浙江万里学院

本科毕业设计(论文)

(2021届)

论文题目 基于机器学习的电商评论情感分析方法研究

(英文) Research on sentiment analysis of ecommerce Reviews based on machine learning

所在学院 大数据与软件工程学院

专业班级 计算机科学与技术 213

学生姓名 谯奕 学号 2021016707

指导教师 毕春跃 职称 教授

指导教师 职称

完成日期 年 月 日

基于机器学习的电商评论情感分析方法研究

谯奕

（浙江万里学院大数据与软件工程学院计算机213班）

2025年4月

诚信声明

本人声明：

所呈交的毕业设计成果和论文是本人经过近四年的基础课程与专业课程学习的基础上，在指导教师的指导下，经过几个月集中的学习、实践和努力工作所取得的。除了论文中特别加以标注和致谢中指出的地方外，毕业设计成果和论文中不包含他人已经发表或撰写过的研究成果。毕业设计成果和论文中的所有内容均真实、可信。

学生签名：（签名拍照插入此处）

签名日期： 年 月 日

摘 要

随着电子商务平台的快速发展，网上在线购物已经成为用户主流购买方式，用户评论数据可以反映出消费者对产品的情感和产品口碑。基于机器学习的情感分析方法能够有效缝隙处大量评论中的情感倾向，对商家优化服务和用户决方面策具有重要价值。因此，本论文以电商评论文本作为数据集，使用人工智能领域的关键技术，对评论文本做情感分析，通过分析可以清洗明了的了解产品情感倾向，是一个具有社会意义和挑战性的研究课题。

随着国内外研究进展迅速，自然语言处理技术在情感分析方向上也取得诸多成就，情感分类技术分为几个大类，分别是基于规则和知识库的方法，基于机器学习的方法，图像与情感细粒度分析的融合和基于情感词典的方法。基于情感词典的方法需要大量的文本来训练，非常消耗资源成本因此研究人员在这个方面进展缓慢，在基于机器学习的情感分析方法中，随机森林，向量机和决策树等经典算法虽然应用广泛，但最新研究也不乏强劲的模型，深度神经网络（DNN）和卷积神经网络（CNN）等更加复杂的模型逐渐成为研究的焦点。图像与情感细粒度分析的融合也是未来一大趋势，许多研究将图像和文本信息结合进行情感分析，细粒度方面通过基于方面（aspect-based）的情感分析来把文本精细分为多个属性来强化情感效果，也是非常厉害的研究方式。本论文选择了在机器学习方面对文本进行分析研究，并设计一个网页前端可视化模型训练成果。

关键词：电商评论；情感分析；机器学习；深度学习；BERT模型

Abstract

With the rapid development of e-commerce platforms, online shopping has become the mainstream purchasing method for users. User review data can reflect consumers' emotions towards products and product reputation. The sentiment analysis method based on machine learning can effectively extract the sentiment tendencies from a large number of reviews, which is of great value for merchants to optimize services and for users to make decisions. Therefore, this paper takes e-commerce review texts as the dataset and uses key technologies in the field of artificial intelligence to conduct sentiment analysis on the review texts. Through the analysis, the sentiment tendencies of products can be clearly understood, which is a research topic with social significance and challenges.

With the rapid progress of research at home and abroad, natural language processing technology has also achieved many achievements in the direction of sentiment analysis. Sentiment classification techniques are divided into several major categories, including rule-based and knowledge-based methods, machine learning-based methods, the fusion of image and fine-grained sentiment analysis, and methods based on sentiment lexicons. The method based on sentiment lexicons requires a large amount of text for training and is very resource-consuming, so researchers have made slow progress in this aspect. Among the machine learning-based sentiment analysis methods, although classic algorithms such as random forests, support vector machines, and decision trees are widely used, there are also strong models in the latest research. More complex models such as deep neural networks (DNN) and convolutional neural networks (CNN) have gradually become the focus of research. The fusion of image and fine-grained sentiment analysis is also a major trend in the future. Many studies combine image and text information for sentiment analysis. In the fine-grained aspect, aspect-based sentiment analysis is used to finely divide the text into multiple attributes to enhance the sentiment effect, which is also a very powerful research approach. This paper selects machine learning-based text analysis for research and designs a web front-end visualization model for training results.

Key words: ecommerce reviews; Sentiment analysis; Machine learning; Deep learning; BERT model

目 录

目录

[本科毕业设计(论文) 1](#_Toc2013432657)

[论文题目 基于机器学习的电商评论情感分析方法研究 1](#_Toc1862840420)

[基于机器学习的电商评论情感分析方法研究 2](#_Toc594849327)

[摘 要 4](#_Toc1096262104)

[Abstract 5](#_Toc1614974315)

[1. 绪论 7](#_Toc827497772)

[1.1. 研究背景 7](#_Toc650956032)

[1.2. 研究现状 7](#_Toc1336332006)

[1.2.1. 国内研究 7](#_Toc1348044516)

[1.2.2. 国外研究 10](#_Toc631704562)

[1.3. 主要工作 11](#_Toc2046906413)

[2. 关键技术简介 12](#_Toc1815541998)

[2.1.1. Word2Vec 12](#_Toc219220163)

[2.1.2. 分类算法 12](#_Toc1498824936)

[3. 系统分析 13](#_Toc767520042)

[3.1. 可行性分析 13](#_Toc1922562012)

[3.1.1. 技术可行性 14](#_Toc1460782922)

[3.1.2. 经济可行性 14](#_Toc1345517550)

[3.1.3. 操作可行性 15](#_Toc1110659940)

[3.2. 系统性能分析 16](#_Toc933751856)

[3.2.1. 模型性能分析 16](#_Toc1904435163)

[3.2.2. 系统界面分析 16](#_Toc1745509653)

[3.3. 主要界面设计 17](#_Toc6636304)

[4. 系统设计 19](#_Toc2014695331)

[4.1. 系统设计目标 19](#_Toc1609765868)

[4.1.1. 数据预处理阶段 21](#_Toc1335958570)

[4.1.2. 模型训练 21](#_Toc1514156605)

[4.1.3. 情感分析层 22](#_Toc748843285)

[4.1.4. 前端显示 23](#_Toc1554919575)

[5 系统实现 25](#_Toc804796682)

[6总结 26](#_Toc1365825568)

[7.参考文献 27](#_Toc977618593)

[附录1 系统总流程图 30](#_Toc438309354)

[附录2 代码 31](#_Toc796403468)

[附录3 毕业设计作品说明书 32](#_Toc2034998572)

[附录4 重复率检测报告 33](#_Toc1396437482)

# 绪论

## 研究背景

随着互联网和电子商务平台的发展，越来越多的用户选择在线购物。在这一背景下，用户的评论数据成为电商平台中不可或缺的部分。这些评论不仅反映了消费者对商品和服务的真实情感，商家提也可以根据评论改进产品和优化服务。随着评论数量快速增长，人工分析处理这些海量数据变得越来越困难。基于规则或词典的情感分析方法虽然可行，但在面对复杂的语言或者网络用语表现不好。例如，同一词语在不同语境下表达的是完全不一样的意思，而传统的规则情感分析方法分辨这种语义上的差别。而且电商评论中包含许多语法冗余的文本，比如用户会添加许多表情，缩写和口语表达，这进一步增加了情感分析的难度。机器学习技术的快速发展为情感分析提供了新的解决方案。通过构建基于统计学习的模型，能够自动标注文本特征实现对评论情感倾向的自动分类。使得模型能够更好地捕捉文本上下文关系，对准确率提示非常有帮助。例如，基于Transformer架构的预训练模型（如BERT）在自然语言处理任务中表现优异，能够处理复杂的语言场景。

本论文使用电商评论做数据集和机器学习的情感分析技术，通过对比不同模型在电商评论数据集上的表现，测评哪种模型可以更有效的分析情感，并设计一个网页前端可视化模型训练成果。

## 研究现状

### 国内研究

#### 基于规则和知识库的方法

之前的国内研究都主要在研究规则驱动的情感分析方法，由情感词典，本体结构和语法规则进行方面提取和情感分类，比如刘丽（2015）提出的一种基于领域情感词典的细粒度情感分析方法，通过建立特定的领域的词典然后结合上下文信息提取显性和隐性的方面情感，为细粒度情感分析提供有效的进步，沈卓和李艳（2020）结合了领域知识和预训练语言模型（PreLM），提出的一种基于 KANO模型的情感分析框架，实现了餐饮领域评论的细粒度情感挖掘。

#### 基于机器学习的方法

在情感分析中，基于机器学习的方法是通过训练分类模型来进行情感的判断（如积极，消极或中立）。这个对于情感词典来说关系不大，因为它是通过训练数据集来学习不同情感的特征。机器学习的方法可以分为几类：支持向量机（SVM），朴素贝叶斯（Naive Bayes），决策树（Decision Trees），随机森林（Random Forest）和KNN算法。比如在数据集中，每一段句子都代表积极消极，向量机的目标是把积极消极的文本分开，越远越好以便更好分类，因为它四合处理高维稀疏的数据，贝叶斯会根据每个词出现的概率来预测文本的情感，决策树和随机森林会根据文本中的关键词来判断情感，比如出现的积极词的数量大于消极词，那么就会把这个句子判断为积极，随机森林是多个决策树的合集，KNN会判断未分类的文本和已经标注的情感的相似度来判断评论的情感。国内的研究人员也尝试将这些算法进行应用，比如敦欣卉（2017）使用微博数据集结合情感本体与SVM模型，对短文本情感进行了分析，在处理负责的语言方面取得很好效果，但是这种方法需要大量人工标注特征。蔡岩（2022）等根据C5决策树的数据挖掘为决策树的应用提供相当大的参考价值，特别是在数据预处理和特征选择方面的应用。还有郭茜（2021）根据决策树研究降雪识别领域的应用，贾晓帆（2021）等朴素贝叶斯和决策树融合的用户评论分类研究在社交媒体评论的情感分类中也很有参考价值，还有马志强（2021）的基于机器学习的微博情感分析，在情感分类的准确率又很大提升。

机器学习的方法有几个优点，第一个是容易理解和实现，一般都通过计算条件概率来进行情感分类，简单而且很高效，第二个对小数据集效果更好，尤其是SVM和贝叶斯往往能提供较好的分类性能。第三训练时间和资源方面比较短和低，依赖的时间和资源相对于其他模型都小很多。然后在缺点方面，许多机器学习的模型都非常依赖与手动特征提取，比如TF-idf和词频，提取的质量影响模型的性能，第二个需要大量的人工数据标注，在实际场景中，获得高质量的数据标注成本很高，而且很有可能标签会有不一致或不准确的问题。

#### 基于深度学习的方法

深度学习的方法在情感分析中也很重要，RNN是一种具有记忆能力的神经网络，能保存输入的信息然后应用在预测中，适合处理依赖与上下文的文本，比如一句话“虽然今天下雨了，但是我还很开心”，RNN知道前半句是负面的，但整体是积极。LSTM是RNN的一种变体，用来解决RNN中的梯度消失问题，很多地方翻译为长短期记忆网络，给人一种歧义，以为是网络一会儿能记很长的内容，一会儿能记很短的内容，但其实正确的翻译应该是长时间的短期记忆网络。它的本质就是能够记住很长时期内的信息，从而准确判断文本的情感。CNN是一种应用在图像处理的神经网络，通过卷积层提取特征，虽然CNN刚开始是为图像设计的，但也可以应用在文本分析的处理，尤其是抓取特征方面，CNN主要由一个滑动窗口来扫描文本，识别信息，就像用放大镜看一样，这样可以判断信息，在短文本的分析中可以快速抓取关键信息然后分类。Transformer更是重量级，是一种基于自注意力机制的神经网络架构，能够抓取整体文本的关系，bert是transformer的预训练模型，使用双向编码方式同时还能考虑到上下文信息，bert目前是情感分析领域的主流，它的双向编码可以很好理解上下文的关系，判断一句话的情感无论是长句子还是短剧在，都可以非常准确的判断。国内也有很多学者发的论文具有深刻影响，张志杰（2020）的基于bert的情感分析重点介绍了bert的预训练技术和上下文建模，赵婷（2020）的情感分析中的CNN与LSTM结合模型研究改善传统深度学习方法，尤其是长句子的情感分类问题。相比于机器学习，深度学习的优点会更明显，自动特征的提取尤其是CNN和RNN网络不需要人工提取特征，模型可以自动从文本中学习特征，减少负担，第二是在处理长文本数据时性能更强大，能有效捕捉文本的上下文关系，避免机器学习中依赖丢失等问题。当然缺点问题也很明显，在资源需求方面比机器学习大很多，特别是bert模型通常需要TPU这样的来进行训练和推理，对于小企业和普通用户来说是一个明显的障碍，第二在模型方面比如CNN，LSTM他们的决策过程难以理解，因为他们的模型构建就很复杂，第三对数据集的需求和训练时间方面都很大很久，当数据不足或训练时间较短时，训练的准确率往往偏低，第四是过拟合问题，当训练时间和epoch次数过大，训练集的准确率会相当高，F1分数和召回率会接近百分百，但验证集和测试集的准确率会下降，需要正则化技术比如调整dropout，l2正则化或者加入早停策略来防止过拟合。第五深度学习对噪声和文本语言复杂度方面更敏感，比如社交媒体的一些网络用语或者一些拼写错误都属于噪声，对于训练准确率又很大影响。

#### 结合知识库的混合模型

国内研究在这几年开始关注知识库在情感分析任务的应用。这种方法通过结合机器学习和深度学习模型，能够有效的增强模型的分析能力，尤其是在理解和处理复杂上下文信息一集情感类别方面。国内学者在结合知识库混合模型的研究主要集合在几个方面，知识图谱的情感分析模型，和与词典知识库结合的情感分析模型，还有基于预训练模型和知识库结合。知识图谱是一种通过事物和其之间的关系，来表示知识库，比如某人是某个公司的老板等等，在情感分析这种，知识图谱可以帮助模型理解文本中实体事物，比如文章提到了苹果，知识图谱可以区分指的是苹果公司还是水果，因为一个词 通常在不同语境下有不同的意思，所以可以提高文本理解。国内学者李惠（2019）提出的基于 知识图谱的情感分析模型，通过增强模型对词汇和上下文的理解来提高分析效果，李艳（2020）将CNN和知识图谱结合，提出的新的情感分析方法作用于事物和文本的情感联系来提高分析准确率。知识图谱有很多优点，比如可以结构化的表示信息，将信息以图的形式表示，可以清洗展示词汇之间的关系，便于模型理解，第二知识图谱可以将不同领域的知识结合在一起，对于需要集合多方面知识的推荐系统来说很有帮助，但是构建知识图谱也需要大量人工和自动标注数据，所以人工标注的工作量也相当大，效率可谓是较低，而且准确提取关系图谱链接到词汇也是一个难点，第三当知识更新时，图谱的数据也需要更新，所以对于一个数据量巨大的知识图谱来说，维护和更新也是一项很复杂的工作。在预训练模型方面，bert和gpt等预训练模型兴起，结合这些模型进行训练成为当下的主流工作，因为预训练模型可以大规模的无监督学习，学习丰富的语言信息，在结合知识库后，模型可以通过学习知识库来获得更精妙的准确率进而提升对特定领域的理解，国内许多学者也在探索结合预训练模型提高情感分析的效果，彭宇翔（2021）提出的bert结合知识图谱的情感分析方法，利用bert提取文本特征，同时再结合知识图谱的实体关系，可以明显提高情感分析的准确度。预训练模型结合知识图谱的优点也是很多，因为无监督学习的关系，可以节省标注数据的时间和人工成本，预训练模型本身已经在大量数据上进行了训练，然后知识图谱再提供知识库，结合后可以减少大量标注的需求，但缺点也很明显，需要高质量的知识图谱，要使两者结合的高效，知识图谱的质量尤为重要，如果图谱中噪点过高或关系不完整，会导致模型推理偏差严重。

#### 多语言与跨领域研究

在多语言的情感分析方面，国内研究还处于起步阶段，因为目前大部分的模型都是在英语语言的层面上训练，中文在准确率和生成能力还比较低，但随着一带一路的进展，2025年就有许多国内模型如Qwen，deepseek，豆包等中文模型引爆全球，deepseek 甚至一度超过国外众多大厂模型达到全世界第一。国内研究人员尝试通过多语言的bert模型结合中英的情感分析任务提升模型的多语言能力，结合预训练的词向量模型来减少不用语言的训练成本。或者将多语言模型与机器学习结合优化分析效果，比如将中英语料混合训练，可以优化训练效果，达到多语言分析的目的。目前国内学者会把重心放在迁移学习的情感分析方面上，比如通过电影评论进行预训练，再迁移到电商领域进行商品评论的分析也能达到很好的效果。赵劲松等（2017）提出的基于多任务学习的情感分析，通过结合多个知识库共享知识。马自强的政务微博情感分析中提出的跨领域机器学习模型通过迁移学习提高分析的准确性。

### 国外研究

#### 基于规则与知识库的方法

基于规则与知识库的方法是自然语言处理的早期技术派别，在1970-1990年在国外有一系列的研究成果，它是通过人为设定的规则和提前构建好的知识库，让计算机没有大量训练的情况下也能进行推理回答问题，基于规则与知识库的方法在国外研究中有广泛应用，特别是产品评论，网络媒体分析和客户反馈等，最著名的例子是wordnet，通过把一个英文单词按照同义词，反义词等关系组织起来的词库，ai可以用它计算词与词之间的相似度来判断句子含义，还有像framenet这样的项目，把语言划分成一个个场景框架，当系统看到一个人花100块买一本书，就能识别出买家，卖家，商品和钱这些角色。Qiuetal.（2011）提出的双传

播（DoublePropagation）算法通过语法依存关系自动扩展情感词汇和目标，解决了数据稀缺问题。这方法对小的标注数据和需要较少资源的场景表现良好，是后续研究的重要基础。

#### 深度学习模型的广泛应用

深度学习技术在国外发展迅速，国外研究者提出了多种基于深度学习的细粒度分析模型，比如循环神经网络（RNN）及其变体，Wangetal.（2016）采用LSTM和GRU模型捕捉长距离依赖信息，这些模型用于处理较长的句子和捕捉上下文信息，特别是在评论分析和情感预测，在细粒度情感分类任务中取得了良好效果。卷积神经网络（CNN）Tangetal.（2016）基于CNN的模型有效捕捉了局部特征，并在多个情感分析任务中表现出色。attention与Transformer模型，Vaswanietal.（2017）提出的Transformer及其变体（如BERT、RoBERTa）通过multi-head attention和大规模预训练文本显著提升了细粒度情感分析的性能。Gaoetal.（2020）开发的一种基于Transformer的多头注意力模型，专门处理方面情感分类任务。国外研究者不仅关注整体情感分类，还扩展到方面级情感分析，判断文本中不同属性方面，比如服务，价格，环境等，Wang et al.（2016）提出 ATAE-LSTM 模型，在 LSTM 基础上将 aspect 向量加入每个词的输入中，引导模型学习与方面相关的情感表达。尽管深度学习中情感分析中取得了很好的发展，但还是有许多缺陷，比如讽刺情绪和反话的识别。

#### 图像与情感细粒度分析的融合

国外研究在细粒度图像分析与情感分类的结合领域也取得了重要成果：

Zhuetal.（2024）在《PatternRecognition》中提出了LocalcontextAttentionLearning

方法，通过设计精细位置注意网络（FLAN）和精细位置标签传递（FLLT），有效解决了场景

图生成任务中的感知模糊和细粒度标注不足问题。

Lu和Han（2024）在《ComputerVisionandImageUnderstanding》中提出MTDSNet，

结合混合掩码教师学生训练策略和双动态选择模块，用于增强细粒度图像分类任务中的特

征表现力。

#### 知识增强方法与跨领域研究

知识库与深度学习的结合是国外研究的另一重要方向。例如，SenticNet知识库通过将情感

知识嵌入深度学习模型中，提高了情感分类的解释性和准确性。Lietal.（2021）提出的

多任务学习混合模型，通过共享知识提升了模型在多个任务中的整体性能。此外，多语言与

跨领域迁移学习技术被广泛探索，通过利用资源丰富语言（如英语）训练模型，并迁移到低

资源语言（如阿拉伯语、印地语）应用场景。

## 主要工作

项目研究分为两个部分，分别是电商评论情感分类算法实现和整体系统实现。算法

实现的目的是训练机器学习电商文本情感分类模型，评估电商评论的情感倾向。系统实

现的目的是制作前端用户交互界面，可以使用已经保存好的电商评论情感分析模型。其中，算法研究是项目的核心内容。

算法实现部分，本论文使用了五种不同的模型来实现情感分析，分别是长短时记忆网络（LSTM），双向长短时记忆网络（Bi-LSTM）， 双向LSTM与自注意力机制结合（Bi-LSTM-Attention），LSTM和自注意力机制结合（LSTM-Attention）和文本卷积神经网络（TextCNN），通过对每个模型进行参数微调来尽量实现一个更高的预测准确度，具体来说优化模型的超参数，比如学习率，批次大小，隐藏层的纬度和卷积核大小等，再加入早停机制来节省训练时间的方式旨在提高模型准确率。

系统实现部分，本论文使用后端flask和前端JavaScript和HTML的方式实现前后端分离，实现用户输入预测文本，前端收到请求后发送给后端Restful api接口，调用模型进行预测，然后返回结果到前端显示页面，前端进行展示。前端界面包括输入文本框和文件上传区域，可以上传多个句子文本以便预测，同时后端也可以接受多文本进行预测前端可以进行分析，数据集方面因为电商平台的反爬虫机制太强，无法进行爬取，使用使用github公开的商品评论数据集。

# 关键技术简介

## 数据预处理

中文文本数据预处理是将原始的中文文本再进行一系列数据清洗和优化后转换成适合机器学习模型处理的格式，通过去除噪音，去除标点符号，表情，非中文的字符，链接等和文本无关的信息，也就是噪音，这样处理的文本可以提高模型预测的性能和效率。由于商品评论原始数据集中有非常多的语气词和网络用语，标点符号，链接字符和表情，这些不能被机器识别，所以原始数据集必须先进行数据预处理，转换成机器可以识别的格式。具体分为，数据清洗，数据分词，文本标准化，去除停用词和数据集划分。

### 中文文本清洗

在原始文本中，比如评论句子中，会包含许多无关文本情感的噪点，比如HTML标签，标点符号表情等对情感分析没有帮助的，去掉这些噪声内容是预处理的第一步，数据清洗。

首先从Excel文件(data.xlsx)加载原始数据，然后重命名列：将"评论"改为"review"，"标签"改为"label"，然后移除重复评论。

df = pd.read\_excel('../data.xlsx')

df.rename(columns={'评论': 'review', '标签': 'label'}, inplace=True)

df = df[[df.columns[1], df.columns[0]]]

df = df.drop\_duplicates(subset=['review'], keep='first')

清洗完毕后的数据从79538条降到了63384条。



### 正则清洗

正则清洗使用正则表达式对文本数据进行处理，移除不必要的内容比如HTML标签，URL，表情标点符号，数字，空格等，移除这些对模型训练无影响的内容可以提升数据的质量提高训练准确率。

首先定义匹配正则表达式的属性，再调用clean\_text函数进行清洗。

re.compile()是re模块的一个函数，将正则表达式变异成一个对象，可以在后续的匹配操作中被重复使用，提高效率，re.sub()用于替换字符串中匹配正则表达式的部分。

本论文中代码预先compile定义属性名，再调用这些属性进行数据清洗，通过两个函数结合使用，可以在多个sub()操作中重复使用这个正则，提高代码效率。

关键代码如下：

# 移除HTML标签

TAG\_RE = re.compile(r'<[^>]+>')

# 匹配URL

URL\_RE = re.compile(r'http[s]?://(?:[a-zA-Z0-9$\-\_@.&+]|[!\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+')

# 匹配表情符号 (包括Unicode范围)

EMOJI\_RE = re.compile(r'[\U0001F300-\U0001F9FF\u2600-\u26FF\u2700-\u27BF]')

# 匹配标点符号 (包括中文标点)

PUNCT\_RE = re.compile(r'[^\w\s\u4e00-\u9fff]')

# 匹配数字

DIGITS\_RE = re.compile(r'\d+')

# 匹配空白字符（包括空格、制表符、换行等）

SPACE\_RE = re.compile(r'\s+')

# 移除英文字符

ENGLISH\_RE = re.compile(r'[a-zA-Z]+')

def clean\_text(review):

"""

对文本进行清洗：

- 去除标签、链接、表情、标点、数字及重复字符

- 去除英文字符

- 去除多余空格

"""

# 处理 NaN 值

if pd.isna(review):

return ""

# 将非字符串类型转换为字符串

if not isinstance(review, str):

review = str(review)

review = TAG\_RE.sub('', review)

review = URL\_RE.sub('', review)

review = EMOJI\_RE.sub('', review)

review = PUNCT\_RE.sub('', review)

review = DIGITS\_RE.sub('', review)

review = REPEAT\_CHARS\_RE.sub(r'\1', review)

review = ENGLISH\_RE.sub('', review) # 去除英文

review = SPACE\_RE.sub(' ', review).strip()

return review

测试clean\_text函数可以看到，标点符号，表情和url等无关因素都被清洗掉了。



### 数据分词

文本分词中，中文是没有空格分隔的一门独特的语言，而情感分析中需要特定词汇来进行分析，所以中文分词是一个关键步骤，分词是把一个句子中的常见的词语划分为单独的词和词组，jieba是一个开源的中文分词工具，通过词典进行分词，THULAC是清华大学推出的中文分词工具，可以自动标注单词的词性，pkuseg是百度推出的分词工具。本论文使用jieba加停用词的方式进行数据分词。

停用词是指在文本分析中出现的频率很高，但没有任何情感的词，对文本分析没有任何帮助，比如‘的’，‘了’，‘是’等词汇。情感分析中，这些词不会对分析结果产生影响，因此去除这些词可以节省训练时间达到节省资源的目的，停用词可以使用现成的停用词表进行匹配，本论文中使用的开源的停用词表进行分词。

关键代码：

# 读取停用词

stopwords = []

with open("../stopword.txt", "r", encoding="utf-8") as f:

for line in f.readlines():

stopwords.append(line.strip())

def tokenize(review, stopwords):

"""

对单个文本进行分词，去除停用词和空白字符，返回以空格分隔的字符串。

"""

words = jieba.cut(review, cut\_all=False) # 使用精确模式进行分词

# 返回一个空格分隔的字符串

return ' '.join([word for word in words if word not in stopwords and word.strip()])

首先加载stopwords，再定义tokenize函数使用jieba加载停用词对文本进行分词。

测试tokenize函数可以看到，句子中的词汇都被空格分隔开了，并返回str格式。

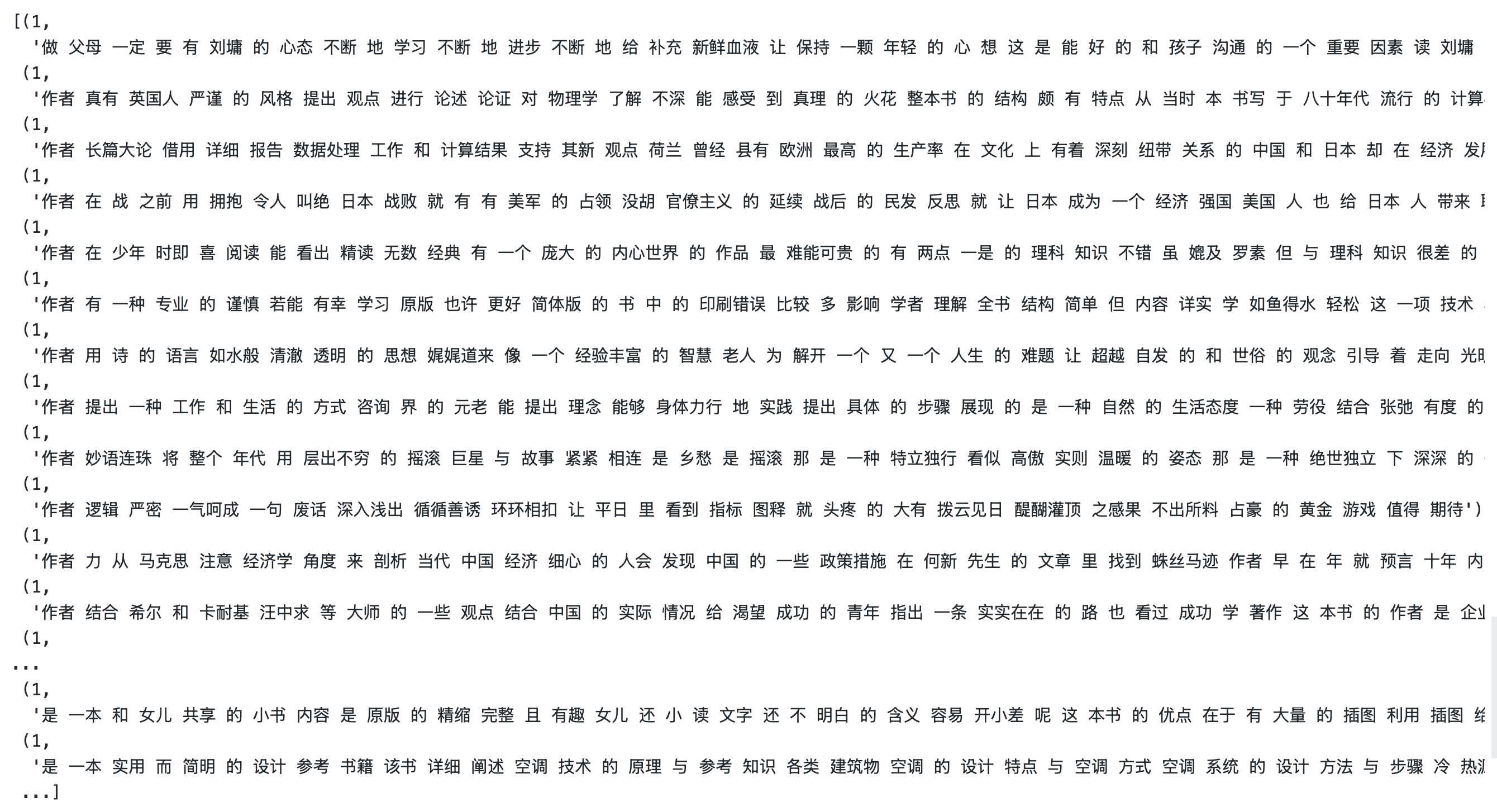


清洗并分词后的文本会写入到trained.txt文件中，以 (label, text) 的格式来保存，但这是未经数据集划分的结果，下一步需要进行划分。

with open('../trained.txt', 'w', encoding='utf-8') as file:

for label, text in combined\_data:

file.write(f"{label} {text}\n")



### Word2vec

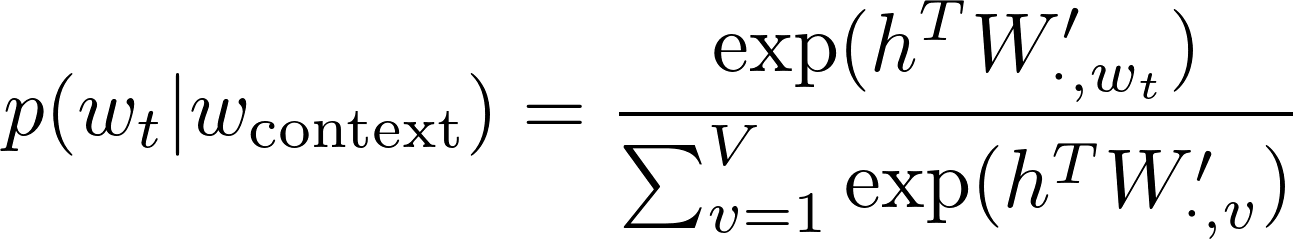
Word2Vec是建模了一个单词预测的任务，通过这个任务来学习词向量。假设有这样一句话Pineapples are spiked and yellow，现在假设spiked这个单词被删掉了，现在要预测这个位置原本的单词是什么。

Word2Vec本身就是在建模这个单词预测任务，当这个单词预测任务训练完成之后，那每个单词对应的词向量也就训练好了。

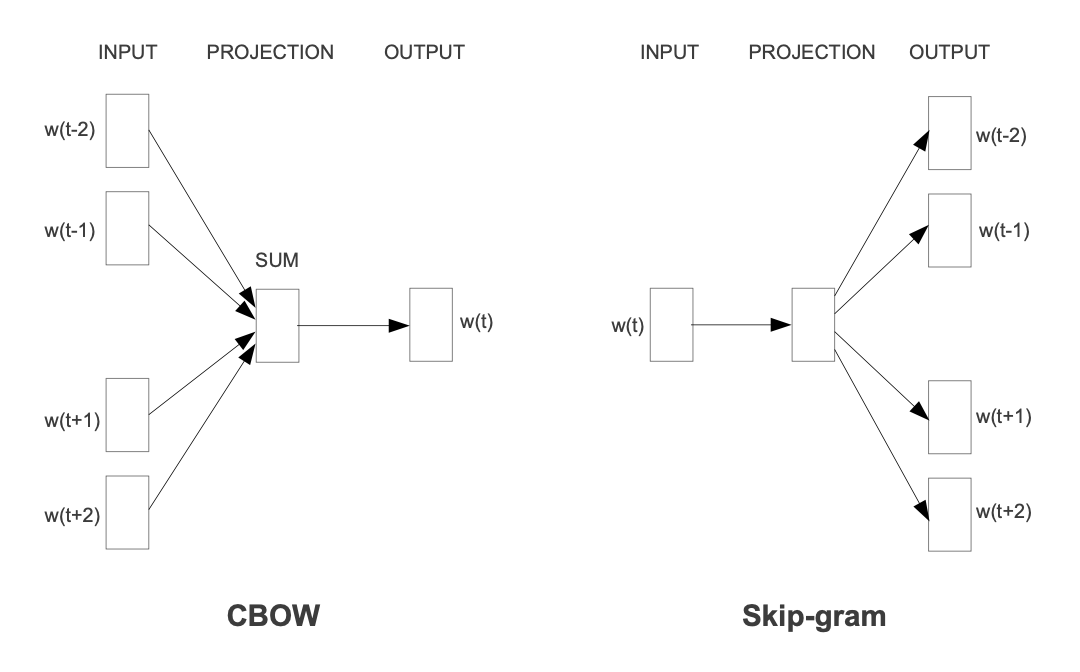
Word2vec是2013年被Mikolov提出来的词向量训练算法，在论文https://arxiv.org/pdf/1301.3781.pdf中作者提到了两种word2vec的具体实现方式：连续词袋模型CBOW和Skip-gram

### CBOW

CBOW(Continuous Bag of Words): 根据上下文wpsoffice,预测当前单词wpsoffice。



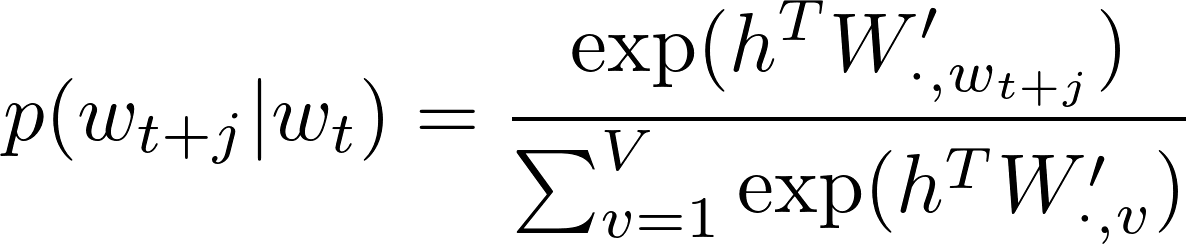
wpsoffice是要预测的中心词，wpsoffice是上下文，wpsoffice是上下文词向量的表示。



一般来说，CBOW比Skip-garm训练快且更加稳定一些，然而，Skip-garm不会刻意地回避生僻词（即出现频率比较低的词），比CBOW能够更好地处理生僻词。

### Skip-gram

Skip-gram: 根据当前单词预测上下文



wpsoffice是中心词，wpsoffice是上下文，wpsoffice是中心词的向量表示

### 文本标准化

文本标准化是nlp处理中的重要环节，它把分词后的文本转换成模型接受的数字格式，在文本分词后，本论文构建了一个word2id词典，将分词的词汇映射成id的格式，模型可以理解，然后对于分词后不在词典中的词汇，需要一个特殊标记进行处理，本论文采用\_PAD\_作为填充，目的是模型处理输入文本时能统一的处理所有文本，就算词汇是模型没见过的。

def build\_word2id(file):

"""

@description: build the dict of 'word2id'

@param {}

- file: str, word2id保存地址

@return {}

返回word2id的字典

"""

stopwords = stopwords\_list() # 获取停用词

word2id = {"\_PAD\_": 0} # 初始化字典，包含PAD标记

path = [Config.train\_path, Config.val\_path]

for \_path in path:

with open(\_path, encoding="utf-8") as f:

for line in f.readlines():

out\_list = []

# 分割文本为单词

sp = line.strip().split()

for word in sp[1:]: # 假设第一列是标签，跳过

if word not in stopwords: # 去除停用词

# 清洗单词

word\_clean = clean\_text(word)

# 如果word经过清洗后不为空，并且不为单字符

if word\_clean and len(word\_clean) > 1:

out\_list.append(word\_clean)

# 将不重复的词加入字典

for word in out\_list:

if word not in word2id:

word2id[word] = len(word2id)

# 将word2id字典保存到文件

with open(file, "w", encoding="utf-8") as f:

for w in word2id:

f.write(w + "\t")

f.write(str(word2id[w]))

f.write("\n")

return word2id

然后使用预训练的词向量模型word2vec，其中的两个主要算法skip-gram和cbow将每个词汇映射为固定维度的向量，预训练词向量会提取词语之间的关系，提升模型的表现。

关键代码：

加载预训练的 word2vec 模型（词向量模型）

model = gensim.models.KeyedVectors.load\_word2vec\_format(fname, binary=True)

创建一个随机初始化的词向量矩阵，大小为 [n\_words, vector\_size]（默认初始化为 -1 到 1 之间的随机值）

word\_vecs = np.random.uniform(-1.0, 1.0, [n\_words, model.vector\_size])

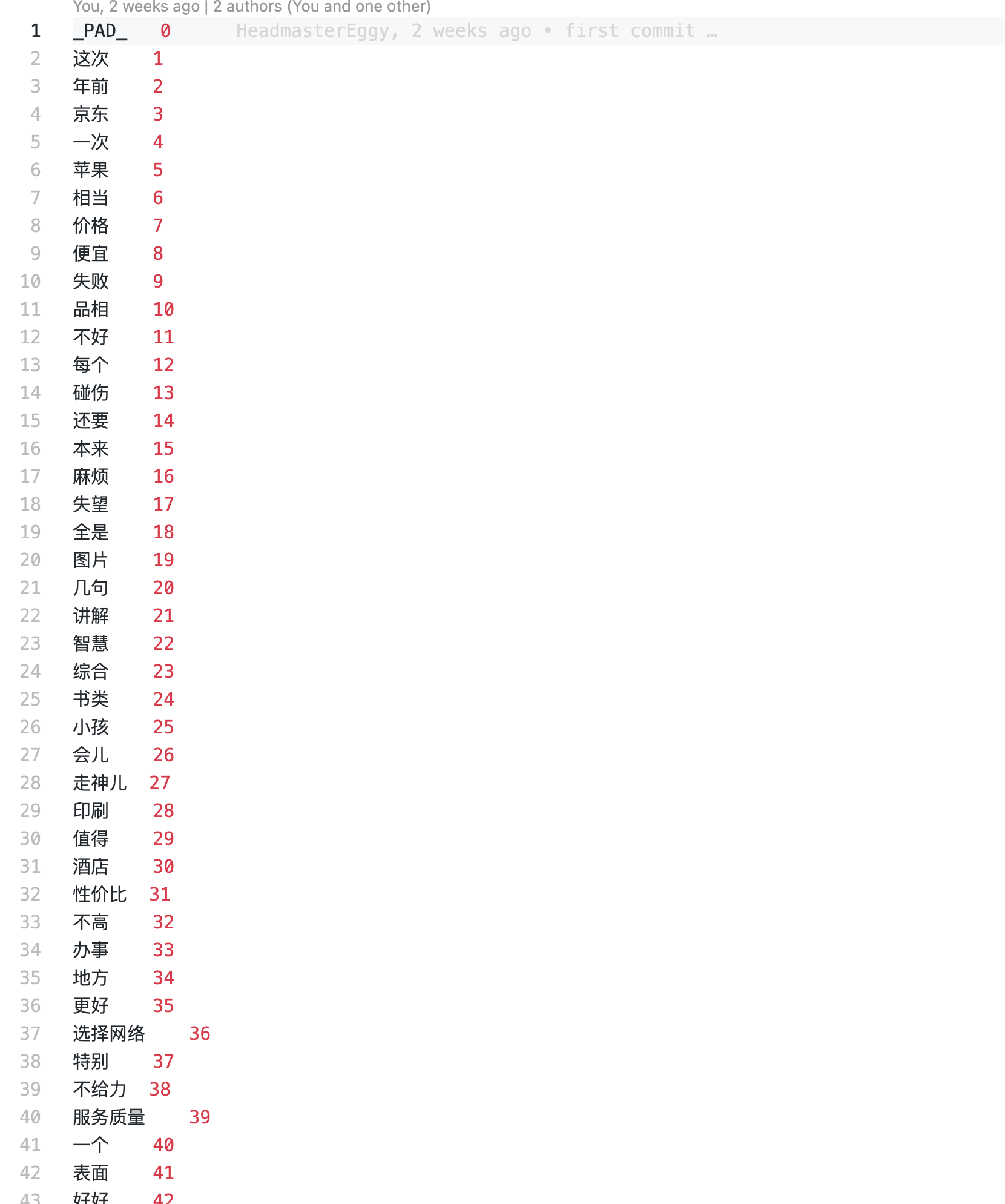


再将文本的长度统一成固定长度，为了使模型处理不定长的文本，本论文设置了max\_senlen=50，再使用<PAD>符号进行填充，最后输出numpy数组这种方式使所有输入文本统一成固定句长，便于批量处理和模型训练。

关键代码:

初始化字典，包含PAD标记

word2id = {"\_PAD\_": 0}



dataloader是pytorch提供的一种数据加载工具，用户可以自定义dataloader，本论文通过创建dataloader来封装数据，将数据分批次加载，每次只处理一个批次的文本，这样可以节省内存提高训练效率。

class Data\_set(Dataset):

def \_\_init\_\_(self, Data, Label):

self.Data = Data

if Label is not None:

self.Label = Label

def \_\_len\_\_(self):

return len(self.Data)

def \_\_getitem\_\_(self, index):

if self.Label is not None:

data = torch.from\_numpy(self.Data[index])

label = torch.from\_numpy(self.Label[index])

return data, label

else:

data = torch.from\_numpy(self.Data[index])

return data

### 数据集划分

数据集划分是机器学习的常见操作，将数据集分为训练集，测试集和验证集，通常训练集占80%，验证集和测试集占20%。

数据集的划分可以通过几种方式实现。随机划分，将数据集随机打乱后，按照预定比例划分成训练集，验证集和测试集，适用与数据量比较大而且分布均匀的数据集中。按时间划分，对于文本数据谁时间变化的数据集，可以按照时间顺序进行划分，最早的数据用于训练，晚的数据用户验证和测试，这个方法可以避免未来数据泄露给模型，提高测试集准确率。交叉验证，通过将数据集划分成多个集合，每个集合进行训练和验证，然后再把结果平均，可以得到更均匀的准确率，这种方法对于小数据集的情感效果极高，能防止过拟合。

本论文中，使用scikit-learn库中的train\_test\_spilt来划分数据集，将原始数据集划分为训练集，测试集和验证集，也就是随机划分，可以保持数据集的随机性，平衡数据集。

关键代码：

将数据划分为训练集、验证集和测试集，80% 训练集，20% 临时集，# 50% 验证集，50% 测试集

train\_df, temp\_df = train\_test\_split(df, test\_size=0.2, random\_state=42)

val\_df, test\_df = train\_test\_split(temp\_df, test\_size=0.5, random\_state=42)

训练集大小: 50674

验证集大小: 6334

测试集大小: 6335

划分完数据集后，分别使用长短时记忆网络（LSTM），双向长短时记忆网络（Bi-LSTM）， 双向LSTM与自注意力机制结合（Bi-LSTM-Attention），LSTM和自注意力机制结合（LSTM-Attention）和文本卷积神经网络（TextCNN）几种模型进行文本情感分类，通过微调超参数和不同的分类算法进行商品评论情感分类，计算分类的准确率，讨论不同方法的优劣。

## 基于长短时记忆网络算法的电商评论分析

### 训练模型

# 系统分析

## **可行性分析**

在“基于机器学习的电商评论情感分析方法研究”中，可行性分析是评估项目是否具备实施条件的重要环节。本节从技术可行性、经济可行性和操作可行性三个方面进行分析，以确定项目的可行性。

### **技术可行性**

技术可行性主要评估项目所需的技术是否成熟、是否能够支持系统的开发和实现。

1. 机器学习技术的成熟性：

机器学习技术，尤其是深度学习技术，近年来在自然语言处理（NLP）领域取得了显著进展。例如，BERT、GPT等预训练模型在情感分析任务中表现出色，能够有效捕捉文本的语义信息。

开源工具和框架（如TensorFlow、PyTorch、Scikitlearn等）为模型的开发和训练提供了强大的支持，降低了技术门槛。

中文分词工具（如Jieba）、词向量模型（如Word2Vec、FastText）以及预训练语言模型（如BERTwwm、RoBERTawwm）在中文文本处理中已得到广泛应用。

1. 数据处理技术的可行性：

数据采集技术（如网络爬虫）已经非常成熟，能够高效地从电商平台获取评论数据。

数据预处理技术（如文本清洗、分词、去停用词）在中文文本处理中已有成熟的解决方案。

特征提取技术（如TFIDF、Word2Vec、BERT）能够满足不同场景下的文本表示需求。

1. 硬件资源的可行性：

深度学习模型的训练和推理对计算资源要求较高，但当前云计算平台（如阿里云、腾讯云、AWS）提供了高性能的GPU实例，能够满足模型训练的需求。

对于中小规模的数据集，普通的工作站或服务器即可完成实验。

综上所述，项目所需的技术已经成熟，且相关工具和资源易于获取，技术可行性较高。

### **经济可行性**

经济可行性主要评估项目的开发成本和预期收益。

1. 开发成本：

人力成本：项目需要具备机器学习、自然语言处理和编程能力的研究人员，但相关人才资源丰富，且开源工具降低了开发难度。

硬件成本：模型训练可能需要高性能计算资源，但可以通过云计算平台按需租用，成本可控。

数据成本：评论数据可以通过爬虫技术免费获取，无需额外购买数据集。

1. 预期收益：

商业价值：电商评论情感分析系统能够帮助商家快速了解用户反馈，优化产品和服务，提升用户满意度，从而增加销售额。

学术价值：研究成果可以发表在相关领域的学术会议或期刊上，提升研究者的学术影响力。

社会价值：通过情感分析技术，消费者可以更快速地获取其他用户的真实评价，辅助购物决策。

1. 成本效益分析：

项目的开发成本相对较低，且具有较高的商业和学术价值，投资回报率较高。

综上所述，项目的经济可行性较高，具有较好的成本效益。

### **操作可行性**

操作可行性主要评估项目的实施难度和使用便利性。

1. 实施难度：

项目的技术路线清晰，相关工具和框架易于使用，开发难度较低。

数据采集、预处理、特征提取和模型训练等环节均有成熟的解决方案，实施过程中不会遇到难以克服的技术障碍。

1. 使用便利性：

系统设计采用模块化结构，便于维护和扩展。

最终的情感分析结果可以通过可视化方式展示，用户无需具备专业知识即可理解和使用。

系统可以集成到电商平台的后台管理系统中，方便商家实时监控用户评论的情感倾向。

1. 用户接受度：

对于电商平台和商家而言，情感分析系统能够提供有价值的用户反馈信息，具有较高的接受度。

对于普通消费者而言，情感分析结果可以帮助其快速了解商品的口碑，提升购物体验。

综上所述，项目的操作可行性较高，实施难度低且用户接受度高。

## **系统性能分析**

系统性能分析是评估系统在实际运行中的表现，主要包括模型性能分析和系统界面分析两部分。通过性能分析，可以了解系统的准确性、效率、用户体验等方面的表现，为后续优化提供依据。

### **模型性能分析**

模型性能分析是评估情感分析模型在分类任务中的表现，主要通过一系列评价指标来衡量。

**准确率（Accuracy）**：衡量所有预测正确的比例，适用于类别分布均衡的数据集。

**精确率（Precision）**：表示模型在预测某类别时，正确的比例，避免误判过多。

**召回率（Recall）**：表示实际属于某类别的数据被正确识别的比例，衡量模型对正例的识别能力。

**F1-score**：精确率和召回率的调和平均，综合评估分类模型的表现。

### 系统界面分析

系统界面分析主要评估用户界面的设计是否合理、操作是否便捷、用户体验是否良好。

1. 界面布局

界面布局应简洁清晰，功能模块划分明确，便于用户快速找到所需功能。

采用直观的导航栏和菜单设计，支持用户在不同功能模块之间快速切换。

2. 交互设计

提供友好的交互提示，如按钮点击反馈、加载进度条等，提升用户操作体验。

支持用户自定义设置，如选择模型、调整参数等，满足不同用户的需求。

3. 可视化效果

情感分析结果以图表形式展示，如情感分布饼图、关键词云等，便于用户直观理解。

支持数据导出功能，用户可以将分析结果保存为文件（如Excel、PDF）。

4. 响应速度

系统应具备较高的响应速度，确保用户操作的实时性。

对于大规模数据处理，提供后台任务管理功能，用户可以在任务完成后查看结果。

## 主要界面设计

系统的主要界面设计包括以下几个核心页面：

1. 首页

显示系统概览信息，如已分析的评论数量、情感分布统计等。

提供快速入口，如“上传数据”、“开始分析”、“查看结果”等。

2. 数据上传页面

支持用户上传评论数据文件（如CSV、Excel）。

提供数据预览功能，用户可以在上传前查看数据格式和内容。

3. 分析设置页面

允许用户选择情感分析模型（如SVM、LSTM、BERT）。

支持用户调整模型参数，如学习率、迭代次数等。

4. 结果展示页面

以图表形式展示情感分析结果，如情感分布饼图、关键词云、情感趋势图等。

提供详细的数据表格，用户可以通过筛选和排序功能查看特定评论的情感分类结果。

5. 任务管理页面

显示当前正在运行的任务及其进度。

支持用户查看历史任务记录及其分析结果。

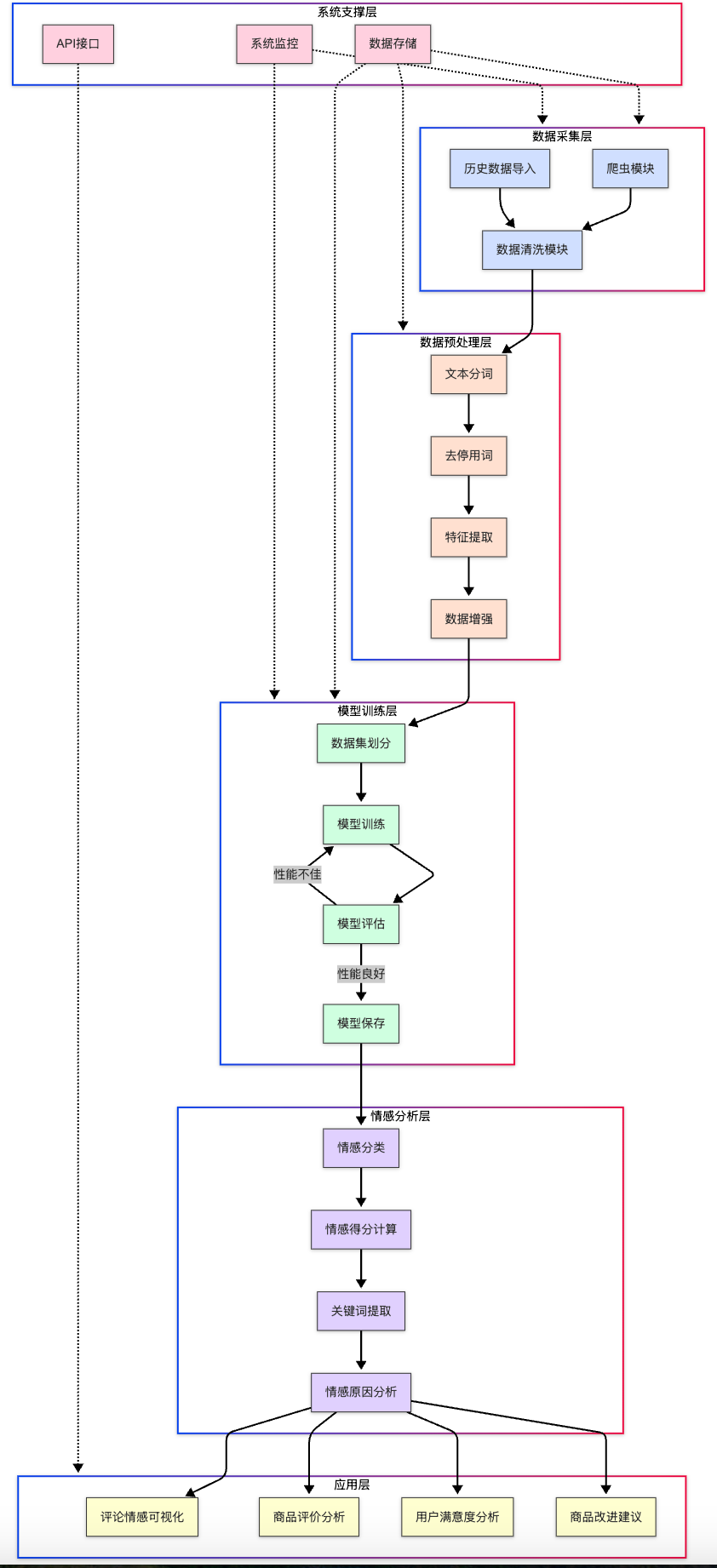
6. 帮助与支持页面

提供系统使用指南和常见问题解答。

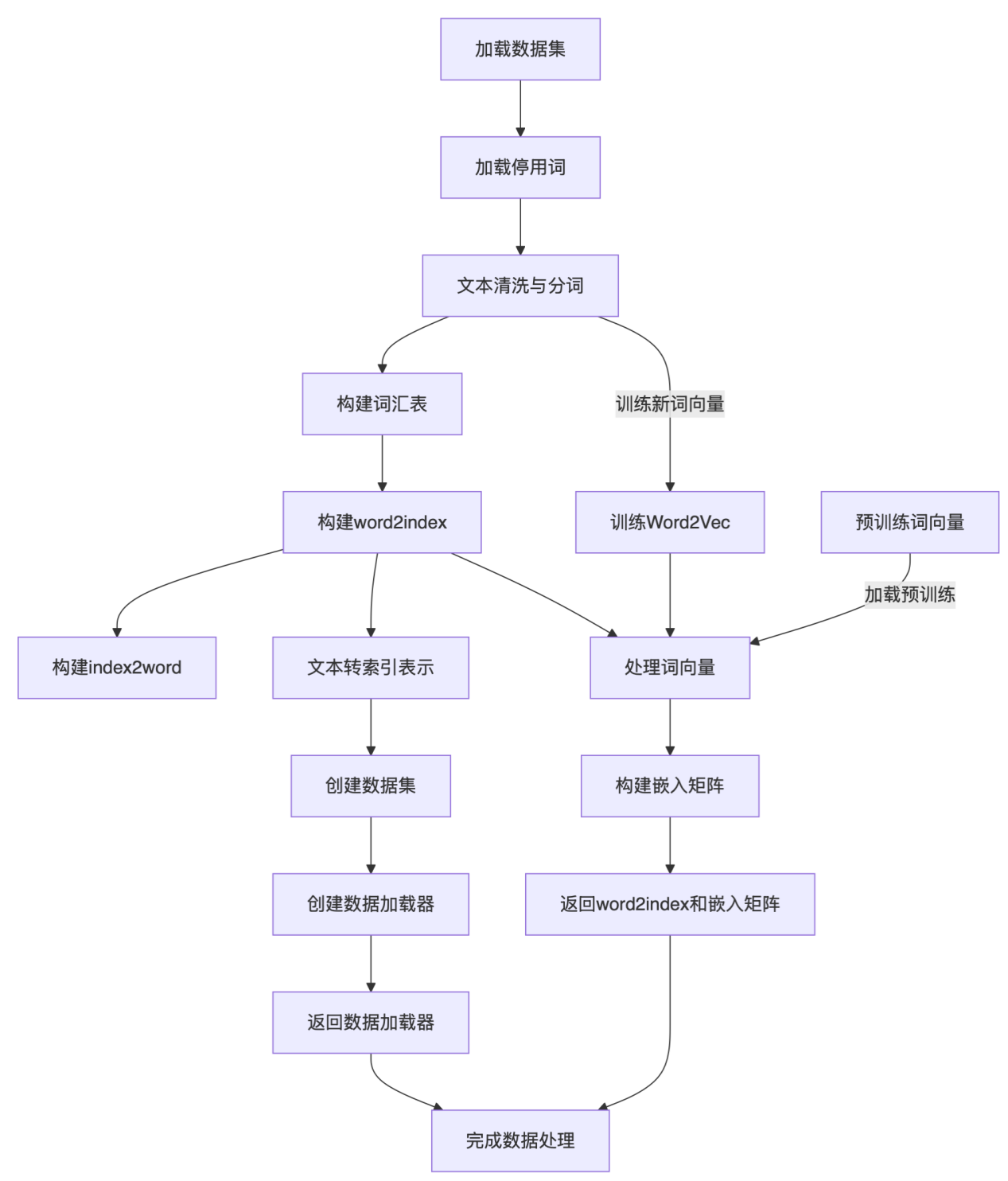
支持用户反馈问题或建议，便于系统改进。

# 系统设计

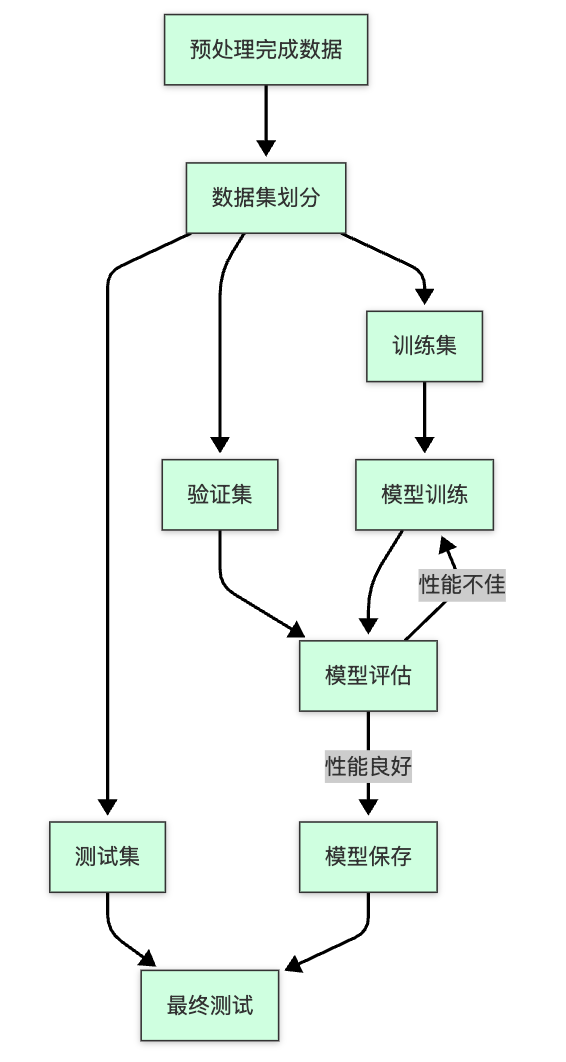
## 系统设计目标



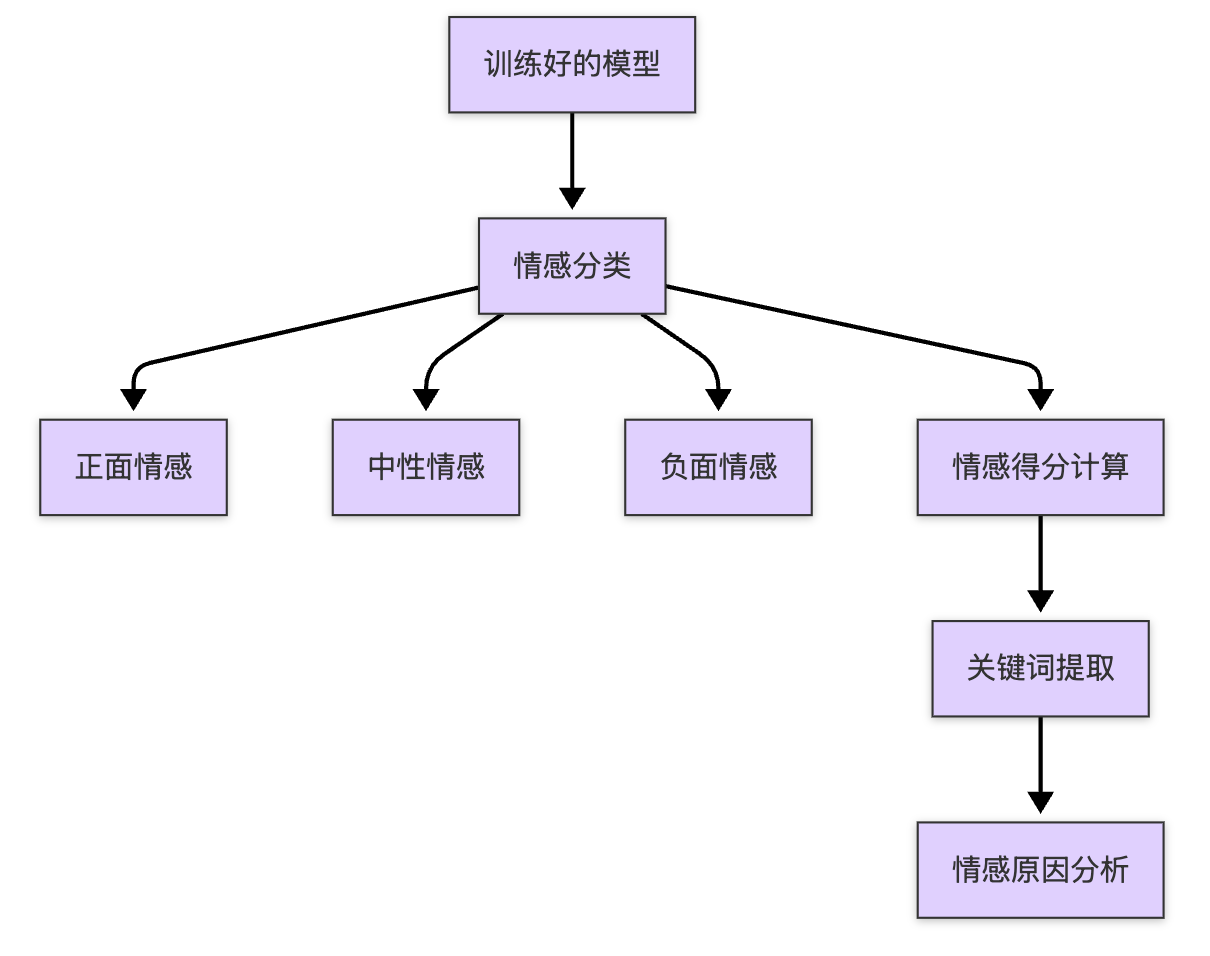
### 数据预处理阶段



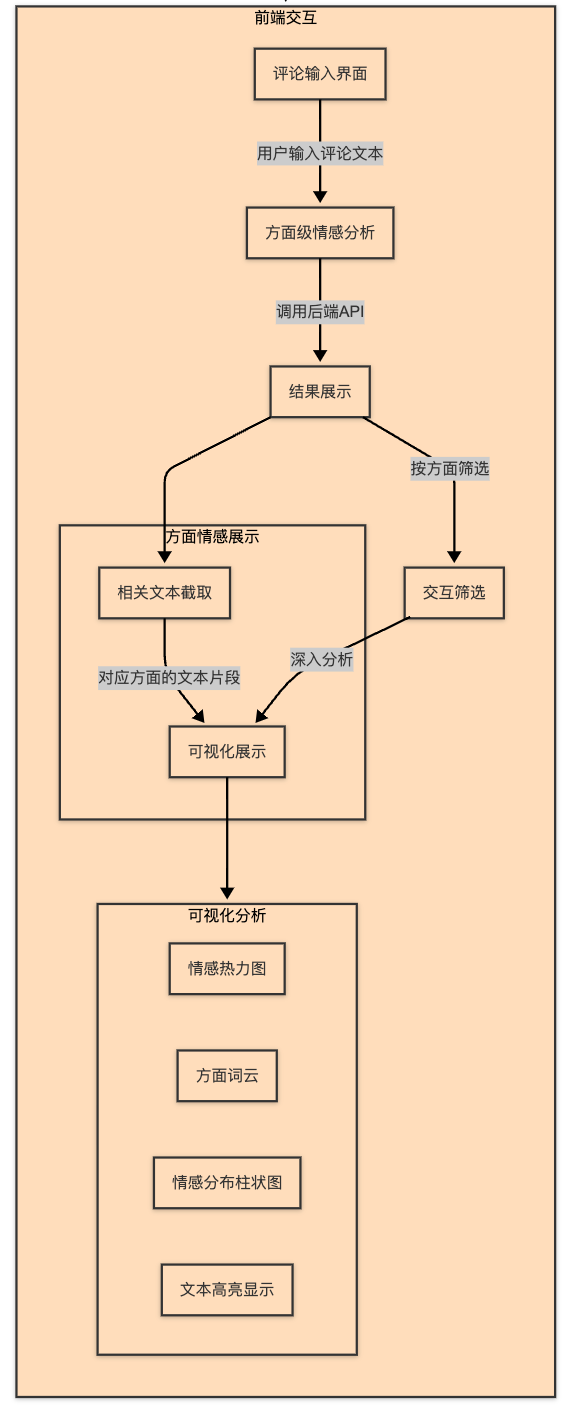
### 模型训练



### 情感分析层



### 前端显示



# 5 系统实现

# 6总结

通过系统的开发与实现，本文完成了基于机器学习的电商评论情感分析方法研究。研究过程中，我们构建了一个完整的情感分析系统，包括数据采集、数据预处理、模型训练、情感分析和应用展示等多个模块。

设计并实现了一套完整的电商评论情感分析系统架构，各模块间协同工作，形成了从数据采集到结果应用的完整流程。对比分析了多种机器学习算法在电商评论情感分析任务中的表现，包括传统机器学习方法和深度学习方法，为电商评论情感分析提供了方法选择依据。提出了针对电商领域的特征提取方法，有效提高了情感分析的准确率，尤其是对于包含专业术语和商品特性描述的评论文本。开发了直观的可视化界面，使分析结果能够为商家提供有价值的决策支持信息，帮助商家了解用户对产品的真实评价和改进建议。

然而，本研究仍存在一些局限性。由于时间和条件的不足，没有进行实际测试，系统在真实环境中的性能和稳定性有待验证。此外，对于复杂的情感表达和隐含的情感倾向，现有模型的识别能力仍有提升空间。这些问题有待在以后的工作学习中不断探索和改进。

未来的研究方向将集中在以下几个方面：一是引入更先进的预训练语言模型，提高系统对复杂情感的理解能力；二是扩大训练数据规模，增强模型的泛化能力；三是结合多模态信息，如图片、视频等，进行更全面的情感分析；四是探索情感分析结果在电商推荐系统中的应用，为用户提供更精准的商品推荐。

# 7.参考文献

1. 卜晓阳,蔡岩,王宗伟,赵郭燚.基于 C5.0 决策树算法的电力营销数据挖掘[J].微型电脑应用,2022,38(01):2326.
2. 彭宇翔,文继芬,李皓,刘涛,唐辟如,郭茜.基于决策树模型的贵州降雹识别研究[J].中低纬山地气象,2021,45(06):99101.
3. 贾晓帆,何利力.融合朴素贝叶斯与决策树的用户评论分类算法[J].软件导刊,2021,20(07):15.
4. 邓晓林,陈毅红,王登辉.大数据环境下决策树的研究[J].太原师范学院学报(自然科学版),2021,20(02):4757.
5. 冷婷,叶仁玉,李沅静.基于因子分析朴素贝叶斯方法的电商文本分类[J].安庆师范大学学报(自然科学

版),2024,30(01):4751.

1. 张财,马自强,闫博.基于机器学习的政务微博情感分析模型设计[J].计算机工程:114.
2. Brauwers G, Frasincar F. A survey on aspectbased sentiment classification[J]. ACM Computing Surveys, 2021, 1(1): 135. DOI: 10.1145/3503044.
3. Pushpam Kumar Sinha. Modifying one of the Machine Learning Algorithms KNN to Make it Independent of the Parameter k by Redefining Neighbor[J]. International Journal of Mathematical Sciences and Computing(IJMSC), 2020, 6(4) : 1225.
4. Jayasree Saha and Jayanta Mukherjee. CNAK : Cluster number assisted Kmeans[J]. Pattern Recognition, 2021, 110.
5. Li Teng and Dou Yong. Representation learning on textual network with personalized PageRank[J]. Science China Information Sciences, 2021, 64(11)
6. Crosssubject emotion recognition using hierarchical feature optimization and SVM with multikernel collaboration[J].Physiological measurement. Volume 44 , Issue 12 . 2023
7. 刘丽, 赵劲松. 面向产品评论的细粒度情感分析[J]. 计算机应用, 2015, 31(12): 481486.
8. Qiu G, Liu B, Bu J, Chen C. Opinion word expansion and target extraction through double propagation[J]. Computational Linguistics, 2011, 37(1): 927. DOI: 10.1162/COLI\_a\_00034.
9. 李慧, 柴亚青. 基于卷积神经网络的细粒度情感分析方法[J]. 数据分析与知识发现, 2019, 31(1): 95103. DOI: 10.11925/infotech.20963467.2018.0158.
10. 沈卓, 李艳. 基于PreLMFT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(4): 6371. DOI: 10.11925/infotech.20963467.2019.0146.
11. 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, 7(7): 6165.
12. 陈炳丰, 郝志峰, 蔡瑞初, 等. 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(3): 814. DOI: 10.12052/gdutxb.170036.
13. Zhu, Z., Wang, H., Zhou, X., et al. SFPL: Samplespecific finegrained prototype learning for imbalanced medical image classification[J]. IEEE Transactions on Medical Imaging, 2024.
14. Stangl, M., Kothari, A., Schuh, B., et al. Reducing the environmental embrittlement effect of TiAl alloys exposed to air at high temperatures by nitrogen strengthening[J]. Acta Materialia, 2024, 252: 11851198.
15. Pan, L., Zhou, Y., Yang, X., et al. Multiscale synchronous contextual network for finegrained urban flow inference[J]. IEEE Transactions on Big Data, 2025.
16. Lu, Z., Han, Y. MTDSNet: Mixmask teacher–student strategies and dual dynamic selection plugin module for finegrained classification[J]. Computer Vision and Image Understanding, 2024.
17. Zhu, Q., Lin, H., Huang, S. Local Context Attention Learning for Finegrained Scene Graph Generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
18. Jin, W., Ho, H. A Novel Lexicalized HMMbased Learning Framework for Web Opinion Mining[C]. Proceedings of the 26th International Conference on Machine Learning, 2009: 465472.
19. Shi, C., Su, Z., Zhang, K., et al. CloudSwinNet: A Hybrid CNNTransformer Framework for Groundbased Cloud Images Finegrained Segmentation[J]. Energy, 2024.
20. Li, G., Yang, M., Wang, J. Detecting Android malware: A multimodal fusion method with finegrained feature. Information Fusion, 2025, 85: 102014. DOI: 10.1016/j.inffus.2025.102014.
21. Y. Zhang, C. Zhu, and Y. Xie, “FineGrained Sentiment Analysis of CrossDomain Chinese ECommerce Texts Based on SKEP\_GramCDNN,” IEEE Access, vol. 11, pp. 74058–74070, 2023. doi: 10.1109/ACCESS.2023.3296447.
22. Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A Chinese Review Dataset Towards Aspect Category Sentiment Analysis and Rating Prediction. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2069–2079, Online. Association for Computational Linguistics.

# 附录1 系统总流程图

一级标题，此页单独。

对于一些不宜放入正文中，但对于毕业设计(论文)有参考价值的内容，或以便他人阅读方便的工具性资料，如调查问卷、公式推演、编写程序、原始数据表、系统实物图、原理图、设计作品说明书等，可以编入附录。附录的篇幅一般不宜超过正文。

没有附录就不要硬凑。有多个附录则按附录1、2、3……依次编排。

# 附录2 代码

没有实验原理图就不要这一附录，后面附录号提前。

# 附录3 毕业设计作品说明书

一、作品名称

如果做的是毕业设计(硬件或软件)，则除了要写出论文，还要有说明书。如果没有设计作品，则不要这项附录

二、作品功能

1、

2、

……

三、运行环境

硬件环境

标题用小四号黑体，正文用5号宋体，所有行间距均为1.5倍。

软件环境

四、操作步骤

1、

2、

……

五、注意事项

1、

2、

……

# 附录4 重复率检测报告

将检测报告用截图或插入对象等方式放入此处，可多页，但要清晰。