果 13 主题挖掘结果，通过对苹果 13 的正负向评论主题挖掘可以发现，用户在购买苹果 13 的过程中，满意的点有：拍照清晰、音效好、客服服务强大、系统运行快、外观漂亮等方面；不满意的点有：发货物流慢、充电不好、会降价等方面，所以对苹果手机厂商的建议是：保持现有的优点，提高充电能力，在物流上进行重点规划，解决物流慢的难题，并提供保价服务。

5.3.2 对用户的购买建议

通过上文的实验结果以及对各个厂商现状的分析，我们可以很清楚的知道其他用户现阶段对各个厂商的评价，它们在哪些方面做得好，哪些做得不好，哪些做得一般，都有结论。此时我们可以把各个手机厂商在手机的主要属性的表现情况列为表格，供用户按需选择，表格按照只出现在正向评价为“好”、只出现在负面评价为“不好”，都不出现为“一般”，都出现为“有争议”的原则进行打分，各手机属性打分如表 5-6 所示：

表 5-6 各手机属性打分

Table 5-6 Scores of mobile phone attributes

	拍照	音效	系统	外观	充电与电量	客服	物流	保价
oppo Find	好	好	不好	好	不好	不好	不好	不好
三星	好	一般	一般	好	不好	一般	一般	不好
华为 P50	好	一般	有争议	好	不好	不好	好	一般
小米 12	不好	好	好	好	好	不好	好	一般
苹果 13	好	好	好	好	不好	好	不好	不好

我们以手机的实用性出发，综合手机论坛等网站网友的意见，将手机的各个属性分别赋予权重：拍照 0.2、音效 0.075、系统 0.3、外观 0.15、充电与电量 0.2、客服 0.025、物流 0.025、保价 0.025；将打分为好设置为 1 分、不好设置为-1 分、有争议与一般设置为 0 分；然后分别对每个手机的最终得分进行计算，分别为 OPPO Find 得-0.2 分、三星 Galaxy S21 得 0.1 分、华为 P50 得 0.15 分、小米 12 得 0.5 分、苹果 13 得 0.45 分。从上述结论可以看苹果 13 与小米 12 在五款手机中脱颖而出，苹果 13 在手机本身的性能方面最为出色，但是因为充电等原因导致最终得分不如小米 12，所以结合众多评论，得出本文结论，即推荐新用户购买小米 12。当然本实验也有不足，最明显的就是对好与不好的定义颗粒度太大，不够精细，这是能造成得分不同的重要因素，有待后续改进；另外还有个因素是本实验暂时无法分析需要用户自己思考的，那就是价格因素。因为本实验所采用的数据是从京东平台的商品评价中获取，此时用户已经购买商品，实际上是对商品价格是认可的，因此除了活动降价外，很少会在评论里讨论商品本身的价格。

第 6 章 总结与展望

6.1 本文工作总结

本文从我国手机使用现状及京东电商平台发展现状出发，指出商品评论挖掘对商家、平台、用户和政府监管的意义，明确了研究目的和方向，随后利用文本挖掘的方法对 oppo、三星、小米、苹果、华为这五家在国内市场较为火热的手机厂商发行的 2021 年旗舰机的商品评论从多维度进行分析研究。主要研究总结如下：

1、简单介绍了本文实验所使用的理论知识和运行平台。

2、利用爬虫技术获取了各手机型号评论共 8010 条，并进行文本预处理工作。

3、分别利用了决策树、情感倾向分析和 LDA 主题模型对电商评论文本进行挖掘，并对挖掘的结果进行分析总结。

4、对 LDA 主题模型中可以改进的步骤进行修改改进，通过添加手机领域分词词库、整合多家情感词、手动增删情感词等方式，使得结果更加准确。

5、对改进后的 LDA 主题模型结果进行分析，并以此为基础，从多维度为各个手机厂商及用户提出建议，基本完成本文的研究目的。

6.2 未来展望

在直播电商火热的年代，线上购物已越来越成为人们的生活必需之一，对电商平台的评论进行挖掘势必在很长一段时间成为热门与刚需。本文以手机评论为研究对象，通过文本挖掘的方式进行研究，虽然取得了一些成果，但在进行实验的过程中，还有可以继续学习改进的地方：

1、建立领域分词词库与领域情感词库，由于本实验所用的都是通用的分词词库与情感词库，有许多手机领域下的名词，被错误分割，许多手机领域的情感词没有被识别到情感，这都是很影响准确率的地方，虽然通过手工添加的方式进行了改进，但这不是长期可行的方式。把目光从手机领域放宽到全领域时，就会发现，在进行任何领域的文本挖掘时，都需要专属的领域词库与情感词库，才会达到更精确的效果。

2、没有情感词的情感表达：在生活用语中，由于种种约定俗成的背景加持，有许多描述中不出现情感词，也可以表达很强烈的情感，比如“这个手机用了 7 天了，还是满电”，这句话没有情感词，但其实表达的也是正向情感。

3、副词的作用：本实验没有重视副词的作用，但其实在现实生活中，副词也可以表达很重要的情感倾向，不同程度的副词，表达出的情感强弱也是不同的，后续可以研究副词组合情感词的权重问题。

4、研究数据相对较少：由于京东平台的反爬机制，导致获取的实验数据量不足，再加上京东展示差评较少、中评较少，差评和中评的实验数据更为缺少。

5、采集的用户信息较少：目前的语义分析只是针对用户发出的评论进行文本挖掘，若想更精细化的进行文本挖掘，更准确的为厂商和用户提供建议，还需要在法律允许的情况下，获得更多的用户信息，如职业、收入、性别、年纪、购物频次等，按人群划分进行文本挖掘。

致　谢

时光匆匆而逝，转眼之间，在长江大学三年求学的日子已接近尾声。其实在进入社会工作，阔别校园三年后重返校园的感觉很奇妙，更明白自己读书想获得什么，也更加懂得要珍惜来之不易的学习机会，所以在这里要首先感谢我的父母和妹妹在我决定辞职读研时的理解与支持，感谢方慧老师在入学时给予的大力支持与鼓励，没有他们的努力与支持，也许就没有后来和长大的缘分。

然而世事无常，刚入学不久后，因为家中接连变故，不得已请假回家多次，宿舍安排到与本科生混住，研一下学期因为疫情在家上网课等原因，导致我与班上的大部分老师、同学都不熟悉，这是十分遗憾的地方；不过倍感幸运的是在倏忽而逝的三年时光中，也有几个人在我求学路上给予了很大的帮助，首先要感谢的就是我的导师周云才老师，周老师是一个学识渊博而且很和蔼的先生，作为导师不仅在学业上帮助我们学习，还从生活上关心我们的成长；感谢付盈老师、黄静老师、李梦霞老师对我学习与生活上的照顾；还要感谢同门艾珍珍、董夏君同学在我不在学校时，忙前忙后的帮忙交各种材料，以及学业上的相互扶持，一起努力与进步；感谢龙雯同学、路云同学为我学习过程中不明白的知识点不厌其烦的进行答疑解惑。

校外锻炼阶段，要感谢老伙计赵科猛强烈邀请我再次前往北京，在首都又开启了一段充实难忘的工作经历；感谢好兄弟丁碧呈和李荣祯在我读研期间的逆向激励，尤其是丁总以表情包的形式见证了我在飞书文档上写完毕业论文的全过程；感谢部门里刘宇航、夏霞、渠青、胡琦、胡婉萍、杨雪婷、李杰、张林、陈琪琪、孙思远等人在一起在工作时的教导，在他们身上我学到了很多技能与知识。感谢刘妍君在我写论文期间的陪伴。

最后祝长江大学越办越好，学弟学妹们都能成为对社会有贡献的人。

参 考 文 献

[1] 莫岱青.上半年网络零售交易规模突破 6 万亿元[J].计算机与网络,2021,47(19):6-7.

[2] 尹晓楠,常敏.电商直播品牌的构建与传播策略--以淘宝直播品牌"薇娅 viya"为例[J].视听,2021(3):164-165.

[3] 曹花蕊.消费者感知网购物流配送服务质量对其态度和购后行为的影响[J].物流技术,2014,33(5):3.

[4] 林敬.国人玩手机每天 5.69 小时[J].就业与保障,2019(15):1.

[5] 崔艳宇[1].你多长时间换一部手机[J].时代邮刊,2019(7):1.

[6] 王超发,孙静春.影响用户换手机的关键因素及其效应研究——基于中国移动西安分公司的样本数据[J].运筹与管理,2018,27(8):8.

[7] 王伟,王洪伟.特征观点对购买意愿的影响:在线评论的情感分析方法[J].系统工程理论与实践,2016,36(1):63-76.

[8] 林丹,刘建明,谷志瑜.一种基于关键词的微博话题聚类算法[J].计算机应用与软件,2018,1:264-268.

[9] Luhn H P.The automatic creation of literature abstracts[J].IBM Journal of research and development,1958,2(2):159-165.

[10] 钱爱兵,江岚.基于改进 TF-IDF 的中文网页关键词抽取——以新闻网页为例[J].情报理论与实践,2008(6):945-950.

[11] 胡学钢,李星华,谢飞,等.基于词汇链的中文新闻网页关键词抽取方法[J].模式识别与人工智能,2010(1):45-51.

[12] 张建娥.基于 TFIDF 和词语关联度的中文关键词提取方法[J].情报科学,2012,30(10):1542-1544.

[13] 王立霞,淮晓永.基于语义的中文文本关键词提取算法[J].计算机工程,2012,38(1):1-4.

[14] Kumari M,Jain A,Bhatia A.Synonyms based term weighting scheme: An extension to TF.IDF[J].Procedia Computer Science,2016,89:555-561.

[15] Zhang K,Xu H,Tang J,et al.Keyword extraction using support vector machine[C]//international conference on web-age information management.Springer,Berlin,Heidelberg,2006:85-96.

[16] 程岚岚,何丕廉,孙越恒.基于朴素贝叶斯模型的中文关键词提取算法研究[J].计算机应用,2005,25(12):2780.

[17] 朱泽德,李淼,张健,等.一种基于 LDA 模型的关键词抽取方法[J].中南大学学报:自然科学版,2015,46(6):2142-2148.

[18] 余本功,张宏梅,曹雨蒙.基于多元特征加权改进的TextRank关键词提取方法[J].数字图书馆论坛,2020(3):41-50.
[19] 梁伟明.中文关键词提取技术[D].上海交通大学,2010.
[20] Chen Y,J Wang,Li P,et al.Single document keyword extraction via quantifying higher-order structural features of word co-occurrence graph[J].Computer Speech & Language,2019.
[21] 郎冬冬,刘晨晨,冯旭鹏,等.一种基于LDA和TextRank的文本关键短语抽取方案的设计与实现[J].计算机应用与软件,2018,35(3):7.
[22] 李跃鹏,金翠,及俊川.基于word2vec的关键词提取算法[J].科研信息化技术与应用,2015(4):6.
[23] Qiu Q,Xie Z,Wu L,et al.Geoscience Keyphrase Extraction Algorithm Using Enhanced Word Embedding[J].Expert Systems with Applications,2019,125(JUL.):157-169.
[24] Fu X,Ouyang T,Chen J,et al.Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks[J].Information Processing & Management,2020,57(4):102236.
[25] 姚天昉,聂青阳,李建超,等.一个用于汉语汽车评论的意见挖掘系统[C]//中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论文集.2006.
[26] Kamps J,Marx M,Mokken R J, et al.Using WordNet to Measure Semantic Orientations of Adjectives[J].national institute for,2004.
[27] Esuli A,Sebastiani F.Determining the semantic orientation of terms through gloss classification[C]//Proceedings of the 14th ACM international conference on Information and knowledge management.2005:617-624.
[28] Andreevskaia A,Bergler S.Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses[C]// EACL 2006,11st Conference of the European Chapter of the Association for Computational Linguistics,Proceedings of the Conference,April 3-7, 2006,Trento,Italy.DBLP,2006.
[29] Liu L,Lei M,Wang H.Combining Domain-Specific Sentiment Lexicon with Hownet for Chinese Sentiment Analysis[J].Journal of Computers, 2013,8(4):195-206.
[30] Lei D ,Zhang L.Method of discriminant for Chinese sentence sentiment orientation based on HowNet[J].Application Research of

Computers,2010,27(4):1370-1372.
[31] 李生琦,田巧燕,汤承.基于《知网》词汇语义相关度计算的消歧方法[J].情报学报,2009(5):706-711.
[32] Qu L,Ifrim G,Weikum G.The bag-of-opinions method for review rating prediction from sparse text patterns[C]//Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).2010:913-921.
[33] Xu Y,Fan X Z,Zhang F.Semantic Relevancy Computing Based on Hownet[J].Journal of Beijing Institute of Technology,2005.
[34] 霍宗凡.基于语义的文本倾向性分析与研究[D].南京邮电大学.
[35] Chen J M,Tang Y,Li J G,et al.Community-Based Scholar Recommendation Modeling in Academic Social Network Sites[C]// International Conference on Web Information Systems Engineering.2013.
[36] 杨立公,朱俭,汤世平.文本情感分析综述[J].计算机应用,2013,033(006):1574-1578,1607.
[37] Wiebe J.Learning subjective adjectives from corpora[J].Aaai/iaai, 2000,20(0): 0.
[38] Wiebe J,Wilson T,Cardie C.Annotating expressions of opinions and emotions in language[J].Language resources and evaluation,2005,39(2): 165-210.
[39] Hatzivassiloglou V,Wiebe J.Effects of adjective orientation and gradability on sentence subjectivity[C]//COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics. 2000.
[40] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.2004:168-177.
[41] Zha Z J,Yu J,Tang J,et al.Product aspect ranking and its applications[J].IEEE transactions on knowledge and data engineering, 2013, 26(5):1211-1224.
[42] Turney P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[J].arXiv preprint cs/0212032, 2002.
[43] 李钝,曹付元,曹元大,等.基于短语模式的文本情感分类研究[J].计算机科学,2008,35(4):3.
[44] Pang B,Lee L,Vaithyanathan S.Thumbs up? Sentiment Classification

using Machine Learning Techniques[J].arXiv,2002.
[45] 刘志明,刘鲁.基于机器学习的中文微博情感分类实证研究[J].计算机工程与应用,2012,48(1):1-4.
[46] 谭松波,程学旗.一种跨领域的文本情感分类器的训练方法和分类方法:,CN101770580A[P].
[47] 黄永文.中文产品评论挖掘关键技术研究[D].重庆:重庆大学,2009.
[48] 张艳辉,李宗伟,赵诣成.基于淘宝网评论数据的信息质量对在线评论有用性的影响[J].管理学报,2017,14(1):9.
[49] 施乾坤.基于LDA模型的文本主题挖掘和文本静态可视化的研究[D].广西大学,2013.
[50] 阮光册.基于LDA的网络评论主题发现研究[J].情报杂志,2014,000(003):161-164.
[51] 彭雨龙.基于VSM和LDA模型相结合的新闻文本分类研究[J].山东工业技术,2016,6:202-203.
[52] 周昭涛,卜东波,程学旗.文本的图表示初探[J].中文信息学报,2005,19(2):37-44.
[53] 孙宏纲,陆余良,刘金红,等.基于HowNet的VSM模型扩展在文本分类中的应用研究[J].中文信息学报,2007,21(6):101-108.
[54] 徐戈,王厚峰.自然语言处理中主题模型的发展[J].计算机学报,2011,34(8):1423-1436.
[55] 钱智明,钟平,王润生.结合非负张量表示与扩展隐Dirichlet分配模型的图像标注[J].国防科技大学学报,2014,36(6):152-157.
[56] 王瑞雷,栾静,潘晓花,等.一种改进的中文分词正向最大匹配算法[J].计算机应用与软件,2011,28(3):195-197.

个 人 简 介

王宁，1993 年生，河北邢台人。2016 年毕业于河北大学工商学院，毕业后先后就职于搜狗科技有限公司、字节跳动科技有限公司，2019 年考入长江大学计算机学院进行研究生学业学习。在校学习认真，积极参加课题活动，英文水平达到研究生英语水平要求，在校期间发表论文《DAST 开发技术设计与分析》于 2021 年 9 月被《电脑知识与技术》收录。

学号：201975088

研究生学位论文原创性声明和版权使用授权书

原创性声明

我以诚信声明：本人呈交的硕士学位论文是在周云才教授指导下开展研究工作所取得的研究成果。文中结论和结果系本人独立研究得出，不包含他人研究成果。所引用他人之思路、方法、观点、认识均已在参考文献中明确标注，所引用他人之数据、图件、资料均已征得所有者同意，并且也有明确标注，对论文的完成提供过帮助的有关人员也已在文中说明并致以谢意。

学位论文作者（签字）：王宁

签字日期：2022 年 4 月 10 日

版权使用授权书

本人呈交的硕士学位论文是本人在长江大学攻读硕士学位期间在导师指导下完成的硕士学位论文，本论文的研究成果归长江大学所有。本人完全了解长江大学关于收集、保存、使用学位论文的规定，即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权长江大学可以将本学位论文的全部或部分内容编入有关数据库，可以采用影印、缩印、数字化或其它复制手段保存论文，学校也可以公布论文的全部或部分内容。（保密论文在解密后遵守本授权书）

学位论文作者（签字）：王宁

签字日期：2022 年 4 月 10 日