

**本科生毕业论文（设计）**



**题 目 基于BERT的中文电影评论情感分析系统设计与实现**

**学 院**

**专 业**

**学生姓名**

**学 号**  **年级 2021**

**指导教师**

**教务处制表**

**二〇二五 年 五 月 二十 日**

**基于BERT的中文电影评论情感分析系统设计与实现**

学生 □□□ 指导老师 □□□，□□□

**摘要：**随着互联网的普及，在线电影评论已成为观众表达观影体验和意见的重要途径。情感分析技术的进步使得自动化分析用户评论的情感倾向成为可能。本研究基于BERT（Bidirectional Encoder Representations from Transformers）模型，构建了一个中文电影评论情感分析系统。系统涵盖数据采集与预处理、BERT模型微调、后端API开发、前端可视化展示及系统部署等关键环节。实验结果表明，基于BERT的情感分析模型在中文电影评论数据集上的准确率优于传统方法。该系统可广泛应用于电影推荐、观众反馈分析及市场趋势预测等领域。

**主题词：**情感分析；自然语言处理；BERT

**Evading Anti-Malware Engines with Deep Reinforcement Learning**

Cybersecurity

Student： Zhiyang Fang Adviser: □□□

**Abstract：**□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□.

**Key Words：**□□□□□；□□□□□；□□□□□；□□□□；□□□□□□；□□□□□□；□□□□□；□□□□□□。

**目 录**

[1 绪论 5](#_Toc195491243)

[1.1 研究背景 5](#_Toc195491244)

[1.2 国内外研究现状 5](#_Toc195491245)

[1.3 论文主要工作 6](#_Toc195491246)

[1.4 论文组织与结构 6](#_Toc195491247)

[2 背景知识简介 8](#_Toc195491248)

[2.1 NLP 8](#_Toc195491249)

[2.2 BERT 8](#_Toc195491250)

[2.3 MongoDB 8](#_Toc195491251)

[2.4 本章小结 9](#_Toc195491252)

[3 系统分析与设计 10](#_Toc195491253)

[3.1 系统需求分析与建模 10](#_Toc195491254)

[3.1.1 系统需求概述 10](#_Toc195491255)

[3.1.2 需求清单 10](#_Toc195491256)

[3.1.3 系统数据分析 11](#_Toc195491257)

[3.1.4 系统功能分析 12](#_Toc195491258)

[3.2 系统设计与建模 12](#_Toc195491259)

[3.2.1 系统设计概述 12](#_Toc195491260)

[3.2.2 系统数据库设计 12](#_Toc195491261)

[3.2.3 系统用户界面设计 13](#_Toc195491262)

[3.2.4 系统API接口设计 13](#_Toc195491263)

[4 系统详细设计与实现 16](#_Toc195491264)

[4.1 实现环境与工具的简要说明 16](#_Toc195491265)

[4.1.1 硬件环境 16](#_Toc195491266)

[4.1.2 实现工具 16](#_Toc195491267)

[4.2 用户登录模块实现 17](#_Toc195491268)

[4.2.1 用户认证实现 17](#_Toc195491269)

[4.2.2 密码加密实现 18](#_Toc195491270)

[4.3 电影展示模块实现 18](#_Toc195491271)

[4.3.1 电影数据集爬取 19](#_Toc195491272)

[4.3.2 数据清洗 20](#_Toc195491273)

[4.3.3 数据展示实现 21](#_Toc195491274)

[4.3.4 电影数据分析实现 22](#_Toc195491275)

[4.4 电影评论分析模块实现 23](#_Toc195491276)

[4.4.1 电影评论数据爬取 23](#_Toc195491277)

[4.4.2 分词处理 23](#_Toc195491278)

[4.4.3 数据格式转换 24](#_Toc195491279)

[4.4.4 评论数据展示实现 25](#_Toc195491280)

[4.4.5 评论数据分析实现 25](#_Toc195491281)

[4.5 情感分析模块实现 25](#_Toc195491282)

[4.5.1 情感分析实现 26](#_Toc195491283)

[4.5.2 情感趋势分析 27](#_Toc195491284)

[4.5.3 词云图 28](#_Toc195491285)

[4.6 用户模块实现 29](#_Toc195491286)

[4.6.1 管理员页面实现 29](#_Toc195491287)

[4.6.2 普通用户页面实现 30](#_Toc195491288)

[4.7 系统测试 31](#_Toc195491289)

[4.7.1 系统部署与优化 31](#_Toc195491290)

[4.7.2 性能优化与负载测试 32](#_Toc195491291)

[4.7.3 功能测试 33](#_Toc195491292)

[4.7.4 用户测试 34](#_Toc195491293)

[4.8 本章小结 34](#_Toc195491294)

[5 实验方法 35](#_Toc195491295)

[5.1 数据集简介 35](#_Toc195491296)

[5.2 BERT模型训练 35](#_Toc195491297)

[5.2.1 模型选择 35](#_Toc195491298)

[5.2.2 超参数调整 35](#_Toc195491299)

[5.2.3 损失函数构建 35](#_Toc195491300)

[5.2.4 模型优化 35](#_Toc195491301)

[5.3 BERT情感分析方法 36](#_Toc195491302)

[5.3.1 情感分类 36](#_Toc195491303)

[5.3.2 情感判别架构 36](#_Toc195491304)

[5.3.3 综合分析 37](#_Toc195491305)

[5.4 BERT模型评估 37](#_Toc195491306)

[5.4.1 混淆矩阵 37](#_Toc195491307)

[5.4.2 交叉验证 38](#_Toc195491308)

[5.4.3 模型指标 38](#_Toc195491309)

[6 工作总结和心得体会 39](#_Toc195491310)

[6.1 工作总结 39](#_Toc195491311)

[6.2 心得体会 39](#_Toc195491312)

[6.3 图片插入和自动题注编号 39](#_Toc195491313)

[6.4 表格插入和自动题注编号 39](#_Toc195491314)

[6.5 公式和自动编号 40](#_Toc195491315)

[参考文献 41](#_Toc195491316)

[致 谢 44](#_Toc195491317)

# 绪论

* 1. 研究背景

随着国民生活水平不断提高，人民对美好生活的向往愈加浓烈。娱乐方式呈现出多样化趋势，其中走进电影院观影成了这其中必不可少的一种娱乐方式[1]。但是如何创作出吸引人的好电影也就成为了令不少电影导演头疼的问题，同时在众多院线大片中如何选择一部好电影进行观看也是观影人员需要思考的问题。随着互联网和社交媒体的快速发展，电影影评已经成为观众表达观影体验的重要渠道[2, 3]。海量的评论文本数据蕴含着丰富的情感信息，对电影制作方、发行方和观众都具有重要的价值。然后人工分析这些评论不仅耗时费力，而且难以实现大规模处理。因此，利用自然语言处理技术自动分析电影评论的情感倾向成为了研究热点[4, 5]。

目前，许多评论分析系统利用自然语言处理技术进行评论情感分析，以实现自动化的检测大众的情绪及意见。其中，利用预训练语言模型进行文本情感分析是比较先进的技术，可以高效地实现自动化的评论理解和分类。其中BERT（Bidirectional Encoder Representations from Transformers）是一种基于Transformer架构的先进语言表示模型，能够通过上下文感知的深度双向编码，有效捕捉文本中的语义和情感特征，完全能够支持中文电影评论情感分析系统的实现。

本项目旨在设计和实现一个基于BERT的中文电影评论情感分析系统，利用自然语言处理技术对中文电影评论进行分析，以实现自动化的检测观众情绪和反馈。该系统将通过对电影影评数据抓取和电影影评情感分析两个模块实现对电影影评的自动化分析。通过本项目研究，可以提高行业人员分析效率和观众人群选择适合自己电影的概率，为电影从业人员提供更高效、更精准的数据分析服务，为观众人群提供更全面的电影分析结果，帮助他们选择适合自己的电影，有助于提升电影从业人员制作更符合大众口味电影的能力和提升观众的观影体验，从而提升人民生活满足感和幸福感。

* 1. 国内外研究现状

在过去的几十年里，机器学习和自然语言处理领域的迅速发展促进了各种文本分析系统的出现。这些系统采用了一系列先进的技术，例如CNN、LSTM和BERT等，以实现对大量文本进行自动化处理和分析。

国外早期情感分析主要基于监督学习和情感词典方法：Pang et al. (2002)首次将机器学习应用于电影评论情感分类，比较了SVM、朴素贝叶斯和最大熵模型在二元情感分类（正面/负面）上的表现。研究发现，使用unigram特征时SVM表现最佳（准确率约82.9%），为后续文本分类任务奠定了基础[6]。情感词典：Taboada et al. (2011)开发了基于SentiWordNet的词典方法，通过计算文本中情感词的极性得分（正/负/中性）来预测整体情感。该方法在商品评论分析中达到约74%的准确率，但面临否定词处理（如"not good"）和上下文依赖的挑战[7]。

随着深度学习兴起，神经网络显著提升了情感分析性能：Kim (2014)提出使用静态词向量（Word2Vec）和卷积神经网络（CNN）提取文本局部特征。在MR电影评论数据集上，单通道CNN达到81.5%的准确率，多通道CNN提升至82.7%[8]。Tang et al. (2015)设计了基于长短期记忆网络（LSTM）的模型，通过建模词序依赖关系提升长文本分析效果。在Stanford Sentiment Treebank数据集上，LSTM比CNN的准确率提高约3%[9]。Vaswani et al. (2017)提出的Transformer架构通过自注意力机制（Self-Attention）动态计算词间权重，在机器翻译任务中取得突破，为后续BERT等模型奠定基础[10]。

最后在大模型时代，情感分析更是实现了质的飞跃：Devlin et al. (2018)提出的双向Transformer模型通过掩码语言建模（MLM）学习上下文相关表示。在SST-2情感数据集上，BERT-base达到92.7%的准确率，较LSTM提升约10%[11]。Zadeh et al. (2018)构建的CMU-MOSEI数据集包含23,500条YouTube视频片段，涵盖文本、音频和视觉模态。实验表明，多模态融合模型（如TFN）比单模态文本分析的F1值提高15%[12]。

总结来说，各种自然语言处理技术为情感分析系统的发展提供了基础。这些技术的综合应用使得自动化评估电影影评成为可能。随着机器学习和自然语言处理领域的进一步发展，情感分析系统的性能和准确性将得到持续提高。与此同时，中文语料和多模态分析也将在未来的研究中得到更多的关注。

* 1. 论文主要工作

本论文的主要工作是实现了一个针对电影评论数据的情感分析系统。在项目中，我构建了一个基于BERT（Bidirectional Encoder Representations from Transformers）模型的中文电影评论情感分析系统。此项目不仅实现了对文本情感倾向性的识别，还通过一系列功能模块的设计与实现，提供了全面的用户交互体验和后台管理能力。

根据不同的功能需求，可以将系统划分为以下六个主要功能模块：

（1）用户管理模块：包括用户注册、用户登录、个人信息管理等功能，新用户通过填写用户名、邮箱和密码进行注册，已注册用户通过用户名和密码进行登录，用户可以查看和修改个人信息，包括头像、邮箱等。

（2）电影管理模块：包括电影信息展示、电影详情等功能，能够展示电影的基本信息，包括标题、类型等，能够展示电影的详细信息，包括简介、评分、评论等。

（3）评论管理模块：包括评论查看、编辑、删除等功能，用户可以对电影发表评论和评分，页面也能够展示电影的所有评论，支持分页和排序。

（4）情感分析模块：包括评论情感分析、情感趋势分析等功能，通过使用BERT模型分析评论的情感倾向，展示电影评论的情感趋势变化。

（5）数据分析模块：包括用户行为分析、电影评分分析、电影数据分析等功能，能够具体分析用户的活跃度、电影数据类型分布等。

（6）系统管理模块：包括用户权限管理、数据管理等功能，能够实现RBAC的页面展示，管理员能够管理用户的权限和状态等。

* 1. 论文组织与结构

本论文主体部分由六部分组成，各部分作用分别为：

第一部分：绪论。本章介绍本项目的研究背景，陈述本项目相关的国内外研究现状，最后对本文的主要工作作出介绍。

第二部分：背景知识简介。主要介绍了项目涉及的相关技术知识背景。

第三部分：系统概要设计。主要介绍了系统设计的概要情况。

第四部分：系统详细设计与实现。主要介绍了系统的详细设计流程和实现的具体方法，包括具体功能的实际效果和代码说明。

第五部分：实验方法。主要介绍了布署微调后的BERT模型以实现对中文电影评论进行情感分析；以及开发Web应用以实现模型的实际应用。

第六部分：讨论和体会。本章写出了项目整个开发过程中的感悟和反思，总结了本篇论文的工作和主要贡献并对现存问题和未来发展发表了自己的看法。

为了方便读者更进一步了解本文设计的系统项目，也便于有兴趣的读者进行二次开发，本文附录部分给出了项目的需求规格说明书、设计文档、软件测试文档、用户文档说明书以及项目的所有源码。

# 背景知识简介

* 1. NLP

自然语言处理（NLP）作为人工智能领域的重要分支，其核心任务主要围绕语言的理解与生成展开。在文本分类方面，Devlin等人（2019）提出的BERT模型通过预训练双向Transformer架构，显著提升了情感分析和主题识别的性能[11]。序列标注任务如命名实体识别和词性标注，则受益于Vaswani等人（2017）提出的Attention机制[10]，该技术被证明在长序列建模中具有显著优势。文本生成领域的重要突破来自Radford等人（2018）的GPT模型[13]，其自回归架构为机器翻译和自动摘要提供了新范式。问答系统的发展则体现在Brown等人（2020）的工作中[14]，他们证明了大语言模型在few-shot问答任务中的卓越表现。

在应用场景方面，Liu等人（2019）优化的RoBERTa模型被广泛应用于搜索引擎的排名优化[15]。智能客服系统通过整合Jurafsky和Martin（2021）提出的语言处理框架[16]，实现了更自然的意图识别和对话生成。金融领域的情感分析应用则借鉴了BERT等预训练模型，有效预测市场趋势。医疗文本处理方面，基于Transformer的模型在临床记录实体抽取任务中展现出显著优势。

未来，NLP技术将继续向多模态、低资源适应等方向发展，同时需要解决模型可解释性等重要挑战。这些发展都建立在Devlin、Vaswani等学者提出的基础架构之上，体现了NLP领域理论与应用的紧密结合。

* 1. BERT

BERT（Bidirectional Encoder Representations from Transformers）是由Devlin等人[11]于2018年提出的突破性预训练语言模型，其核心创新在于采用双向Transformer编码器结构和掩码语言建模（MLM）任务，使模型能够同时利用上下文信息进行表征学习。与传统的单向语言模型不同，BERT通过随机遮盖输入文本中的部分词汇（通常为15%）并要求模型预测这些被遮盖词，从而学习深层的上下文相关词表征[1]。研究表明，BERT-base（12层）和BERT-large（24层）在11项NLP基准任务上均取得state-of-the-art性能，特别是在GLUE基准测试中平均提升7.7%[11]。后续研究如Liu等人[15]提出的RoBERTa通过优化训练策略（如动态掩码、更大批次训练）进一步提升了模型性能。在应用层面，BERT及其变体已广泛应用于问答系统[17]、文本分类[18]和命名实体识别[19]等任务。然而，BERT仍存在计算资源消耗大[20]和领域适应能力有限[21]等挑战，这促使研究者探索知识蒸馏[22]和参数高效微调[23]等优化方法。当前，基于BERT的架构仍是NLP领域的重要基础，其设计理念深刻影响了后续大语言模型的发展[24]。

* 1. MongoDB

MongoDB作为一款主流的NoSQL数据库系统，其文档型数据存储架构在应对非结构化数据管理需求方面展现出显著优势[25]。该数据库采用BSON（Binary JSON）格式存储数据，通过灵活的文档模型支持嵌套数据结构，有效解决了传统关系型数据库在模式变更方面的局限性[26]。MongoDB的核心特性包括分片集群架构实现水平扩展、基于WiredTiger存储引擎的高性能读写操作，以及丰富的聚合管道功能支持复杂数据分析[27]。在索引优化方面，MongoDB支持多种索引类型（如单字段、复合、多键和地理空间索引），显著提升了查询效率[28]。安全机制上，其提供了基于角色的访问控制（RBAC）和传输层加密（TLS）等企业级安全特性[29]。实际应用中，MongoDB在物联网时序数据存储、内容管理系统和实时分析场景中表现突出，但面对复杂事务处理时仍存在性能挑战。最新版本通过引入分布式事务支持和增强的聚合操作符持续优化其功能体系，使其在当今多模态数据管理领域保持竞争力[30]。

* 1. 本章小结

本章介绍了系统中使用的算法和模型。

# 系统分析与设计

* 1. 系统需求分析与建模
     1. 系统需求概述

基于BERT的中文电影评论情感分析系统旨在通过BERT预训练模型对中文电影评论进行情感分析，在通过爬取互联网上公开的电影数据后，系统需要能够对中文电影评论数据进行数据清洗、评论分词，随后通过BERT模型进行情感分析，计算评论情感指数，将评论数据可视化，最后通过结合电影数据综合展示。

* + 1. 需求清单

表 3‑1 需求功能清单

|  |  |  |
| --- | --- | --- |
| 模块 | 功能 | 功能描述 |
| 用户管理 | 用户注册 | 新用户通过填写用户名、邮箱和密码进行注册 |
| 用户登录 | 已注册用户通过用户名和密码进行登录 |
| 个人信息管理 | 用户可以查看和修改个人信息，包括头像、邮箱等 |
| 电影管理 | 电影信息展示 | 展示电影基本信息，包括标题、导演、演员、类型等 |
| 电影搜索 | 支持按标题、类型等条件搜索电影 |
| 电影详情 | 展示电影的详细信息，包括简介、评分、评论等 |
| 评论管理 | 发表评论 | 用户可以对电影发表评论和评分 |
| 评论列表 | 展示电影的所有评论，支持分页和排序 |
| 评论删除 | 用户可以删除自己的评论 |
| 情感分析 | 评论情感分析 | 使用BERT模型分析评论的情感倾向 |
| 情感趋势 | 展示电影评论的情感趋势变化 |
| 数据分析 | 用户行为分析 | 分析用户活跃度和行为模式 |
| 电影评分分析 | 分析电影评分分布和趋势 |
| 评论情感分析 | 分析评论情感分布和变化 |
| 系统管理 | 用户权限管理 | 管理员可以管理用户权限和状态 |
| 数据备份 | 支持系统数据的备份和恢复 |

* + 1. 系统数据分析

1. 用户数据模型
2. 电影数据模型
3. 评论数据模型
   * 1. 系统功能分析

系统功能点关系，流程图

* 1. 系统设计与建模
     1. 系统设计概述

系统采用前后端分离架构，主要包含以下组件：

1. 前端：Vue.js 3 + Ant Design Vue

2. 后端：FastAPI + MongoDB

3. 机器学习：BERT模型

4. 部署：Docker + Nginx

* + 1. 系统数据库设计

数据库设计采用MongoDB文档型数据库，实现了灵活的数据模型。用户表设计实现了用户基本信息和权限管理，支持用户数据的快速查询。电影表设计实现了电影信息的存储，包括基本信息、评分、评论等，支持多维度查询。评论表设计实现了评论数据的存储，包括内容、情感分析结果等，支持评论的快速检索。系统还实现了数据库索引优化、查询性能优化等，提高了数据访问效率。

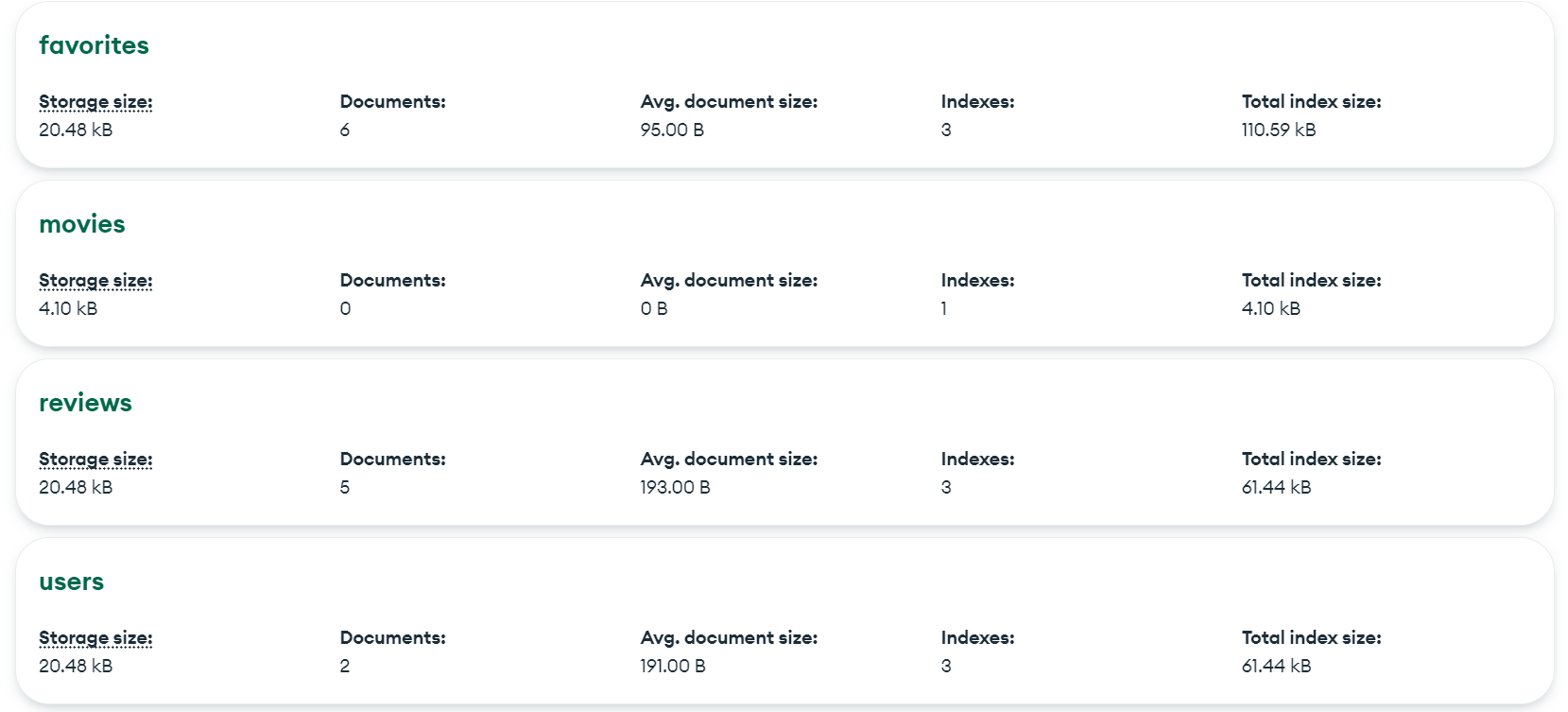


图 3‑1 MongoDB集合列表

包括如下集合：

用户集合 (users)：\_id, user\_id, username, email, password, avatar, role, status, created\_at。

电影集合 (movies)：\_id, title, director, actors, genres, release\_date, rating, poster, description, created\_at。

评论集合 (reviews)：\_id, movie\_id, user\_id, content, rating, sentiment, created\_at。

喜好集合 (favorites)：\_id, user\_id, movie\_id, created\_at。

* + 1. 系统用户界面设计

系统用户页面主要包括四个部分：

（1）登录界面：用户名/密码输入框、登录按钮、注册链接选项

（2）电影列表界面：电影卡片网格布局、搜索和筛选功能、分页控件、排序选项

（3）电影详情界面： 电影海报、基本信息展示、评分和评论、评论表单

（4）数据分析界面：情感分析图表、用户行为分析、评分分布图、评论趋势图

* + 1. 系统API接口设计

系统接口采用FastAPI设计，主要接口包括：

（1）用户相关接口



图 3‑2 用户相关接口

（2）电影相关接口

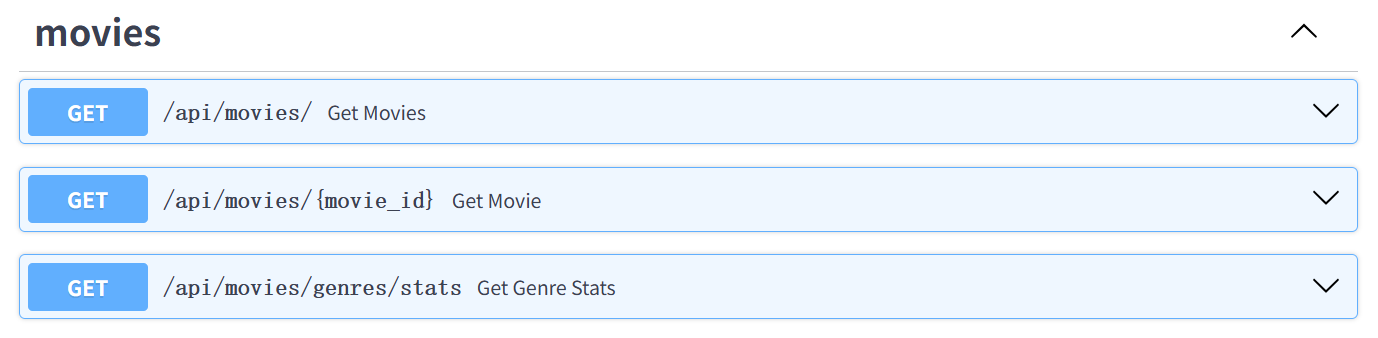


图 3‑3 电影相关接口

（3）评论相关接口

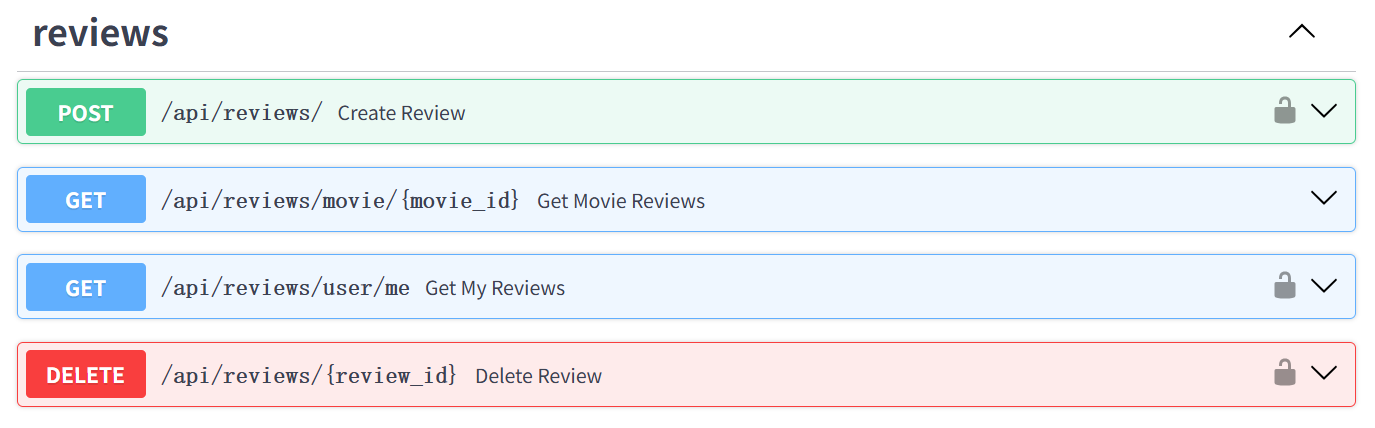


图 3‑4 评论相关接口

（4）喜好相关接口

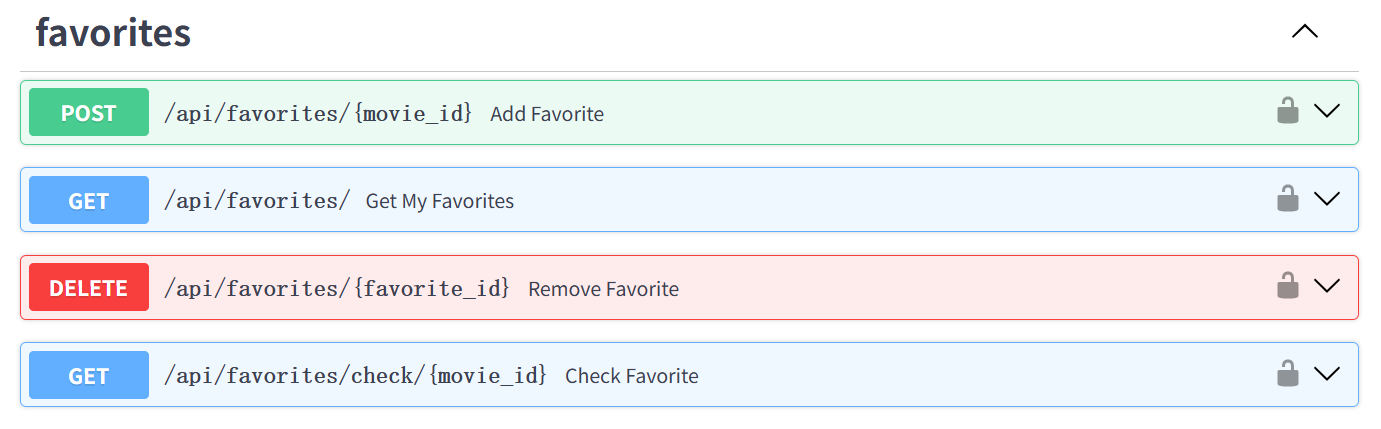


图 3‑5 喜好相关接口

（5）分析相关接口



图 3‑6 分析相关接口

# 系统详细设计与实现

* 1. 实现环境与工具的简要说明
     1. 硬件环境

表 4‑1 硬件环境

|  |  |  |  |
| --- | --- | --- | --- |
| CPU | RAM | GPU | OS |
| Intel ® Core(TM) i7-12650H 2.30GHz | 32.0 GB | NVIDIA RTX 4060 6G | Windows 10 64位操作系统 |

* + 1. 实现工具

表 4‑2 实现工具列表

|  |  |  |
| --- | --- | --- |
| **工具名称** | **版本** | **用途** |
| WebStorm | 2024.3 | 前端开发和设计主要IDE |
| MongoDB | 8.0.6 | 数据库 |
| PyCharm | 2024.1 | 后端开发以及网页爬虫主要IDE |
| Python | 3.11.9 | 后端开发主要使用的编程语言 |
| MongoDB Compass | 1.46 | 数据库管理工具 |
| Node.js | 22.12.0 | 前端开发所需环境 |
| TypeScript | 5.8.3 | 前端开发主要使用的编程语言 |
| Docker | 28.0.1 | 容器化部署 |
| Git | 2.41 | 版本控制 |

* 1. 用户登录模块实现

用户管理模块作为系统的核心组成部分，实现了完整的用户生命周期管理。在用户注册环节，系统采用Pydantic进行数据验证，确保用户信息的完整性和有效性，同时通过bcrypt算法对密码进行加密存储，保障用户数据安全。登录认证采用JWT（JSON Web Token）机制，实现了无状态的用户会话管理，提高了系统的可扩展性。个人信息管理功能支持用户头像上传、邮箱修改等操作，并实现了实时数据同步。系统还实现了基于角色的访问控制（RBAC），区分普通用户和管理员权限，确保系统资源的安全访问。在用户体验方面，实现了自动登录等便捷功能，同时提供了完整的错误处理和友好的用户提示。

* + 1. 用户认证实现

使用JWT（JSON Web Token）进行身份认证，在请求头中携带token，验证token的有效性，从token中提取用户信息，查询用户是否存在，返回用户对象或抛出异常。

* + 1. 密码加密实现

密码加密使用bcrypt算法进行密码加密，自动生成salt，支持密码强度验证，防止彩虹表攻击，支持密码更新。

* 1. 电影展示模块实现

电影管理模块实现了电影数据的全生命周期管理。系统通过异步爬虫技术从爱奇艺等平台获取电影数据，包括电影基本信息、演员阵容、剧情简介等。数据清洗环节采用正则表达式和自然语言处理技术，对爬取的数据进行标准化处理，确保数据质量。电影展示采用响应式网格布局，支持按类型、评分等多维度筛选，并实现了电影海报的懒加载和渐进式加载，优化了页面性能。电影详情页面集成了评论系统、评分功能，并提供了电影推荐算法，基于用户历史行为和电影特征实现个性化推荐。系统还实现了电影数据的实时更新机制，确保数据的时效性。

* + 1. 电影数据集爬取

数据爬取过程如图所示，具体流程如下（此处以爬取爱奇艺影评为例）：

1、 从接口获取电影的基本信息。这里以地区分类，分语种（如华语、英语等）获取当下热门电影的信息。

2、 电影编号一般是视频播放软件对一部电影的唯一识别码，通过电影编号。可以找到该电影的评论接口。

3、 通过评论接口，可以请求评论数据，但是一次请求并不能获取足够多的数据，需要更改接口的lastId参数（不同视频放映软件该参数不同，爱奇艺中是lastId）多次请求以获取足量数据。

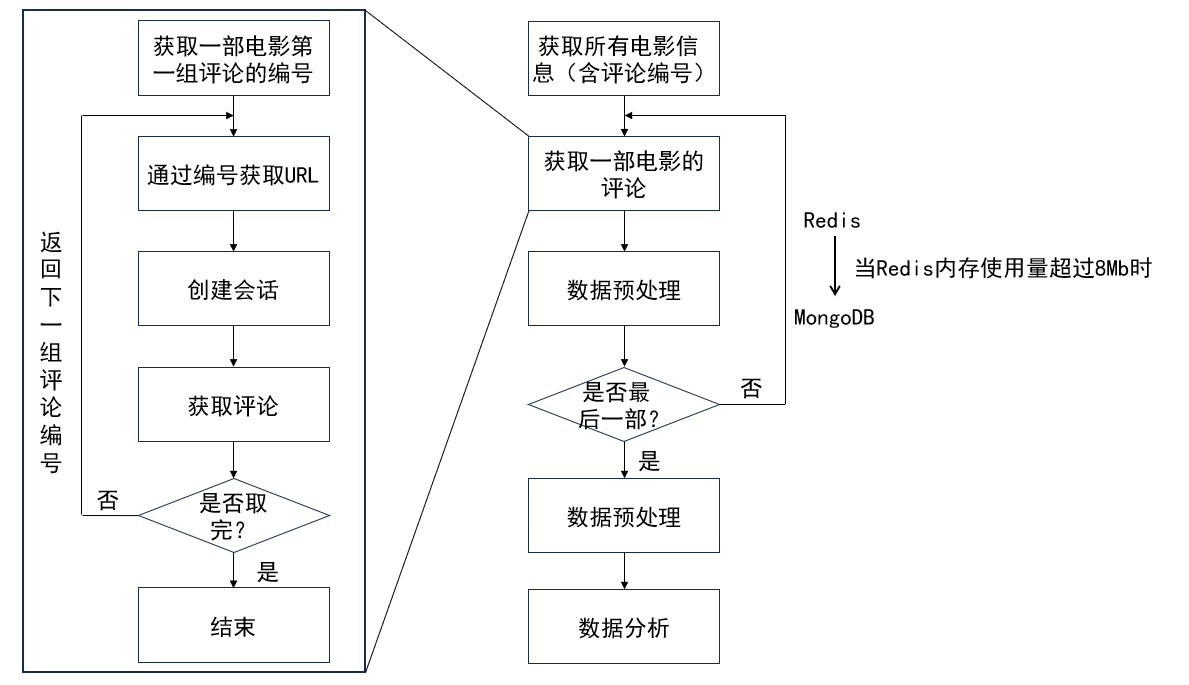


图 4‑1 爬虫流程

步骤2和步骤3的代码实现如下，其中lastId可以在当前接口请求到的数据中找到，并通过lastId可以找到下一组评论的接口。此处用异常的方式结束爬取，当全部评论爬取完后，会自动中断while循环。

在上述代码中，创建了requests的session对象，该对象在while循环外持久创建，并以参数的形式传递给getMovieinfo函数（具体实现见下述代码）。这样做可以防止过多建立连接，减少请求次数，并加快爬取速度。

最后将数据封装在字典内，该字典可作为mongoDB的一个文档直接存储，实现如下：

爬虫使用aiohttp进行异步HTTP请求，设置User-Agent模拟浏览器，支持分页爬取，异常处理和重试机制，数据去重和验证。

* + 1. 数据清洗

在数据获取阶段，取得了个数如图4-2所示的一组数据，在本阶段，会对这些数据进行预处理。

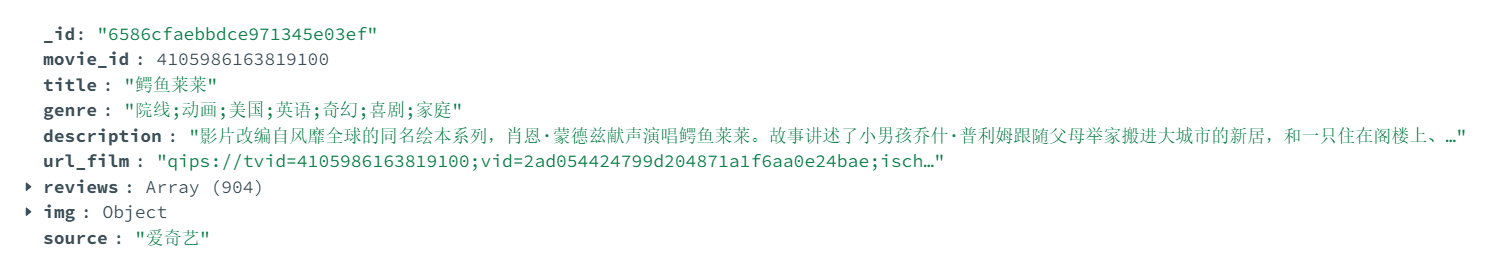


图 4‑2 数据预处理展示

数据清洗包括去除多余空格、标准化分隔符、处理缺失值、统一数据格式、验证数据有效性等。

* + 1. 数据展示实现

电影列表展示使用Ant Design Vue的Card组件、响应式网格布局、图片懒加载、卡片hover效果、 点击跳转详情。

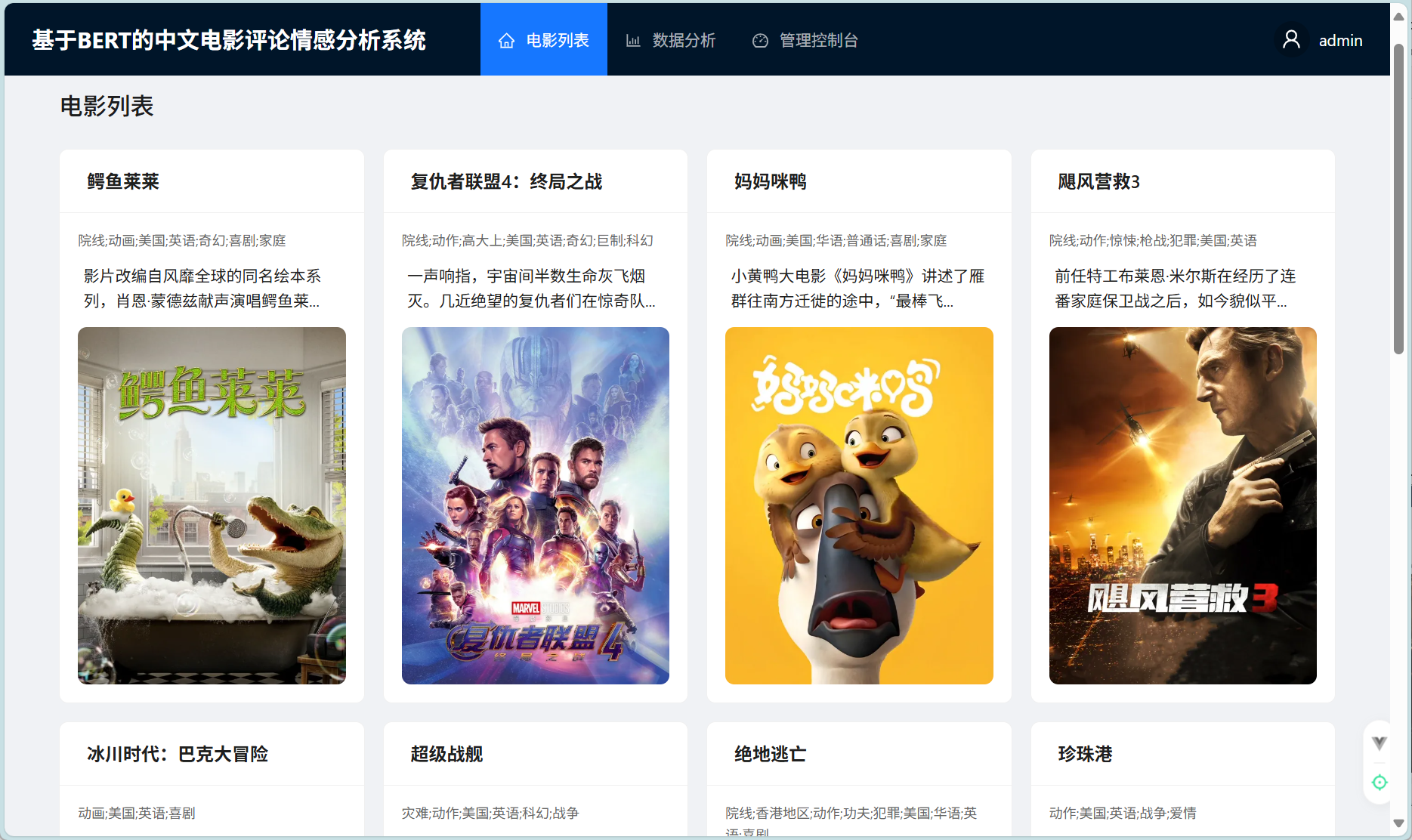


图 4‑3 电影数据展示页面

* + 1. 电影数据分析实现

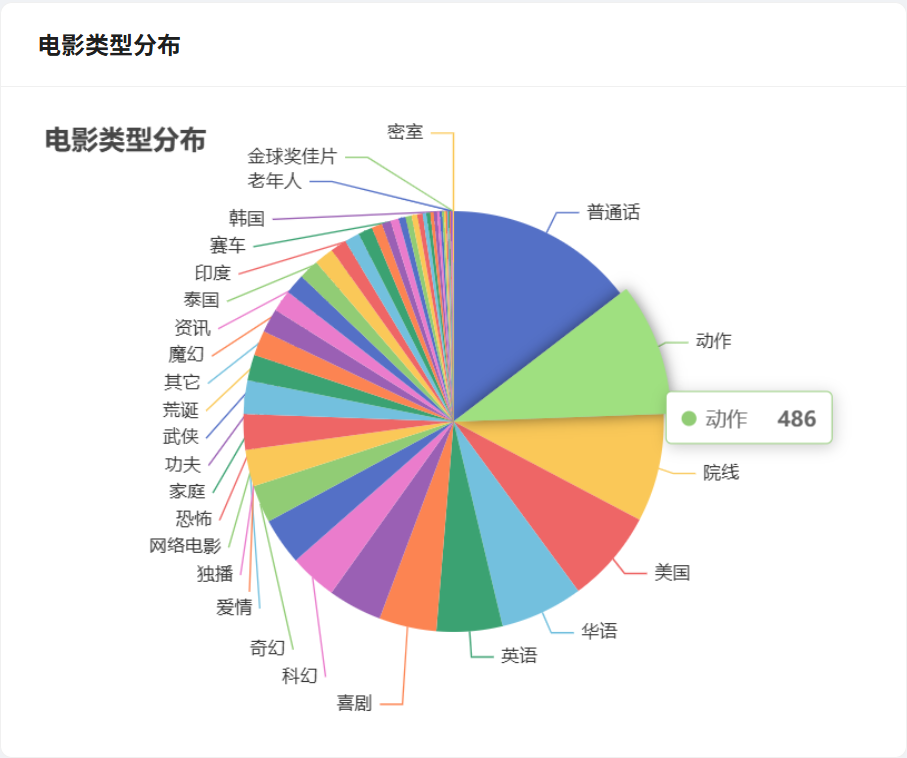


图 4‑4 电影类型分布可视化

* 1. 电影评论分析模块实现

评论管理模块实现了用户评论的完整生命周期管理。系统采用BERT模型对用户评论进行实时情感分析，将评论分为积极、中性和消极三类，并计算情感置信度。评论展示采用分页加载和虚拟滚动技术，优化了大数据量下的性能表现。在数据安全方面，实现了评论的审核机制和举报功能，确保评论内容的质量和安全。

* + 1. 电影评论数据爬取

评论爬取支持分页爬取、异步请求处理、自动停止机制、数据去重、异常处理。

* + 1. 分词处理

分词处理使用jieba分词器，加载停用词表、去除标点符号、保留关键词、支持自定义词典。

部分分词结果：

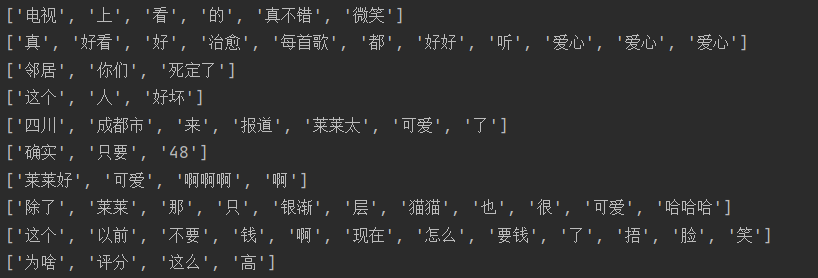


图 4‑5 分词效果展示

* + 1. 数据格式转换

首先加载预训练tokenizer，文本编码、填充和截断、转换为PyTorch张量、批处理支持

同时在预处理时将数据格式统一。例如，某些电影的类别用字符串表示，这里将其改为列表。效果如图：

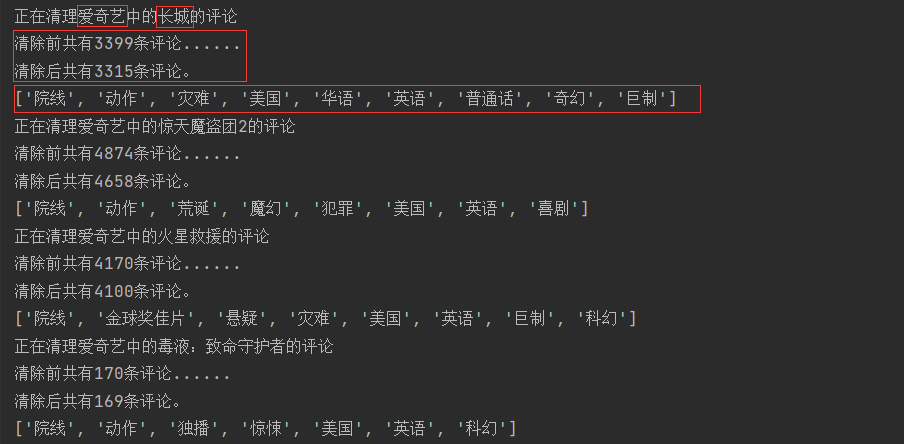


图 4‑6 数据格式处理示例

* + 1. 评论数据展示实现



图 4‑7 评论数据展示页面

* + 1. 评论数据分析实现

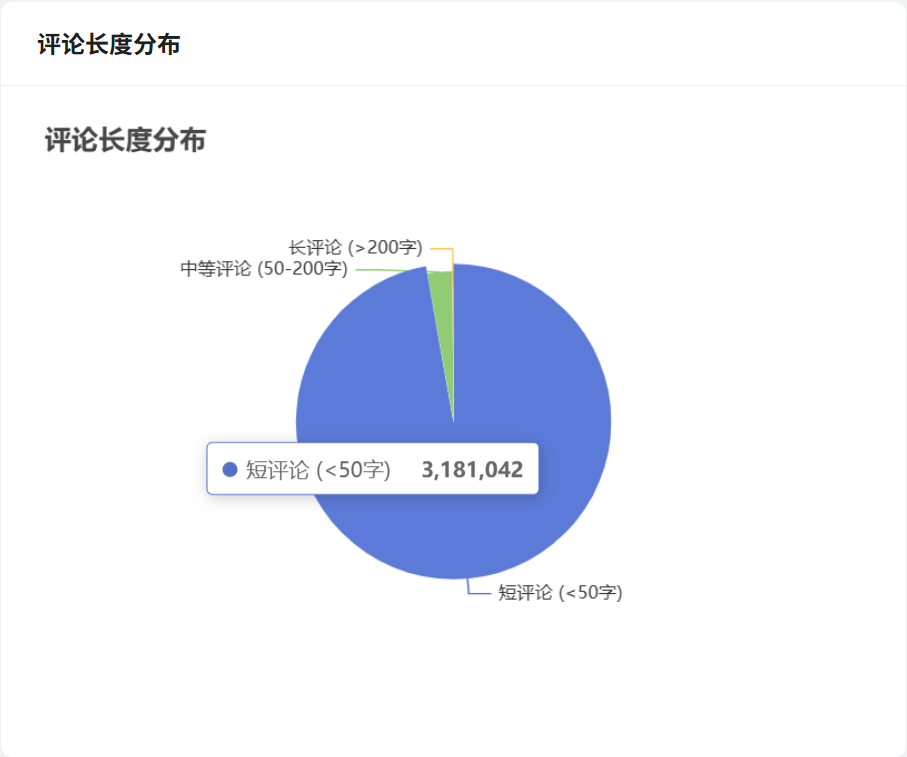


图 4‑8 评论数据分析实现

* 1. 情感分析模块实现

情感分析模块是系统的核心技术组件，采用BERT预训练模型进行中文文本情感分析。系统对BERT模型进行了针对性的微调，使用电影评论数据集进行训练，提高了模型在电影评论领域的表现。情感分析过程包括文本预处理、分词、特征提取和分类预测等步骤，实现了对评论情感的精准识别。系统还实现了情感分析的可视化展示，包括情感分布饼图、情感趋势折线图等，帮助用户直观理解评论情感走向。在性能优化方面，实现了模型推理的批处理机制和缓存策略，提高了系统的响应速度。

* + 1. 情感分析实现

情感分析包括实现加载预训练模型、模型推理模式、批处理支持、情感标签映射、置信度计算。

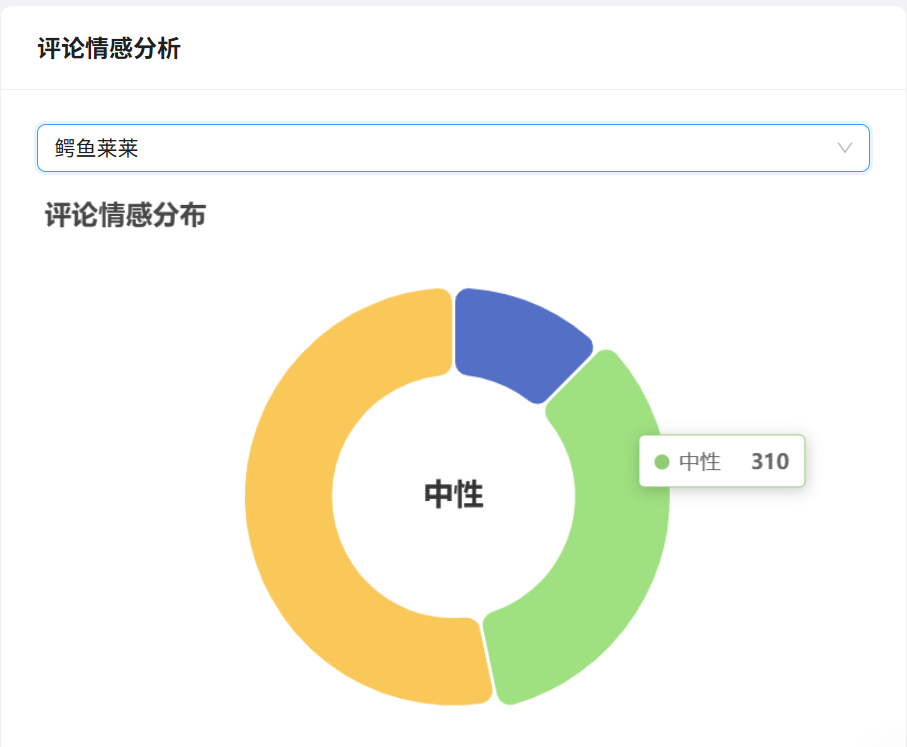


图 4‑9 情感分析可视化

* + 1. 情感趋势分析

情感趋势步骤包括时间序列分组、情感计数统计、数据排序、趋势可视化、异常检测。

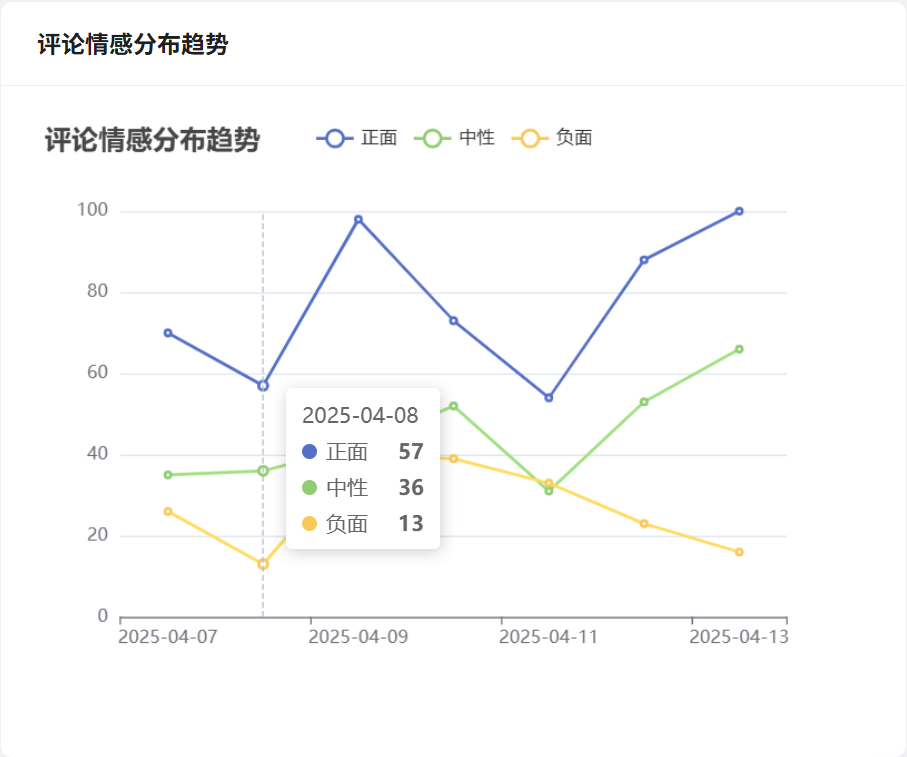


图 4‑10 情感分布趋势可视化

* + 1. 词云图

词云图生成包括文本合并、分词处理、词频统计、自定义样式、图片生成。



图 4‑11 评论词云图可视化

* 1. 用户模块实现
     1. 管理员页面实现

管理员页面使用Ant Design布局，包括侧边栏导航、图标集成、路由视图、响应式设计。页面包括用户管理，用户活跃度分析，电影数据编辑等。



图 4‑12 管理员控制页面

* + 1. 普通用户页面实现

普通用户页面实现了表单布局、数据绑定、文件上传、表单验证、响应式设计等功能，页面包括个人信息，账号设置，我的收藏，我的评论。

* 1. 系统测试
     1. 系统部署与优化

系统部署定义了系统所用的服务，包括端口映射、环境变量、依赖关系、数据持久化等部署。

* + 1. 性能优化与负载测试

性能测试说使用并发请求模拟，实现了响应时间统计、错误率计算、资源监控。

* + 1. 功能测试

功能测试包括单元测试、集成测试、接口测试、边界测试、异常测试。

* + 1. 用户测试

用户测试主要关注以下几个方面：界面易用性、功能完整性、响应速度、错误处理、数据准确性

* 1. 本章小结

本章详细介绍了系统的实现过程，包括 开发环境的搭建、核心功能模块的实现、系统测试和优化、部署方案等内容。系统实现了预期的所有功能，并通过测试验证了其可靠性和性能。

# 实验方法

* 1. 数据集简介

数据集来自从爱奇艺爬取的1018部电影数据，其中包含4324956条影评数据等。在数据存储和预处理结束后，通过聚合查询对数据进行初步分析。得到以下信息：从爱奇艺获取3272529评论，超过2000条评论的电影有683个，占电影总数的33.8%。数据的格式可参照图5-1，以JSON数据的格式存储。其中，电影的评论信息是一个列表，在进行了数据变换后，评论也以列表的形式存储。图片在经过转换过后，也可以以文本的形式存储，具体情况如图5-2，其中type是图片编码前的格式。



图 5‑1 数据集结构

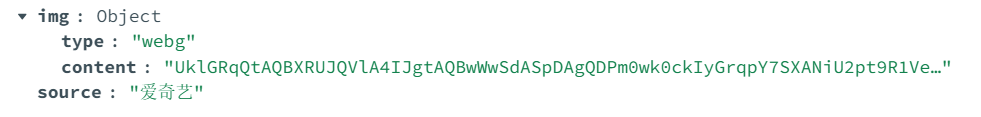


图 5‑2 图片存储格式

* 1. BERT模型训练
     1. 模型选择

选择BERT-base-chinese作为基础模型，支持中文处理、预训练效果好、计算资源需求适中。

* + 1. 超参数调整

超参数调整包括训练轮数、批次大小、学习率、权重衰减、日志记录

* + 1. 损失函数构建

损失函数包括交叉熵损失、标签处理、输出转换、损失计算、梯度更新

* + 1. 模型优化

模型优化包括学习率调整、批量大小优化、早停策略、模型集成等策略。

1. 学习率调整

预热阶段、衰减策略、动态调整

（2）批量大小优化

内存限制、训练速度、收敛效果

（3）早停策略

验证集监控、耐心值设置、模型保存

（4）模型集成

多模型训练、投票机制、权重融合

* 1. BERT情感分析方法
     1. 情感分类

将情感分为三类：

1. 积极 (positive)：表达喜爱、推荐观看、正面评价
2. 中性 (neutral)：客观描述、无明显倾向、混合评价
3. 消极 (negative)：表达不满、批评建议、负面评价
   * 1. 情感判别架构

（1）BERT模型：预训练参数、特征提取、上下文理解

（2）分类器：全连接层、激活函数、输出转换

（3）前向传播：输入处理、特征提取、分类预测

* + 1. 综合分析

结合以下因素进行综合分析：

（1）情感得分：模型输出、置信度、阈值判断

（2）上下文信息：评论内容、用户历史、电影信息

（3）时间趋势：评论时间、情感变化、异常检测

* 1. BERT模型评估
     1. 混淆矩阵

混淆矩阵评估包括模型评估模式、预测结果收集、真实标签收集等

* + 1. 交叉验证

交叉验证部分包括数据集划分、模型训练、性能评估、结果统计、方差分析等。

* + 1. 模型指标

主要评估指标：准确率 (Accuracy)，精确率 (Precision)，召回率 (Recall)，F1分数，ROC曲线等

工作总结和心得体会

* 1. 工作总结

本项目成功实现了基于BERT的中文电影评论情感分析系统，并且独立完成了从系统设计、开发到测试的整个过程。项目涵盖了从前端用户界面设计到后端服务逻辑实现的全方位内容，包括用户管理、电影信息管理、评论管理和数据分析等功能模块。通过实验验证了BERT模型在处理中文电影评论情感分析任务上的高效性和准确性，同时也展示了该系统在实际应用中的潜力和价值。

* 1. 心得体会

在整个项目的开发过程中，我深刻体会到自然语言处理技术的重要性以及其在提升信息理解和处理效率方面的巨大潜力。在面对复杂的中文语境时，选择合适的模型和算法尤为关键。同时，我也认识到数据预处理的质量直接影响到模型的效果，为此，在数据清洗、特征提取等前期准备工作中投入足够的时间和精力是非常必要的。此外，作为一个个人项目，虽然没有团队合作的挑战，但这也要求我在时间管理和自我监督方面更加严格，确保项目的顺利推进。这次经历让我学到了很多宝贵的知识和技能，也增强了我的独立解决问题的能力。

参考文献

1. 李佳, 实现人民对美好生活向往的路径. 党课参考, 2023(06): 第65-70页.
2. 王宜文与张云, 互联网时代的电影评论. 艺术评论, 2015(08): 第14-22页.
3. 司若, 黄莺与许婉钰, 近年来我国网络影评发展脉络研究. 当代电影, 2020(03): 第137-143页.
4. 张璐, 从Python情感分析看海外读者对中国译介文学的接受和评价：以《三体》英译本为例. 外语研究, 2019. 36(04): 第80-86页.
5. 刘玲玉与邓燕燕, 基于Python情感分析和批评隐喻的网络话语分析——以影片《流浪地球》中美德影评为例. 江苏大学学报(社会科学版), 2022. 24(03): 第76-88页.
6. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. EMNLP.
7. Taboada, M., Brooke, J., Tofiloski, M., et al. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics.
8. Kim, Y. (2014). Convolutional neural networks for sentence classification. EMNLP.
9. Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. EMNLP.
10. Vaswani, A., et al. (2017). Attention is all you need. NeurIPS.
11. Devlin, J., et al. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL.
12. Zadeh, A., et al. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion. ACL.
13. Radford, A., et al. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI.
14. Brown, T., et al. (2020). Language Models are Few-Shot Learners. NeurIPS.
15. Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv.
16. Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Pearson.
17. Rajpurkar P, et al. SQuAD: 100,000+ questions for machine comprehension of text. EMNLP 2016.
18. Sun C, et al. How to fine-tune BERT for text classification? CCL 2019.
19. Yamada I, et al. LUKE: Deep contextualized entity representations with entity-aware self-attention. EMNLP 2020.
20. Strubell E, et al. Energy and policy considerations for deep learning in NLP. ACL 2019.
21. Gururangan S, et al. Don't stop pretraining: Adapt language models to domains and tasks. ACL 2020.
22. Sanh V, et al. DistilBERT, a distilled version of BERT. NeurIPS 2019.
23. Houlsby N, et al. Parameter-efficient transfer learning for NLP. ICML 2019.
24. Raffel C, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR 2020.
25. Chodorow K. MongoDB: The Definitive Guide. O'Reilly, 2013.
26. Banker K. MongoDB in Action. Manning, 2011.
27. Sharma R. Mastering MongoDB Aggregation Framework. Packt, 2021.
28. Bradshaw S. MongoDB Indexing Strategies. Apress, 2015.
29. MongoDB Inc. Security Reference Architecture. Technical Report, 2022.
30. Sadalage P. NoSQL Distilled. Addison-Wesley, 2012.

**声 明**

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，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，也不包含为获得四川大学或其他教育机构的学位或证书而使用过的材料。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谢意。

本学位论文成果是本人在四川大学读书期间在导师指导下取得的，论文成果归四川大学所有，特此声明。

作者签名： 导师签名：

年 月 日

**学位论文使用授权书**

本学位论文作者完全了解四川大学有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或相关机构送交论文的原件、复印件和电子版，允许论文被查阅和借阅。本人授权四川大学将本学位论文的全部或部分内容编入有关数据库进行信息技术服务，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，并用于学术活动。

（涉密学位论文在解密后适用于本授权书）

作者签名： 导师签名：

年 月 日

致 谢

首先