分类号 C93/65　　密级 公开
U D C 0004259　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 基于 BERT-LDA 的在线评论细粒度情感分析--以手机产品为例

研 究 生 姓 名：丁申宇

指导老师姓名、职称：王玉珍 教授

学 科、专 业 名 称：管理科学与工程

研 究 方 向：电子商务

提 交 日 期：2022 年 5 月 31 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：丁申宇 签字日期：2022年5月31日

导师签名：王志刚 签字日期：2022年5月31日

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：丁申宇 签字日期：2022年5月31日

导师签名：王志刚 签字日期：2022年5月31日

Fine-grained emotion analysis of online reviews based on BERT-LDA
——Take smartphone products for example

Candidate: Ding Shenyu

Supervisor: Wang Yuzhen

摘 要

情感分析可以在文本数据中辨识出用户所表达的情感意向，是自然语言处理中的热门领域。但传统的情感分析大多是基于粗粒度层面的研究，以探讨产品整体情感极性为主，因此无法掌握产品特征及对应的细粒度情感。为弥补传统情感分析的不足，本文从产品特征和情感分类两个角度展开研究。其中，产品特征角度的研究是对产品具体特征进行分析，明确消费者重视的产品属性；情感分类角度的研究是将情感词典进一步细分，不再沿袭传统的情感极性二分类的方法，而是将情感分类划分的更加细致，并计算情感词的情感强度，明确消费者情感的细分倾向。将产品特征与情感细粒度结合，能够更加清楚的了解消费者对产品特征的具体要求，从而帮助商家更好地提供个性化服务。

本文的主要研究内容包括以下四部分：

（1）构建了手机产品关键短语集合。本文基于关键短语结构，以手机特征词典为基础，构建了手机关键短语集合，为手机特征和情感词的高精度提取创造了前提条件。

（2）构建了 BERT-LDA 模型。首先对 BERT（Bidirectional Encoder Representations from Transformers）模型中相似度值域进行训练；然后应用 BERT 模型进行相似短语的提取；最后将提取结果输入到 LDA 模型中，获得评价对象和评价词。高质量的相似短语在降低 LDA 模型困惑度上发挥了极大作用，也有利于提高 LDA 模型主题提取的准确率。

（3）建立了基于手机特征词的细粒度情感词典。为实现对情感词的分类与情感强度的计算，采用由大连理工大学信息检索研究室整理的中文情感词汇本体库（DUTIR），以本文所建立的手机关键短语集合和手机特征词典为基础，构建了基于手机产品评价特征的细粒度情感词典，并增加了代表“疑惑”的“疑”的情感，构成 8 类情感词，同时对修饰词与否定词进行扩充。

（4）应用了细粒度情感分析模型。以某品牌的手机评论文本为例，应用 BERT-LDA 模型，找出消费者关注度较高的特征及情感词，并应用细粒度情感计算方法，计算特征的情感分类与情感强度，从而详细了解消费者对不同特征的情感倾向，有利于店铺提升口碑、增加销量，更好地提供给个性化服务。

关键词：情感分析 细粒度情感分析 BERT-LDA 模型 关键短语

Abstract

Emotion analysis can identify the emotional intention expressed by users in text data, and is a popular area in natural language processing.However, the traditional emotion analysis is mostly based on the coarse-grained research, mainly to explore the overall emotional polarity of the product, so it is impossible to grasp the product characteristics and the corresponding fine-grained emotion.In order to make up for the deficiency of traditional emotion analysis, this paper studies from the perspectives of product characteristics and emotion classification.Among them, the study of product characteristics is to analyze the product attributes that consumers value; the study of emotion classification is to further subdivide the traditional method of emotion polar classification, but to divide the emotion classification more carefully, calculate the emotion intensity of emotion words, and clarify the segmentation tendency of consumer emotion.The combination of product characteristics and emotional grancan have a clearer understanding of consumers' specific requirements for product characteristics, so as to help businesses better provide personalized services.

The main research content of this article includes the following four parts:

(1)Build a collection of key phrases of mobile phone products.Based

on the structure of key phrases and the dictionary of mobile phone features, this paper creates the preconditions for the high-precision extraction of mobile phone features and emotional words.

(2)The BERT-LDA model was constructed.First, the similarity value domain in BERT (Bidirectional Encoder Representations from Transformers) model is trained; then BERT model is applied to extract similar phrases; finally, the extraction results are input into LDA model to obtain the evaluation object and evaluation words.High-quality similarity phrases play a great role in reducing the confusion of the LDA model, and also help to improve the accuracy of the LDA model topic extraction.

(3)A fine-grained emotion dictionary based on mobile phone feature words is established.For the calculation of the classification of emotional words and emotional strength, organized by Dalian university of technology information retrieval Chinese emotional vocabulary ontology library (DUTIR), on the basis of the phone key phrases and mobile features dictionary, built based on the evaluation of mobile product emotion dictionary, and increase the "doubt" of "doubt" emotion, constitute eight types of emotional words, and expand the modifier and negative words.

(4)A fine-grained emotion analysis model was applied.With a brand of mobile phone review text, for example, the application of BERT-LDA model, find out the characteristics of consumer attention and emotional

words, and apply fine-grained emotion calculation method, calculate the characteristics of emotional classification and emotional intensity, to understand the emotional tendency of different characteristics, to store word of mouth, increase sales, better provide personalized service.

Keywords: Sentiment analysis;Fine-grained emotion analysis; BERT-LDA model; Key phrases

目录

1 绪论

1.1 研究背景及意义

1.1.1 研究背景

近年来，电子商务模式日渐成熟，手机产品借助于成熟的线上渠道销量庞大。2021 年双十一，仅小米品牌的手机，在电商销售全渠道中消费者支付金额就已经破 70 亿[1]。通常消费者在线上购买手机后，会将购物、使用体验等写成评价，形成评论文本，这些数量庞大且内容丰富的手机评论文本包含消费者对产品的态度、看法和情感倾向等，因此具有重要的研究价值。

对评论文本进行情感分析的目的是分析用户主观评论中针对主题、事件等的情感、观点、态度，并根据分析结果指导具体行动。常见的分析方法有自动化分析方法（如词典法）以及半自动化分析方法（如机器学习、深度学习方法）。文本情感分析较传统的数据挖掘算法增加了对文本内容的理解能力，因而能够解决很多传统方法所不能解决的问题，对现实生活及生产具有很强的应用指导意义。如为消费者购买产品提供参考，帮助商家找到消费者的喜好、产品的优缺点等，因此，基于情感分析挖掘产品及其特征的情感、观点、态度具有重要的研究意义。

目前电商平台产品的情感分析往往是将情感粗略划分为好评与差评，并不能准确地获得产品具体的信息及对应的细粒度情感。然而，随着消费者对商品品质重视程度的不断提高，粗粒度层面的情感分析已经不能有效的帮助商家挖掘消费者对产品的具体细节情感、多样化感情。因此，本文以手机评论文本为例，对其进行细粒度情感分析，提取手机特征及对应情感信息，并加以分析，从而准确判断出消费者对手机不同特征的喜好程度，帮助商家有效的节约时间、精准的把握顾客需求，更好地提供个性化服务。

1.1.2 研究意义

近年来，电子商务发展迅速，淘宝、京东、拼多多等各大电商平台拥有巨大的购物群体，网络购物已经成为了人们主要的消费途径之一，由此产生了海量的在线评论。这些评论信息中包含了消费者整个购物过程的体验与感受，所以挖掘这些在线评论对了解消费者的消费行为、商品销售情况等具有极大的价值，而对

[1] 电商报.小米:双 11 首日支付金额破 70 亿,手机品类全平台销量第一,(2021-11-2) [2022-3-6]. https://www.sohu.com/a/498704117_115865

其进行细粒度情感分析是获得更多有价值信息的基础。

（1）理论意义

从理论角度讲，传统的情感分类方法对于情感极性通常都是积极和消极的二分类法，或是正向、中度、负向的三分类法，因此对情感词的分类不够精细，也并未考虑到到副词、否定词等对情感词的影响，且不同的情感词其情感强度也存在差异，而目前的研究成果对情感强度的分析还较少；同时现有的研究成果大多基于英文评论文本，而对中文评论文本的研究较少；此外，缺乏各领域专属的情感词典，因此在进行细粒度情感分析时，需要针对不同领域构建专属的情感词典。可见，目前对中文文本细粒度情感分析的研究仍有许多不足之处，还存在较大的研究空间。因此，本文在已有手机特征词典的基础上对其加以扩建，并结合关键短语结构，构建了手机领域的关键短语集合，同时以扩建后的手机特征词典和手机领域的关键短语集合为基础，构建了基于手机特征的细粒度情感词典，从而解决了特征研究不足、情感分类笼统以及缺乏专业领域情感词典的问题；同时，本文的研究拓展了情感分析的研究领域，丰富了情感分析的研究方法，可为其他研究者提供参考。

（2）现实意义

从实际应用角度讲，对在线评论进行细粒度分析，惠及消费者、商家、电商平台、政府等多个主体。对消费者而言，将产品评论文本进行细粒度分析，可以从其他消费者购买、使用后反馈的评论中获得所关注的信息，为了解商品提供了更加行之有效的方法，从而便于购买到真正需要的产品；对于商家而言，评论文本是了解消费者需求、掌握商品与服务优势与不足的重要基础，对这些文本进行细粒度分析，有利于及时地为消费者提供个性化服务，进而提高市场竞争力，以期获得更大的市场占有率；对电商平台而言，将细粒度情感分析应用于个性化推荐方面，即基于特征或特征的情感倾向为消费者进行产品推荐，能够更大程度上符合消费者对产品属性或特征的个性化需求，精准的满足不同消费者的消费需求，从而推动消费者购买产品，提高平台的成交量；对于政府而言，对评论文本的细粒度情感分析可以更详尽地掌握消费市场现状，及时发现破坏消费市场的不法行为，从而构建健康的网络消费环境。

总之，对在线评论文本进行细粒度情感分析具有重要的价值，通过本文的研

究不仅解决了现有研究的不足，拓展了研究领域，而且使消费者、商家等主体获益良多。

1.2 国内外研究现状

1.2.1 细粒度情感分析研究现状

近年来，细粒度情感分析的研究在国内外已成为热点问题，并取得了一定的研究成果。将研究成果按照对情感分析的细粒度研究角度，可以分为产品特征和对应情感词的抽取、情感分类研究两大方向[1]。

产品特征和对应情感词的抽取方向是细粒度情感分析研究的一大重要方向，按照研究方法可以细分为传统研究方法、机器学习方法、深度学习方法三大类，具体分析如下：

传统的研究方法是基于统计和规则，如薛福亮[2]使用 IF-IDF 算法抽取餐饮评论数据的评价对象及相应情感观点，并基于具有情感强度的情感词典计算具体特征的情感强度，有效地给出了消费者对于餐饮各个方面的具体情感强度；Li Ji 和 Lowe Dan 等人[3]从单词级和句子级两方面研究了基于词典的情感分析方法。在单词级情感分析上，使用高频情感词与评论词作对照；在句子级情感分析上，计算出用户对产品各特征的情感极性进行分类；胡飞菊[4]基于运用综合词语相似度的特征聚类方法获得民宿评论中的特征，构建情感词典及评价体系，并以评价体系为依据，对抽取的情感评价单元进行分类，在此基础上计算每一类别的平均情感值，该方法在提高情感判断的准确率上起到一定作用。

随着机器学习的发展，机器学习的众多方法被应用在细粒度情感分析研究中，如陈炳丰等人[5]以汽车评论数据为研究对象，提出双层结构 CRF 模型，成功解决了忽略实体与情感分类联系的问题，并提高了情感分析的准确性；I Titov[6]基于对标准主题建模方法（如 LDA 和 PLSA）的扩展，构建了基于多粒度 LDA 的情感模型，以诱发多粒度主题，解决了标准模型往往只产生对应于对象的全局特征（例如，产品类型的品牌）的主题，而不是对象基于特征的细粒度主题的问题，并完成了对多颗粒模型的定性和定量的评估，证明它们在标准主题模型上有了显著的改进；Emmanuel Awuni Kolog 等人[7]概述了通过支持矢量机器学习系统检测文本中情感的机器学习应用程序，将系统的分类器与 WEKA 的多项式朴素

贝叶斯和 J48 决策树分类器的性能进行了比较，得出机器学习算法可以跟踪文本中的情感，特别是来自学习生成的内容。

近年来，深度学习方法受到广大学者的青睐，在细粒度情感分析中大放异彩，常见的算法有卷积神经网络（CNN）、循环神经网络（RNN）、长短时记忆网络（LSTM）、深度神经网络（GRU）以及 BERT 等，具体分析如下：

在基于卷积神经网络（CNN）的细粒度情感分析方面，蔡庆平、马海群[8]为实现基于特征的产品评论的聚类，以华为手机的评论为例，在 Word2Vec 提取的产品评论特征词的基础上，使用 CNN 进行细粒度情感分类；Lapata[9]提出了一个用于细粒度情感分析的神经网络模型——多实例学习神经模型(MILNET)，该模型由三个组件组成：一个 CNN 段编码器、一个软最大段分类器和一个基于注意的预测加权模块，模型将注意力与极性评分法融合，来识别正、负文本片段，以及一个新的数据集，来评估多实例学习神经模型，实验结果表明，该模型有更优的细粒度分析性能。在基于循环神经网络模型（RNN）的细粒度情感分析方面，贾川等人[10]利用循环实体网络能够为指定的特征做情感分类的功能，实现基于特征的细粒度情感分析效果的提升，经实验证明该方法能够更好地针对不同特征类别抽取特定的情感特征。在基于长短时记忆网络模型（LSTM）的细粒度情感分析方面，Jin Zheng 等人[11]提出了一种深度学习的方法，以电影影评为例，并尽可能地恢复用户的真实情感，该方法使用 Word2Vec 模型将单词转换为向量、长短期记忆网络(LSTM)以学习语义依赖性，以及利用逻辑回归分类器进行了分类，使得正负细粒度情感分类准确率得以提高；张津[12]为了判定不同评教维度的情感，将双向 LSTM 与注意力机制融合，同时增加了概率分布、嵌入情感语义，通过分析大量的真实学生评教数据，该模型在精确率和召回率上表现优异；薛福亮、刘丽芳[13]为更好地分析用户对产品特征的偏好，首先运用 CRF 进行特征词的抽取，然后利用基于注意力机制的 LSTM 做特征情感分析，最后基于 Word2Vec 将特征词聚集为特征面，从而分析产品特征面的情感，该方法通过对特征词的提取、情感分析以及特征面聚类，可较好地解释用户对产品的特征偏好；D Tang 等人[14]提出每一阶段序列处理中都考虑属性特征的 TC-LSTM 模型，实现了在提取情感特征时，将特征词与上下文词区别对待，且准确率提高明显；C Sindhu[15]等人提出 EIE-CBiL-Att 模型，该模型在嵌入阶段使用标记、方向、语法函数、

场和强度分量来丰富输入嵌入，利用卷积核细化模式提取，不仅可以用于方面和字检测，而且能用于极性和强度分析，实验结果表明，与 Bi-LSTM-ATT-G、IAN 等模型相比，该方法增强性能更优；李鸿宇[16]聚焦于解决方面级情感分类问题，提出基于多级注意力机制的模型，即在第三层中使用两个 BiLSTM，该模型不仅无需人工设计特征而且可以构建词与句之间的关联，从而在预测情感极性的准确率上得到很大改进；马攀[17]将语言模型学习层加入 BiLSTM-CRF 网络模型中，该模型不但可以实现多任务的学习，而且考虑文本的语境，并通过实验证明了该模型的优越性。在基于深度神经网络（GRU）的细粒度情感分析方面，刘测[18]为了提高评价对象提取的精确度提出了由 LDA（主题模型）模型与 GUR（神深度经网络）模型结合的 GLDA 模型，并将该模型应用于酒店评论领域，构建了酒店领域的评价对象集合，在此基础上准确分析用户对不同特征的情感倾向，模型在提取评价对象和评价词的效果上更优；程艳等人[19]将双向 GRU 与卷积神经网络融合，不仅提高了文本特征提取的全面性，而且实现了情感信息的充分利用。在基于 BERT 的细粒度情感分析方面，沈卓、李艳[20]提出 PreLMFT（Pretrained Language Model Fine-Tuning）模型，该模型基于 ELMo、ULMFiT、BERT 三种语言模型，进行细粒度情感分析，并挖掘用户偏好，该方法可以不依赖大规模训练集，提高情感分析的效率；轩刘亦[21]提出 EN-BERT 模型，通过改进文本的表示方法和分类方法，提高了基于对象特征的情感分类的精度。

情感分类是细粒度情感分析研究的另一重要方向，Munikar M 等人[22]使用了预先训练好的 BERT 模型，并将其用于 SST 数据集上的细粒度情感分类任务，该模型能够超越复杂的架构，如递归、递归和卷积神经网络，有更高的准确性；Yejin Tan[23]等人为了解决随着用户评论数量的增加，情感分析方法高成本和高错误率的问题提出了一种基于依赖树和图神经网络的细粒度情感分析方法，经 SemEval 2014 Task4 数据验证，该模型优于大多数基线模型，能更好地编码语义关系信息，在捕获情感分析的重要语法结构方面的有效性显著提高；Georgios Balikas[24]等人提出传统的情感分析方法通过分别学习任务来解决三元（3 类）和细粒度（5 类）分类等问题，认为这种分类任务是相关的，并提出了一种基于递归神经网络的多任务方法——biLSTM+Multitask，改进了在细粒度情感分类问题中的分类结果，从而为解分析多语言和不同情感粒度级别的长短不一的评论创造

了可能；侯艳辉[25]等人将 TF-IDF（词频逆文档频率）和 TextRank 算法用于中文影评的特征词提取上，在此基础上将 Bi-LSTM 和鲁契克多维度情感模型融合，实现了基于特征层面的情感强度的 8 分类；曹雪[26]采用加权平均和门控单元机制对词语级、短语级和句子级三种粒度情感强度进行结合，进一步提升了情感分类的准确性；万岩、杜振中[27]通过建立微博领域的情感词典及情感分析模型，提高了细粒度情感分类的准确率，有助于政府部门实现舆情的应急管理和有效控制。

综上所述，细粒度情感分析的成果主要集中在基于特征角度的细粒度分析和情感分类的细粒度化等方面，其中，基于特征角度的细粒度分析的研究成果远多于情感分类的研究成果，BERT 模型以其特征提取精度高成为细粒度情感分析的新热点。

1.2.2 特征挖掘研究现状

自然语言处理作为人工智能的一个重要分支，文本特征的提取是计算机准确理解与感知文本数据的基础与关键[28]。近年来，学者们在文本特征提取领域研究成果颇丰，其主要研究可分为算法研究与实践应用两大方向。

在算法研究方面，Liu 和 Hu 等人[29]最先对产品的隐性特征展开提取，并提出了基于产品评论的细粒度情感分析的框架，从而实现了对细粒度情感分析的系统性研究；Lidong Bing[30]开发了一个无监督的学习框架，从不同的 Web 产品描述页面中提取流行的产品特征，该模型不仅能够检测有关产品的流行特性，也能将这些流行的特性映射到相关的产品特征，同时从描述页面提取这些特征；Zairan Li 等人[31]为了提高 KANSEI 知识提取的准确性，将可变精度粗糙集与基于规则的系统相结合，并应用于鞋类 KANSEI 评价系统，成功预测了某个形容词对应产品的主要形式特征；Changxuan Wan 等人[32]提出了 AC-LDA——受关联约束的 LDA，该模型在共生关系的捕获上发挥了重要作用，不仅解决了 LDA 模型无法识别共生关系的问题，而且在提取准确性获得了很大提高；Fabiano M 等人[33]认为直接从与目标对象相关的文本或类似相关对象中提取候选术语，并使用单词出现的次数进行词频统计，不足以有效地对候选标签进行排序，从而分析了与 Web2.0 对象相关联文本的各种语法模式（例如，句子中的单词之间的语法依赖关系），并利用这些模式来识别和推荐标签，在此基础上提出了新的标签质量特征，根据特征描述和特征重要性分析结果表明，该方法可以帮助区分相关标签

和非相关标签，是对其他更传统的标签质量特征的补充，特别是对于文本特征短或呈现低质量的数据集；Jian Liao 等人[34]不仅关注学习方面实体之间的语义结构，而且考虑到语言表达特征，提出了融合重构嵌入式表示学习(FREERL)框架，将语义结构和语言表达式特征融合到对象实体和特征实体的嵌入中，并在 COAE2014 和 COAE2015 数据集上实验，结果表明该模型效果更好；李伟卿，王伟军[35]通过对同义词的人工扩充和词向量的训练及语义分类，成功构建了特征词典，并提高了大数据量文本特征识别的全面性、准确性；余琴琴等人[36]通过将词语序列化（即考虑词语的顺序、重复性、同义性）、结合词语角色的方法与加权关联规则挖掘方法融合的策略，成功获得了特征短语，在 Map Reduce 算法扩展后，提高了算法的效率及准确性；王瑞等人[37]以解决 LDA 模型特征提取准确率低的问题为目标，对 LDA 模型与 Labled 模型加以结合，提出了一种既能够挖掘隐含特征又能够明确分类的新模型，解决了 LDA 模型的不足；赵勤鲁等人[38]将 LSTM 与 Attention 网络融合，形成 LSTM-Attention 模型，并通过对文本特征进行提取的实验，证明了该方法不仅能够更好地把握词与词、句与句之间的结构信息，而且能够有效地提高特征提取的准确性；徐冠华[39]结合大数据 Spark 技术，对文本特征进行有效降维，从而实现了文本分类算法准确性的提高；周源等人[40]为解决传统 IF-IDF 特征提取方法不能考虑上下文环境和特征词之间的分类分布状况的问题，将 IF-IDF 算法与文本网络、PageRank 算法相结合，通过计算不同节点的重要程度来实现对上下文结构的考虑，并通过对文本集中度的衡量来划分分类标准，该方法在提高文本特征词提取准确性上效果显著；陈可嘉、骆佳艺[41]提出一种产品隐式特征的提取方法——GT-PLSA 模型，并以手机评论文本为例，对评论中的隐式特征进行提取，经实验证明，该模型在准确地提取评论文本中的隐式特征上效果显著，从而提高了产品特征提取的全面性；周诗嘉[42]以消费者购买茶叶的在线评论文本为例，不仅对显性特征评论进行训练，而且将容易忽略的隐性特征通过词向量训练的方法加以补充，并将提取的隐性特征显性化，成功地提高了特征提取地全面性；陶娅芝[43]为了提高特征提取地全面性，提出了一种复合规则，即将统计规则、依存句法和条件概率等规则结合到一起，该方法在隐式特征地提取上效果显著，因此提高了特征提取的全面性。

在实践应用方面，Yuwei Zou 等人[44]针对医疗网站的特征通常是将实体和特

征值存储在不同页面中的问题，提出了通过有效注释多个页面中特征的关系，然后生成数据记录提取规则，形成与独立存储的实体和特征关系（属性和值）相关的自动提取系统，实验表明，该系统不仅可以完成分离存储提取的特征关系，而且可以与规则关系提取兼容，同时保持较高的精度；Jean-Baptiste Lamy 等人[45]针对标志性语言中图形组件的不一致、组合不一致，对给定图标的不同解释，很难将图标映射到现有术语本体资源的概念等问题，描述了一种用本体来形式化标志性语言的语义的方法，并证明它可以推广到其他标志性的语言；Guoqing Chen 等人[46]以给用户提供一个可以很好地反映原始信息语料库的代表性子集（即小集）为目标，从不同的代表性角度讨论各种指标，然后提出一系列相关的代表性提取方法，解决“小”如何通过测量和提取来反映“大”的问题；纪雪等人[47]为了更加清楚地了解用户关注的产品特征，提出了层次主题模型聚类方法——hLDA，使用该方法对评论文本进行挖掘，得到具有层次结构的主题层次树，并由领域专家对其加以分析处理，形成最终的产品特征层次结构树，从而进一步研究消费者对产品特征的喜好；黄磊[48]为全面了解客户对产品的评价，对商品评论中的特征词做聚类或分类处理，进而获得了客户对产品各个特征维度的质量、口碑等更加全方位的评价结果；李可悦[49]将 BERT 预训练语言模型应用于社交电商文本，通过向量的表示方法识别出句子层面的特征，并实现有针对性地分类，能够较为高效准确地判断文本所描述商品的类别，为从海量信息中提取有价值的信息提供了方法；王涛、李明[50]以电商评论文本为例，将主题模型（LDA）与语义网络模型（LTC-SNM）结合，并对评论文本进行挖掘，从而获得全方位的用户评论主题，方便用户快速浏览评论，同时还可以帮助商家实现酒店推荐服务；贺珂[51]将 OPPO 手机 R15 系列作为研究对象，挖掘评论文本中用户对特征的需求状况，从而找到偏好侧重点，为产品研发提高效率、降低成本提供了依据；孙琳[52]以汽车领域在线评论数据为例，运用 HF-CRFs 模型有效抽取出汽车评论对象，实验结果表明该方法明显提高产品特征抽取的准确率和召回率，搭建了一个面向汽车领域的用户意见分析系统。

综上所述，对文本特征提取方法的研究主要集中在提高特征词提取的全面性与准确性上，如方法改进、隐性特征提取等领域；文本特征提取在实践方面的应用领域较广泛，但目前的研究主要集中在电商评论领域。

1.2.3 文献述评

经国内外研究现状分析表明,尽管已有的研究结果对细粒度情感分析的实践运用有着一定的借鉴意义，但目前看来还存在着许多问题：（1）由于中英文文本在句式和语法结构中存在很大不同，英文文本的细粒度情感分析方法并不完全适用于中文文本；（2）细粒度情感分析多以基于特征角度的研究为主，同时考虑特征提取与细粒度情感分类的研究比较少；（3）目前，中文文本的细粒度情感分析更多在于方法的研究上，对于应用领域的研究比较少。基于此，本文基于中文语料集，在产品特征的基础上建立细粒度情感词典，构建适用于手机领域的细粒度情感词典，对在线手机评论文本进行细粒度情感分析，为商家提供产品上新与销售建议，进一步拓宽了细粒度情感分析的应用领域。

此外，现有文本特征提取的研究主要集中在方法研究与实践应用两个方面。对文本特征提取方法的研究主要集中在提高特征词提取的全面性与准确性上，如方法改进、隐性特征提取等领域；文本特征提取在实践方面的应用领域较广泛，但目前的研究主要集中在电商评论领域。通过对国内外研究现状进行分析发现，虽然现有的研究成果对于特征提取的实际应用具有一定的参考意义，但目前来说还面临着诸多问题：（1）特定评论领域的特征提取不全面；（2）缺少特定领域的特征集合。因此本文以手机评论文本为研究对象，依据语料对特征进行扩充，提高特征词提取的全面性与准确性，并建立手机领域的特征集合。

综上所述，本文以中文手机评论文本为研究对象，同时考虑特征提取与细粒度情感分类，对手机评论领域进行细粒度情感分析研究。

1.3 论文研究内容

基于上文研究现状，本文的研究工作如下：

第一，构建了手机关键短语集合，为手机特征和情感词的高精度提取创造了前提条件。

第二，构建了 BERT-LDA 模型，有利于降低 LDA 模型的困惑度、提高主题提取的准确率。

第三，构建了基于手机特征词的细粒度情感词典，为细粒度情感分析奠定了基础。

第四，以某品牌的手机评论文本为例，应用细粒度情感分析模型，帮助店铺

提升口碑、增加产品销量，提供个性化服务。

1.4 论文组织结构

文章共分为七章，文章结构及各章节内容简介如下:

第一章：绪论。介绍了在线评论细粒度情感分析的提出背景与研究意义，论述了细粒度情感分析及特征提取的研究现状，同时阐述了本文的主要内容及创新点。

第二章：相关理论研究。介绍了情感分析相关的理论、LDA 模型和 BERT 模型的基本思想和原理。

第三章：关键短语集合构建。基于现有研究中的手机特征词典进行扩建，形成本文的手机特征词典，按照关键短语的结构，对分词的短句文本与手机特征词典进行匹配，最终构建手机关键短语集合。

第四章：BERT-LDA 模型构建。阐述了 BERT-LDA 模型构建的思想与过程，并运用手机在线评论对模型进行训练与检验。

第五章：基于细粒度词典的情感分析模型构建。依据大连理工大学信息检索研究室整理的 7 大类中文情感词汇本体库（DUTIR），将关键短语集合中情感词按照特征进行归类，获得手机领域基于特征的细粒度情感词典，并补充敦欣卉等人[1]构建的表示疑惑的“疑”情感，构建 8 类情感分类的情感词典，同时增加情感修饰词等，扩充基于特征的细粒度情感词典。

第六章：细粒度情感分析模型在手机产品评价领域的应用。运用经过训练且检验过的模型，对某品牌手机的评论进行细粒度情感分析，并对结果进行研究，在准确了解消费者对手机特征的喜好程度的基础上，为商家上新和销售手机提出相应的对策建议。

第七章：总结与展望。对文章的主要内容加以概括总结，同时结合现状，对后续工作进行展望。

文章结构图如图 1.1 所示：

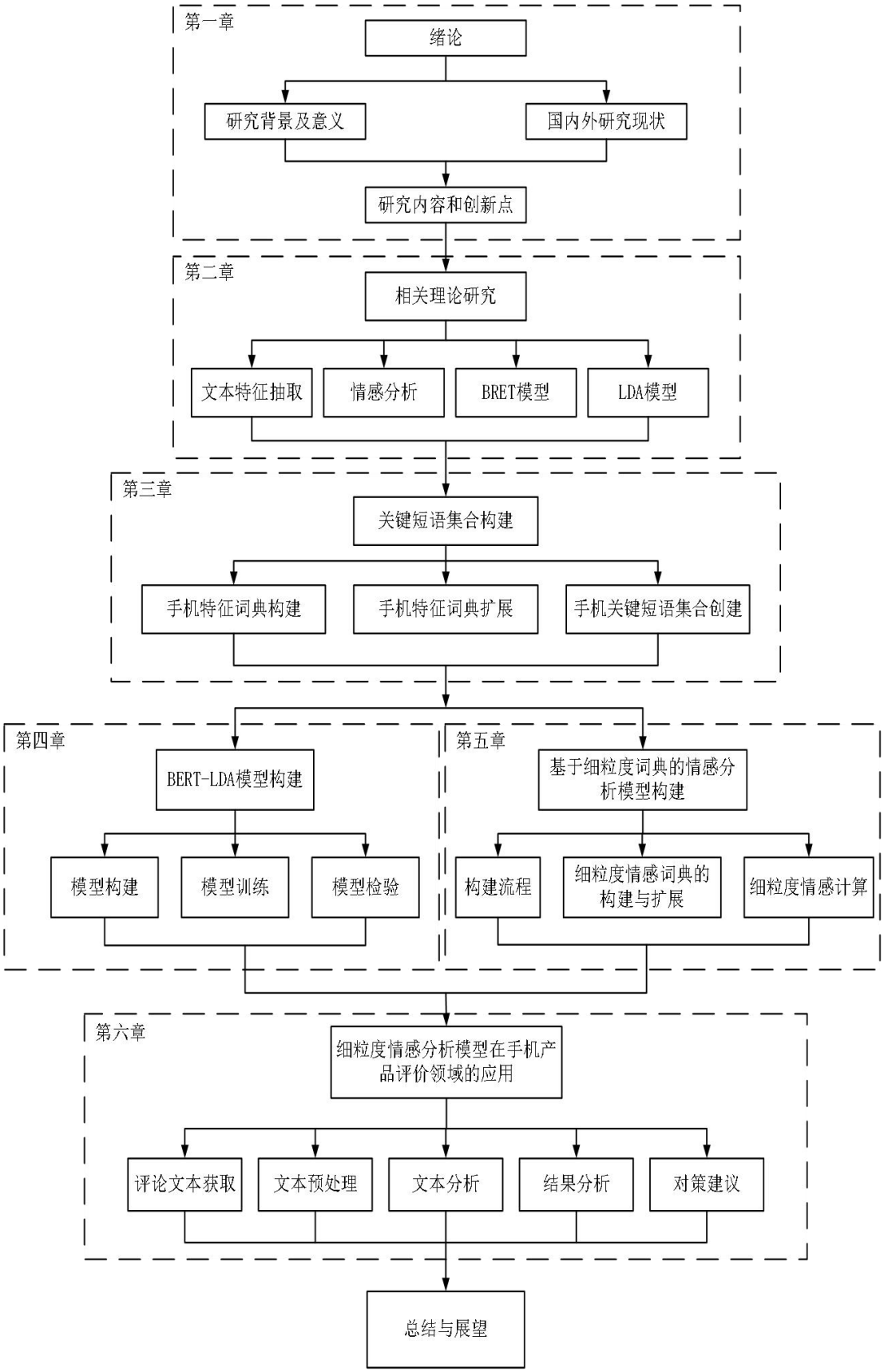

第一章
绪论
研究背景及意义
国内外研究现状
研究内容和创新点
第二章
相关理论研究
文本特征抽取
情感分析
BRET模型
LDA模型
第三章
关键短语集合构建
手机特征词典构建
手机特征词典扩展
手机关键短语集合创建
第四章
BERT-LDA模型构建
模型构建
模型训练
模型检验
第五章
基于细粒度词典的情感分析模型构建
构建流程
细粒度情感词典的构建与扩展
细粒度情感计算
第六章
细粒度情感分析模型在手机产品评价领域的应用
评论文本获取
文本预处理
文本分析
结果分析
对策建议
总结与展望

图 1.1 文章结构

1.5 创新点

本文的创新点主要包括以下两个方面：

第一，构建了手机领域关键短语集合和细粒度情感词典。文章对现有研究的手机特征词典进行扩建，形成本文的手机特征词典，在此基础上按照关键短语的结构，构建了手机关键短语集合。现有情感词典的情感分类以 2 分类或 3 分类为主，缺少更细致的分类词典；另外由于分析不同领域的评论文本，其情感词典之间也有所不同，且缺乏基于产品特征的情感词典。因此，为了准确地计算出评论文本的情感值，本文构建了手机领域产品特征的细粒度情感词典。

第二，应用领域方面的创新。本文构建了 BERT-LDA 模型，并运用于手机评论领域，拓展了 BERT-LDA 模型的应用领域。以 BERT-LDA 模型为基础，进一步分析手机产品不同的特征信息和用户复杂的情感倾向，获得消费者对手机产品具体特征的喜好程度，以此为依据，帮助商家及时掌握顾客的喜好、发现潜在热销产品的特点，从而实现店铺口碑提升、产品销量增加、盈利水平提高。

1.6 本章小结

本章介绍了在线评论细粒度情感分析的背景及意义，论述了细粒度情感分析及特征提取的研究现状，同时阐述了本文的主要内容及创新点。

2 相关理论研究

2.1 文本特征抽取相关理论研究

在文本分析中，理解文本内容是重要的分析方向，对文本进行特征抽取（也叫做关键词抽取）能够很好的反应文档中的重要词汇或短语。通过对信息量更少更精确的词汇或者短语的分析，可以对文本内容的进一步理解起到辅助作用，故而文本特征抽取受到学者们的广泛关注，成为自然语言处理的主要任务之一。在文本特征抽取工作中，可以按照抽取粒度分为基于关键词、关键短语和关键句子的抽取[45]。关键词抽取是对单一词语进行抽取，割裂了词语之间的联系。关键短语与关键词相比，在信息的表达上更为丰富与准确，辨识能力更高，合理性更好[45]。关键句子抽取由于文本长度过长、核心词语过多而在实际应用中受限，因此关键句子的抽取意义不高。综上所述，学者们更关注对关键短语的抽取。

2.1.1 关键短语

关键短语能够起到对文本主题内容的简单概括作用，具有覆盖性、纯度、短语性、完整性的特征[53][54]。其中，覆盖性是指词语重要程度高且出现次数频繁；纯度是指根据词语能够清晰的判断出所属的主题；短语性是指两个具有关联的词语多次同时出现；完整性是指短语具有代表性，是词语集合的全集，覆盖性高。

关键短语是短语的一种。通常将短语分为自由短语、固定短语和类固定短语（或半固定短语），要找出评论文本中的关键短语首先要判定其分类，三类短语的分析如下[55] [56]：

自由短语也叫做非固定短语，其特点是具有很强的临时性，在短语中如果符合语义和句法的要求可以由其他词语替换，如“解决问题”、“商讨建议”、“等一下”、“过一会”等，但是这些自由短语不能代表文本的特征，也不具备统计意义，所以，关键短语不是自由短语。

固定短语与自由短语相对立，其特点是具有稳定性，不能被随意替换，主要包括成语，如“亡羊补牢”、“坚持不懈”等，还有一些惯用的俗语，如“扯后腿”、“卖关子”等。固定词语内容精炼，形象生动，往往具有修辞效果，会产生一定联想，且存在整体意义与字面意义不一致的现象，因此与关键短语纯度、完整性相悖，所以将其排除。

半固定短语介于固定短语与自由短语之间，且有规律可循，能够用规则加以描述，具有完整性与单一性。与自由短语相比，半固定短语具有稳定性，与固定短语相比，半固定短语具有语义单一性的特征，相较而言，更加适合用来表达文本特征内容。根据关键短语的界定可知，产品的特征词是关键短语的一部分，基于特征词，按照一定的规则可以构建领域关键短语集合。

2.1.2 关键短语抽取方法

通常将特征抽取方法分为基于语法规则的方法和基于统计信息的方法两大类[45]。

（1）基于语法规则的方法。早期的基于语法规则的方法是学者们研究确立的语言规则，现在更多的是通过对词性序列规则做出设定，按照规则生成短语。基于语法规则的特征提取方法适用于句式简单、语法规则简单的文本，应用条件较为苛刻，应用范围小。

（2）基于统计信息的方法。常见的基于统计信息的方法有 TF-IDF 方法、词频方法（Word Frequency）、文档频次方法（Document Frequency）等。这类算法通过构造的评估函数对特征进行评估打分，按照分值从高到低取一定数目的特征子集，进而进行数据统计。基于统计信息方法的文本提取效果依赖于评估函数，而基于统计学的评估函数要在庞大的训练集下才会起到较好的提取作用，过于庞大的数据集的构建与计算不仅消耗巨大而且效率低。如果数据量不足，会导致权值高但是对分类无用的特征被选中，对分类有用的特征因为权值低而被忽略，从而影响分类的效果。

通过上述分析可知，基于语法规则的方法适用于句式、语法规则简单地文本，基于统计信息的方法适用于数据集庞大的文本。因为本文的语料句式结构简单，所以采用基于语法规则的方法进行特征抽取。

2.2 情感分析的相关理论研究

近年来，随着淘宝、京东、拼多多等各大电商平台购物群体规模的持续增长，在线评论数量越来越大，对这些评论文本进行挖掘和使用已成为当前文本情感分析的研究热点。

2.2.1 情感分析简介

评论文本情感分析是文本分析的重要分支之一，其目的是分析用户主观评论中针对主题、事件等的情感、观点、态度，并根据结果指导具体行动。通常评论文本情感分析可以分为粗粒度情感分析和细粒度情感分析两种[8]。

粗粒度的情感分析主要是基于篇章级和句子级[1]，利用分析情感词对文章的总体情感加以研究。粗粒度情感分析主要通过对具有情感倾向性的词加以细分，一般分为积极和消极两类，来进行情感情感分析。然而过粗的粒度的情感分析会忽略文本信息中许多细节，造成有价值信息的浪费。

细粒度的情感分析一般是指对词汇级文本的情感分析，目前关于细粒度情感分析的研究主要分为两大方向: 一是对文本中产品特征和对应情感词进行抽取，二是对情感进行分类[1]。基于特征分析的细粒度情感分析将研究单元进一步细化，为获得更加精细的、有价值的信息提供了可能；该方法对情感的分类不再是正负二分类或者加入中性的三分类，而是将情感分类的粒度进一步细分。

细粒度情感分析既要对更细化的单元进行挖掘，又要计算情感强度，以期对评论文本中的情感更加细腻地分析。例如，句子“买给朋友的，她很喜欢，蓝色很正，手机运行很快，薄薄的手感很好，充电速度也很快，用了一会基本没有发热，物美价廉 ”。在这句话中，我们能抽取到的细粒度信息有：产品特征“颜色、运行速度、充电速度”，对应的情感词“好看、很快、很快”等等。对更多评论进行如上分析可以获得各种特征出现的频率以及对应的特征的评价词的情感强度，可知消费者对于手机的哪些特征重视程度更高以及相应的情感倾向，从而帮助店铺提升口碑、增加产品销量、提高盈利水平。

2.2.2 情感分析方法

在文本情感分析领域中，目前常用的情感分析方法可以分为基于情感词典的方法和基于机器学习的方法两大类，对两种方法的详细说明如下：

（1）基于词典的情感分析的方法

基于词典的情感分析的方法最关键的工作是情感词典的构建，情感词典的质量将会对情感分析的结果产生直接的影响。情感词典构建好后，便可以基于情感词典计算情感值，对情感值求和，确定情感强度。情感词典包括基础情感词典和情感词典的扩建，下面对两部分情感词典进行详细介绍。

基础情感词典：基础情感词典为已经应用的、被学者们广泛认可的、具有代

表性的词典，如知网整理的 Howent 情感词典、大连理工大学情感词汇本体库、停用词词典等。将这些专业情感词词典组合起来作为基础情感词典，能够提高情感词典的专业性和准确性。

情感词典的扩建：对情感词典的扩建主要是为了增加时效性以及考虑修饰词与否定词。时效性是指随着学者们研究的深度增加以及社会的发展，会出现新的情感词，为提高情感词的覆盖面，需要将新出现的情感词加入到情感词典中。修饰词对情感强度起到加强或者削弱的作用，否定词则会改变情感倾向，因此也需要考虑到情感词典的扩建中。

（2）基于机器学习的情感分析的方法

基于机器学习的情感分析方法是一种有监督的方法，以解决二分类问题为主，常用的方法有支持向量机（SVM）、朴素贝叶斯、最大熵等。用机器学习的方法进行情感分析的步骤为：首先，将文本进行训练集与测试集的划分，并对训练集进行标注；其次，对标注好的训练文本做文本清洗、分词、词性标注等处理；然后使用模型抽取特征、训练分类等；最后对训练好的模型使用测试集进行测试，从而评估模型的性能。

（3）情感分析方法对比

通过比较基于词典的情感分析方法与基于机器学习的情感分析方法，可以发现，虽然基于词典的情感分析方法精确度较低，但是在分类类别上更加灵活，而机器学习的方法只能处理情感的二分类问题，对情感的分类问题限制多。另外，基于机器学习的情感分析方法中文本标注的效果对模型训练结果起到重要的影响甚至决定作用，而文本标注工作需要人工手动完成，工作量巨大且存在很大主观性，标注效果没有固定的衡量标准，存在很大的不确定性，对模型的的训练存在不利影响，而基于词典的情感分析方法无需进行文本标注工作，以词典为依据进行情感值的计算，准确性与可靠性高。因为细粒度情感的分析方法中对情感的分析不止二分类，所以更适合用基于词典的情感分析方法。

2.2.3 情感强度计算与分析

传统的情感分析以对情感极性的分类为主，即确定情感词的正向性与负向性的二分类，常采用的计算方法为点互信息法。点互信息法确定情感倾向是计算正向情感词与待计算情感词的点互信息与负向情感词与待计算情感词的点互信息

差，根据差值来确定情感倾向。由于细粒度情感分析不止对情感进行二分类，而是多个情感分类，所以该方法不适合用于细粒度情感分类。

鉴于上面所提到的问题，参考敦欣卉等人[1]细粒度情感计算的方法，即为对情感词进行正确的分析，先要确定情感分类，然后在分类的基础上计算情感强度。情感强度计算的具体步骤为：首先计算情感词的情感值，获得该词的对应情感得分；然后判断是否有修饰词、否定词，并把相应的权值计算出来，与情感得分复合；接着按照这个过程循环遍历完所有的该特征的情感词为止，并把遍历获得的情感值得分加和，获得该特征的情感得分。对于未出现在情感词典中的情感词可以采用语义相似度计算的方法使用余弦相似度计算方法找出与其语义最为相近的情感词的对应情感值，作为该情感词的情感值。

2.3BERT 模型的相关理论研究

2.3.1BERT 模型的基本思想

基于变换器的双向编码器表示技术——BERT，是用于自然语言处理（NLP）的预训练技术，由谷歌 AI 于 2018 年 10 月提出的一种基于深度学习的语言表示模型[57]。BERT 采用双 transformer 的结构，是一个能实现自编码语言的模型，并以 Mask 语言模型做预训练，以实现同时利用前后两个方向的信息，做到综合考虑上下文信息，即充分考虑语义环境，从而提高信息提取的精确度。BERT 模型使用简单，可直接调用训练好的模型。

2.3.2BERT 模型建立的基本步骤

建立 BERT 模型的具体步骤如下：

步骤 1，预处理后的文本经 BERT 模型进行词嵌入，并使用公开的预训练好的词向量进行词向量化转换。BERT 模型与循环神经网络相较，只运用了前馈神经网络和多头注意力机制，因此不能直接学习到词的位置信息，鉴于此，在 BERT 模型中加入学习词位置信息的位置编码向量(Positional Encoding)，用来判断给定词在序列中的具体位置，如公式（2.1）和（2.2）所示：

$$PE_{(pos,2i)} = \sin(pos / 1000^{2i/d_{\mathrm{model}}}) \tag{2.1}$$

$$PE_{(pos,2i+1)} = \cos(pos / 1000^{2i/d_{\mathrm{model}}}) \tag{2.2}$$

其中，pos 表示词位置，d_{model} 表示词向量维度。输入向量为词向量与位置

编码向量的拼接组合。

步骤 2，多头自注意力机制(Multi-Head Attention)是 Transformer 编码器的关键，其主要的功能是增强关注力，从而提高特征提取的准确性。横向拼接 Self-attention 矩阵，并与附加权重的矩阵 W^o 相乘，然后将乘积矩阵进行压缩，其维度要与输入序列维度一致的矩阵，计算公式如式（2.3）所示：

$$\begin{aligned} MultiHead\ (Q,K,V) &= Concat\ (head_1, head_2, ..., head_k)W^o \\ Where\quad \text{head}_i &= Attention\ (Q_i, K_i, V_i) \end{aligned} \tag{2.3}$$

其中，Q、K、V 用来表示输入序列中词对应的 query、key 和 value 向量，由输入特征向量 X 与对应的权重矩阵 W^Q、W^k、W^v 相乘得到，head 的个数为超参数，其具体数量可以人为进行设定，也可以选择默认值，Attention 的计算公式如式（2.4）所示。

$$Attention(Q,K,V) = Soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{2.4}$$

其中，d_k 为 query 和 key 的向量维度，向量维度的具体取值可以根据具体情况确定。

Softmax()为归一化激活函数，其计算公式如式（2.5）所示。

$$Soft\ \max(Z)_j = \frac{e^{z_j}}{\sum_{k=1}^{N} e^{Z_k}} \tag{2.5}$$

其中，z 表示 N 维的行向量。

步骤 3，将前两个步骤的处理结果在输入前馈网络、残差网络，从而获得最终的全局语义特征向量，并使用余弦相似度公式计算向量之间的相似度，余弦相似度计算公式如公式（2.6）所示。

$$COS\ （\theta）= \frac{\sum_{i=1}^{n}(X_i * Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} * \sqrt{\sum_{i=1}^{n}(Y_i)^2}} \tag{2.6}$$

其中，X、Y 分别为对应的特征向量，通过该可以计算出特征向量的余弦值，两个向量相似度越高，向量的夹角越接近 0，余弦值越接近 1，当夹角为 0 时，余弦值为 1，两个向量相等。

2.3. 3BERT 模型的衍生模型

由于 BERT 在自然语言处理诸多领域中表现十分突出，所以受到了学者们的高度重视，因此产生了很多衍生模型，如 RoBERTa、ALBERT、BERTopic 等，由于本文主要研究内容为特征提取，所以对主题模型 BERTopic 进行介绍，为后文应用该模型奠定理论基础。

BERTopic 模型是 BERT 模型的衍生模型，以 BERT 为嵌入，能够实现主题的自动聚类。BerTopic 算法可以分为如下三个阶段：

1.文本数据嵌入。该步骤中常见的的方法是直接使用 BERT 算法提取文档进行嵌入，也可以将 BERT 算法替换成其他嵌入算法。

2.文档聚类。首先降低嵌入文本的维数，然后通过聚类来减少嵌入文本量，在此基础上创建语义相似的文本。

3.确定主题文本。通过 TF-IDF 算法对主题进行提取，生成不同主题的文本。由于 BERTopic 模型属于无监督模型，在根据文本内容自动进行主题文本提取时会将复合特征归成新主题，导致主题数量庞大，为减少主题数量，提高分类准确性需要对主体数量进行调整。调整方法有手动调节、自动调节和训练调节三种方式。其中，手动调节需要知道文本主题分类的结果，在实际操作中实现困难；训练调节则需要花费大量时间成本，因此选择节约成本且具有可实际操作意义的自动调节方法，即将"nr_topics"设置为"auto"。

2.4LDA 模型的相关理论研究

2.4.1LDA 模型的基本思想

LDA 模型由 Blei 等人于 2003 年提出，是一种基于贝叶斯模型的、无监督的机器学习方法，由文档、主题和词 3 层构成，不需要对文本进行人工标注，通过指定主题数量，调整迭代次数，依靠词与词之间的关联构成的词袋模型，根据概率来获得主题分类结果。LDA 模型降维能力强，拓展性好，具有良好的文本挖掘能力，因而受到广大学者的喜爱。

2.4.2 建立 LDA 模型的基本步骤

主题模型（LDA）是一种能够以概率的形式给出每篇文章各主题的模型。模型会首先假设文档的主题数，并将单词依据概率划分到不同主题中；然后通过反复计算，找到最优结果的词组合；最后将其输出，输出的每一个包含概率的词组

合都对应一个话题。具体步骤如下：

步骤 1，对文档集合中的每篇文档 d 进行分词，并过滤掉分词结果中无意义词，从而可以得到语料集合 $W=\{w_1,w_2,\cdots,w_x\}$；

步骤 2，对每篇文档 d 中的词做统计，得到 $p(w_i|d)$；

步骤 3，为语料集合 W 中的每个 w_i，确定一个随机的初始主题；

步骤 4，使用 Gibbs Sampling 公式多次计算每个 w 的所属初始主题，直到 Gibbs Sampling 公式收敛为止。使用 Gibbs Sampling 公式计算，首先计算每一个主题下每一个词项的主题概率，概率值计算公式如公式（2.7）所示。

$$p(\omega|\alpha,\beta)=\int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(Z_n|\theta)p(\omega_n|Z_n,\beta)\right)d\theta \tag{2.7}$$

公式（2.7）中，ω是文本中的单词，α是评论——主题分布的 Dirichlet 超参数，β是主题——单词分布的 Dirichlet 超参数，θ是评论——主题分布，z 是评论中单词的主题。

当每个词的主题都确定后，就要计算词频来对参数做出估计，即计算词列表的主题序列的对应条件概率，计算公式如公式（2.8）所示。

$$p(z_i=k|\overrightarrow{z_{\neg i}}\overrightarrow{\omega})=\frac{p(\overrightarrow{\omega}\overrightarrow{z})}{p(\overrightarrow{\omega z_{\neg i}})}\alpha\frac{n_{k_i-1}^{\omega}+\beta_{\omega}}{\sum_{t=1}^{v}n_{k_i-1}^{\omega}+\beta_{\omega}}\left(n_{m_1\neg i}^{t}+\alpha_k\right) \tag{2.8}$$

其中，$\neg i$ 表示不包括第 i 项，$z_{\neg i}$ 为不包括第 i 词项词的主题变量，n_k^{ω} 表示第 k 主题中词ω出现的次数，β_{ω} 表示词ω的 Dirichlet 先验，α_k 表示主题 k 的 Dirichlet 先验。当每个单词获得主题标号之后，通过公式（2.9）、（2.10）计算所需的参数：

$$\theta_{k,\omega}=\frac{n_m^k+\beta_{\omega}}{\sum_{k=1}^{k}n_k^{\omega}+\beta_{\omega}} \tag{2.9}$$

$$\theta_{m,k}=\frac{n_m^k+\alpha_k}{\sum_{k=1}^{k}n_m^k+\alpha_k} \tag{2.10}$$

其中，$\theta_{m,k}$ 为文本 m 中主题 k 的概率，$\phi_{k,\omega}$ 为主题 k 中词ω的概率，n_k^{ω} 为文本 m 中出现主题 k 的次数。

步骤 5，收敛以后可以获得主题——词的概率矩阵，即 LDA 矩阵，同时可

以获得文档——主题的概率矩阵，经统计后，最后可得文档——主题的概率分布结果。

2.5 本章小结

本章首先介绍了文本抽取的相关理论，对关键短语的定义和抽取方法做了详细概述；然后介绍了情感分析的相关理论，包括对粗粒度情感分析和细粒度情感分析，及情感分析的方法和细粒度情感的计算方法与步骤；最后对 BERT 模型和 LDA 模型的相关理论进行了分析。

3 关键短语集合构建

细粒度情感分析的关键在于正确识别出属性特征及其情感观点。因为关键短语由属性特征、情感特征以及情感程度特征组成[4]，其组合表示如表 3.1 所示，这样的结构正是细粒度情感分析的关键，所以本章以手机产品的评论为例，构建关键短语集合，从而为进行细粒度情感分析打下重要基础。

表 3.1 关键短语结构

属性特征	情感特征	情感程度特征
屏幕	好看	非常
电池	耐用	特别
摄像	行	不
...	...	...

3.1 构建流程

构建手机产品关键短语集合的基本流程图如图 3.1 所示，具体可以分为 6 个步骤。

步骤 1：手机特征词典扩建。在现有研究的基础上对手机特征词典进行扩充，构建手机特征词典。

步骤 2：短句集合构建。将手机评论语料按照 7：3 的比例划分为 train 文本与 test 文本，将 train 文本进行预处理，并使用 python 中 Scikit-Learns 包的 CountVectorizer 方法进行分词，形成短句集合。

步骤 3：手机特征短句集合构建。使用 for 循环将短句集合与手机特征词典进行特征词匹配，得到含有手机特征的短句集合。

步骤 4：单一特征短句集合构建。去除含有多个特征的短句，形成单一特征的短句集合。

步骤 5：手机产品关键短语集合构建。去除不含情感词的单一特征短句，构建手机产品关键短句集合。

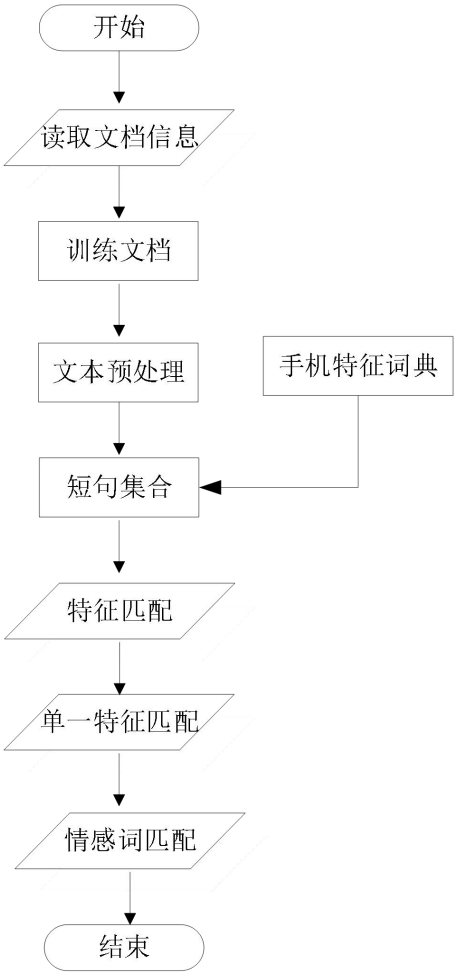


图 3.1 手机关键短语集合构建流程

3.2 手机特征词典构建

产品特征词典是产品特征的集合，属于半固定短语中的固定部分。构建手机领域的特征词典，就是找出手机关键短语中的固定部分，在此基础上通过语法结构将短语补充完整，使其更大程度上覆盖该领域的数据特征，从而提高特征提取的全面性。目前该领域已有多位学者构建了手机特征词典，汇总结果如表 3.2 所示。

表 3.2 手机特征词典

类型	手机特征词
屏幕	屏幕 桌面 触摸屏 显示屏 弧面 触屏 曲屏 分辨率 亮度 显示 界面
电池	电池 待机 容量 充电 快充 电量 耗电量 用电 时长

续表 3.2 手机特征词典

摄像	摄像 拍照 照相 像素 自拍 柔光 清晰度 色彩 摄影 镜头 画面感 相素 画质 美颜
内存与处理	性能 速度 运行速度 兼容性 不卡 卡顿 处理器 反应速度 延迟 网速 卡机 内存 死机
配件	玻璃膜 保护壳 壳子 钢化膜 手机壳 手机套 贴膜
系统与软件	功能 软件 智能 系统 操作
游戏	游戏 娱乐 玩游戏 手游
多媒体	多媒体 收音机 声音 语音 耳机 音质 视频 音响 电影 蓝光 音量 播放器 音箱 听歌 播放 画质
外形	外形 机身 手感 体积 外观 缝隙 重量 质感 工艺 颜色 触感 外表 线条 机身 样子 造型 设计 个性 颜值
服务	售后 服务 物流 客服 态度 口碑 顺丰 卖家
价格	价格 价钱 性价比 降价 打折 定价 标价

从表 3.2 可以得到手机特征词典分为 11 类，共有 93 个特征词。该手机特征词典是由语料训练获得的，会受语料数量和时效性的影响，所以难免存在特征词不够全面的问题，因此需要对手机特征词典进行扩建。

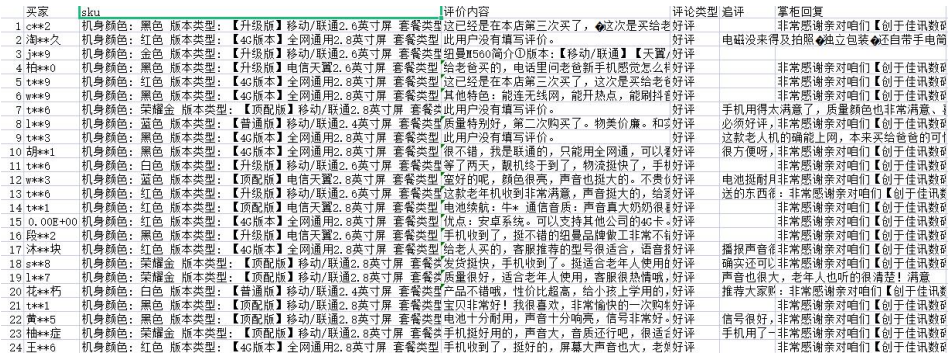
3.3 手机特征词典扩展

为了构建更加全面的手机特征词典，本文使用思睿智训平台[2]获取的手机评论语料，基于基础词典对手机特征词典进行扩充。

3.3.1 数据来源

实验数据来源于思睿智训平台，该平台数据为淘宝平台真实数据的汇总。所采用的数据包括产品信息、买家、评价内容、追评等字段的内容，数据总量为 41438 条，部分数据内容如图 3.2 所示。

[2] 思睿智讯.电商数据化运营综合实训平台,(2019-11-2) [2021-3-6]. https://dsjx.srzx.com/



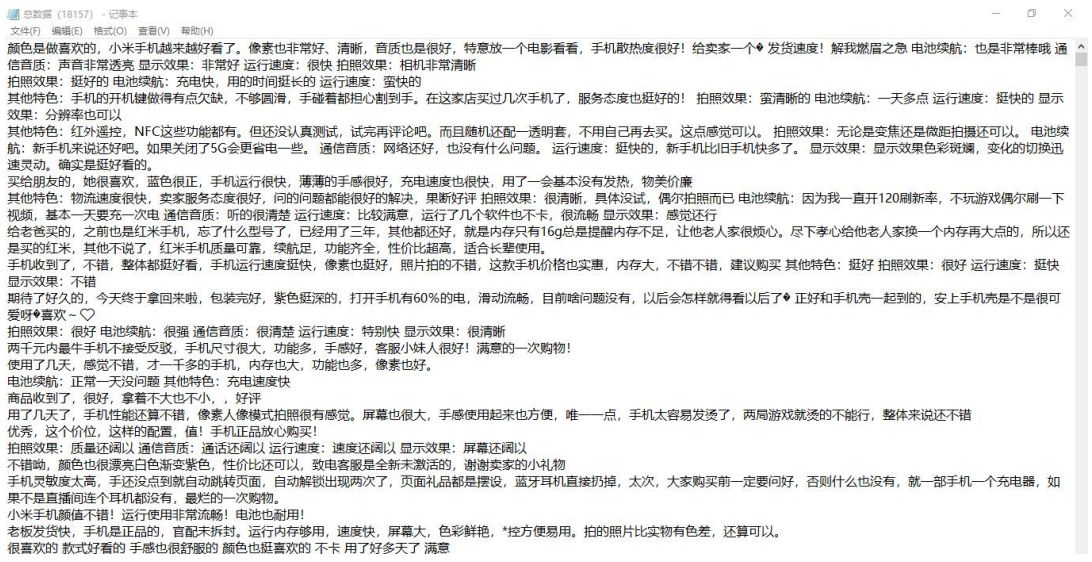
	买家	sku	评价内容	评论类型	追评	掌柜回复
1	c**2	机身颜色：黑色 版本类型：【升级版】移动/联通2.6英寸屏 套餐类型	这已经是在本店第三次买了，❀这次是买给老	好评		非常感谢亲对咱们【创于佳讯数码
2	淘**久	机身颜色：红色 版本类型：【4G版本】全网通用2.8英寸屏 套餐类型	此用户没有填写评价。	好评	电磁没来得及拍照❀独立包装❀适自带手电筒	
3	j**9	机身颜色：金色 版本类型：【升级版】移动/联通2.6英寸屏 套餐类型	纽曼M560简介①版本：【移动/联通】【天翼/	好评		
4	柏**0	机身颜色：黑色 版本类型：【升级版】电信天翼2.6英寸屏 套餐类型	给老爸买的，电话里问老爸新手机感觉怎么样	好评		非常感谢亲对咱们【创于佳讯数码
5	t**9	机身颜色：红色 版本类型：【4G版本】全网通用2.8英寸屏 套餐类型	这已经是在本店第三次买了，这次是买给老爸	好评		非常感谢亲对咱们【创于佳讯数码
6	w**9	机身颜色：黑色 版本类型：【4G版本】全网通用2.8英寸屏 套餐类型	其他特色：能连无线网，能开热点，能刷抖音	好评		非常感谢亲对咱们【创于佳讯数码
7	t**6	机身颜色：荣耀金 版本类型：【顶配版】移动/联通2.8英寸屏 套餐类型	此用户没有填写评价。	好评	手机用得太满意了，质量颜色也非常满意、	
8	l**9	机身颜色：蓝色 版本类型：【普通版】移动/联通2.4英寸屏 套餐类型	质量特别好，第二次购买了。物美价廉。和实	好评	必须好评，非常感谢亲对咱们【创于佳讯数码	
9	t**3	机身颜色：黑色 版本类型：【4G版本】全网通用2.8英寸屏 套餐类型	此用户没有填写评价。	好评	这款老人机的确能上网，本来买给爸爸的可	
10	胡**1	机身颜色：黑色 版本类型：【4G版本】全网通用2.8英寸屏 套餐类型	很不错，我是联通的，只能用全网通，可以看	好评	很方便呀，非常感谢亲对咱们【创于佳讯数码	
11	t**6	机身颜色：白色 版本类型：【升级版】移动/联通2.6英寸屏 套餐类型	等了两天，靓机终于到了，物流挺快了，手机	好评		非常感谢亲对咱们【创于佳讯数码
12	w**3	机身颜色：蓝色 版本类型：【顶配版】电信天翼2.8英寸屏 套餐类型	蛮好的呢，颜色很亮，声音也挺大的。不贵的	好评	电池挺耐用非常感谢亲对咱们【创于佳讯数码	
13	t**6	机身颜色：红色 版本类型：【升级版】移动/联通2.6英寸屏 套餐类型	这款老年机收到非常满意，声音挺大的，给爸	好评	送的东西多：非常感谢亲对咱们【创于佳讯数	
14	t**1	机身颜色：红色 版本类型：【顶配版】电信天翼2.8英寸屏 套餐类型	电池续航：*** 通信音质：声音真大奶奶很喜	好评		非常感谢亲对咱们【创于佳讯数码
15	0.00**00	机身颜色：红色 版本类型：【4G版本】全网通用2.8英寸屏 套餐类型	优点：安卓系统。可以支持其他公司的4G卡。	好评		非常感谢亲对咱们【创于佳讯数码
16	段**2	机身颜色：黑色 版本类型：【升级版】电信天翼2.6英寸屏 套餐类型	手机收到了，挺不错的纽曼品牌做工非常不错	好评		非常感谢亲对咱们【创于佳讯数码
17	冰**块	机身颜色：红色 版本类型：【4G版本】全网通用2.8英寸屏 套餐类型	给老人买的，客服推荐的型号很适合，语音报	好评	播报声音很非常感谢亲对咱们【创于佳讯数码	
18	s**8	机身颜色：荣耀金 版本类型：【顶配版】移动/联通2.8英寸屏 套餐类型	发货挺快，手机收到了。挺适合老年人使用的	好评	确实还可以非常感谢亲对咱们【创于佳讯数码	
19	l**7	机身颜色：荣耀金 版本类型：【顶配版】移动/联通2.8英寸屏 套餐类型	质量很好，适合老年人使用，客服很热情哦，	好评	声音也很大，老年人也听的很清楚！满意	
20	花**朽	机身颜色：白色 版本类型：【普通版】移动/联通2.4英寸屏 套餐类型	产品不错哦，性价比超高，给小孩上学用的，	好评	推荐大家购：非常感谢亲对咱们【创于佳讯数	
21	t**1	机身颜色：黑色 版本类型：【顶配版】移动/联通2.8英寸屏 套餐类型	宝贝非常好！我很喜欢，非常愉快的一次购物	好评		非常感谢亲对咱们【创于佳讯数码
22	黄**5	机身颜色：黑色 版本类型：【顶配版】移动/联通2.8英寸屏 套餐类型	电池十分耐用，声音十分响亮，信号非常好。	好评	信号很好，非常感谢亲对咱们【创于佳讯数码	
23	柚**症	机身颜色：荣耀金 版本类型：【顶配版】移动/联通2.8英寸屏 套餐类型	手机挺好用的，声音大，音质还行吧，很适合	好评	手机用了一非常感谢亲对咱们【创于佳讯数码	
24	王**6	机身颜色：红色 版本类型：【4G版本】全网通用2.8英寸屏 套餐类型	手机收到了，挺好的，屏幕大声音也大，老人	好评		非常感谢亲对咱们【创于佳讯数码

图 3.2 部分数据内容

3.3.2 文本预处理

文本预处理是进行文本情感分析的第一步，主要是对文本中没有价值以及价值低的信息进行清洗与过滤，从而提高文本信息的有效性。文本预处理主要包括文本规范化、分词、去停用词等。

（1）文本规范化。消费者的评论中往往存在很多无意义的评论，如“给爸爸买的”、“好评”、“不错”等，为减少这些无意义文本的干扰，应将其过滤掉。文本规范化处理后共剩余 18157 条有效文本，部分内容如图 3.3 所示。将有效文本按照 7:3 的比例划分为 12710 条 Train 文本和 5448 条 Test 文本。



总数据（18157）- 记事本

文件(F)　编辑(E)　格式(O)　查看(V)　帮助(H)

颜色是做喜欢的，小米手机越来越好看了。像素也非常好、清晰，音质也是很好，特意放一个电影看看，手机散热度很好！给卖家一个❀ 发货速度！解我燃眉之急 电池续航：也是非常棒哦 通信音质：声音非常透亮 显示效果：非常好 运行速度：很快 拍照效果：相机非常清晰

拍照效果：挺好的 电池续航：充电快，用的时间挺长的 运行速度：蛮快的

其他特色：手机的开机键做得有点欠缺，不够圆滑，手碰着都担心割到手。在这家店买过几次手机了，服务态度也挺好的！ 拍照效果：蛮清晰的 电池续航：一天多点 运行速度：挺快的 显示效果：分辨率也可以

其他特色：红外遥控，NFC这些功能都有。但还没认真测试，试完再评论吧。而且随机还配一透明套，不用自己再去买。这点感觉可以。 拍照效果：无论是变焦还是微距拍摄还可以。 电池续航：新手机来说还好吧。如果关闭了5G会更省电一些。 通信音质：网络还好，也没有什么问题。 运行速度：挺快的，新手机比旧手机快多了。 显示效果：显示效果色彩斑斓，变化的切换迅速灵动。确实是挺好看的。

买给朋友的，她很喜欢，蓝色很正，手机运行很快，薄薄的手感很好，充电速度也很快，用了一会基本没有发热，物美价廉

其他特色：物流速度很快，卖家服务态度很好，问的问题都能很好的解决，果断好评 拍照效果：很清晰，具体没试，偶尔拍照而已 电池续航：因为我一直开120刷新率，不玩游戏偶尔刷一下视频，基本一天要充一次电 通信音质：听的很清楚 运行速度：比较满意，运行了几个软件也不卡，很流畅 显示效果：感觉还行

给老爸买的，之前也是红米手机，忘了什么型号了，已经用了三年，其他都还好，就是内存只有16g总是提醒内存不足，让他老人家很烦心。尽下孝心给他老人家换一个内存再大点的，所以还是买的红米，其他不说了，红米手机质量可靠，续航足，功能齐全，性价比超高，适合长辈使用。

手机收到了，不错，整体都挺好看，手机运行速度挺快，像素也挺好，照片拍的不错，这款手机价格也实惠，内存大，不错不错，建议购买 其他特色：挺好 拍照效果：很好 运行速度：挺快 显示效果：不错

期待了好久的，今天终于拿回来啦，包装完好，紫色挺深的，打开手机有60%的电，滑动流畅，目前啥问题没有，以后会怎样就得看以后了❀ 正好和手机壳一起到的，安上手机壳是不是很可爱呀❀喜欢～♡

拍照效果：很好 电池续航：很强 通信音质：很清楚 运行速度：特别快 显示效果：很清晰

两千元内最牛手机不接受反驳，手机尺寸很大，功能多，手感好，客服小妹人很好！满意的一次购物！

使用了几天，感觉不错，才一千多的手机，内存也大，功能也多，像素也好。

电池续航：正常一天没问题 其他特色：充电速度快

商品收到了，很好，拿着不大也不小，，好评

用了几天了，手机性能还算不错，像素人像模式拍照很有感觉。屏幕也很大，手感使用起来也方便，唯——点，手机太容易发烫了，两局游戏就烫的不能行，整体来说还不错

优秀，这个价位，这样的配置，值！手机正品放心购买！

拍照效果：质量还阔以 通信音质：通话还阔以 运行速度：速度还阔以 显示效果：屏幕还阔以

不错呦，颜色也很漂亮白色渐变紫色，性价比还可以，致电客服是全新未激活的，谢谢卖家的小礼物

手机灵敏度太高，手还没点到就自动跳转页面，自动解锁出现两次了，页面礼品都是摆设，蓝牙耳机直接扔掉，太次，大家购买前一定要问好，否则什么也没有，就一部手机一个充电器，如果不是直播间连个耳机都没有，最烂的一次购物。

小米手机颜值不错！运行使用非常流畅！电池也耐用！

老板发货快，手机是正品的，官配未拆封。运行内存够用，速度快，屏幕大，色彩鲜艳，*控方便易用。拍的照片比实物有色差，还算可以。

很喜欢的 款式好看的 手感也很舒服的 颜色也挺喜欢的 不卡 用了好多天了 满意

图 3.3　部分有效文本

（2）分词及去停用词。在每一条消费者对产品的评论文本中都包含多个产

品特征，而且每个产品特征的情感评论词也不尽相同。比如“电池续航非常好，屏幕够大，外形很美观，接收信号确实挺快的，但是像素不是很给力”，这条文本中包含“屏幕”、“外形”、“信号”、“像素”五个特征，这五个特征的情感词分别为“够大”、“很美观”、“确实挺快”、“不是很给力”，五个不同的情感评论词。使用 jieba 分词法会将特征与对应评价切开，影响细粒度情感分析的准确性，因此本文使用 python 中 Scikit-Learns 包中的 CountVectorize 方法进行分词，该分词方法将评论文本划分为包含特征与对应评价的短句，如[电池续航非常好]、[屏幕够大]、[外形很美观]。另外 CountVectorize 分词方法还可以去掉常见的停用词，如，“我们”、“了”，“而且”、“如果”等。停用词是指在文本中高频率出现，但是与产品特征和情感词无关或相关性小的词，将停用词去掉可以将注意力更多的停留在产品特征与情感词上，从而提高短句信息的纯度与有效性。最终形成了由特征与对应评价组成的短句集合，部分结果如图 3.4 所示，为下文手机特征词典的扩建打下基础。

新手机的运行速度确实好',
'手机机身颜色漂亮手感很好',
'手机颜值很好',
'手机质感超好',
'手机运行速度也很nice',
'手机很好用颜色很漂亮',
'这个手机性价比很高是正品很不错充电又快',
'性能好手机很流畅送的蓝牙耳机也挺好',
'手机质感很不错的',
'收到了质量不错照相清晰电池耐用',
'手机不错很流畅就是顺丰慢了点价格很实惠',
'手机颜值高挺喜欢的',
'手机很不错运行速度很快',
'手机本身机型和配色都很好看',
'手机功能很齐全到货很快很满意拍照效果',
'手机质感非常好',
'手机很好看呀喜欢的快冲',
'手机拍照色彩很好',
'手机的质感很好',

图 3.4 部分短句集合

3.3.3 手机特征词典扩建

基于词典的情感分析的关键是构建准确且全面的特征词典，为此本文对手机基础词典进行扩建。手机特征词典扩建的算法流程如图 3.5 所示，具体可以分为

3 个步骤。

步骤 1：确定文本数据。train 文本预处理后获得的短句集合与上文构建的手机词典进行匹配，将无特征匹配的短句作为训练文本。

步骤 2：使用 python 中的 Jieba 包对无特征词匹配的短句分词并标注词性，将名词短语进行汇总。

步骤 3：计算名词短语与特征词典之间的余弦相似度，将相似度大于 0.8 的词语作为新的的特征词，加入到手机特征词典中。

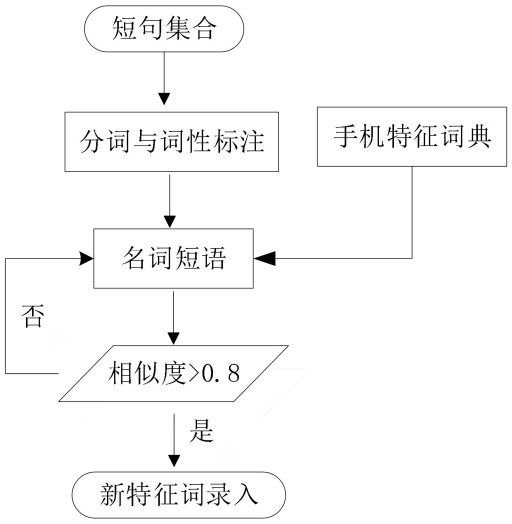


图 3.5 手机特征词典扩建流程

经过计算后分别将“全面屏”与“折叠屏”、“续航”、“变焦”、“刷新率”与“高刷”、“价位”特征加入到“屏幕”、“电池”、“摄像”、“内存与处理”、“价格”类型中对应的手机特征词典中，结果如表 3.3 所示。

表 3.3 手机特征词典扩展

类型	手机特征词
屏幕	屏幕 桌面 触摸屏 显示屏 弧面 触屏 曲屏 分辨率 亮度 显示 界面 全面屏 折叠屏
电池	电池 待机 容量 充电 快充 电量 耗电量 用电 时长 续航
摄像	摄像 拍照 照相 像素 自拍 柔光 清晰度 色彩 摄影 镜头 画面感 相素 画质 美颜 变焦
内存与处理	性能 速度 运行速度 兼容性 不卡 卡顿 处理器 反应速度 延迟 网速 卡机 内存 死机 刷新率 高刷

续表 3.3 手机特征词典扩展

类型	手机特征词
配件	玻璃膜 保护壳 壳子 钢化膜 手机壳 手机套 贴膜
系统与软件	功能 软件 智能 系统 操作
游戏	游戏 娱乐 玩游戏 手游
多媒体	多媒体 收音机 声音 语音 耳机 音质 视频 音响 电影 蓝光 音量 播放器 音箱 听歌 播放 画质
外形	外形 机身 手感 体积 外观 缝隙 重量 质感 工艺 颜色 触感 外表 线条 机身 样子 造型 设计 个性 颜值
服务	售后 服务 物流 客服 态度 口碑 顺丰 卖家
价格	价格 价钱 性价比 降价 打折 定价 标价 价位

3.4 手机关键短语集合创建

对手机评论文本进行某一特征的细粒度情感分析，首先，要找到带有倾向性的评价观点组合，即关键短语。所以本文基于扩建的手机特征词典构建手机关键短语集合，具体方法如下：

（1）特征识别。特征为关键短语中的一部分，且特征的识别以手机特征词典为依据，具有固定性。使用 for 循环将 train 文本的短句集合与手机特征词典进行特征词匹配，并按照手机特征词典的类型进行分类，共划分为 11 类，部分结果如表 3.4 所示。

表 3.4 特征识别

属性特征	举例
['电池']	05 起床充电 0 到 100 只要 23 分钟
[]	0 点下单上午 10 点就拿到
['系统与软件']	0 版本后少了很多功能
[]	1000 元性价最高的手机
['价格']	1000 多块钱我也懒得跟他们生嫌气留着当备用机
[]	100 实测 20 分钟
['屏幕']	100 息屏状态下 25 分钟就能充满
['多媒体']	100 的电上班路上刷个视频
[]	100 米外人脸基本清晰可辨
['价格']	1050 的价格
['屏幕']	1080p 屏幕不错
['摄像'， '价格']	10p 拍照被吊打华为价格和配置不对等
['系统与软件'， '内存与处理'，'电池']	120w 充电器充电功能速度超级快
...	...

（2）识别单一特征。关键短语应具有高纯度的特点，即属性特征为单一的。从表 3.4 可以看出，分类结果既有不含手机特征的语料，也有含有单一特征、多个特征的语料，因此需要剔除不含手机特征以及多个特征的文本。将不含手机特征以及多个手机特征的短语筛选并删除，保留仅含单一特征的文本，部分结果如表 3.5 所示。

表 3.5 单一特征文本

属性特征	举例
['屏幕']	屏大
['屏幕']	屏保
['电池']	不怕电池损耗
['电池']	不想浪费时间了
['摄像']	1080 像素
['摄像']	10p 像素太高
['内存与处理']	16g 内存就是流畅
['内存与处理']	128g 内存
['配件']	不带壳很难用
['配件']	三件套
['系统与软件']	功能上
['系统与软件']	功能不错
['游戏']	一直玩文字游戏
['游戏']	打游戏相对差一点
['多媒体']	一边看视频一边聊天
['多媒体']	声音不够厚实
['外形']	手感丝滑
['外形']	外形外观
['服务']	qq 客服
['服务']	客服态度都挺好的
['价格']	1679 的价格
['价格']	价格不贵
...	...

（3）构建手机关键短语集合。由表 3.5 可以看出，虽然文本中只含有单一特征，但是像“qq 客服”、“功能上”、“屏保”、“128g 内存”等文本，没有情感特征词和情感程度词，不符合关键短语结构，因此还需要进一步对其进行处理。为提高处理结果的准确性采用人工识别方法，剔除不含有情感词的单一特征的短语，从而构建手机关键短语集合，部分结果如表 3.6 所示。

表 3.6 部分手机关键短语集合

属性特征	举例
屏幕	屏幕很棒 屏幕很流畅 显示好 调节亮度不流畅 分辨率不好
电池	充电很舒服 充电杠杠的 电量很耐用 待机强大 电池续航给力
摄像	像素太高 摄像头真是超赞 拍照很强大 清晰度不错 画面色彩鲜亮
内存预处理	速度很快 内存流畅 处理器还可以 一点不卡 反应速度流畅 开机快
配件	不带壳很难用 套餐可以 手机膜是烂的 后壳太浮夸 送运动手环
系统与软件	功能不错 操作还算简单 程序比较慢 运行软件秒进 下载软件很快
游戏	玩游戏非常舒服 看小说发烫 飞车游戏感觉不是很流畅
多媒体	声音清楚 双扬声器效果很棒 手机的视频防抖也很厉害 画质好
外形	外观很漂亮 手感特好 制作工艺很精美 商品设计完美 外表好看
服务	物流发货慢到货快 客服差的很 服务周到 售后没话说 顺丰快
价格	性价比高 价格很值 降价太快 亲民价钱 1000 块钱都不值

3.5 本章小结

本章在现有手机词典的基础上，通过 train 文本对其进行扩展，并在此基础上根据关键短语结构，构建了手机关键短语集合，为 BERT-LDA 模型的构建奠定基础。

4 BERT-LDA 模型构建

4.1 模型的构建

传统的主题模型 LDA，虽然能够实现评价对象（特征，属性）和评价词（情感词）的同时提取，但是存在语义环境考虑不充分、提取精度不高等问题。为此，本文引入 BERT 模型，构建 BERT-LDA 模型，从而提高产品评论文本中评价对象和评价词提取的精确度。

4.1.1 模型构建思想

BERT-LDA 模型的基本思路：由于 LDA 模型提取精度不高，而 BERT 模型基于词语相似度进行信息识别提取，具有准确率高的优点，且 LDA 模型会把评价对象（特征，属性）和评价词（情感词）同时提取出来，使得特征提取结果可读性强。因此，首先，将语料集输入 BERT，进行向量化，并基于词语相似度进行识别提取，通过改变相似度阈值来控制提取的相似度词语的质量[54]；然后，将经过相似度提取的语句输入到 LDA 模型，获得评价对象和评价词，提高特征提取的精确度与可读性，为建立基于特征的细粒度情感词典打下基础。

4.1.2 模型构建过程

BERT-LDA 模型的基本原理是先用 BERT 模型将文本词向量化，使用训练获得的相似度值对相似词识别提取，将结果输入 LDA 主题模型进行训练，获得评

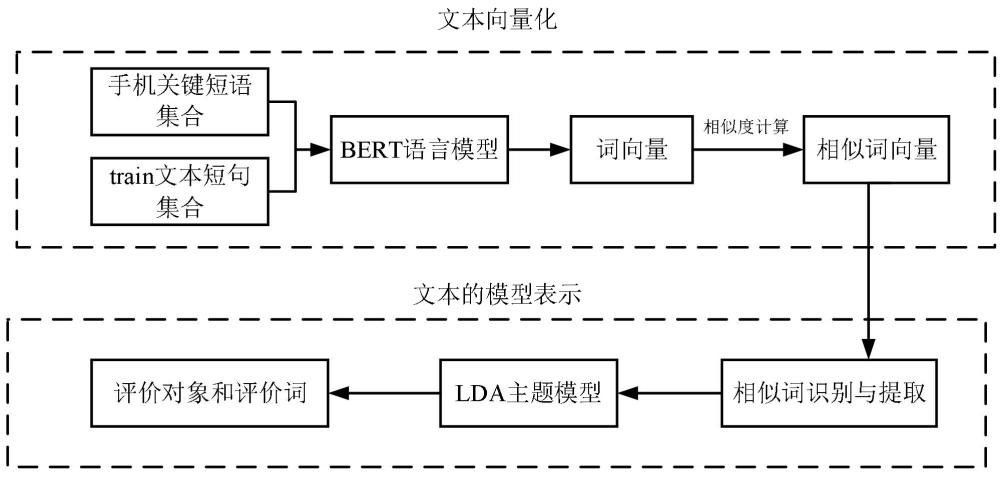


图 4.1 BERT-LDA 模型构建基本流程

价对象和评价词，从而提高特征提取的精确度与可读性。BERT-LDA 模型构建基本流程如图 4.1 所示，算法描述如 Algorithm 4.1 所示。

Algorithm 4.1 BERT-LDA 算法程序

输入：手机关键短语集合
　　train 文本分割成的短句集合
输出：主题结果

```
文本长度限定在 126，加上【BOS】和【EOS】两个特殊 token，序列长度为 128，文本长度不足 126 的用 0 补齐，超过的截断
加载 bert 分词器
加载 bert-base-chinese 预训练模型
将文本转化为数字序列
获得输入的文本序列、位置编码、mask 矩阵
获取 bert 预训练模型的文本嵌入向量
def get_cos_similar(v1, v2)#计算余弦相似度
    num = float(np.dot(v1, v2))  # 向量点乘
    denom = np.linalg.norm(v1) * np.linalg.norm(v2)  # 求模长的乘积
    return 0.5 + 0.5 * (num / denom) if denom != 0 else 0   #归一化
获得指定相似度下相似文本结果
分词
生成词矩阵
训练迭代次数与主题数量
获得主题结果
```

4.2 模型训练及检验

衡量主题模型性能的常用方法为困惑度（Perplexity），所以本文将对 BERT-LDA 模型进行训练，从而降低其困惑度，并与 BERTopic-LDA、LDA 模型进行困惑度比较。

4.2.1 模型性能评价标准

困惑度早在 2003 年由 Blei 等提出并应用于衡量 LDA 模型的优劣，其基本思想是模型的困惑度越低，模型的性能越好[57]。困惑度的计算公式[58]如式（4.1）所示：

$$Perplexity\ (D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log\ p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \qquad (4.1)$$

其中，D 是测试集，M 是测试语料集的文本数量，N_d代表第 d 篇文本的大小（即单词的个数），$p(w_d)$是文档中每个词的概率，词概率的计算方法如公式（4.2）所示：

$$p(w) = p(z/d) * p(w/z) \tag{4.2}$$

其中，p(z|d)是一个文档中每一个主题出现的概率，p(w|z)表示词典中的每个单词在某个主题下出现的概率。

4.2.2 模型训练

（1）BERT 模型的训练

由于评论文本中本身可能含有完全一样的描述语言，所以对训练评论文本进行编码，给与唯一确定的编码号，用来区分同样的评论文本，因此重复文本是指同一编号多次出现的文本，其数量统计方法为同一编码号初次出现不计数，第二次出现记为 1，以此累计。无关文本指不符合关键短语结构的文本。

在 BERT 模型中，主要应用其词向量化及相似度计算，获得相似短句，因此要确定相似度的取值。在训练过程中发现，相似度取值过高与过低都会影响相似短句的质量。当相似度取值过高时，被识别的文本与关键短语集合内容要高度一致才能被提取，这在实际应用中会导致很多特征不能被识别，不仅降低了文本的利用率，而且导致识别结果没有说服力；当相似度取值过低时，不仅会导致同一文本与多个关键短语具有相似关系而被多次识别，即出现重复文本，而且会识别出与关键短语无关的文本，即获得不具备手机特征及相应情感词的无关文本。相似度计算方法如公式（4.3）所示。为了方便进行比较，将相似度取值做归一化处理，将相似度取值范围变为（0，1），归一化计算方法如公式（4.4）所示。

$$COS（\theta）= \frac{\sum_{i=1}^{n}(X_i * Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} * \sqrt{\sum_{i=1}^{n}(Y_i)^2}} \tag{4.3}$$

$$\left\{ \begin{array}{c} 0.5 + 0.5\dfrac{\sum_{i=1}^{n}(X_i * Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} * \sqrt{\sum_{i=1}^{n}(Y_i)^2}}, \sqrt{\sum_{i=1}^{n}(X_i)^2} * \sqrt{\sum_{i=1}^{n}(Y_i)^2} \neq 0 \\ 0, \sqrt{\sum_{i=1}^{n}(X_i)^2} * \sqrt{\sum_{i=1}^{n}(Y_i)^2} = 0 \end{array} \right\} \neq 0, \tag{4.4}$$

因为改变相似度阈值可以控制提取的相似度词语的质量，所以通过 train 文本对 BERT 模型进行训练，确定最优的相似度取值范围。对不同的相似度取值下的相似文本的结果进行统计，获得部分相似度与无关文本、重复文本之间数量的关系，结果如表 4.1 所示。

表 4.1 相似度、无关文本、重复文本之间的关系

相似度值	无关文本比重（%）	重复文本比重（%）
1	0	0
(0.999，1)	1.52	2.09
(0.998，1)	3.68	4.14
(0.997，1)	6.35	8.21
(0.996，1)	9.63	15.46

由表 4.1 可知，当相似度取值逐渐减小时，无关文本与重复文本的比重逐渐递增，当相似度的值域取(0.997，1)时，无关文本、重复文本比重开始快速增加。所以为了识别出更多相似文本，并保证文本的质量，相似度的值域取(0.998，1)。获得的部分相似文本如表 4.2 所示。

表 4.2 相似文本

显示效果也很棒
显示效果很棒
显示画面很好
机身很舒服
屏幕大
流畅度也非常好
清晰度确实很好
画面很清晰
.....

（2）LDA 模型训练

为减少重复文本与无关文本对 LDA 结果的干扰，将表 4.2 中的无关文本与重复文本清理后输入到 LDA 模型中。由于影响 LDA 困惑度的因素为迭代次数与

主题数[59]，所以通过对迭代次数和主题数量、困惑度进行计算，获得迭代次数、主题数、困惑度关系，结果如表 4.3 所示，同时为了能够更加清楚明了的观察迭代数与困惑度、主题数与困惑度之间的关系，分别绘制迭代数-困惑度、主题数-困惑度三维折线图。由于主题数为 1-5 时困惑度大且呈下降趋势，因此分析价值小，另外主题数为 16-19 时困惑度变化幅度大，因此参考价值小，所以取主题数为 6-15，绘制迭代数-困惑度、主题数-困惑度图，结果如图 4.2、4.3 所示。

表 4.3 迭代次数、主题数、困惑度

迭代次数 / 主题数	50	100	150	200	250	300	350
1	304.84	309.42	298.07	299.87	312.98	350	296.93
2	315.55	285.14	283.15	300.86	318.18	294.84	286.49
3	279	272.84	279.7	304.32	278.89	270.43	268.22
4	284.94	264.79	257.87	266.07	256.74	289.19	292.7
5	258.81	257.99	255.75	247.3	282.91	254.5	251.13
6	241.7	248.19	233.98	250.37	273.85	241.71	255.9
7	266.12	251.45	249.43	232.26	227.95	243.55	257.42
8	237.09	235.26	237.25	239.03	238.07	227.77	222.66
9	228.81	219.92	217.78	244.91	223.69	255.54	219.03
10	256.31	237.77	227.27	223.31	254.67	230.52	227.76
11	228.15	232.43	222.42	206.86	216.66	205.83	238.56
12	220.9	210.7	216.54	211.64	220.99	213.98	226.21
13	220.52	218.91	220.29	223.51	227.88	226.2	197.99
14	186.76	227.66	223.58	212.95	210.48	197.89	202.35
15	222.87	226.65	208.05	196.1	189.67	195.41	221.12
16	197.22	210.38	200.13	205.77	220.92	201.09	185.65
17	175.26	190.74	210.63	209.35	180.56	163.03	215.29
18	304.84	309.42	298.07	299.87	312.98	350	296.93
19	315.55	285.14	283.15	300.86	318.18	294.84	286.49

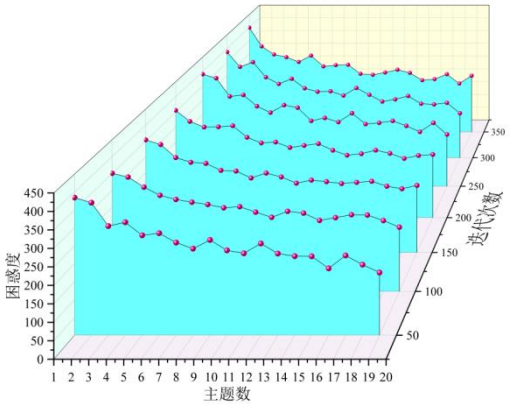


图 4.2　迭代数-困惑度

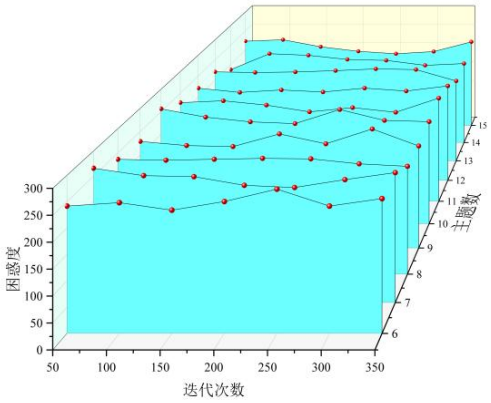


图 4.3　主题数-困惑度

由图 4.2、4.3 可知，模型的困惑度受迭代次数的影响较小，当迭代数在 100-200 时困惑度趋于收敛。主题数较少时，困惑度波动较大，主题数为 11 时，模型的困惑度收敛，并随着主题数量的增多困惑度呈整体下降趋势。

（3）主题提取结果

一个主题为在同一编号下包含了属于该主题的从高到低的 top n 个词概率与词的组合，通常有几个编号则有几个主题。取 5 个主题编号下词概率分布排名前 10 的词，结果如表 4.4 所示：

表 4.4　主题结果

主题结果
(0,'0.536*"超好" + 0.153*"电池" + 0.123*"续航" + 0.024*"挺大" + 0.000*"这家" + 0.000*"没人会" + 0.000*"后边" + 0.000*"原以为" + 0.000*"生态系统" + 0.000*"中要"')
(1, '0.446*"总体" + 0.297*"好看" + 0.128*"手机" + 0.067*"不错" + 0.018*"效果" + 0.014*"清晰" + 0.000*"价格便宜" + 0.000*"补充" + 0.000*"原以为" + 0.000*"冬天"')
(2, '0.447*"真的" + 0.321*"感觉" + 0.133*"很快" + 0.073*"棒" + 0.006*"处理器" + 0.000*"原以为" + 0.000*"生态系统" + 0.000*"这家" + 0.000*"一步" + 0.000*"后边"')
(3, '0.384*"整体" + 0.359*"手感" + 0.115*"算" + 0.081*"手机" + 0.024*"外表" + 0.000*"刷新率" + 0.000*"原以为" + 0.000*"后边" + 0.000*"传输" + 0.000*"这家"')
(4, '0.526*"不错" + 0.258*"手机" + 0.160*"挺" + 0.042*"一款" + 0.007*"很大" + 0.002*"显示" + 0.000*"生态系统" + 0.000*"原以为" + 0.000*"这家" + 0.000*"补充"')

从主题强度排名前五的主题提取结果可以看出，提出的主题中保留了特征和情感词、情感修饰词，与关键短语结构一致，为下文进行情感分析奠定了良好的基础。

4.2.3 模型检验与对比分析

为验证 BERT-LDA 模型的优越性，在 test 文本上对 BERT-LDA 模型、BERTopic-LDA、LDA 模型做性能评价，评价标准为主题困惑度，主题困惑度越低，模型的性能越好。

将验证集分别输入 BERTopic-LDA、LDA 模型，获得模型主题数、迭代次数、困惑度关系表，结果如表 4.5 和 4.6 所示。为方便观测，分别绘制相应的迭代数—困惑度、主题数—困惑度三维折线图，结果如图 4.4、4.5 和 4.6、4.7 所示。

表 4.5 BERTopic-LDA 模型主题数、迭代次数、困惑度

迭代数 / 主题数	50	100	250	200	250
1	501.76	494.23	500.37	486.7	512.91
2	439.99	436.83	452.1	455.31	457.99
3	409.4	405.32	416.55	396.08	440.2
4	404.71	395.12	402.05	420.79	420.26
5	392.85	396.45	374.43	395.25	401.4
6	383.46	381.21	384.41	372.57	398.29
7	371.08	377.72	384.69	385.67	378.77
8	357.67	353.22	356.33	361.79	384.54
9	347.35	355.2	339.1	356.34	346.49
10	355.2	345.33	367.35	364.68	360.77
11	358.38	345.81	346.78	345.69	349.09
12	352.48	351.93	344.61	334.22	350.26
13	326.24	339.6	346.73	352.32	342.3
14	336.28	336.55	339.02	347.79	312.68
15	332.15	332.16	328.22	328.1	347.01
16	328.65	334.93	342.55	322.65	346.13
17	323.57	338.45	314.03	319.97	328.81
18	320.07	307.3	318.2	323.29	333.76
19	307.8	300.82	308.66	316.36	302.0

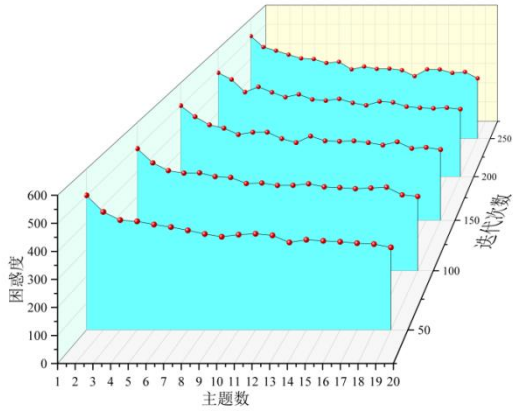


图 4.4　BERTopic-LDA 迭代数-困惑度

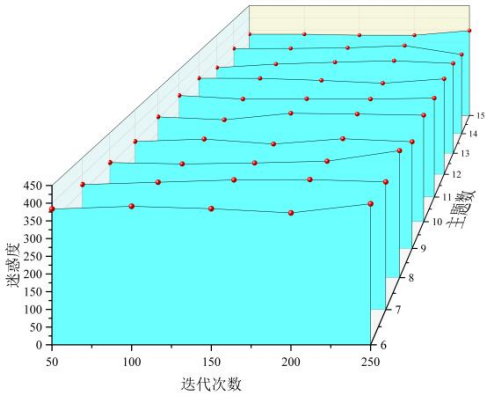


图 4.5　BERTopic-LDA 主题数-困惑度

表 4.6　LDA 模型主题数、迭代次数、困惑度

迭代数 / 主题数	50	100	150	200	250
1	1081.49	1100.42	1088.03	1076.01	1117.87
2	941.16	935.06	939.77	932.76	937.55
3	780.35	777.79	820.89	818.99	779.94
4	701.23	732.54	722.78	715.94	697.66
5	646.39	622.29	654.92	609.51	633.93
6	590.02	578.73	579.15	597.92	571.84
7	519.01	524.86	551.81	539.52	547.56
8	506.16	505.81	497.63	504.39	505.54
9	455.25	456.16	467.54	479.41	454.92
10	429.06	437.14	434.57	423.66	450.42
11	416.73	422.9	412.21	431.01	424.01
12	426.86	410.6	411.8	421.83	412.36
13	385.53	398.63	393.04	402.34	415.78
14	370.82	403.1	396.1	370.58	397.07
15	377.4	369.04	377.79	369.85	377.61
16	367.47	374.01	359.29	340.9	365.96
17	337.83	344.75	339.42	338.73	351.41
18	317.2	323.21	323.28	333.0	332.46
19	328.85	327.51	327.56	305.73	334.8

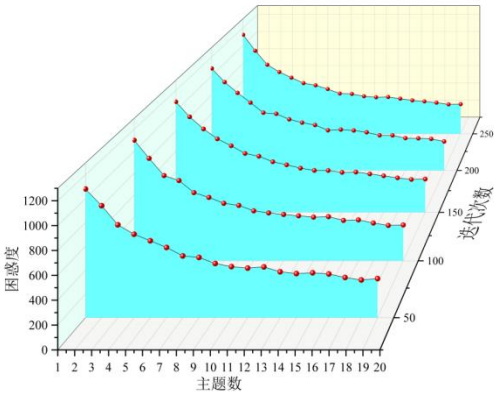


图 4.6　LDA 迭代数-困惑度

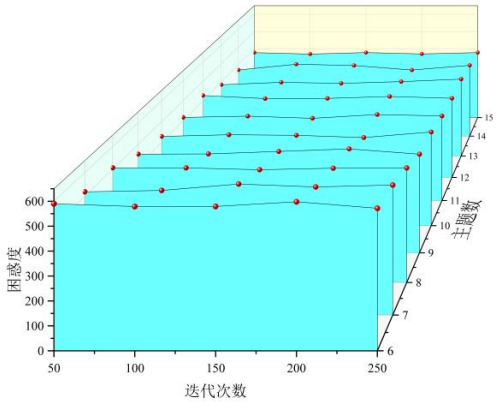


图 4.7　LDA 主题数-困惑度

通过图 4.4 和 4.6 可知，BERTopic-LDA 模型、LDA 模型的困惑度受迭代次数的影响较小。迭代次数在 100-200 时，困惑度趋于收敛。由于 BERT-LDA 模型、BERTopic-LDA 模型、LDA 模型的困惑度受迭代次数的影响较小，当迭代数在 100-200 时困惑度趋于收敛，为节约运算次数，取迭代次数为 100，比较 BERT-LDA 模型、BERTopic-LDA、LDA 模型的主题数与困惑度的关系，并画出主题数与困惑度关系图，结果如图 4.8 所示。

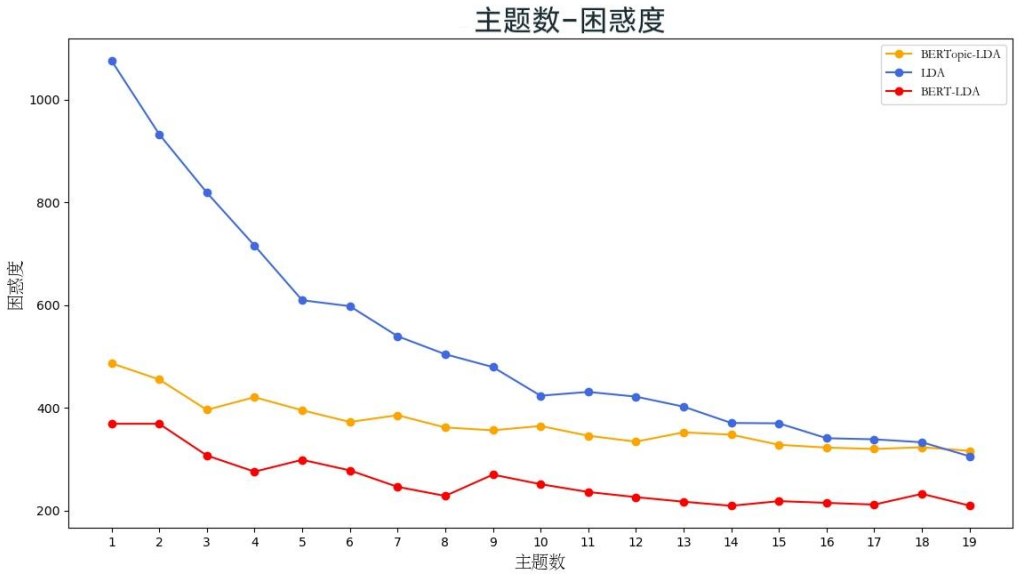


图 4.8　主题数与困惑度的关系（迭代次数为 100）

由图 4.8 可知，BERT-LDA 模型的困惑度值始终小于 BERTopic-LDA 和 LDA 模型的困惑度，说明 BERT-LDA 模型优于 BERTopic-LDA 模型和 LDA 模型。由于主题模型的困惑度随主题数的增大逐渐减小，所以当困惑度在主题数更小时收敛，说明模型效果更好。BERT-LDA 模型在主题数为 11 时困惑度趋于收敛，BERTopic-LDA 模型在主题数为 15 时困惑度趋于收敛，LDA 模型在主题数为 16 时困惑度趋于收敛，说明 BERT-LDA 模型优于 BERTopic-LDA 模型和 LDA 模型。综上所属，本文构建的 BERT-LDA 模型性能更优越。

4.3 本章小结

本章介绍了 LDA 算法的改进模型 BERT-LDA 模型的改进思路与构建过程，并使用训练语料进行模型训练，获得训练后的模型参数，并与 BERTopic-LDA 模型、LDA 模型进行对比分析，证明 BERT-LDA 模型性能优于 BERTopic-LDA 模型和 LDA 模型。

5 基于细粒度词典的情感分析模型构建

基于词典的情感分析的关键是构建全面、正确的词典，所以本文先基于大连理工大学情感词汇本体库构建细粒度情感词典，并进行词典的扩充，在此基础上构建情感分类与情感值计算的模型，从而构建基于细粒度词典的情感分析模型。

5.1 构建流程

基于细粒度词典的情感分析模型的工作流程如图 5.1 所示，可分为三个步骤，具体如下：

步骤 1：确定训练文本。为减少无效信息、提高文本的规范化，将手机关键短语集合作为训练文本。

步骤 2：细粒度情感词典构建与扩展。基于训练文本计算手机关键短语中的情感与大连理工大学情感词汇本体库（DUTIR）中情感词的相似度，将相似度大于等于 0.8 的词作为基础细粒度情感词典，同时按照手机特征词典进行分类，并进行第 8 类“疑”情感词、修饰词扩展。

步骤 3：基于基础情感词典计算文本情感分类与得分。

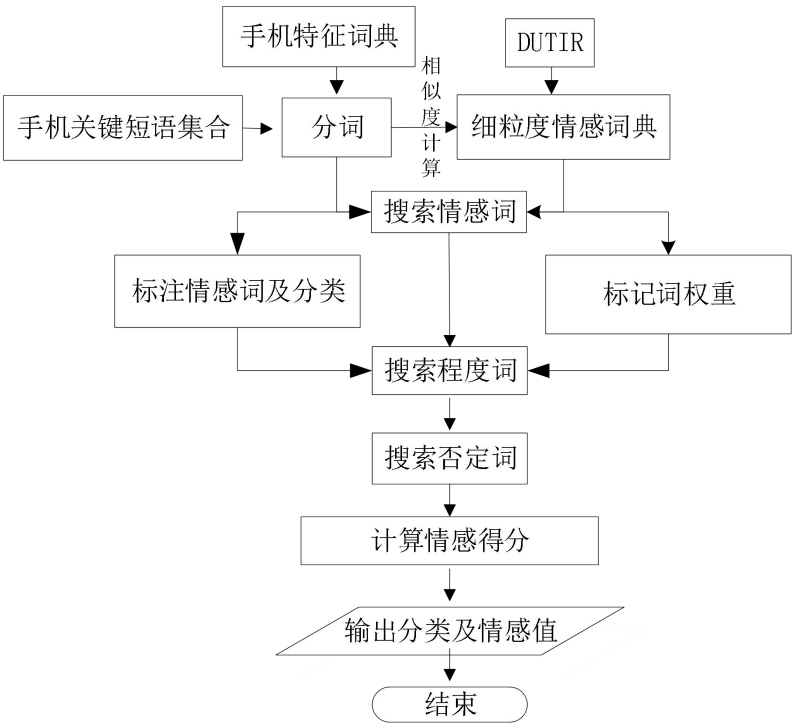


图 5.1　基于细粒度词典的情感分析模型流程

5.2 细粒度情感词典的构建与扩展

本文参考大连理工大学情感词汇本体库（DUTIR）的“乐、好、怒、哀、惧、恶、惊”7 类情感，以本文所建立的手机关键短语集合和手机特征词典为基础，构建基于手机产品评价特征词的细粒度基础情感词典，增加敦欣卉等人[1]构建的表示疑惑的“疑”情感，扩充为一个手机领域的 8 类细粒度情感词典，同时对修饰词进行扩充。

5.2.1 细粒度情感词典的构建

目前在情感分析领域，并没有形成统一的情感词典，各学者往往都是自行构建情感词典进行情感分析。另外对于不同领域的情感词典，其情感词的构成往往具有差别，情感词典质量越高，情感分析的有效性也就越高，因此，基于训练文本计算手机关键短语中的情感与 DUTIR 中情感词的相似度，将相似度大于等于 0.8 的词作为基础细粒度情感词典，同时按照手机特征词典进行分类。

大连理工大学信息检索研究室整理的中文情感词汇本体库（DUTIR），其情感强度分为 1，3，5，7，9 五档，9 表示强度最大，1 为强度最小，每个词在每一类情感下都对应了一个极性，其中，0 代表中性，1 代表褒义，2 代表贬义，3 代表兼有褒贬两性，结果如表 5.1 所示。情感共分为 7 大类 21 小类，结果如表 5.2 所示。

表 5.1 情感词汇本体格式举例

词语	词性种类	词义数	词义序号	情感分类	强度	极性	辅助情感分类	强度	极性
发人深思	idiom	1	1	PH	1	1			
卖人情	verb	1	1	NN	2	2			
不干不净	idiom	1	1	NN	3	2			
惹事生非	idiom	1	1	NN	4	2			
盛誉	noun	1	1	PH	5	1	PA	1	1
沁人心肺	idiom	1	1	PA	6	1			
千姿百态	idiom	1	1	PH	7	1	PB	5	1
敲锣打鼓	idiom	1	1	PA	8	1			
立国安邦	idiom	1	1	PH	9	1			

表 5.2 情感分类

编号	情感大类	情感类	例词
1	乐	快乐	喜悦、欢喜、笑眯眯、欢天喜地、兴高采烈
2		安心	踏实、宽心、定心丸、问心无愧、服服帖帖
3	好	尊敬	恭敬、敬爱、毕恭毕敬、肃然起敬、另眼相看
4		赞扬	英俊、优秀、通情达理、实事求是、异彩纷呈
5		相信	信任、信赖、可靠、毋庸置疑、可以信赖
6		喜爱	倾慕、宝贝、一见钟情、爱不释手、梦寐以求
7		祝愿	渴望、保佑、福寿绵长、万寿无疆、心仪已久
8	怒	愤怒	气愤、恼火、大发雷霆、七窍生烟、令人生气
9	哀	悲伤	忧伤、悲苦、心如刀割、悲痛欲绝、不胜其苦
10		失望	憾事、绝望、灰心丧气、心灰意冷、不可企及
11		疚	内疚、忏悔、过意不去、问心有愧、高不成低不就
12		思(PF)	思念、相思、牵肠挂肚、朝思暮想、梦魂萦绕
13	惧	慌(NI)	慌张、心慌、不知所措、手忙脚乱、最后一根稻草
14		恐惧	胆怯、害怕、担惊受怕、胆颤心惊、七上八下
15		羞	害羞、害臊、面红耳赤、无地自容、黯然失色
16	恶	烦闷	憋闷、烦躁、心烦意乱、自寻烦恼、忍气吞声
17		憎恶	反感、可耻、恨之入骨、深恶痛绝、为所欲为
18		贬责	呆板、虚荣、杂乱无章、心狠手辣、手脚不干净
19		妒忌	眼红、吃醋、醋坛子、嫉贤妒能、犯红眼病
20		怀疑	多心、生疑、将信将疑、疑神疑鬼、不知所以
21	惊	惊奇	奇怪、奇迹、大吃一惊、瞠目结舌、叹为观止

参照大连理工大学信息检索研究室整理获得中文情感词汇本体库（DUTIR），计算手机关键短语中的情感与 DUTIR 中的情感词的相似度，将相似度大于等于 0.8 的词作为基础细粒度情感词典，并按照手机特征词典进行分类，构建基于手机属性特征的细粒度情感词典，共 1213 个情感词，以屏幕特征为例，基于屏幕特征的细粒度情感词典如表 5.3 所示。

表 5.3 细粒度词典

特征	情感大类	情感类	例词
屏幕	乐	快乐(PA)	超爽、舒服、顺畅、幸运、令人满意的、带劲
		赞扬(PH)	不错、美好、极好、流畅、快速、完美、细腻
	好	相信(PG)	可以信赖、可靠、合格、值得、十分重要
		喜爱(PB)	爱不释手、顶了、棒儿香，滑溜、喜欢、新鲜

续表 5.3 细粒度词典

特征	情感大类	情感类	例词
屏幕	怒	愤怒(NA)	气愤、可气、恼火、失落、窝火、令人生气
	哀	失望(NJ)	报废、不行、无语、心灰意冷、遗憾、绝望
	惧	恐惧(NC)	有毒、灾难性、担忧、岌岌可危、吓人、厉害
	恶	贬责(NN)	慢吞吞、未成熟、黄色、模模糊糊、不良、迟
	惊	惊奇(PC)	惊人、奇妙、good、晕、不得了、邪门儿

5.2.2 细粒度情感词典的扩展

提高基于词典的文本分析准确性的关键在于构建全面且准确的情感词典。由于不同领域的评论文本存在很大差异，因此情感词典也会具有一定的差别，所以为保证特定领域的评论文本情感分析的准确性需要根据具体领域的评论文本对情感词典进行扩展，从而提高情感分析的准确性。

（1）疑问词扩充。由于在手机评论中经常出现“行不行”、“神马情况”等疑问词，如“电池蓄电能力行不行？”、“拍照效果是神马情况？”，这些疑问词也涵盖了消费者对手机特征的疑问情感。因此，参照敦欣卉等人[1]构建的表示疑惑的“疑”情感，构建并扩展疑问词表，结果如表 5.4，构成 8 类情感。

表 5.4 疑问词词表

序号	疑问词	强度值	极性值
1	怎么样、怎么着、如何、为什么、难道、'呢？'、'吧？'、'啊？'、啥、为何、怎么办、哪些、问题、请问、为神马、神马情况、为啥、干嘛、能否、何时、求问	7	1
2	谁、何、什么、神马、几时、怎么、怎的、怎样、岂、何尝、吗、么、多大、有没有、会不会、好不好、能不能、可不可以、行不行	5	1
3	几、多少、怎、难怪、反倒、何必、你知道	3	1
4	居然、竟然、究竟	1	1

（2）修饰词扩充。在评论语句中，表达情绪的词除情感词还有修饰词，修饰词由程度词与否定词组成。为了能够更加细致的识别出情感及强度，对细粒度情感进行扩充。程度词对于情感强度具有加强或削弱的作用，否定词则会改变情

感词的极性，因此修饰词的加入能够帮助我们更细致的识别出情感及其强度。参照知网 HowNet 副词词典、敦欣卉等人[1]构建的修饰词典及手机评论词典中常用的修饰词，构建了 55 个程度副词, 47 个否定词，结果如表 5.5、5.6 所示。

表 5.5 程度词词表

序号	程度词	权值大小
1	极、极为、极其、透顶、极端、顶、顶尖、最、最为、绝顶、无比、无敌	2.00
2	多、很、非常、甚至、十分、十足、太、分外、特别、万分、尤其、真、格外、何等、过于、多么、更加、更为、更、超、越加、越发、愈加、愈、相当、不凡、好	1.50
3	颇、还、挺、比较、较、较为、较比	1.2
4	怪、有点、有点儿、有些、稍、稍稍、稍微、稍许、少许、略、略微	0.50

表 5.6 否定词词表

否定词
白白、甭、别、并非、不、不必、不曾、不可、不要、不用、不是、从不、从未、非、毫不、毫无、何必、何曾、何尝、何须、决不、绝不、绝非、绝无、没、没有、莫、难以、切勿、尚未、徒、徒然、枉、未、未必、未曾、未尝、未有、无、无从、无须、无庸、毋须、毋庸、勿、否

5.3 细粒度情感计算

通过情感强度值能够直观的对文本情感进行分析的结果做出判断。因为细粒度情感词典中既有情感分类又有情感强度，所以首先要判断出情感词的分类，然后再计算出具体的情感强度值。因此，本文参照文献[1]并对其进行改进，结果如公式（5.1）、（5.2）所示。

情感词的情感分类公式如公式（5.1）所示：

$$E_{pi}=\begin{Bmatrix} M,|\alpha_{k1}\beta_{k1}|<|\alpha_{k2}\beta_{k2}| \\ N,|\alpha_{k1}\beta_{k1}|\geq|\alpha_{k2}\beta_{k2}| \end{Bmatrix} \quad (5.1)$$

其中，M、N 分别为情感词的第一个情感分类和第二个情感分类，通过表 5.2 可以查到具体的情感类别，α为情感词的情感强度，β为情感词的极性值，

$\alpha_{k1}\beta_{k1}$ 、$\alpha_{k2}\beta_{k2}$分别为 M、N 情感分类对应的情感强度值。

情感词的强度值的计算公式如公式（5.2）所示：

$$E_j = \sum_{i=1}^{n}(-1)^{o_i}\alpha_i E_{pi} m \quad (5.2)$$

其中，n 为情感词的数量，E_{pi}为所属情感分类的情感强度，o_i为短句中包含否定词的数量，α_i为短句中程度副词的强度，m 为组合的权值[1]，如表 5.7 所示。

表 5.7 组合权值表

序号	类型	示例	权值
1	情感词	流畅	1
2	否定词+情感词	不 流畅	1
3	程度副词+情感词	太 流畅	1
4	否定词+程度副词+情感词	不 太流畅	0.4
5	程度副词+否定词+情感词	太 不 流畅	1
6	否定词+否定词+情感词	没有 不 流畅	1

通过上述步骤获得的情感词典作为基准词，对于不包括在基准词当中的情感词采用语义相似计算的方法找到最为接近的基准词的对应情感分类和情感强度，从而使得情感分析更加完整准确。

5.4 本章小节

本章构建并拓展了细粒度情感词典，并对细粒度情感计算方法进行介绍，构建了基于细粒度词典的情感分析模型，为其应用奠定了基础。

6 细粒度情感分析模型在手机产品评价领域的应用

本章对构建的细粒度情感分析模型进行实际应用，首先对 BERT-LDA 模型输出的主题结果依据特征词进行分类，然后基于特征对情感词进行情感分类与情感值计算，确定该特征的细粒度情感分类及情感强度，从而确定消费者对不同特征的关注程度与评价倾向，明确消费者对手机特征的不同要求，最后依此为商家提供手机产品上新、销售的建议。细粒度情感分析工作流程图如图 6.1 所示。

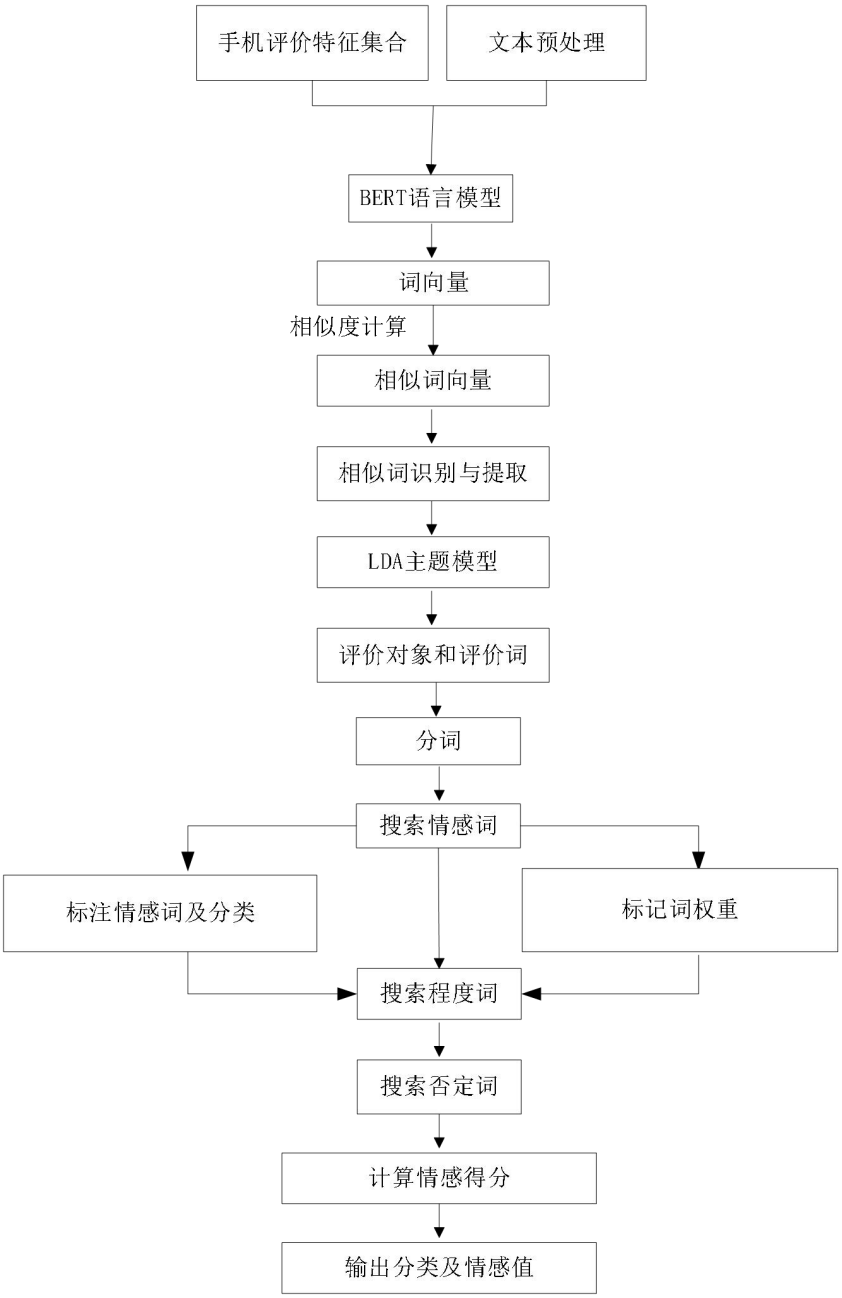


图 6.1 工作流程

6.1 评论文本获取

本文中的数据来自于思睿大数据平台，共下载某品牌手机 6008 条评论数据，所获取的信息包括产品信息、买家、评价内容、追评等字段内容。

6.2 文本预处理

对文本规范化、分词、去停用词处理。首先对评论文中无意义的文本进行清洗并剔除；然后使用 Scikit-Learns CountVectorizer 方法进行分词与去停用词，将评论文本划分为短句，构成短句集合，最大程度保留特征与情感词。

6.3 文本分析

本文以某品牌手机的评论文本为例，对其进行细粒度情感分析，主要包括对手机不同特征的关注度和关注靠前的特征细粒度情感，从而明确消费者对该品牌手机不同特征喜好的具体情况。

（1）主题提取。将预处理后的文本输入到 BERT-LDA 模型中，最终输出的主题如表 6.1 所示：

表 6.1 主题输出结果

主题结果
(0, '0.293*"东西" + 0.142*"超好" + 0.086*"够用" + 0.060*"挺舒服" + 0.033*"屏幕" + 0.027*"分辨率" + 0.010*"质量" + 0.006*"亮度" + 0.000*"传输" + 0.000*"胜价"')
(1, '0.458*"满意" + 0.336*"机身" + 0.085*"好看" + 0.035*"挺不错" + 0.033*"款" + 0.000*"崩" + 0.000*"传输" + 0.000*"价格便宜" + 0.000*"胜价" + 0.000*"冬天"')
(2，0.561*"不错" + 0.141*"质量" + 0.092*"棒" + 0.086*"还好" + 0.034*"颜色" + 0.021*"系统" + 0.012*"视频" + 0.000*"胜价" + 0.000*"传输" + 0.000*"越发"）
(4, '0.231*"很大" + 0.076*"续航" + 0.019*"电池" + 0.000*"时不时" + 0.000*"价格便宜" + 0.000*"胜价" + 0.000*"崩" + 0.000*"较卡顿" + 0.000*"越发" + 0.000*"冬天"')
(5, '0.368*"很快" + 0.325*"挺" + 0.113*"总体" + 0.075*"还行" + 0.061*"行" + 0.001*"清晰度" + 0.000*"较卡顿" + 0.000*"越发" + 0.000*"时不时" + 0.000*"崩"')
(6, '0.691*"舒服" + 0.039*"稍微" + 0.013*"屏" + 0.000*"时不时" + 0.000*"价格便宜" + 0.000*"胜价" + 0.000*"崩" + 0.000*"较卡顿" + 0.000*"越发" + 0.000*"冬天"')
(7, '0.490*"购买" + 0.173*"体验" + 0.171*"力" + 0.052*"看着" + 0.010*"不太" + 0.001*"运行速度" + 0.001*"挺大" + 0.000*"便重" + 0.000*"性价" + 0.000*"时不时"')
(8, '0.429*"点" + 0.195*"一款" + 0.029*"真不错" + 0.024*"内存" + 0.020*"细腻" + 0.009*"理想" + 0.000*"冬天" + 0.000*"性价" + 0.000*"价格便宜" + 0.000*"价廉物美"')
(9, '0.337*"真的" + 0.335*"感觉" + 0.079*"差评" + 0.068*"不错" + 0.051*"效果" + 0.042*"清晰" + 0.007*"显示" + 0.000*"胜价" + 0.000*"传输" + 0.000*"越发"')

为了解主题中所涉及的特征，根据构建的手机特征词典对评价特征进行提取，

并跟据与词典内相同或者相近的词判断主题的评价特征，提取结果如表 6.2 所示。

表 6.2 提取的评价特征输出结果

主题编号：0	评价特征：屏幕
主题编号：1	评价特征：外形
主题编号：2	评价特征：屏幕
主题编号：3	评价特征：外形
主题编号：4	评价特征：电池
主题编号：5	评价特征：摄像
主题编号：6	评价特征：屏幕
主题编号：7	评价特征：内存与处理
主题编号：8	评价特征：内存与处理
主题编号：9	评价特征：屏幕

由表 6.2 可以看出，主题编号 0、2、6、9 的评价特征为屏幕，主题编号 1、3 的评价特征为外形，主题编号 4 的评价特征为电池，主题编号 5 的评价特征为摄像，主题编号 7、8 的评价特征为内存与处理，可以看出对于该品牌手机消费者看重的特征为屏幕、外形、电池、摄像、内存与处理，想要进一步了解消费者对该品牌手机的这些关注度高的特征的具体评价需要进行细粒度情感分析。

（2）特征细粒度情感分析。为了解消费者对重视特征的具体评价，将屏幕、外形、电池、摄像、内存与处理 5 个特征的主题按照关键短语结构进行提取，情感特征强度与情感特征在顺序上对应，没有情感程度的对应位置用其权值“1”来表示，结果如表 6.3 所示。对同一特征结果进行合并，结果如表 6.4 所示。

表 6.3 基于关键短语结构的主题结果分析

主题标号-属性特征	情感特征	情感程度特征
0-屏幕	好、舒服	超、挺
1-外形	满意、好看、不错	1、1、挺
2-屏幕	不错、棒、还好	1、1、1
3-外形	特别	1
4-电池	大	很
5-摄像	快、行、行	很、还、1
6-屏幕	舒服	1
7-内存与处理	大	挺
8-内存与处理	不错、细腻、理想	真、1、1
9-屏幕	差评、不错	1、1

表 6.4 基于关键短语结构的特征结果分析

属性特征	情感特征	情感程度特征
屏幕	好、舒服、不错、棒、还好、舒服、差评、不错	超、挺、1、1、1、1、1、1
外形	满意、好看、不错、特别	1、1、挺、1
电池	大	很
摄像	快、行、行	很、还、1
内存与处理	大、不错、细腻、理想	挺、真、1、1

按照细粒度情感计算的步骤计算 5 个特征的情感分类与情感强度，对于情感词典中没有的情感词通过基于 BERT 进行相似度进行计算，例如“手机好”跟“手机不错”相似度在 0.998~1，细粒度情感分析结果如表 6.5 所示。

表 6.5 细粒度情感分析

属性特征	情感词类别	情感词	情感强度值
屏幕	好（PH）	好、不错、棒、好、不错	35.5
	乐（PA）	舒服	3.6
	恶（NN）	差评	3
外形	好（PB、PH）	满意、好看、不错、特别	24.6
电池	好（PH）	大	7.5
摄像	好（PH）	快、行、不错	8.1
内存与处理	好（PH）	大、不错、细腻、理想	23.6

由表 6.5 可知，屏幕特征的情感类别最多，有三个情感类别，说明了消费者对屏幕特征的情感最为丰富，其中“好”跟“乐”的情感强度都大于“恶”，且“好”的情感词最多，所以消费者整体上对屏幕特征的情感是偏爱的。外形、电池、摄像、内存与处理特征只有一个情感类别，都是“好”这个情感类别，其中外形特征的情感类别包括两个小类情感，比其他 3 个情感更为丰富，且情感强度值较其他 3 个特征最大，所以对外形这个特征消费者的满意度很高；电池、摄像、内存与处理特征只有一个情感类别、一个小类情感，情感值最大的是内存与处理特征，摄像特征次之，最小的是电池特征，所以，消费者对内存与处理、摄像、电池三个特征都是满意的，内存与处理特征满意度较高，摄影与电池特征的满意度稍小。

6.4 结果分析

在表 6.5 中的五个特征中都有的情感类别是“好”，屏幕特征中的“乐”同样也是代表满意的情感，由此可知消费者最为重视的屏幕、外形、电池、摄像、内存与处理五个特征的评价都是满意的，其满意程度由高到低依次为屏幕特征>外形特征>内存与处理特征>摄像特征>电池特征，可知该品牌手机的优势特征为屏幕与外形。手机屏幕的情感类别最为丰富，说明消费者对屏幕的需求层次多，是区分不同消费需求的重要依据。

6.5 对策建议

通过上文对某品牌手机评论文本的细粒度分析可知，虽然消费者对该品牌手机的整体认可度比较高，但是其对不同的特征的满意度之间存在很大差异，为满足消费者需求，提高产品好评率，实现店铺口碑提升、产品销量增加，从产品上新、销售以及售后全过程提出如下建议：

（1）选择上新的该品牌手机产品，其屏幕、外形、电池、摄像、内存与处理性能要优越。从细粒度分析结果可知，消费者最为重视的该品牌手机的特征为屏幕、外形、电池、摄像、内存与处理五个特征，因此在选择上架的该品牌手机时首先要对这五个产品特征的性能进行仔细查验，确保各方面的性能良好，从而达到留住老客户、吸引新客户的目的。

（2）产品详情介绍页面中着重突出屏幕与外形优势。由于该品牌手机的优势特征为屏幕和外形，对消费者的吸引力度最强，基于此，商家在上新该品牌手机时务必在详情页突出屏幕与外形的独特之处，吸引消费者目光，从而激发消费者的购买热情。

（3）优化产品标题，将屏幕的个性化特点信息加入其中。由于消费者对该品牌手机屏幕特征的情感评论最为丰富，且情感倾向为两级分化，因此，消费者对屏幕的要求具有极强个性化的需求，为了满足消费者的不同需求，可按照手机屏幕的具体属性特征进行分别销售，将屏幕的个性化特点信息融入到产品标题中，如“折叠屏”、“曲屏”等。通过有针对性的产品标题，帮助消费者快速找到自己心仪的手机，提高顾客的满意度，达成销售目的，同时也可以避免因为屏幕性能不符合消费者需求带来的差评。

（4）加强对客服的培训。网店的客服质量对店铺的购买转化率十分重要，

因此加强对客服的培训，提高客服的专业性与服务水平是十分必要的。网店客服根据消费者消费的阶段可以分为售前、售后两种，针对不同阶段的客服培训的侧重点也要加以区别。针对售前客服，培训的侧重点在于对产品详情、优势特征的把握，从而引导消费者找到自己需求的产品，促进销售的达成；对于售后客服培训的侧重点是在于解决顾客销售过程中的各种异常问题，平复消费者的不满情绪，减少产品的差评数量，维系与消费者之间情感，维护店铺的口碑。

6.6 本章小结

本章以某品牌手机评论文本为例，应用了本文所构建的细粒度情感分析模型，得出消费者最为重视的某品牌手机的特征为屏幕、外形、电池、摄像、内存与处理五个特征，并基于对五个特征的细粒度情感分析结果为商家提供上新、销售、售后建议。

7 总结与展望

7.1 总结

本文从产品特征与情感分类两个角度对评论文本进行了细粒度情感分析。其中，产品特征角度是对产品的分析从整体评价到具体特征，明确消费者对产品具体特征的重视程度；情感分类角度则是将情感词典进一步进行细分，不再沿袭传统的情感极性二分类的方法，而是将情感分类划分为多个分类，并计算情感词的情感强度，明确消费者情感的细分倾向。将产品特征与情感细粒度结合，能够更加清楚的明确消费者对产品特征的具体要求，从而更好地帮助商家了解消费者的详细需求，为实现店铺口碑提升、产品销量增加提供针对性建议。

本文的主要研究成果：

（1）构建了手机关键短语集合。以手机评论文本为例，扩展了手机领域的特征词典。在拓展特征词典的基础上，依据关键短语结构，构建了手机关键短语集合。同时包含特征词与情感词的关键短语，为手机特征和情感词的高精度提取创造了前提条件。

（2）构建了 BERT-LDA 模型。基于手机关键短语集合，通过调整 BERT 模型中相似度值域，提取相似短语，并将提取结果输入到 LDA 模型，获得特征与情感词。高质量的相似短语在降低 LDA 模型困惑度上发挥了极大作用，也有利于提高 LDA 模型主题提取的准确率。

（3）构建了基于手机特征词的细粒度情感词典。为实现对情感词的分类与情感强度的计算，借助于大连理工大学信息检索研究室整理的中文情感词汇本体库（DUTIR），以手机关键短语集合和手机特征词典为基础，构建基于手机产品评价特征词的细粒度情感词典，增加了代表“疑惑”的“疑”情感，构成 8 类情感词，同时进行修饰词与否定词的扩充。

（4）以某品牌的手机评论文本为例，对本文所构建的细粒度情感分析模型进行了具体应用。根据分析结果，为实现店铺口碑提升、产品销量增加提供对策建议。

7.2 展望

本文基于 BERT-LDA 模型进行细粒度分析，虽然在一定程度上降低了模型

的困惑度，并在手机评论文本上进行适用，但仍然存在一些不足之处。对此，针本文中存在的不足和局限，提出以下几方面需要在未来进行进一步研究：

（1）电子商务正如火如荼的发展着，消费者的评论习惯和用语也随着时代的变化而变化。基于情感词典的情感分析需要不断完善情感词典，高质量的情感词典是情感分析准确性的保障，因此细粒度情感词典的研究也要不断的丰富与完善。

（2）信息化进程不断发展，5G 时代也迅猛发展，消费者对手机的产品特征提出了更多的要求，手机功能也进一步发展更新，因此评论文本中也会出现很多以往未出现的新内容，对手机特征的细粒度情感分析也有待进一步研究。

参考文献

[1] 敦欣卉,张云秋,杨铠西.基于微博的细粒度情感分析[J].数据分析与知识发现,2017,1(07):61-72.

[2] 薛福亮,刘丽芳.基于 TF-IDF 和情感强度的细粒度情感分析——餐饮评论为例[J].信息系统工程,2020(03):83-84+86.

[3] Li Ji, Lowe Dan, Wayment Luke, Huang Qingrong. Text mining datasets of β-hydroxybutyrate (BHB) supplement products' consumer online reviews.[J].Data in brief, 2020, 30(06):23-30.

[4] 胡飞菊. 基于文本挖掘的民宿评论细粒度情感分析[D].江西农业大学,2019.

[5] 陈炳丰,郝志峰,蔡瑞初,温雯,王丽娟,黄浩,蔡晓凤.面向汽车评论的细粒度情感分析方法研究[J].广东工业大学学报,2017,34(03):8-14.

[6] TITOV I， MCDONALD R． Modeling online reviews with multigrain topic models［C］//LIZ． Proceeding of the 17th international con-ference on World Wide Web.New York: ACM，2008(04): 111－120．

[7] Emmanuel Awuni Kolog,Samuel Nii Odoi Devine,Kwame Ansong-Gyimah,Richard Osei Agjei. Fine-grained affect detection in learners' generated content using machine learning[J]. Education and Information Technologies,2019,24(06):3767-3783.

[8] 蔡庆平,马海群.基于 Word2Vec 和 CNN 的产品评论细粒度情感分析模型[J].图书情报工作,2020,64(06):49-58.

[9] Stefanos Angelidis,Mirella Lapata. Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis[J].Transactions of the Association of Computational Linguistics,2018(06):17-31.

[10] 贾川,方睿,浦东,康刚.基于循环实体网络的细粒度情感分析[J].中文信息学报,2019,33(09):123-128.

[11] Jin Zheng,Limin Zheng,Lu Yang. Research and Analysis in Fine-grained Sentiment of Film Reviews Based on Deep Learning[J].Journal of Physics: Conference Series,2019,1237(02):149-152.

[12]张津. 基于学生评教文本的细粒度情感分析研究[D]. 华南理工大学, 2020.

[13]薛福亮, 刘丽芳. 一种基于 CRF 与 ATAE-LSTM 的细粒度情感分析方法[J]. 数据分析与知识发现, 2020, 4(Z1):207-213.

[14]D Tang，B Qin，X Feng，et al. Effective LSTMs fortarget-dependent sentiment classification[J].Coling ,2016,9(09):11-17.

[15]C Sindhu,Vadivu G. Fine grained sentiment polarity classification using augmented knowledge sequence-attention mechanism[J].Microprocessors and Microsystems,2021,81(03):37-42.

[16]李鸿宇. 基于图卷积记忆网络的细粒度情感分析研究[D].辽宁工程技术大学,2020.

[17]马攀. 面向科技资源文本评论的细粒度情感分析方法研究[D].电子科技大学,2020.

[18]刘测. 产品评论的细粒度情感分析研究[D].西安石油大学,2019.

[19]程艳,孙欢,陈豪迈,李猛,蔡盈盈,蔡壮.融合卷积神经网络与双向 GRU 的文本情感分析胶囊模型[J].中文信息学报,2021,35(05):118-129.

[20]沈卓,李艳.基于 PreLM-FT 细粒度情感分析的餐饮业用户评论挖掘[J].数据分析与知识发现,2020,4(04):63-71.

[21]刘亦轩. 商品评论细粒度情感分析系统设计与实现[D].西南大学,2020.

[22]Munikar M, Shakya S, Shrestha A. Fine-grained sentiment classification using BERT[C]//2019 Artificial Intelligence for Transforming Business and Society (AITB). IEEE, 2019, (01): 1-5.

[23]Yejin Tan,Tan Yejin,Guo Wangshu,He Jiawei,Liu Jian,Xian Ming. A Fine-grained Sentiment Analysis Method Based on Dependency Tree and Graph Attention Network[J]. Journal of Physics: Conference Series,2020,1651(1):21-23.

[24]Balikas G, Moura S, Amini M R. Multitask learning for fine-grained twitter sentiment analysis[C]//Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. 2017,(01): 1005-1008.

[25]侯艳辉,董慧芳,郝敏,崔雪莲.基于本体特征的影评细粒度情感分类[J].计算机应用,2020,40(04):1074-1078.

[26]曹雪.多粒度的汉语情感极性分类方法研究[D].黑龙江大学,2020.

[27]万岩,杜振中.融合情感词典和语义规则的微博评论细粒度情感分析[J].情报探索,2020(11):34-41.

[28]曾明睿,袁梦奇,邵曦,鲍秉坤,徐常胜.文本特征提取的研究进展[J].南京信息工程大学学报(自然科学版),2019,11(06):706-715.

[29]Hu M,Liul B. Mining and summarizing customer reviews[C].Proceedings of the tenthACM SIGKDD international conference on Knowledge discovery and data mining.ACM,2004(01):168-177.

[30]Bing L, Wong T L, Lam W. Unsupervised extraction of popular product attributes from e-commerce web sites by considering customer reviews[J]. ACM Transactions on Internet Technology (TOIT), 2016, 16(02): 1-17.

[31]Li Z, Shi K, Dey N, et al. Rule-based back propagation neural networks for various precision rough set presented KANSEI knowledge prediction: a case study on shoe product form features extraction[J]. Neural Computing and Applications, 2017, 28(03): 613-630.

[32]Wan C, Peng Y, Xiao K, et al. An association-constrained LDA model for joint extraction of product aspects and opinions[J]. Information Sciences, 2020, 519(01): 243-259.

[33]Belem F M, Heringer A G, Almeida J M, et al. Exploiting syntactic and neighbourhood attributes to address cold start in tag recommendation[J]. Information Processing & Management, 2019, 56(03): 771-790.

[34]Liao J, Wang S, Li D, et al. FREERL: Fusion relation embedded representation learning framework for aspect extraction[J]. Knowledge-Based Systems, 2017, 135(01): 9-17.

[35]李伟卿,王伟军.基于大规模评论数据的产品特征词典构建方法研究[J].数据分析与知识发现,2018,2(01):41-50.

[36]余琴琴,彭敦陆,刘丛.大规模词序列中基于频繁词集的特征短语抽取模型[J].小型微型计算机系统,2018,39(05):1027-1032.

[37]王瑞,龙华,邵玉斌,杜庆治.基于 Labeled-LDA 模型的文本特征提取方法[J].电子测量技术,2020,43(01):141-146.

[38]赵勤鲁,蔡晓东,李波,吕璐.基于 LSTM-Attention 神经网络的文本特征提取方法[J].现代电子技术,2018,41(08):167-170.

[39]徐冠华.基于 Spark 的文本特征提取方法研究[D].曲阜师范大学,2018.

[40]周源,刘怀兰,杜朋朋,廖岭.基于改进 TF-IDF 特征提取的文本分类模型研究[J].情报科学,2017,35(05):111-118.

[41]陈可嘉,骆佳艺.中文网络评论的隐式产品特征提取方法研究[J].福州大学学报(哲学社会科学版),2020,34(01):59-65.

[42]周诗嘉.基于隐性特征提取的产品评论挖掘[D].西南财经大学,2019.

[43]陶娅芝.基于隐式产品特征的网络商品评论情感分析研究[D].重庆邮电大学,2017.

[44]Zou Y, Gu J, Fu H. Medical entity and attributes extraction system based on relation annotation[J]. Wuhan University Journal of Natural Sciences, 2016, 21(02): 145-150.

[45]Lamy J B, Soualmia L F. Formalization of the semantics of iconic languages: An ontology-based method and four semantic-powered applications[J]. Knowledge-Based Systems, 2017, 135(01): 159-179.

[46]Chen G, Wang C, Zhang M, et al. How “small” reflects “large”?—Representative information measurement and extraction[J]. Information Sciences, 2018, 460(01): 519-540.

[47]纪雪,高琦,李先飞,高菲.考虑产品属性层次性的评论挖掘及需求获取方法[J].计算机集成制造系统,2020,26(03):747-759.

[48]黄磊. 基于电商网站商品评论的商品属性提取及其情感的可视化表示[D].北京邮电大学,2017.

[49]李可悦,陈轶,牛少彰.基于 BERT 的社交电商文本分类算法[J].计算机科学,2021,48(02):87-92.

[50]王涛,李明.基于 LDA 模型与语义网络对评论文本挖掘研究[J].重庆工商大学学报(自然科学版),2019,36(04):9-16.

[51]贺珂. 基于关联规则的用户产品属性偏好变化挖掘[D].山东大学,2019.

[52]孙琳. 在线评论中产品属性识别及其应用研究[D].武汉理工大学,2018.

[53]张艳丰,李贺,彭丽徽.基于模糊情感计算的商品在线评论用户品牌转换意向研究[J].现代图书情报技术,2016(05):64-71.

[54]刘华.基于关键短语的文本内容标引研究[D].北京语言大学,2005.

[55]Danilevsky M, Wang C, Desai N, et al. Automatic construction and ranking of topical keyphrases on collections of short documents[C]//Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2014,1(01): 398-406.

[56]刘华.基于关键短语的文本分类研究[J].中文信息学报,2007(04):34-41.

[57]关鹏,王曰芬.科技情报分析中 LDA 主题模型最优主题数确定方法研究[J].现代图书情报技术,2016(09):42-50.

[58]Zhestiankin B, Ponomareva M. Zhestyatsky at SemEval-2021 Task 2: ReLU over Cosine Similarity for BERT Fine-tuning[J].Computer Science, 2021.4(13):163-168

[59]Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(01): 993-1022.

致谢

转眼间，三年的研究生学习生活临近尾声，刚到学校的场景依然历历在目，犹如昨日，万般的不舍都汇聚成短短的致谢。

在此我要特别感谢我的导师王玉珍教授，她通才硕学，恪尽职守。在学业上，王老师诲人不倦，要求严格；在生活上，给予我极大的关怀与帮助，使我可以顺利度过每一个关口。

此外，要感谢父母对我的付出，在父母的支持下我才能够心无旁骛的完成学业；其次，要感谢信工学院的领导、老师、同学、舍友们的陪伴，让我拥有温馨、快乐的校园生活；再次，要感谢男朋友的鼓励，让我有面对困难的勇气与力量。

最后，由衷感谢审阅论文的每位老师！

攻读硕士学位期间发表的论文及科研情况

[1]丁申宇,王玉珍,秦精俏.乡村旅游电商发展的影响因素研究——基于SWOT-AHP 模型[J].洛阳师范学院学报,2021,40(08):24-27+38.

[2]王玉珍,丁申宇.K-means 算法在农资网站客户管理中的应用[J].枣庄学院学报,2020,37(05):45-51.

[3]Y. Wang and S. Ding, "Research on Consuming Behavior Based on User Search Data : A Case of Xiaomi Mobile Phone," 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), 2020,82 (10):391-394.

[4]李佳儒,王玉珍,丁申宇.在线评论情感分析的影院推荐[J].宁德师范学院学报(自然科学版),2020,32(03):253-258.

[5]李佳儒,王玉珍,丁申宇.基于逻辑回归的在线评论情感分类方法研究[J].东莞理工学院学报,2020,27(05):50-54.

[6]参编清华大学出版社出版的《电子商务概论》，参编共计 8 万字.

[7]参与兰州财经大学项目：甘肃省农产品电子商务发展模式与对策研究；项目编号：Lzu fe2018B-04.

[8]参与兰州财经大学项目：甘肃省县域电商发展问题研究；项目编号：2019SL05.