湖州师范学院

2023 届毕业设计(论文)

课 题 名 称：基于深度学习的文本情绪分析及其应用

课 题 名 称（英文）：Text emotion analysis based on deep learning and its application

学 生 姓 名：缪岱烨 学 号：2019082215

专 业 名 称：计算机科学与技术

指 导 教 师：唐琦哲 职 称：讲师

所 在 学 院：信息工程学院

完 成 日 期：2023 年 3 月 1 日

基于深度学习的文本情绪分析及其应用

摘要：随着社交媒体和互联网的发展普及，网络文本数据快速增长，对文本进行情绪分析具有重要意义。而基于深度学习的文本情绪分析方法则可以通过学习大量的数据和语义信息，自动学习文本中的情感信息，情感分类的准确度已经得到极大的提升，但是基于情绪分析结果的应用确较为稀缺。

本项目通过获取知乎用户的文本浏览记录，并通过训练 LSTM 情绪分析模型来进行分析，并使用 LSTM 的文本情绪分析的结果作为 Deep Crossing 推荐系统算法的输入提高对用户点击率预测的准确度，主要工作包括：（1）使用爬虫工具对知乎用户的浏览信息进行爬取。该过程使用 selenium 爬虫框架通过模拟浏览器操作，从而进入用户个人主页，对其近期浏览的问题、问题内容等一系列信息进行爬取来作为原始数据。（2）对 LSMT 文本情绪分类模型进行训练。利用开源的带有情绪标签的文本数据对模型进行训练，使该模型可以对长文本中所隐含的情绪因素进行较为准确的判断。（3）使用 Deep Crossing 推荐系统算法来预测知乎用户对某一回答的点击率。使用 LSTM 对文章进行情绪分析，同时用文章分析结果与使用 TF-IDF 算法提取出的文章关键字进行合并，将合并后的数据作为 Deep Crossing 算法的输入，对该算法模型进行训练。

本项目在开源情绪文本数据集和知乎用户浏览记录数据集上进行试验，试验结果表明，LSTM 情绪分析模型的准确为81%。而加入了情绪分析结果作为输入后，deep crossing 算法的预测准确率提升了0.9%。

关键词：文本情感分析，自然语言处理，LSTM 文本情绪分类模型，deep crossing 算法

Text emotion analysis based on deep learning and its application

Abstract: With the popularity of social media and the Internet, people use text communication more and more frequently in daily life, so the demand for text emotion analysis is also increasingly urgent. Traditional text emotion analysis methods are mainly based on rules and statistical methods, which have the problems of low accuracy and inability to adapt to complex semantics. The text emotion analysis method based on deep learning can automatically learn the emotion information in the text by learning a lot of data and semantic information. The accuracy of emotion classification has been greatly improved. However, applications based on emotion analysis are scarce.

This project obtains Zhihu users' text browsing records, trains LSTM emotion analysis model to analyze, and uses LSTM analysis results as the input of Deep Crossing algorithm to analyze the user click rate. The main work includes: (1) crawling Zhihu users' browsing information using crawler tools. This process uses the selenium crawler framework to simulate the browser operation, so as to enter the user's personal home page and crawl a series of information such as the recently browsed problems and problem contents as the original data. (2) Train the LSMT text emotion classification model. The model is trained with open source text data with emotional labels, so that the model can accurately judge the emotional factors implied in the long text. (3) Use the Deep Crossing recommendation system algorithm to predict the click-through rate of users to a certain answer. Use LSTM to analyze the emotion of the article, and combine the article analysis results with the article keywords extracted using TF-IDF algorithm. Use the combined data as the input of the Deep Crossing algorithm to train the algorithm model.

This project is tested on the open-source emotional text dataset and Zhihu user browsing record dataset. The test results show that the accuracy of LSTM emotional analysis model is 81%. After adding the result of emotion analysis as input, the prediction accuracy of the deep crossing algorithm improved by 0.9%

Keywords: Text emotion analysis, Natural language processing, LSTM text emotion classification model, Deep Crossing algorithm

目 录

第一章 绪 论

1.1 选题的意义

随着互联网普及率到达 71.60%[1]，互联网上的文本数量呈现爆炸式增长，网民已经习惯于在网络上表达意见和建议，如何从这些文本中获得有价值的信息成为新的难题[2]。

比如知乎、微博等文本网站上所发布的文本内容、讨论等。这些文本中蕴含着大量的隐藏信息。比如对大量文本进行情绪分析，可以以此为基础推测用户对文章关键词的情绪，从而推测用户对社会热点、政策、事件等情绪占比。对文本进行关键词提取，可以获取网络热词和网络风向。

文本情绪分析是一种利用自然语言处理和机器学习技术对包含情绪信息的文本进行细粒度的分类，通过分析得到文本中所隐含的情绪成分，如“正面”、“负面”、“中立”等，一般分为文档级的情绪分类[3]和句子级的情绪分类[4]

随着深度学习成为强大的机器学习技术，文本情绪分析的准确率得到了大幅的提升，但对情绪分析结果的应用却十分稀少。

该项目通过将情绪分析结果融入舆情分析和用户行为预测，尝试发掘文本情绪分析的应用价值。

1.2 研究现状与发展趋势

1.2.1 研究现状

情感分析是人工智能一直以来的重要课题之一。为了能解析文本中所隐含的情绪信息，学者们提出了许多方法：基于情感词典和规则的方法[5]、基于机器学习的方法[6-8]以及基于深度学习的方法[9-15]。

文本情绪分析一直是自然语言处理领域的一个重要的研究方向，并且得到了广泛的应用。随着技术的不断发展，文本情绪分析将会在更多的领域得到应用，如营销、舆情监测、社交媒体分析等。文本情感分析的任务主要包括情感信息抽取、情感分类以及情感检索与归纳[16]，文本情绪分析的任务主要包括情绪识别和情绪分类[17]。

1.2.2 发展趋势

文本情绪分析是一个不断发展和改进的领域。以下是一些可能的文本情绪分析发展趋势：

文本情绪分析的准确度将进一步提升：随着基于 transfromer 模型[18]的 Bert[19]和 GPT[20]算法的出现，对深度学习网络对文本特征的提取能力得到了飞跃式的提升，而基于此的文本情绪分析也响应的快速发展。

多模态情感分析[21]：多模态情感分析是结合了不同的信息源（如语音、图像和视频）来进行情感分析的方法。它可以提供更全面的情感分析结果，因此在各种应用场景中都有很大的潜力。

跨语言情感分析[22]：随着全球化的发展，越来越多的公司和组织需要对不同语言的文本进行情感分析。因此，跨语言情感分析将会成为一个重要的研究领域。

情感分析的实时性：在一些应用场景中，如社交媒体监测，需要对文本进行实时的情感分析。因此，开发实时情感分析算法将会变得更加重要。

1.3 本章总结

本章说明了推荐系统的重要性，介绍了在大数据时代的背景下针对信息过剩提出的解决方案：推荐系统，针对音乐推荐领域存在的问题，也都介绍了具体的方法解决。最后简单分析了推荐系统的发展趋势。

第二章 相关理论及技术介绍

2.1 文本情绪分析

文本情绪分析是指使用自然语言处理和机器学习技术，对一段文本或文档中的情感进行自动分析和识别的过程。这种分析可以帮助我们了解文本中所表达的情感、情绪和态度，例如喜悦、愤怒、悲伤等等，从而帮助我们更好地理解文本的含义。

深度学习自身的概念与人类实际存在的神经元运行方式息息相关，他由多层的非线性单元组成，并且将前一层的输出作为后一层的输入，从大量输入数据中获取有效的特征表示信息。深度学习作为机器学习领域中的热门研究领域，越来越多的人使用深度学习算法对文本进行情绪分析。本节介绍常见的两种深度学习算法。

2.1.1 循环神经网络

循环神经网络是深度学习算法中非常经典的算法分支，在处理序列数据方面有着较好的性能。RNN基本结构如图 2-1 所示

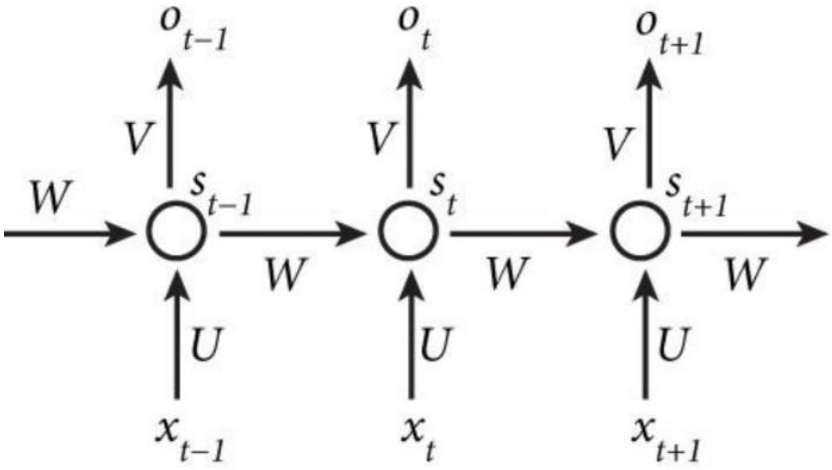

图 2-1 RNN 结构图

在 RNN 中，句子中存在的每个单词都隐含的蕴藏了其所处时间信息。每个时间步骤都会将之前所提取的时间信息与当前时间步所存在的，构成隐藏状态向量 h(t)。从某一方面来讲，这个向量蕴含了到 t 时刻位置的所有时间和文本信息。h(t)的计算公式如式(2.1)所示

$$h_t = \sigma(W_H S_{t-1} + W_X X_t) \tag{2.1}$$

该公式中 W_H 代表 RNN 层自传输的网络权重，W_X 代表输入层到隐藏层的网络权重，H_{t-1} 为 RNN 层上一次训练时的输入，X_t 为本次训练时的输入。

2.1.2 长短期记忆人工神经网络

长短期记忆人工神经网络是循环神经网络的一种，最早由 Sepp Hochreiter 和 Jurgen Schimidhuber 与 1997 年提出。是 20 世纪深度学习研究中被应用最多的论文。与普通的 RNN 不同，LSTM 在每个时间步上都有一个记忆单元和三个门（输入们、输出门、遗忘门）来控制信息的流动和保留。LSTM 结构图如图 2-2 所示

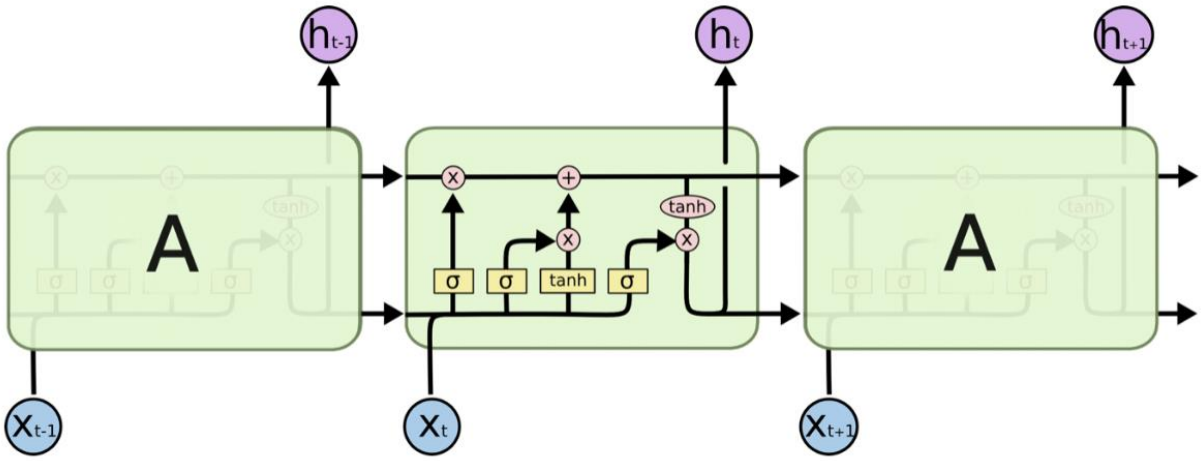

图 2-2 LSTM 结构图

在 LSTM 中，记忆单元负责存储历史信息，三个门则决定了当前信息的保留或舍弃，从而有效地解决了普通 RNN 中的梯度消失问题。

遗忘门用来学习之前存在的信息是否需要被遗忘。该遗忘门会将来自前一个隐藏状态的信息和当前输入的信息进行选择加权组合传递到 sigmoid 函数中去，从而实现遗忘的效果，输出值介于 0 和 1 之间，越接近 0 意味着越应该丢弃，越接近 1 意味着越应该保留。遗忘门的公式如式（2.2）所示

$$f_t = \sigma\left(W_f \cdot \left[H_{t-1}, X_t\right] + b_f\right) \tag{2.2}$$

W_f 为遗忘门的权重，而$[H_{t-1},X_t]$为输入数据和 LSTM 层上一次训练时的输出的合并向量，b_f 为偏移量。

通过增加遗忘门让信息选择性通过，从而缓解长序列模型训练过程中梯度消失和梯度爆炸的问题。因此 LSTM 与 RNN 相比，可以更好的处理较长的序列

2.2 网络爬虫

网络爬虫是一种利用计算机程序或脚本对万维网信息进行抓取。根据系统结构和实现方式网络爬虫可分为以下几种：增量式网络爬虫、聚焦网络爬虫、深层网络爬虫、通用网络爬虫。 而用于实践的爬虫技术通常由上述多种爬虫技术组合而成。简单的网络爬虫框架如图 2-3 所示

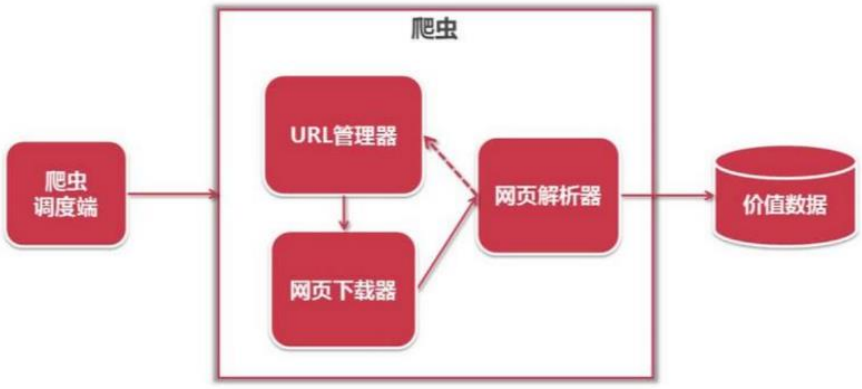

图 2-3 网络爬虫框架

2.2.1 selenium 爬虫工具

Selenium 是爬虫技术中的一种技术。Selenium 在运行时通过控制浏览器进行操作，从而达到模拟用户操作的效果。支持的浏览器包括 Google Chrome，Safari，IE，Mozilla Firefox， Opera，Edge 等。Selenium 的主要功能包括：测试兼容性——测试程序在不同浏览器上的适用性。测试系统功能——通过调用系统结构来测试系统功能。

因为知乎拥有强大的反爬虫机制，通过 selenium 模拟用户操作，较好绕过知乎的反爬虫机制，从而获取用户在一段时间内浏览的文章内容。Selenium 无法像 requests 等爬虫框架与目标服务器进行直接的 HTTP 交互，从而直接获得原始数据，而是必须等待知乎用户页面在完全加载之后，利用 selenium 框架提供的 broswer 类对页面内容进行读取，故信息的爬取效率较低。

2.3 数据库技术

数据库使用大量的存储空间对大量的数据进行存储。同时数据库对大量的数据按照一定的规则进行存放，以增加数据库管理系统的的查找效率。随着互联网的数据收集能力和数据产生能力大幅度提升，对数据进行存储的需要也随之提高。从某一方面讲，互联网就是一个巨大的数据世界。数据的来源很多，比如购物记录、聊天记录、形成记录、浏览行为等等。除了文本类型的数据，图像、音乐、声音都是数据。

数据库管理系统是数据库最重要的部分，主要用于对数据库中数据的操作和管理，包括数据库对象的创建、数据库存储数据的查询、添加、修改与删除操作和数据库的用户管理、权限管理等。

2.4 MySQL 数据库

MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，属于 Oracle 旗下产品。MySQL 是最流行的关系型数据库管理系统之一，在 WEB 应用方面，MySQL 是最好的 RDBMS (Relational Database Management System，关系数据库管理系统) 应用软件之一。

MySQL 是一种关系型数据库管理系统，关系数据库将数据保存在不同的表中，而不是将所有数据放在一个大仓库内，这样就增加了速度并提高了灵活性。MySQL 所使用的 SQL 语言是用于访问数据库的最常用标准化语言。MySQL 软件采用了双授权政策，分为社区版和商业版，由于其体积小、速度快、总体拥有成本低，尤其是开放源码这一特点，一般中小型和大型网站的开发都选择 MySQL 作为网站数据库。

2.5 基于深度学习的推荐算法

人们一直试图在大量数据中检测模式。机器学习通过精确检测此类模式加速了这种探索，这也是构成机器学习模型的一项技能。然后通过机器学习算法应用这些模式，以便在将一组新数据输入到算法时预测结果。

2.5.1 Deep Crossing 算法

Deep crossing 模型是 CTR 预估的深度学习模型，将类别特征与数值特征融合送入残差网络中进行点击的预估。

Deep Crossing 模型的应用场景是微软搜索引擎 Bing 中的搜索广告推荐场景。用户在搜索引擎中输入搜索词之后，搜索引擎除了会返回相关结果，还会返回与搜索词相关的广告。尽可能地增加搜索广告的点击率，准确地预测广告点击率，并以此作为广告排序的指标之一，是非常重要的工作，也是 Deep Crossing 模型的优化目标。Deep crossing 的模型结构如图 2-4 所示

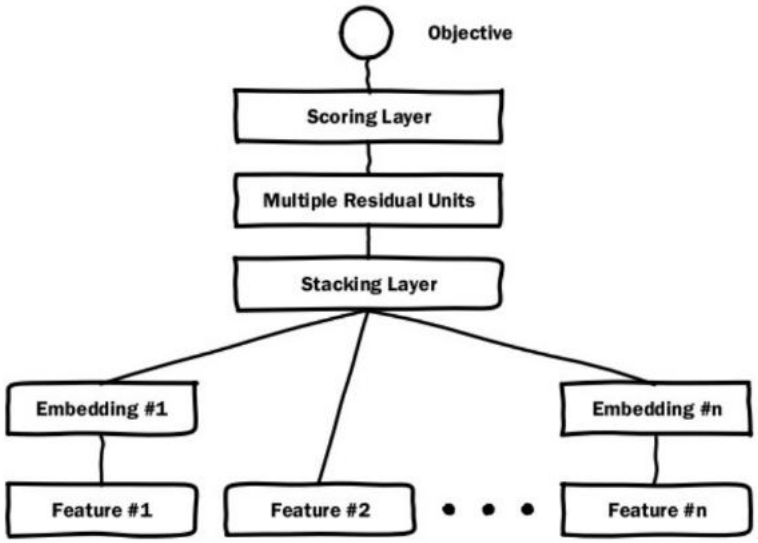

图 2-4 Deep Crossing 模型结构

Deep Crossing 算法将多个特征通过 embedding 层转换为向量形式，在 Stacking 层将特征向量进行拼接，然后让数据多层残差网络进行学习，最后通过 sigmoid 函数等函数对结果进行预测。

Deep Crossing 对多种特征进行拼接组合的方式可以非常方便的将问情绪分析结果进行运用，故选用该模型来探索文本情绪对点击率的影响。

第三章　系统总体设计

3. 1 开发技术简介

本节介绍本项目开发过程中所用到的一些技术和开发环境。用到的开发技术包括：爬虫框架 selenium、数据库 MySQL、深度学习开发框架 Pytorch

系统开发环境：开发工具 visual code、操作系统 Win10、CPU i7-8700、内存 8G

3.2 系统总体流程图

本项目总体结构如图 3-1 所示

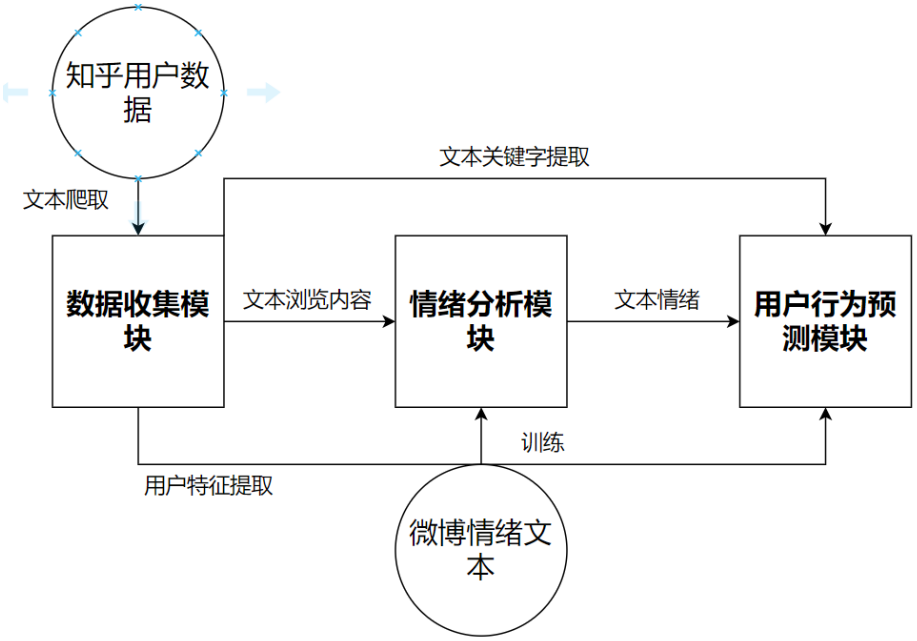

图 3-1 系统总体流程图

3. 3 数据库设计

本项目创建了三个表对从知乎用户爬取的信息进行存储，分别用户表、问题表、回答表，具体设计如表 3.1、3.2、3.3 所示

表 3.1 用户表

序列	列名	类型	是否为空	注释
1	id	INT(32)	NOT NULL	主键，用户 ID
2	username	VARCHAR(45)	NULL	用户名
3	time	VARCHAR(45)	NULL	记录用户时间

表 3.2 问题表

序列	列名	类型	是否为空	注释
1	user_id	INT(32)	NOT NULL	用户 ID
2	question_id	INT(32)	NULL	问题 ID
3	headline	TEXT	NULL	问题标题
4	content	TEXT	NULL	问题内容
5	visit_nul	INT	NULL	问题浏览人数
6	time	VARCHAR(45)	NULL	用户浏览时间

表 3.3 回答表

序列	列名	类型	是否为空	注释
1	user_id	INT(32)	NOT NULL	用户 ID
2	question_id	INT(32)	NULL	问题 ID
3	creater_name	VARCHAR(32)	NULL	回答作者 ID
4	Content	TEXT(45)	NULL	回答内容
5	Time	VARCHAR(45)	NULL	回答时间
6	Thump_up_num	INT(32)	NULL	回答点赞数
7	location	INT(32)	NULL	回答网页位置
8	love	INT(32)	NULL	用户是否点赞

3.4 本章总结

本章具体描述了本项目所示用到的具体技术与环境，并且展示了项目总体的流程图以及数据库的设计，对本项目的结构进行了大体的描述。

第四章 系统具体实现及其展示

本章介绍了该项目的的详细设计与实现。本章通过系统流程、系统运行图对该项目的数据收集和存储、情感分析、用户点赞三个模块进行展示。

4.1 数据收集模块

4.1.1 流程图

数据收集模块使用了 3 个工具类，分别是浏览器操作工具类，用户页面数据收集工具类、数据存储工具类。三个类皆为自己实现。浏览器操作类通过 selenium 包的调用实现用户主页的访问和关闭网页等辅助操作，对 chrome 浏览器进行模拟人工操作，绕过了知乎大量的防爬机制，但增加了代码的复杂度，降低了信息获取的速率。用户页面数据收集工具类，通过 selenium 包进行调用，收集用户页面中所需要的数据。数据存储类工具类使用 MySQL Connector，该包为开发人员提供了数据库应用编程接口，通过对该包的调用，可以将爬取到的数据永久存储在本地的 MySQL 数据库中。

该项目利用爬虫工具对知乎用户的浏览记录进行了爬取，并将提取的数据存入 MySQL 数据库。爬取的信息主要有用户浏览的问题、用户点赞的回答、用户未点赞的回答、文章发布时间等信息。

同时该数据收集模块记录每次爬取用户文章的时间点，当再次爬取相同用户浏览记录时，只需更新该时间点之后的记录，保证了数据的实时性。数据收集模块的流程图如图 4-1 所示

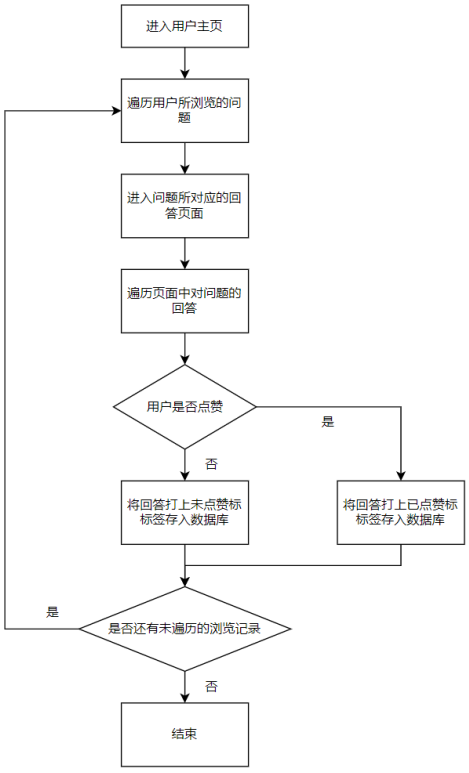

图 4-1 数据收集模块流程图

主要的思想为通过用户名构建用户主页的 URL，遍历用户近期浏览记录，通过用户是否对浏览记录进行点赞从而生成正负文本。

4.1.2 浏览器操作类

该类主要通过 selenium 提供 browser 类，对浏览器页面进行操作，从而实现对浏览器的控制，具体功能如图 4-2 所示

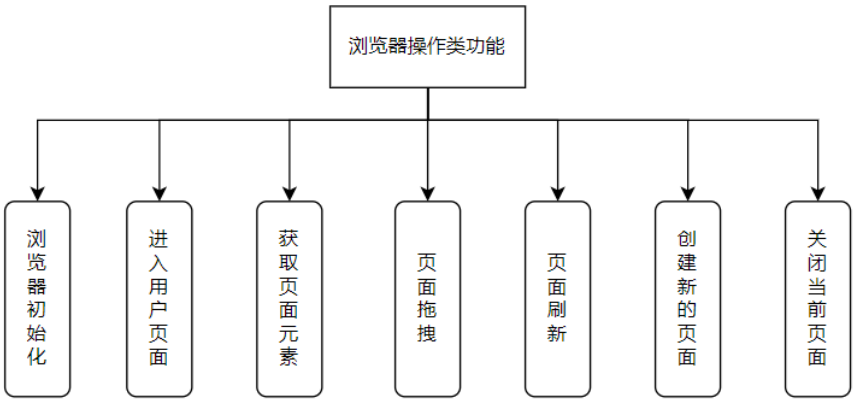

图 4-2 浏览器操作类具体方法

（1）深层网络爬虫：指爬取不能通过静态资源获取的信息，往往是用户输入某些关键词才会出现的页面。

（2）浏览器初始化：用于初始化浏览器操作类，通过 selenium 提供的 webdriver 类获得 Chrome 浏览器的控制器 broswer，并存入知乎用户主页和问题回答主页的前缀 URL。

（3）进入用户页面：通过用户的用户名构造用户主页的 URL，并通过 broswer 进入该页面。

（4）页面拖拽：根据输入的某一 html 元素，来时浏览器当前页面滚动到该元素的位置。

（5）页面刷新：当操作过程出现错误时，刷新当前页面

（6）创建新的页面：通过输入的 url 创建新的页面，并将控制权转移到该页面。

（7）关闭当前页面：将当前页面关闭，将控制权转移回之前的页面

4.1.3 页面数据收集类实现

该类对用户的浏览记录进行遍历，并提取相应浏览记录对应的问题内容、回答内容、回答发布、用户点赞时间、问题浏览中人数等信息，并调用数据存储类将数据进行存储。页面数据收集类的具体方法如图 4-3 所示

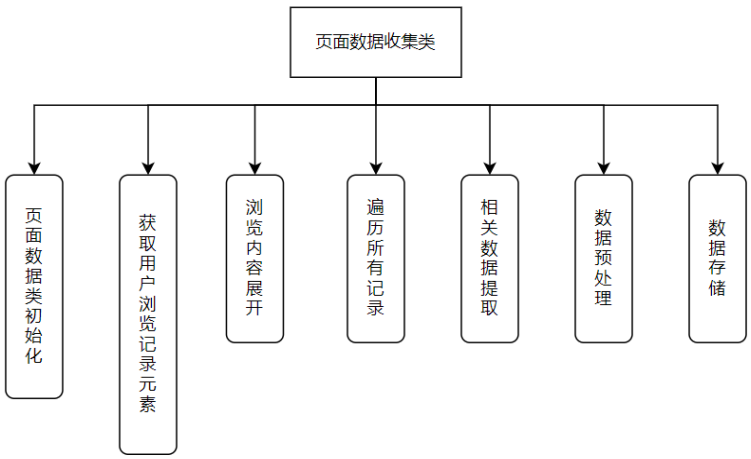

图 4-3 页面数据收集类具体方法

（1）页面数据类初始化：将浏览器操作类和数据存储类的实例传入初始化函数。

（2）获取用户浏览记录元素：通过浏览器操作类获取用户所有的浏览记录所在 html 元素

（3）浏览内容展开：用户浏览内容初始状态为简略版，需要通过点击展开元素将内容展开

（4）遍历所有记录：对所有的浏览记录元素进行遍历，并通过浏览内容展开方法显示出用户浏览的所有内容

（5）相关数据提取：对每一浏览记录相关的数据进行提取

（6）数据预处理：对数据进行初步处理，使其转换成可以存入对应数据库中的数据。

（7）数据存储：调用数据存储类提供的方法，将经过数据预处理后的数据存储数据库。

4.1.4 数据存储工具类实现

该类基于为开发人员提供了数据库应用编程接口 mysql.connector，通过 python 连接 MySQL 数据库，对用户信息表、问题表、回答表进行选择、更新、删除等操作。数据存储工具类的具体实现如图 4-4 所示

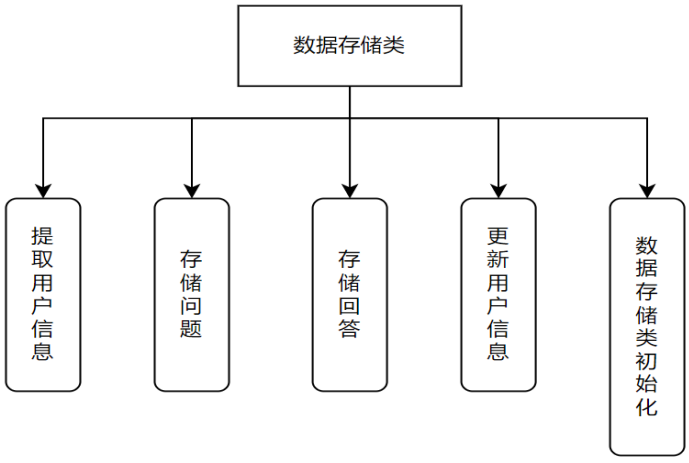

图 4-4 数据存储类具体方法

（1）数据存储类初始化：通过 mysql.connector.connect()方法获取数据库的操作句柄

（2）提取用户信息：通过 select 方法获取用户信息表中所有用户的信息

（3）存储问题：将问题基本信息存储数据库中

（4）存储回答：将问题所对应回答的信息存入数据库中

（5）更新用户信息：更新用户的最新浏览时间、浏览问题总数等信息

4.1.5 数据呈现

通过该爬虫爬取的问题内容和回答内容部分数据如图 4-5、4-6 所示

user_id	question_id	headline	content	visit_num	time
112	584576636	日本留学文硕毕业纠结去留，求分析?	先介绍条件 可能一般人都不如 男 27 独身子…	620	2023-02-17 15:21
112	584606393	大三，对未来感到迷茫?	本人普通211大三，日语零基础，本科专业…	291	2023-02-17 15:19
112	584499180	日本的大学院出愿后还能不能联系教授拿非…	已经出愿了东大综合文化，超域文化科学专…	727	2023-02-16 22:14
112	581860424	丰田重申不支持全面电动化，应发展更多混…	Automotive News 近日报道，丰田汽车首席…	748759	2023-02-16 22:04
112	578501634	想去芬兰读高中，芬兰留学有什么缺点吗?	NULL	458	2023-02-16 18:13
112	584410160	各位已经赴日留学的知友们大家好!?	小弟我今年23岁，参加工作六年(厨房)，中…	787	2023-02-16 17:52
112	584354318	现在去日本开办语言学校或私塾还有钱途吗?	现在日本（尤其是东京）的语言学校和私塾…	1244	2023-02-16 12:54
112	584153178	该怎么在日本宣传保录大学计划?	和泰国的大学有合作，保录项目，不限三校…	1251	2023-02-15 21:16
112	584145354	漫改真人电影《圣斗士星矢 The Beginning》…	漫改真人电影《圣斗士星矢 The Beginning》…	16679	2023-02-15 17:19
112	584186675	赴日留学，日币应该怎么取更合适?	大家好，我是今年四月生，然后最近看了些…	4410	2023-02-15 16:52
112	584205768	不想去语言学校了，我这种情况怎么办?	但是中介非要等再留下来后才联系学校说我…	714	2023-02-15 16:32
112	268565724	全世界最瘦的为什么是日本人?	世界卫生组织的报告显示，日本是全世界肥…	4861970	2023-02-14 21:16
113	584646196	与多人发生不正当性关系，云南 80 后正厅…	2月17日，云南省纪委监委网站通报了去年8…	544031	2023-02-18 12:33
113	579795917	为什么说有人说"刚需2023年不买房必后悔"?	请理性讨论，接纳不同的意见和声音。	4204236	2023-02-16 14:49

图 4-5 部分问题内容

user_id	question_id	creater_name	content	time	thump_up_num	location	love
101	579908180	an-de-sen-59	两万多年了，我们仍然在吃大米，而且两千…	2023-01-23 19:16	4260	1	0
101	579908180	tang-lang-zai-hou	我是做可乐的，我冬眠了200年了，我现在…	2023-01-22 19:24	886	2	0
101	579908180	rnavision	三大饮料中的咖啡可可都喝了几百年，茶已…	2023-01-22 08:34	735	3	0
101	579908180	maomaobear	你吃的豆腐，已经2000年了。	2023-01-23 19:33	377	4	0
101	579908180	xiao-ming-15-27	那你要是知道我还喝3000年前发明的米酒你…	2023-02-13 17:39	170	5	0
101	579908180	yyy-40-50-86	有什么可怕的，一千年这么短的时间区间根…	2023-02-18 14:21	0	5	1
101	578110161	mai-cui-ya-96	我还以为你妹妹拿着盆儿一边敲一边唱歌呢	2023-01-15 02:03	3679	2	0
101	578110161	chen-yang-23-96	点进来以前我还以为她是跑出去嘣迪或者泡…	2023-01-15 14:57	16765	1	1
101	562477513	zhang-san-2-6	54的先贤早就总结了，奴才，主子，吃人，…	2022-12-10 23:24	2417	1	0
101	562477513	bao-lai-wang	《几何原本》是古希腊数学家欧几里得创作…	2023-02-18 03:49	30	5	1
101	564154819	a-yo-41	一代卷王，卷死了无数人。现在她卷不动了…	2023-02-14 09:32	2986	1	0
101	564154819	shark-74-46	这？？？依稀记得当年，在央视节目中大…	2023-02-12 20:40	2379	2	0
101	564154819	RoseofVersailles	她的意思是她不能太累，要么就限制她实现…	2023-02-13 09:18	1022	3	0
101	564154819	zhouleiwang.bio....	我的天，她居然到现在才发现，难道是因为…	2023-02-14 10:59	1078	4	0

图 4-6 部分回答内容

本项目总共收集知乎问题 1440 条，知乎回答 6770 条。

通过使用 jieba 包对知乎回答内容进行分词操作，并使用 wordcloud 进行词频统计，词频统计结果如图 4-7 所示

图 4-7 词频统计

本图片根据字符的大小来反应词频数。

4.1.6 数据收集模块性能测试

因 selenium 模块是模拟浏览器操作从而对数据进行爬取，其性能影响因素有计算机 CPU 运行速度，网络数据传输速率，浏览器渲染速度、服务器性能等。在大量因素的影响下，该数据收集模块性能较差且不稳定。

4.2 情绪分析模块

情绪分析模块是整个项目的核心部件。获取互联网上开源的微博文本情绪数据(xml 格式)，通过 xml.dom.minidom 对文本进行读取，在用 jieba 包对文本进行分词操作，将文本变为一个词组数组，词组元素的先后顺序与其在文本中出现的位置相对应。基于所有文本中出现的词组和词组出现的次数构建字典，同时为每个词组赋予唯一编号。将每个文本对应的词组数组中的元素转化为该元素所对应的编号，最后将转换后的数组输入 LSTM 模型中进行训练。情绪分析模块主要由数据加载操作和模型训练操作构成。

4.2.1 数据预处理

该过程主要将开源的情绪文本数据进行数据预处理和特征提取。数据处理过程如图 4-8 所示

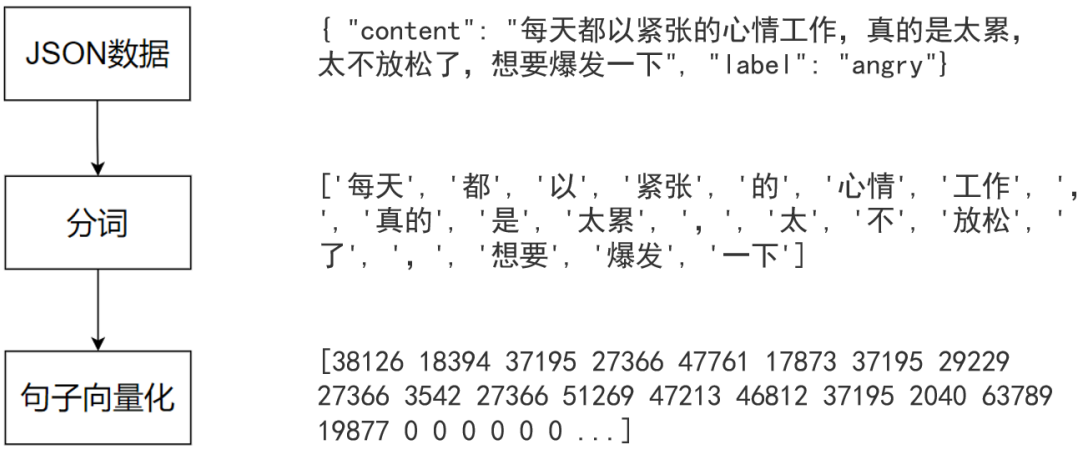

图 4-8 数据处理过程

微博情绪文本的初始数据以 JSON 的形式给出，我们通过 jieba 包提供的分词操作对 content 中的内容进行分词操作，最后通过已经构建好的字典将分词数组中的词语转换为对应的编号，最终实现句子向量化。

4.2.2 LSTM 文本情绪训练

该 LSTM 文本情绪分析模型基于深度学习框架 pytorch，使用上节已经预处理过的数据作为输入，对该框架进行训练。该 LSMT 框架结构图如 4-9 所示

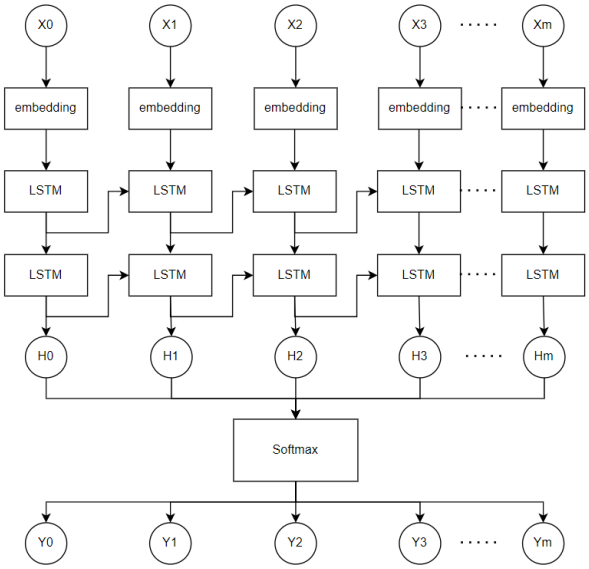

图 4-9 LSTM 框架图

该模型基于 RNN 基础模型思想，将本次训练时 LSTM 层的输出作为下一次 LSTM 层输入的一部分来实现对长文本特征的读取。

我们将已经经过向量化的数据输入 embedding 层进行转换，将转换后的数组输入到 LSTM 层进行句子特征提取，并将多层 LSTM 的结果输入到 Softmax 函数中得出最终的结果。Softmax 的公式如式（4.1）所示

$$P(Y_i|X) = \frac{e^{h(X,Y_i)}}{\sum_{j=1}^{n} e^{h(X,Y_i)}} \tag{4.1}$$

函数 $h(X,Y_i)$代表向量 X 在 Y_i 所对应维度上的值，$P(Y_i|X)$为向量 X 的前提下，该向量属于 Y_i 类的概率。

Softmax 函数在将向量实现归一化的同时不会改变输入向量各个维度的相对大小，我们选择 Softmax 输出向量数值最大数所在维度作为输入数据的类别。

4.2.3 模型性能测试

本节利用不同大小的训练集来对 LSTM 情绪分类模型进行训练，并统计其在测试集上的准确度，最后选定较为适合的训练集大小来得到较高的分类准确率。

情绪分析文本数据集和大小为 20000 条，情绪种类分为为负面情绪 5000 条、正面情绪 5000 条、无情绪 10000 条，情绪分析模型将数据以 9:1 的比例将数据分为训练数据和测试数据。LSTM 模型情绪分析准确率如图 4-10 所示

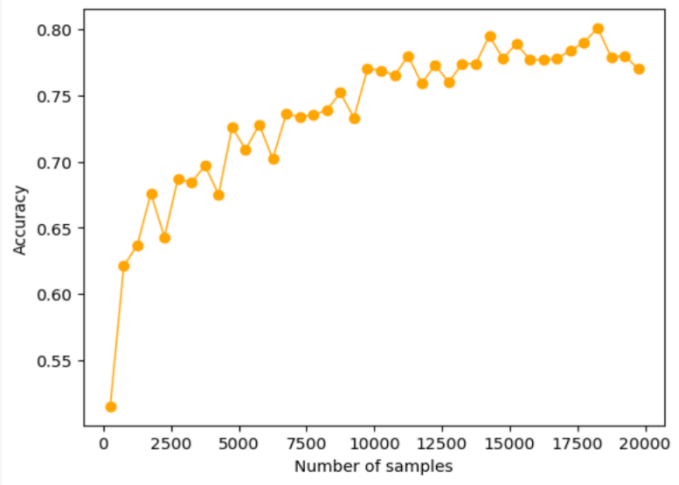

图 4-10 LSTM 情绪分析模型准确率

可知在数据样本量超过 15000 份之后，LSTM 情绪分析模型的准确率可达到 80%。

4.3 Deep Crossing 算法改进

Deep Crossing 算法的优点是能够自动学习特征之间的交互作用，不需要手动设计特征工程。根据这个特性，本项目将用户近期情绪和文本情绪作为其输入的一部分，通过情绪输入和其他特征相结合来尝试提高其对用户文本点击率预测的正确性。

4.3.1 关键字提取

我们使用 TF-IDF 算法对用户近期浏览的文本和回答内容进行关键字的提取。

通过 jieba 包对所需要分析的文本进行分词操作，并利用容器对文本中词语出现的次数进行统计，并根据文本的总词数计算 TF 值，TF 值的计算入式（4.2）所示

$$词频(TF) = \frac{某个词在文章中出现的次数}{文章总词数} \tag{4.2}$$

TF 值主要展现某个词在文本中出现的比重从而推测其重要程度。

其次，我们通过统计爬取到的文档总数，同时统计某一词语对应的包含该词语文档数量，从而家计算 IDF 值。IDF 公式如式（4.3）所示

$$逆文档频率(IDF) = \log\left(\frac{词料库的文档总数}{包含该词的文档数 + 1}\right) \tag{4.3}$$

IDF 通过计算词语出现的稀缺程度来推测其在所有文本之中的重要程度。

最后通过 TF 和 IDF 相乘计算 TF-IDF 的的值。其公式如式（4.4）所示

$$TF - IDF = 词频(TF) \times 逆文档频率(IDF) \tag{4.4}$$

最后我们通过比较用户近期浏览内容不同单词的 TF-IDF 值的大小来确定其浏览内容的关键字，最

终实现对关键字的提取。

4.3.2 输入数据构建

该模块主要基于爬虫模块所爬取的知乎用户的浏览的回答，及回答的相关信息来提取出关键特征，并组成 Deep Crossing 的输入。输入组成如表 4-1 所示

表 4-1 输入数据组成

字段位置	数据类型	是否允许为空	含义
0	Int	否	用户是否点赞
1	float	否	回答点击率
2	Int	否	用户点赞时间与文章发布时间差值
3	Int	是	该回答在问题回答列表的位置
4	Int	否	用户近期情绪
5	Int	否	回答情绪
6-15	String	否	用户近期关键词
16-20	String	否	问题关键词
21-30	String	否	回答关键词

Deep Crossing 实际输入数据如图 4-11 所示

```
('yyy-40-50-86', 101, '2023-03-12 11:50')
[1, 0.0, 0, 5, 0.17857142857142858, 0.5714285714285714, 0.25, 0.0, 0.0, 1.0, '手表', '苹果', '手腕', 'pxj', 'apple',
 'watch', '日本', '凹陷', '为什么', '肥胖率', '网友', '可乐', '皮肤', '凸起', '颜宁', '哔哩', '如何', '运动', '客服',
  '祖父', '可乐', '人类文明', '一代', '可怕', '1000', '儿起', '古柯', '人类', '可口可乐公司', '糖水', '一千年', '任何',
  '可乐', '口味', '预见']
[1, 76.42705239572483, 0, 1, 0.17857142857142858, 0.5714285714285714, 0.25, 0.2, 0.2, 0.6, '手表', '苹果', '手腕',
'pxj', 'apple', 'watch', '日本', '凹陷', '为什么', '肥胖率', '网友', '可乐', '皮肤', '凸起', '颜宁', '哔哩', '如何', '运动',
 '客服', '祖父', '祖父', '娱乐活动', '平板', '游戏', '妹妹', '守灵', '屋子里', '爷爷', '蹦迪', '手机', '可玩', '晚上', '泡吧',
  '吹拉弹唱', '前半夜']
[1, 0.13191886989501456, 0, 5, 0.17857142857142858, 0.5714285714285714, 0.25, 0.5, 0.16666666666666666, 0.3333333333333333,
'手表', '苹果', '手腕', 'pxj', 'apple', 'watch', '日本', '凹陷', '为什么', '肥胖率', '网友', '可乐', '皮肤', '凸起', '颜宁',
 '哔哩', '如何', '运动', '客服', '祖父', '文明', '超越', '近代', '东方', '西方', '13', '欧几里得', '几何', '原本', '平面几何', '立体几何',
  '讨论', '300', '所著', '共分']
```

图 4-11 Deep Crossing 实际输入数据

4.3.3 Deep Crossing 模型结构

Deep Crossing 模型的主要包括两个部分：连接层和多层的残差网络。特征交叉部分将输入特征进行组合，以捕捉特征之间的交互作用；多层感知器部分利用这些组合特征作为输入，通过多个全连接层来学习这些特征之间的复杂非线性关系。Deep Crossing 的模型如图 4-12 所示

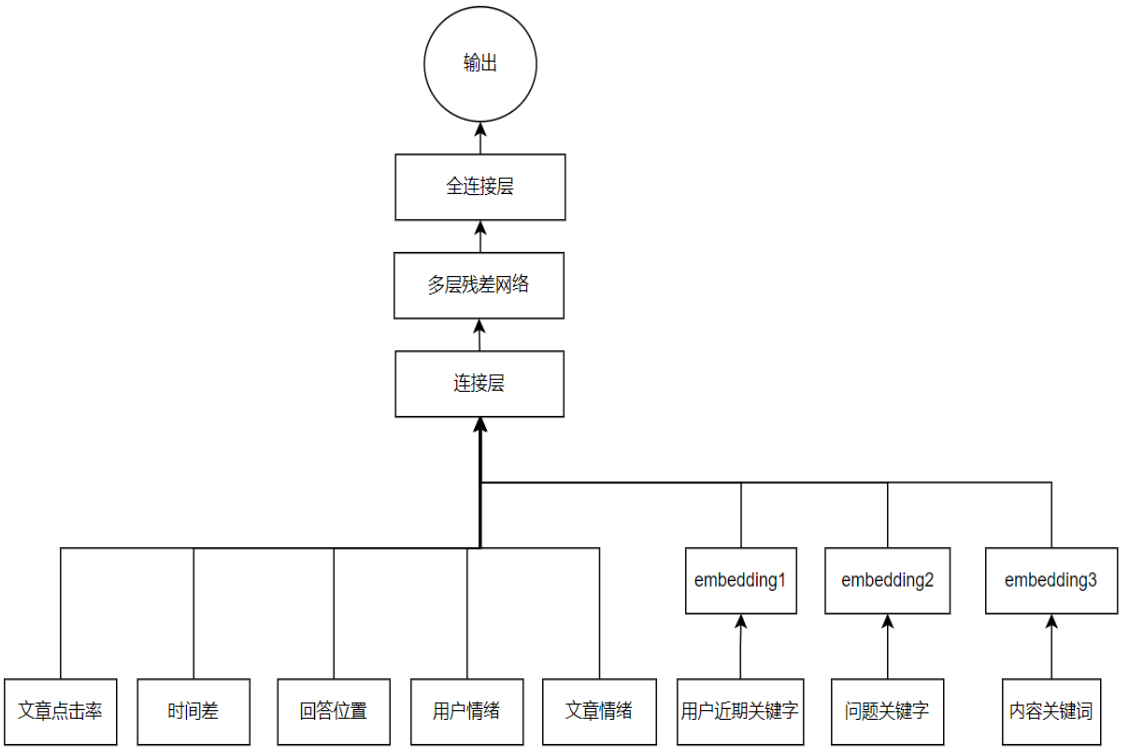

图 4-12 Deep Crossing 结构

Embedding 层该将数值型特征不做变换直接传给 Stack 层，将文本特征转化一定维度的 Embedding 向量。

Stacking 层将数值型特征和经过 Embedding 层转换的文本特征进行拼接，形成新的包含全部特征的特征向量。

Multiple Residual Units 层使用多层的残差网络使特征向量的各个维度进行充分的交叉组合，使模型能够抓取到更多的非线性特征和组合特征的信息。

最后使用逻辑回归模型预测用户是否会对该回答点赞。

4.3.4 Deep Crossing 模型性能测试

本项目将没有情绪特征输入与附加了情绪特征输入的 Deep Crossing 算法在相同数据集上进行测试，然后经过多次运行对正确率去平均值。测试结果如图 4-13 所示

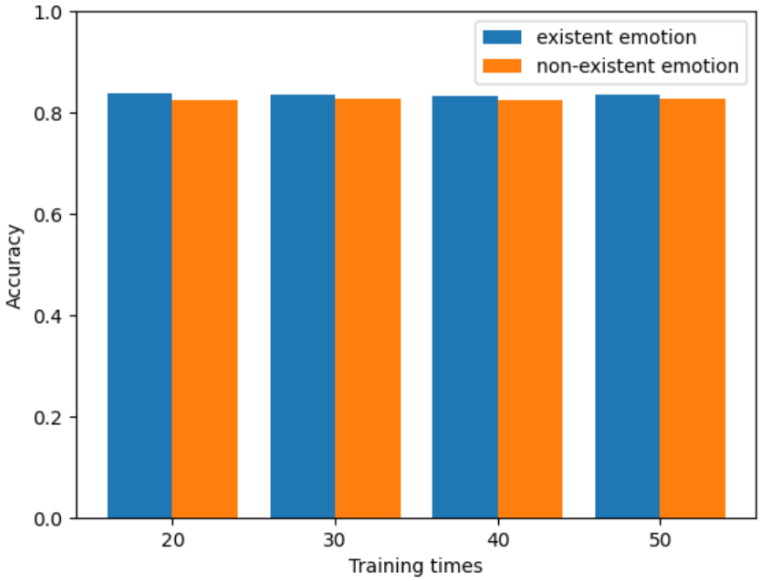

图 4-13 有无情绪输入对比

可知将情绪分析作为 Deep Crossing 输入的一部分可以提高约 0.9%的准确性

4.4 本章总结

本章主要介绍了本项目的具体实现和性能测试，首先对知乎用户的浏览记录进行爬取，再使用开源的情绪文本数据集合对 LSTM 情绪分类模型进行训练，最后利用 Deep Crossing 模型对用户是否点赞进行预测。

总结与展望

随着互联网的飞速发展，网络中的文本数量呈现指数倍的增加，而这些文本中存在者巨量的信息要素和隐藏价值。而对这些文本的潜在价值进行发掘和利用具有重要意义。经过多年的探索和研究，随着 RNN、LSTM、GRU 等深度学习算法被相继提出，计算机对文本情绪分析的准确度到达了惊人的高度，但对情绪分析结果的应用还处于空白阶段，本项目对情绪分析结果进行了尝试性应用。本文主要工作如下：

（1）介绍了文本情绪分析的研究背景及发展现状，简单介绍了本项目所用的文本情绪分析所用到的算法，和预测用户点击率的推荐系统算法 Deep Crossing。

（2）提出了将文本情绪分析结果作为 Deep Crossing 算法输入的一部分，从而提高其预测准确度。

（3）实现了知乎文本爬取、LSTM 情绪分析框架训练、Deep Crossing 预测用户点击率，并将结果和实现进行说明。

因作者能力有限，只能将文本情绪分析结果进行有限的利用，在情绪分析准确率的提升和应用领域有如下展望：

（1）基于深度学习的情感分析方法仍存在一些挑战，如情感识别的主观性和不确定性，情感分析中多样性的考虑。

（2）可以探索不同领域和语言下的情感分析应用，以满足更多场景的需求，比如用户使用聊天软件，使用情绪分析对好友发送的消息进行情绪分析等。

参考文献

[1] 第 48 次中国互联网络发展状况统计报告[EB/OL]. 中国互联网络信息中心.[2021-09-15].http://www.cnnic.net.cn/hlwfzyj

[2] 季泓宇.互联网生大爆炸[J].互联网周刊,2012(19):24-25.

[3] Li C., Wu H., and Jin Q.Emotion Classification of Chinesse Microblog Text via Fusion of Bow and eVector Feature Representations[C]. In: Proceedings of NLP&CC-14, 2014, 217-228.

[4] Xu J, Xu R, Lu Q, et al. Coarse-to-fine Sentence-level Emotion Classification Based on the Intra-setence Features and Sentential Context[A]. In: Preceedings of the 21st ACM international conference on Infromation and knowledge management[C].USA:ACM,20-12:2455-2458.

[5] Cho H, Kim S, Lee J, et al. Data-driven Integration of Multiple Sentiment Dictionaries for Lexicon-based Sentiment Classification of Product Reviews[J]. Knowledge Based Systems, 2014, 71(nov.): 61-71.

[6] Catal C, Nangir M. A sentiment classification model based on multiple classifiers[J]. Applied Soft Computing, 2017, 50:135-141.

[7] Alkubaisi G A A J, Kamaruddin S S, Husni H . Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers[J]. International Journal of Engineering & Technology, 2018,11(1):52.

[8] Yang K, Liao C, Zhang W. A Sentiment Classification Model Based on Multiple Multi-classifier Systems[M]. Springer, Cham, 2019.

[9] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014,3(1):1746-1751.

[10] 梁军，柴玉梅，原慧斌，等.基于深度学习的微博情感分析[[J].中文信息学报，2014,28(5).

[11] Wang X, Jiang W, Luo Z, et al. Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts.[C]. In: International Conference on Computational Linguistics, 2016:2428-2437.

[12] Xu D, Tian Z, Lai R, et al. Deep learning based emotion analysis of microblog texts[J].Information Fusion, 2020, 64.

[13] 朱烨，陈世平. 融合卷积神经网络和注意力的评论文本情感分析[J].小型微型计算机系统, 2020,041(003): 551-557.

[14] 胡德敏，褚成伟，胡晨，等.预训练模型下融合注意力机制的多语言文本情感分析方法[J].小型微型计算机系统，2020, 041(002):278-284.

[15] Qian Q, Huang M, Lei J, et al. Linguistically Regularized LSTMs for SentimentClassification[C]. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017,1,1679-1689.

[16] 赵妍妍，秦兵，刘挺.文本情感分析[J].软件学报, 2010, 21(8): 1834-1848.

[17] Aman S, Szpakowicz S. Identifying Expressions of Emotion in Text[C]. In: Proceedings of the 10th International Conference (TSD 2007). 2007: 196-205.

[18] Vaswani, Ashish, et al. Attention is all you need[J]. *Advances in neural information processing systems* 30 (2017).

[19] Jacob Devlin, Chang Ming-wei, Kenton Lee, et al. BERT: pre-training of deep bidirectional transformersfor language understanding [C].Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, 2019:4171-4186.

[20] Floridi, Luciano, and Massimo Chiriatti. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and Machines 30 (2020): 681-694.

[21] 张亚洲，戎璐，宋大为，等多模态情感分析研究综述[J]．模式识别与人工智能，2020，33(5)：426-438.

[22] 徐月梅，曹晗，王文清，等跨语言情感分析研究综述[J]．数据分析与知识发现，2023，7(1)：1-21.