附件 2

2017 级理科

毕业论文（设计）排版模式

教务处编制

分类号 _____ 论文选题类型 非师范类应用研究

U D C _____ 编号_____

本科毕业论文（设计）

题　　目	中文语料的情感分析与可视化
学　　院	计算机学院
专　　业	计算机科学与技术
年　　级	2017 级
学生姓名	张乐艺
学　　号	2017211668
指导老师	马长林

二〇二一 年 四 月

华中师范大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下独立进行研究工作所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

学位论文作者签名：　　日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士学位论文评选机构将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密 □，在_____年解密后适用本授权书。

2、不保密 □。

（请在以上相应方框内打“√”）

学位论文作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

摘　要

随着互联网的蓬勃发展，全世界诞生了很多聊天社交软件，也出现了很多购物、外卖、预定的软件，如淘宝、饿了么、美团等。在这些软件里会出现很多评论，有时候可以达到上百上千条。如果可以尽可能高效地利用这些数据，可以迅速对大众舆情进行分析和处理，为市场分析等多个方面的应用领域提供帮助。此类数据往往数量十分庞大，如果仅仅通过安排员工来进行阅读与分析这些文字，这需要消耗大量人力物力和时间，而且准确率并不能得到很好的保障，如何通过自然语言处理的方法利用计算机有效地在海量文本数据中获取和分析舆情成为了十分有实际意义的问题。

主流的研究方向为基于监督学习，通过深度学习的办法来完成文本情感分析，用在自然语言处理领域的深度学习网络模型主要有卷积神经网络和长短时记忆网络两种。可以通过训练一个长短时记忆网络模型来获得一个具有 positive， neutral，negative 三种情感的一个分类器，输入为句子的序列。对于大多数评价信息中，情感只有喜欢、不喜欢和中性三种，因此可以训练出具有较好预测能力的模型。

本文的任务之一是基于 Keras 框架来搭建一个神经网络模型，训练出准确率可靠的中文情感分析工具。针对文本情感分析需要考虑到文本序列、情感特征的重要度等，本文模型包含四层，其中第一层为 embedding 层，对神经网络进行降维；第二层为 BiLSTM 层，其第二个维度是 sequence，对于 32 个单元，由于使用双向 LSTM，故设置 64 个单元；第三层为单向 LSTM 层，将返回的 sequence 导入单向的 LSTM 中，并且可以将一个结果返回到模型中保存；最后一层为全连接层，从 Sigmoid 中输入一个数值，并根据此判断输入的文本是正向情绪还是负向情绪。

本论文同时对情感分析模型制作了一个可视化工具，让使用者可以自订测试的文本，点击预测按钮即可进行情感预测，简化了使用难度。可视化工具利用相应方法实现多个窗口相互切换，以及在窗口内调用中文情感分析项目运行和结果返回。

关 键 词：情感分析；双向长短时记忆网络；卷积神经网络；文本情感可视化

Abstract

As the Internet developed in a rapid trend, there are many social applications in the market, as well as shopping, takeout, and reservation software, such as Taobao, Ele.me, Meituan, etc. There are a lot of comments in these applications, sometimes there are even thousands of them. If we can use these data as efficiently as possible, we can quickly analyze and process public opinions, and provide help for market analysis and other applications. This kind of data is often very large in scale, if only rely on manual reading and analysis, it must cost a bunch of time and energy, and the accuracy can not be well guaranteed. How to use the Natural Language Process method to effectively obtain and analyze public opinion in the massive text data has become a very practical problem.

The mainstream research direction was based on supervised learning, especially Deep Learning for Text Sentiment Analysis. Convolutional Neural Network (CNN) and Long-term and Short-term Memory Network (LSTM) were two kinds of Deep Learning Network Models used in the field of Natural Language Process. Using the LSTM model, we could train a three-level classifier for emotional analysis with positive, neutral, and negative, and input it as a sequence of sentences. For the major parts of the evaluation information, there are three typical emotions: positive, negative, and neutral, so we can train a model with good prediction ability.

In this project, we use the Keras to build a Neural Network model. For text sentiment analysis, we need to consider the importance of text sequence and sentiment features. This biLSTM model consists of four layers. The first layer is the embedding layer, to reduce the dimensions of the neural network; the biLSTM layer is the second, and its second dimension is a sequence. For 32 units, 64 units are set because of using bidirectional LSTM; the third layer is the biLSTM layer to import the returned sequence a One-way LSTM, and return a result; the last layer is the full connection layer, to input a value from sigmoid, and judge whether the input text is a positive emotion or negative emotion.

In the aspect of visualization, Pyqt5 is used to design GUI. Pyqt5 is developed by Python 3.8, and the interface of the button and label provided by pyqt5 is used. When creating a new

window, Widget is used for development; PushButton is used for a button, and its size, text, and other attributes are set; Label is used as an input text box for sentiment analysis, and adjusts the font color, size; the final output results, based on the output results of three categories, are transfered into positive evaluation, negative evaluation, and neutral evaluation.

In this paper, a visual tool is made for the emotional analysis model, which allows users to customize the text of the test, click the 'forecast' button to predict emotion, which simplifies the difficulty of using it. The visualization tool uses the corresponding method to realize the switching of multiple windows and calls the Chinese emotion analysis project to run and return the results in the window.

Key words：Chinese Text Sentiment Analysis; biLSTM; CNN; Text Sentiment Visualization

目　录

第1章　引言

本章主要介绍论文的研究背景、研究意义、研究现状、论文总体概述、和本论文的组织结构。

1.1 研究背景与意义

随着互联网的蓬勃发展，全世界诞生了很多聊天社交软件，也出现了很多购物、外卖、预定的软件，如淘宝、饿了么、美团等。在这些软件里会出现很多评论，有时候可以达到上百上千条。除此之外，在做金融产品量化交易时，需要利用新闻、舆论等分析市场情绪变化和进行投资建议等。如果可以尽可能高效地利用这些数据，可以迅速对大众舆情进行分析和处理，为市场分析等多个方面的应用领域提供帮助。此类数据往往数量十分庞大，如果仅仅通过人力进行阅读与分析，这需要消耗大量人力物力和时间，而且准确率并不能得到很好的保障，如何通过自然语言处理的方法利用计算机有效地在海量文本数据中获取和分析舆情成为了十分有实际意义的问题。

目前，文本情感分析是中文自然语言处理方向研究中比较热门且有实际意义的方向，情感分析往往可以有效地帮助人们快速定位一段句子的评价是正面或者负面。通常来讲，情感分析是对带有主观性评价语句的文本进行分析、处理、归纳和判断。面对大量文本数据时，通过进行情感分析，可以很快地了解到大众舆论对某件事情的看法和态度。

酒店预订平台如爱彼迎、美团等面世多年来，已经产生了相当大的信息量，其中有很多信息可以为我们利用。此类平台常常围绕一个特定话题（关于酒店的居住感受、价格等）进行大量阐发意见、相互讨论，因此在评论里蕴含着大量的用户语言特点、需求特点、观点与情感。

针对网络上对于订单的评价，各个平台一般都只是利用了一些简单的方法，比如打分模式。目前的评价机制一般是采取打分+评论两部分，但是有时候也会出现打分与评论背离的情况，比如用户给了五星好评，但是评价内容确实负面的，这往往会给消费者带来困扰。通常来讲，评论的信息更能够反应用户的真实感受，一般是主体对于客体的主观看法。因此，基于评论文本本身进行分析，更能够对大众舆情获得准确的认识，分析出来的情感倾向可以广泛应用于针对用户对产品的评价的调查，也可以

对其他消费者进行积极的指引。

文本情感分析是一种利用算法来分析文本中表达的情感的方法，可以做到快速判断、定位和归纳出词语、句子、篇章表达的情感属于正向或者负向[1]。面对如今互联网时代，每天出现的很多条评论、文章，如果商家或者管理员可以快速地处理这些文字，可以节省很多时间，这是很有必要的。在人类社会的各个方面，例如在市场营销、金融、通讯行业、生物医疗等，文本情感分析都是很有意义，全世界也关注到了其重要的商业意义。例如，现在某个人如果想要购买一件之前没买过的产品，询问家人朋友来知道商品的品质等属性已经不是第一选择了，而通过查看各个 APP 上的评论和评分往往就足以了解产品的优缺点。因此，有效的文本情感分析可以收集公众的意见，监控公众情绪变化和舆论方向，可以避免恶性事件发生，也可以快速了解产品的状态。

文本情感分析的基本流程如图 1.1 所示，主要包含了原始文本获取、文本预处理、语料库和情感词典构建、情感分析、情感抽取等。文本预处理主要包括了分词、去除停用词等[2]。经过几十年的发展，原始文本获取通过 Python 及其它语言的爬虫技术实现已经比较完善，分词也可以通过调用 jieba 库的 API 来完成。目前阶段，情感分析的重难点主要在语料库和情感词典构建、情感的分析和抽取这几个阶段。

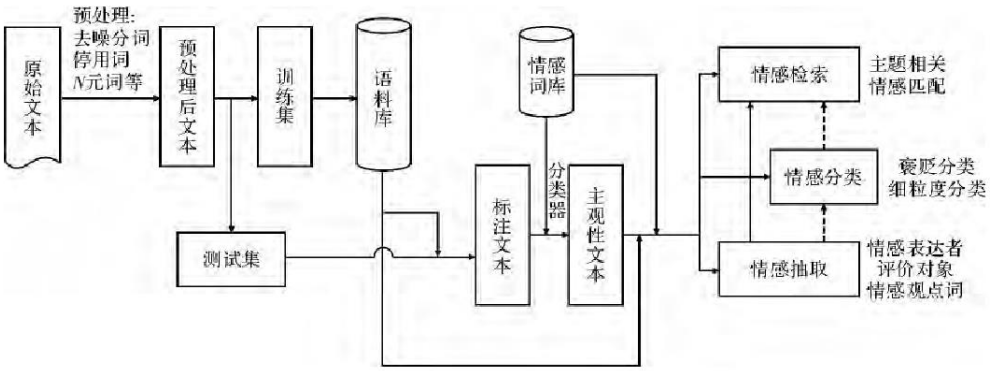

图 1.1 文本情感分析流程图

通常互联网里的中文文本主要由主观性评价构成，这是因为网络文本，尤其是评论文本，是用来表达作者对某个事物的看法的工具。主观性评价是写文本的人对某个事物产生的看法和评价的描述，带有其本人对此明确的喜好或者厌恶的情感。本项目利用酒店评论，则评论内容主要是用户本人的主观性评价。在美团、爱彼迎等平台上，许多热门酒店会有成千上万条评论，如果手动分析的话很费时间和精

力。通过利用机器学习的方式来进行自动的文本情感分析，则可以提高处理的效率，得到较为准确的结果。

文本情感分析是数据挖掘与分析的较为新兴诞生的一种领域，它的研究和应用的价值显得尤为重要。根据王洪伟教授的定义，情感分析的定义是对用户在互联网上发表的内容，如评论、评价等，进行监测和分析，目的是识别出文本背后带有的情感信息，并且得到大众普遍的情感信息与看法，发现用户的情感变化规律[3]。因此，情感分析中包括了非常多的难度大、有挑战的任务，不过如果可以完成这些任务可以让商家获得顾客需求和反馈意见，也可以找到潜在的购买者。

1.2 研究现状

尝试理解人类的情感、行为并作出判断一直是人工智能领域的一个重要目标和任务。在当今时代，受益于各大网络社交平台的蓬勃发展，这一领域有了极其活跃的发展。情感分析往往可以有效地帮助人们快速确定观点、喜好和情感，可以用于网络言语监控，商家查看与了解消费者们对产品的喜好或厌恶，或者实时监控网上评论中的涉及差评的信息。文本情感分析是多学科交叉的研究方向，主要涉及了自然语言处理、计算语言学、人工智能、机器学习等。

文本的规模和处理文本的长度往往是不同的，较短的例如一个或几个词语，较长的则是一段文字或者一篇文章，情感分析与可视化因此可以分为词语级、句子级和篇章级[4]。对于不同的级别，文本情感分析的处理方式也有些许不同，如词语级可以考虑词语本身的意思，而篇章级则需要考虑句子之间的关系。本篇论文以谭松波酒店评论语料为训练样本，单个文本规模常为多个句子，故研究层次主要是篇章级。

针对情感分析任务，最早的研究是在上世纪九十年代开展的，那时已经出现了构建语义词典的研究。而后人们发现了一些特征词，如连词、形容词等，对情感词会产生影响，于是利用互信息扩展形成了正向和负向情感词典。基于早期情感词典分类的方法预测情感可以达到 74%的预测率。

目前，文本情感分析的方法主要可以分为基于无监督学习方法和基于监督学习方法两种。基于无监督学习方法进行文本情感分析的相关研究成果较少，目前主要有朱嫣岚提出的用 HowNet 对中文语料进行情感倾向计算。基于知识库的方法是通过现有

的知识库进行语义分析，判断语句的情感倾向。

现如今有一些做好的工具包面世，如斯坦福 CoreNLP、Alias-i 的 Lingpipe（使用 Logistic 回归进行文档分类）、SentiWordNet 和来自不同来源的综合库，包括其他几种用于文本挖掘的技术，通过给句子打分来评估特征选择对整体情感分析的影响使用不同的评分方法进行评价[5]。

主流的研究方向为基于监督学习，利用深度学习进行文本情感分析，在自然语言处理领域里的神经网络模型主要有卷积神经网络（Convolutional Neural Networks, CNN）和长短期记忆网络（Long Short-Term Memory，LSTM）两种。CNN 非常擅长处理分类，一般来说，这是由若干个卷积和池化操作组成。CNN 的每个单元只会与上一层的部分单元相连接，卷积层尝试分析神经网络的每个块以获得更多的特征。利用长短时间记忆网络 LSTM 模型可以训练出一个具有 positive, neutral, negative 的情感三分类器，输入为句子的序列[6]。CNN 在处理高维数据时，因为其具有池化的功能，效率十分高并且准确率较高；同时可以自动提取特征。但是，CNN 的缺点在于其采用了梯度下降算法，导致其结果很容易收敛在局部最小值而非全局最小值，导致在面对较为复杂的情况时，不太能够得到正确的结果；另外池化层往往会忽略局部与整体之间的联系，因而丢失大量与上下文有关的有价值信息。LSTM 模型可以解决循环神经网络中存在的相互依赖问题，因为它是由非线性单元构成，所以可以构造较大的神经网络。但是 LSTM 具有的缺点是每一个 LSTM 单元都需要 4 个全连接层，对于大型数据其处理时间非常慢；对于文本分析，往往需要同时考虑上下文，这同样是 LSTM 无法做到的。对于大多数评价信息中，情感只有喜欢、不喜欢和中性三种，因此可以训练出具有较好预测能力的模型，也可以采用分等级的方法，例如将情感分为 1-5 级，本文采用前者[7]。

1.3 论文研究内容

本文主要采用双向长短时记忆网络模型(Bi-directional Long Short-Term Memory，biLSTM)来对中文文本的语料的数据进行训练和预测，在利用 jieba 对语料进行预处理后，利用 gensim 将词语转化为词向量，预训练词向量选用"chinese-word-vectors"，根据 keras 要求建立一个维度为 300 的词向量矩阵。对建立好的词向量矩

阵利用biLSTM模型进行训练,从神经网络全连接层的Sigmoid中返回一个概率数值，根据结果评价语句为正向、负向或者中立。最后，利用PyQt5工具设计一个可视化界面，实现语句输入窗口与评价结果的返回。本项目全部代码均用Python3.8实现，实验采用的工具包主要包括numpy, jieba, Tensorflow, gensim等，神经网络框架选用Keras。

本文采用jieba库和gensim库来对文本进行预处理，jieba 是一个基于python实现的中文分词组件,而本项目的预处理部分主要流程是利用jieba结合自建词典进行分词，通过读取数据源，将停用词、自建词表载入到数组中，通过使用TF-IDF算法来对文本进行分析和处理,再声明一个数组用来保存分词后的词语集合。结合textrank，对词频进行统计与排序。在进行舆情分析的时候，用词云这个工具可以很好地进行可视化，实现词频统计的显示，这里我用到了wordcloud模块。舆情分析方面，本项目主要基于biLSTM模型实现舆情分析。调取前期分词结果保存的数组，对每个情感词进行评分，并记录匹配到的情感词分值，最后，对系统得出的分值进行判断，如果分值大于0.6，表示认为情感倾向为积极的，即positive；如果小于0.4，则表示认为情感倾向为消极的，即negative；介于0.4到0.6之间则认为是中性情感,即neutral。

1.4 论文组织结构

本文分为五个章节，按照总体介绍、相关技术与模型、实验过程与结果、可视化部分、总结的顺序撰写。第一章为引言，总体介绍本文章的研究背景和现状，分为研究背景、研究意义、研究现状、论文总体概述、论文组织结构五个部分。通过第一章的描述，可以知道本项目研究在世界上的研究进展、主流研究方法和存在的缺点，作者对此进行了相应改进并得到本实验成果。

第二章为系统开发的相关技术与相关模型。主要相关技术包括：编程语言Python、神经网络框架Keras和自然语言处理工具jieba,本文对此进行了简单的描述。此外，运用到的一些相关模型有：biLSTM双向长短时记忆网络、词向量矩阵和BERT语言模型。

第三章主要介绍了本实验的完成过程和对结果进行分析,首先介绍本实验的思路，

如何利用 biLSTM 进行文本情感分析的大体过程和可行性；然后是对文本的处理过程，在利用 jieba 进行处理过程时需要注意哪些事项和最后呈现的数据类型；接着描述仿真实验过程，包括机器学习过程、实验数据与环境、实验评价指标，知道仿真实验的细节描述和评价方式；最后是对结果进行分析，从索引长度、训练过程、训练结果 accuracy 指标评价三个方面分析。

第四章为可视化，主要介绍了本文在可视化方面所使用的技术与方法。

第五章为论文总结与展望。

第 2 章 系统开发相关技术与模型

本章主要介绍系统开发的相关技术与相关模型，相关技术主要包括 Python、Keras 神经网络框架、Jieba 中文分词库，相关模型主要包括 biLSTM 双向长短时记忆网络、词向量矩阵、BERT 预训练语言模型。

2.1 相关技术

2.1.1 Python 介绍

Python 是一种解释性和面向对象设计语言。这是一种开源的语言，它有非常多的库可以用来高效开发各种应用软件。Python 常被用于制作操作系统、进行科学计算、GUI 的开发、和软件设计。Python 的设计风格简洁明了，这使得 Python 成为一种可读性和可维护性很强的语言。

Python 在完成科学计算和进行机器学习上显得尤其高效，这受益于它有许多开源的库和软件。Numpy 是用来提供进行数据分析和科学计算的接口的库，它支持完成矩阵操作和线性代数。Pandas 提供简单数据结构和数据分析的工具，帮助使用者了解数据和索引之间的关系。Jieba 是用于中文自然语言处理的库，可以高效完成分词、词频统计、实体识别等功能，具体内容见 2.1.3 节。此外，Tensorflow 是用于机器学习的平台，它允许将神经网络的模型部署在 CPU 或者 GPU 上，并提供相应的 API，这样使得机器学习变得简单。

Python 具有简单易学、可移植、解释性、面向对象、可扩展、拥有丰富的库等特点。机器学习是通过将数据提供给计算机，并转换成相应的决策模型，再对未来出现的数据进行预测。因此，利用 Python 里的库，可以快速高效搭载相应的框架，配置参数并完成训练和预测。

2.1.2 Keras 介绍

Keras 是神经网络模型框架的一种，基于 Python 语言编写，并且是搭载在 Tensorflow 上，可以作为 Tensorflow, Theano 等的接口，主要是用在深度学习模型的设计、调试、应用等，这是一个为人们操作简化而设计的 API。Keras 遵循的是减

少人们认知压力，减少操作规模，因此 Keras 提供了一套一致且简单的 API，减少常见用例所需的用户操作数量，并提供清晰且可操作的错误反馈消息，它还具有全面的文档和开发人员指南。

Keras 模型是采用面向对象程序设计方法来编写的，它的模块化和可扩展性具有较高的水平。它提供了统一的 API，并记录在了文档里，这样做可以将用户的操作量降到了尽可能最小，并在用户出现错误时候给予清晰明确的反馈。Keras 模型尽可能独立地构成，具有高内聚低耦合特点，特别是在神经网络层、损失函数等地方，它们可以构成新的模块。Keras 是完全通过 Python 语言进行编写和开发的，在 Python 中也定义了它的模型，可以直接加载到项目里，具有简单且易操作的特点。

Keras 是目前最受研究人员喜爱的模型之一，在 arXiv.org 的科学论文中，引用量达到了第二位。目前，Keras 拥有个人用户超过了 25 万人，在各个行业中都有较高的使用率。此外，Keras 还可以在 Netflix, Uber 等网站上使用。

Tensorflow 上所有的开发功能都被部署了，可以将 keras 模型导出为 JavaScript 来直接在浏览器中运行，也可以将其导出为 TF-Lite 以在 IOS、Android 和嵌入式设备上运行。

2.1.3 jieba 介绍

jieba 库是最优秀的第三方中文分词工具库之一，它是由一位百度工程师设计完成的。除了中文分词外，它还可以进行关键词抽取、词频统计等任务。

Jiba 提供了四种分词方式，精准模式主要是用于将词语最精确地分开，但不会去考虑上下文关系，只是采取最高概率的情况。搜索索引模式主要用于在精确模式的基础上，对较长的词汇进行分词，这个模式更多是应用在搜索引擎的分词功能里。全模式是把一个句子里面所有的可能形成词语的部分都挖掘出来，这个模式处理的过程时间较长且步骤较复杂但是比较精确。Paddle 模式是利用了 Paddle 深度学习框架，训练出来了一个模型用来进行机器学习分词，它也支持词性标注 POS Tagging。

除分词以外，jieba 库也可以完成词频统计的任务，首先利用键值对的形式储存词语和词语出现的次数，再根据次数进行排序得到词频序列。

在关键词抽取方面，jieba 主要通过利用 TF-IDF 算法和 TextRank 算法进行关键词抽取。在关键词提取过程中运用到的逆向文件频率能够被转到自定义语料库的路径

里面；停用词语料库也可以使用自定义的语料库，通过修改停用词语料库的路径转到这个语料库里面。基本思想是首先将待抽取的关键词文本进行分词，然后找到词语之间的关系并且将这些关系构建成图来表达，最后计算节点之间的 PageRank。

2.2 相关模型

2.2.1 biLSTM 双向长短时记忆网络

在处理中文文本时，当前词语同时与上下文都产生联系，所以本文采用双向长短时记忆网络来同时分析上下文的信息。如图 2.1 所示，biLSTM 是通过构建一个正向 LSTM 和一个负向 LSTM 叠加而成的模型，可以同时连续分析上下文信息，找到词语与上下文之间、上下文之间的依赖关系。biLSTM 模型是对 LSTM 的一种优化，LSTM 无法对从后到前的信息进行编码，即无法处理上下文信息，而通过 biLSTM 则可以很好找到词语与前后文之间的相互作用关系。尤其在对于高粒度的分类时，如对于正向、中性、负向情感分类时候，需要注意上下文、程度词、否定词的交互，这时候利用 biLSTM 会更具有优势。

当分析句子时，往往会把句子看作是词语的叠加，例如当作相加，则句子里的情感便是一些特定词如“不”的叠加。但是这种方法不能够处理到词语在句子中的前后关系，即上下文位置的顺序不同会导致错误，如“我不认为这样好”和“我认为这样不好”，其实在情感上是有细微差距的。使用 LSTM 可以利用其通过训练学习和遗忘一些知识，从而很好地捕捉到词语的距离对情感的依赖关系。而 biLSTM 因为是两个 LSTM 模型的叠加，可以更好地注意到情感词、程度副词、否定词的关系，更好捕捉双向语义的情感。

LSTM 模型是由在 t 时刻的输入词 Xt，细胞状态 Ct，隐层状态 ht，遗忘门 ft，记忆门 it 和输出们 ot 组成。biLSTM 是通过组合了正向的 LSTM 和负向的 LSTM 来构成，通过两个 LSTM 的叠加，可以使句子的表示里包含前向和后向的信息。

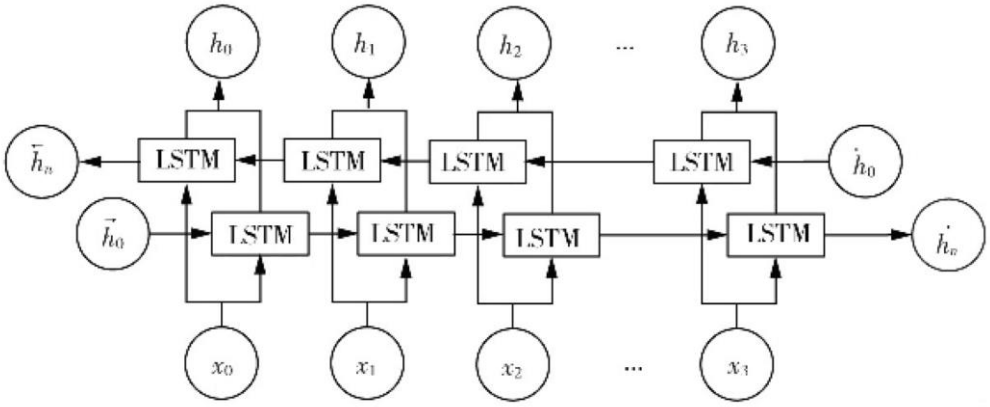

图 2.1 双向长短时记忆网络

2.2.2 词向量矩阵

词向量是我们在增加分析词、句子、文章的关系上的能力的一个重大飞跃，通过使用词向量，机器可以了解有关词语之间的关系，这比传统的词语表示法能够得到更多的信息。词向量矩阵是含有词语信息的向量矩阵。

在传统的 NLP 方法里，词语编码往往会使用 one-hot 和 bag-of-words 模型，这对机器学习处理 NLP 问题有一定的作用，但是不能捕获和词语含义、上下文关系的信息。例如，“狗”和“猫”都是家庭常见宠物，它们之间有许多相似的地方，但它们同样属于不同物种。因此，在一些维度上二者十分相近，但在如物种、体型、习性等属性上却相差很远。词向量则可以很好解决高维度的词语含义问题。

词向量矩阵可以利用两种模型体系结构中的一种来生成单词的分布式表示形式：连续词袋（Continuous Bag-Of-Words Model，CBOW）或连续跳过语法(Skip-gram)。在 CBOW 体系结构中，该模型从周围上下文词的窗口中预测当前词的意思和情感，在这种情况中，上下文词的顺序不会影响预测的结果。在连续跳过语法体系结构中，该模型使用当前单词来预测上下文单词的周围。连续跳过语法架构对附近上下文词的权重更大。

根据 keras 的要求，准备一个词向量维度为 300 的词向量模型，如图 2.2 所示，并选择前 50k 词频最高的词进行预训练。在这个模型里，每个词成为一个索引，对应长度为 300 的向量，这是因为 biLSTM 不能够直接处理汉字。词向量我选择使用的是北京师范大学中文信息处理研究所与中国人民大学 DBIIR 实验室的研究者合作开发并且开源的"chinese-word-vectors"。

```
array([    0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
         290,  3053,    57,   169,    73,     1,    25, 11216,    49,
         163, 15985,     0, 55319,    30,     8, 50197,     1,   228,
         223,    40,    35,   653,     0,     5,  1642,    29, 11216,
        2751,   500,    98,    30,  3159,  2225,  2146,   371,  6285,
         169, 27396,     1,  1191,  5432,  1080, 20055,    57,   562,
           1, 22671,    40,    35,   169,  2567,     0, 42665,  7761,
         110,     0, 65932, 41281, 80655,   110,     0, 35891,   110,
           0, 28781,    57,   169,  1419,     1, 11670, 80655, 19470,
           1, 58932, 76788,   169, 35071,    40,   562,    35, 12398,
         657,  4857])
```

图 2.2 预处理之后的词向量矩阵

2.2.3 BERT 预训练语言模型

基于转换器的双向编码表征(Bidirectional Encoder Representations from Transformers, BERT) 是 Google 开发的基于变压器的机器学习技术，用于自然语言处理预训练。

BERT 的一项关键创新是将 Transformer 的双向的训练应用在语言建模当中。经过双向训练的语言模型能够比单向的语言模型更具有深层次的语言背景和过程意识。BERT 使用了一个 Transformer，如图 2.3 所示。这是一种用于学习文本中词语之间的上下文关系的学习机制。最初形式的 Transformer 包括了两种独立的机制：读取文本输入的编码器、产生任务预测需要的解码器。

与定向模型按照一定顺序读取并输入文本内容（从左到右或者从右到左）不同的是，Transformer 编码器一次读取的是整个句子。因此，普遍将它认为是双向的。这种特征使得这个模型可以根据单个词语词的所有周围情况（词语的上下文）的排列组合来学习整个文本，并找到上下文之间的相互依赖关系。

BERT 使得通过利用机器学习进行自然语言处理方面有了长足的进步。目前，它具有容易使用、并且可以快速对参数进行微调，它的优势能够在未来得到广泛的推广。

由于传统语言模型在处理多义词、上下文关系的问题上表现不好，在本项目进行篇章级文本情感分析时，采用 BERT 模型可以有效获取上下文关系。和传统的 LSTM 与

CNN 做对比，可以发现，双向的 Transformer 抽取器能够获得更多的上下文互信息[8]。

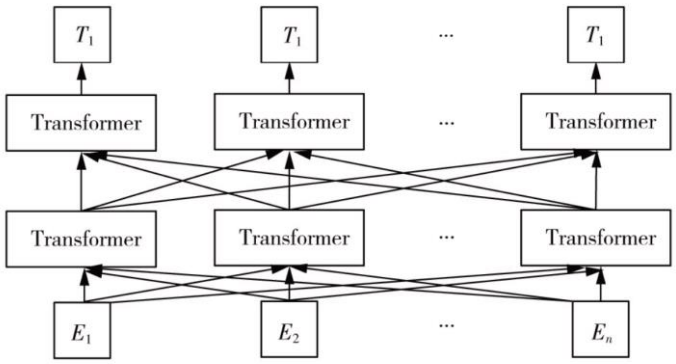

图 2.3 BERT 预训练模型

2.3 本章小结

本章主要介绍了本文开发系统所使用的相关技术和相关模型，相关技术主要包括 Python、Keras 神经网络框架、Jieba 中文分词库，相关模型主要包括 biLSTM 双向长短时记忆网络、此向量矩阵、BERT 预训练语言模型。通过以上技术和模型可以有效地进行自然语言处理训练，获取上下文关系。

第 3 章 中文情感分析神经网络模型

本章主要描述中文情感分析神经网络模型的研究思路、模型描述、仿真实验的流程、实验数据与环境、结果分析。

3.1 模型描述

为了较好地对评论文本里的中文文本实现情感分析，本文采用谭松波酒店评论语料作为训练样本，并且训练出一个具有 positive, neutral, negative 的情感三分类器。词向量这里使用的是北京师范大学中文信息处理研究所与中国人民大学 DBIIR 实验室的研究者合作开发并且开源的"chinese-word-vectors"。

本项目中，首先对文本进行预处理。首先，因为标点一般对于文本情感产生不了影响，所以去掉每个样本的标点符号。然后用 jieba 分词，并将它索引化。由于 LSTM 神经网络模型不能够直接处理汉字，所以还需要将词转化为词向量。

利用 Keras 搭建神经网络模型。针对文本情感分析需要考虑到文本序列、情感特征的重要度等，本 BiLSTM 模型包含四层， embedding 层是第一层的模型，对神经网络进行降维；第二层为 BiLSTM 层，其第二个维度是 sequence，对于 32 个单元，由于使用双向 LSTM，故设置 64 个单元；第三层为单向 LSTM 层，将返回的 sequence 导入单向的 LSTM 中，并将一个结果进行输出返回；最后一层为全连接层，从 Sigmoid 中输入一个数值，并根据此判断输入的文本是正向情绪还是负向情绪。

可视化方面，采用 pyqt5 制作 GUI，使用 Python3.8 语言开发 PyQt5,利用 PyQt5 提供的 button, label 等接口。在新建窗口时候，采用 Widget 进行开发；按钮选用 pushButton，并设置其大小、文字等属性；情感分析输入文本框采用 label,并调整字体颜色、大小；在最后输出结果中，根据输出结果进行三分类 ，分为“正面评价”、“负面评价”、和“中性评价”，并设计相应图片显示出来。

3.2 处理过程

本论文采用 biLSTM 和神经网络模型对中文文本语料进行情感分析，其具体过程如下。本项目采用 biLSTM 模型进行对中文文本语料的训练，在通过使用 jieba 对语

料的文本进行预处理后，利用 gensim 将词语转化为词向量，预训练词向量选用"chinese-word-vectors"，由于 keras 对词向量矩阵的规模有要求，本文建立了一个维度为 300 的词向量矩阵。对建立好的词向量矩阵利用 biLSTM 模型进行训练，从神经网络全连接层的 Sigmoid 中返回一个概率数值，根据结果评价语句为正向、负向或者中立。最后，通过使用 PyQt5 工具，本文设计了一个可视化的图形界面，实现语句输入窗口与评价结果的返回。本项目全部代码均用 Python3.8 实现，实验采用的工具包主要包括 numpy, jieba, Tensorflow, gensim 等,本文的神经网络框架选用 Keras。

首先为模型准备词向量矩阵，根据 Keras 要求，需要准备一个维度为 300 的矩阵，词汇的数量设置为 100000，每一个词汇都用一个长度为 300 的向量来表示。这个预训练模型里一共有 260 万个词汇量，选取前 100000 个词频最高的词，这样可以提高效率，在训练样本很小情况下效果也比较好。

对训练样本进行分割，采用前 90%的样本用来训练，后 10%样本用来测试。针对文本情感分析需要考虑到文本序列、情感特征的重要度等,本 BiLSTM 模型包含四层，如图 3.1 所示，embedding 层是本模型的第一层，主要目的是对神经网络进行降维；第二层为 BiLSTM 层,其第二个维度是 sequence,对于 32 个单元,由于使用双向 LSTM，故设置 64 个单元；第三层为单向 LSTM 层，将返回的 sequence 导入单向的 LSTM 中，并返回一个结果；最后一层为全连接层，从 Sigmoid 中输入一个数值，并根据此判断输入的文本是正向情绪还是负向情绪[9]。

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 236, 300)	30000000
bidirectional (Bidirectional	(None, 236, 128)	186880
lstm_1 (LSTM)	(None, 16)	9280
dense (Dense)	(None, 1)	17

Total params: 30,196,177
Trainable params: 196,177
Non-trainable params: 30,000,000

图 3.1 BiLSTM 模型层次结构

主要代码如下：

```
model = Sequential()
```

```
    model.add(Embedding(num_words,
                        embedding_dim,
                        weights=[embedding_matrix],
                        input_length=max_tokens,
                        trainable=False))
    model.add(Bidirectional(LSTM(units=64, return_sequences=True)))
    model.add(LSTM(units=16, return_sequences=False))
    model.add(Dense(1, activation='sigmoid'))
    optimizer = Adam(lr=1e-3)
    model.compile(loss='binary_crossentropy',
                  optimizer=optimizer,
                  metrics=['accuracy'])
    path_checkpoint = 'sentiment_checkpoint.keras'
    checkpoint = ModelCheckpoint(filepath=path_checkpoint,
monitor='val_loss',
                                 verbose=1,
save_weights_only=True,
                                 save_best_only=True)
    earlystopping = EarlyStopping(monitor='val_loss', patience=5,
verbose=1)
    lr_reduction = ReduceLROnPlateau(monitor='val_loss',
                                     factor=0.1, min_lr=1e-8,
patience=0,
                                     verbose=1)
    callbacks = [
        earlystopping,
```

```
    checkpoint,
    lr_reduction
]
model.fit(X_train, y_train,
          validation_split=0.1,
          epochs=20,
          batch_size=128,
          callbacks=callbacks)
result = model.evaluate(X_test, y_test)
print('Accuracy:{0:.2%}'.format(result[1]))
model.save('./model/my_model.h5')
model.save_weights('./checkpoints/my_checkpoint')
```

3.3 仿真实验

3.3.1 实验数据与环境

为了较好地对评论文本里的中文文本实现情感分析，本文采用谭松波酒店评论语料作为训练样本，并且训练出一个具有 positive, neutral, negative 的情感三分类器。本文的词向量采用的是北京师范大学中文信息处理研究所与中国人民大学 DBIIR 实验室的研究者合作开发并开源的"chinese-word-vectors"。实验采用工具包为 Anaconda，编程语言这里选择使用的是 Python3.8，并且将项目搭建在 Jupyter Notebook 和 PyCharm 上。使用的框架是 Keras。

在做可视化时候，本文采用 pyqt5 框架进行制作 GUI，利用 WordCloud 库制作词云。

本文采用准确率 Accuracy，损失率 Loss。和其他的项目里一样，其指标是通过 TP、TN、FP、FN 计算。TP 表示预测值为真，实际值也为真；TN 为预测值为真，实际值为假；FP 为预测值为假，实际值为真；FN 为预测值为假，实际值也为假[10]。

Accuracy = (TP+TN)/(TP+TN+FP+FN)

3.3.2 实验过程

原始文本中常含有数字、字母、特殊符号、标点等，因为标点一般对于文本情感产生不了影响，所以本文需要去掉所有样本里的全部标点符号。对于特殊的符号，比如在评论里可能出现的表情、颜文字等，也是应当去掉。以及对于大量重复性评论，类似“默认好评”等，也需要在数据预处理时候清洗掉。否则会造成文本模型过于稀疏，造成过拟合，影响分析结果。然后用 jieba 分词，并将它索引化。由于编码的缘故，使得 LSTM 神经网络模型不能够直接处理汉字，所以还需要将词转化为词向量。

主要代码如下:

```
train_tokens = []
for text in train_texts_orig:
    text = re.sub("[\s+\.\!\/_,$%^*(+\"\']+|[+——！，。？ 、~@#￥%……&*（）]+", "",text)
    cut = jieba.cut(text)
    cut_list = [ i for i in cut ]
    for i, word in enumerate(cut_list):
        try:
            cut_list[i] = cn_model.vocab[word].index
        except KeyError:
            cut_list[i] = 0
    train_tokens.append(cut_list)
```

预处理之后，得到的文本是索引化后的数字，如图 3.2 所示。

对于预处理后的文本，对其进行实体抽取，这是因为实体抽取过程中可以获得清楚、明确的信息。抽取实体，即进行命名实体识别，包括了实体的检测和分类。获得有效的实体后，再进行关系抽取，获得实体 A-关系-实体 B 的三元组。通过实体识别得到的序列，在进行文本分类和情感分类时候可以获得更好的表现[11]。

本项目重难点在于如何搭建有效的 biLSTM 模型。利用模型训练出来的情感分析器得到的结果必须是偏向 0 或者 1 的，而不能够在 0.5 左右徘徊。据此，进行文本分

类预处理、构建合适的词向量矩阵是有用的。并且还要选择适当的评价标准，我选用Tensorflow 的 Accuracy 进行评价。

对预处理后的文本，本文将其加载进 biLSTM 模型中进行训练，通过 20 批的训练得到了准确率为 88.50%的训练模型。利用这个训练模型可以进行文本的情感分析与预测，结果是可信的。

主要代码如下：

```
def predict_sentiment(text):
print(text)
# 去标点
    text = re.sub("[\s+\.\!\/_,$%^*(+\"\']+|[+——！，。？、~@#￥%……&*（）]+", "",text)
# 分词
    cut = jieba.cut(text)
    cut_list = [ i for i in cut ]
# tokenize
    for i, word in enumerate(cut_list):
try:
            cut_list[i] = cn_model.vocab[word].index
if cut_list[i] >= 50000:
                cut_list[i] = 0
        except KeyError:
            cut_list[i] = 0
    # padding
    tokens_pad = pad_sequences([cut_list], maxlen=max_tokens,
                           padding='pre', truncating='pre')
# 预测
    result = model.predict(x=tokens_pad)
    coef = result[0][0]
if coef >= 0.5:
```

```
    print('是一例正面评价','output=%.2f'%coef)
else:
    print('是一例负面评价','output=%.2f'%coef)
return coef
```

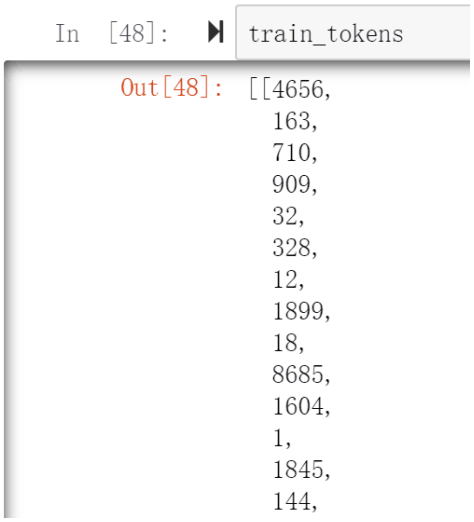

图 3.2 预处理后的文本

3.3.3 结果分析

对于切分和索引化后的词语，索引长度不应太短，否则会丢失很多信息；同时也不应太长，否则会浪费掉许多计算和训练相关的资源，包括对 GPU 的使用。如图 3.3 中，本项目中大部分索引样本长度都在 0 到 100 之间，对其取对数 log 后，变成类似于正态分布(图 3.4)。因此，经过计算可知，将本项目中最大长度设置为 236，可以覆盖置信区间为 95%的样本。

在训练样本中，本文设置的参数为 validation_split=0.1,epochs=100,batch_size=128，并且设置了 early stopping，当 3 个 epoch 内 val_loss 没有改善的话，则停止训练。图 3.5 为本模型训练的过程。

经过训练后，将训练模型应用于验证集，得到结果如图 3.6 所示，val_loss 为 0.3356，acc 为 0.8600。由此可见，BiLSTM 凭借其对于处理较长文本具有记忆功能，能够达到 88.50%左右的准确率。

如图 3.7 所示，BiLSTM 模型对于酒店评论的分析中，其输出结果大多偏向 0 或

者 1，即有较明显的正面或者负面的情感倾向，而不会取到 0.5 左右的模糊值。因此，可以认为，利用 BiLSTM 训练出来的中文文本情感分析模型输出的结果是健壮的，可以对其有信心。我们的研究发现，利用 BiLSTM 对情绪分类可以产生很好的影响（0.932）。

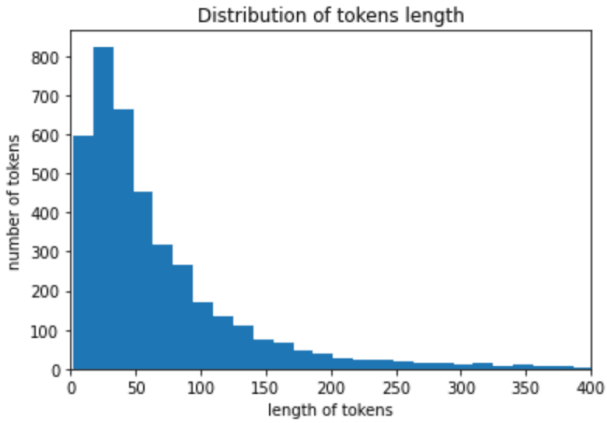

图 3.3 引长度分布

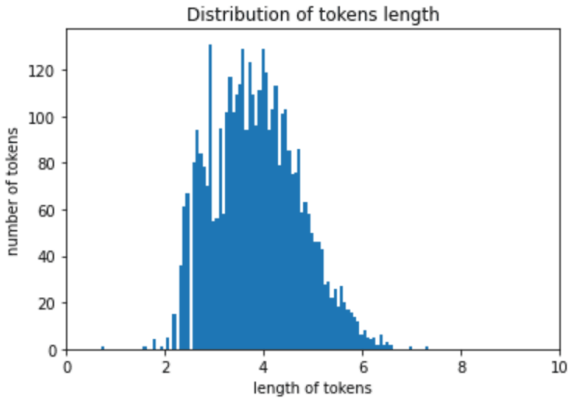

图 3.4 取对数后的索引长度分布

```
Epoch 5/100
3200/3240 [============================>.] - ETA: 1s - loss: 0.2128 - acc: 0.9219
Epoch 00005: val_loss did not improve from 0.29723

Epoch 00005: ReduceLROnPlateau reducing learning rate to 1.0000000656873453e-06.
3240/3240 [==============================] - 118s 36ms/sample - loss: 0.2124 - acc: 0.9222 - val_loss: 0.3009 - val_acc: 0.8889
Epoch 6/100
3200/3240 [============================>.] - ETA: 1s - loss: 0.2045 - acc: 0.9253
Epoch 00006: val_loss did not improve from 0.29723

Epoch 00006: ReduceLROnPlateau reducing learning rate to 1.0000001111620805e-07.
3240/3240 [==============================] - 117s 36ms/sample - loss: 0.2061 - acc: 0.9241 - val_loss: 0.2999 - val_acc: 0.8889
Epoch 7/100
3200/3240 [============================>.] - ETA: 1s - loss: 0.2047 - acc: 0.9247
Epoch 00007: val_loss did not improve from 0.29723

Epoch 00007: ReduceLROnPlateau reducing learning rate to 1.0000000082740371e-08.
3240/3240 [==============================] - 117s 36ms/sample - loss: 0.2058 - acc: 0.9238 - val_loss: 0.2998 - val_acc: 0.8889
Epoch 8/100
3200/3240 [============================>.] - ETA: 1s - loss: 0.2049 - acc: 0.9244
Epoch 00008: val_loss did not improve from 0.29723

Epoch 00008: ReduceLROnPlateau reducing learning rate to 1e-08.
3240/3240 [==============================] - 111s 34ms/sample - loss: 0.2057 - acc: 0.9241 - val_loss: 0.2998 - val_acc: 0.8889
Epoch 9/100
3200/3240 [============================>.] - ETA: 1s - loss: 0.2074 - acc: 0.9231
Epoch 00009: val_loss did not improve from 0.29723
3240/3240 [==============================] - 109s 34ms/sample - loss: 0.2057 - acc: 0.9241 - val_loss: 0.2998 - val_acc: 0.8889
```

图 3.5 训练过程

图 3.6 测试结果

酒店设施不是新的，服务态度很不好
是一例负面评价 output=0.14
酒店卫生条件非常不好
是一例负面评价 output=0.10
床铺非常舒适
是一例正面评价 output=0.82
房间很凉，不给开暖气
是一例负面评价 output=0.27
房间很凉爽，空调冷气很足
是一例正面评价 output=0.75
酒店环境不好，住宿体验很不好
是一例负面评价 output=0.06
房间隔音不到位
是一例负面评价 output=0.32
晚上回来发现没有打扫卫生
是一例负面评价 output=0.29
因为过节所以要我临时加钱，比团购的价格贵
是一例负面评价 output=0.11

图 3.7 BiLSTM 输出结果

3.4 本章小结

本章主要描述了中文情感分析神经网络模型的研究思路、模型描述、仿真实验的流程、实验数据与环境、结果分析。研究思路主要提出了构建本 biLSTM 模型的思路和进行文本情感分析的思路。模型描述主要介绍 biLSTM 模型和词向量矩阵模型的思路和搭建。仿真实验介绍了实验的数据与环境、实验过程、实验结果分析。

第 4 章 情感分析可视化

本章主要介绍本文运用到的可视化工具 PyQt5，以及可视化结果的分析和主要代码展示。

4.1 可视化工具

本文在可视化方面选用的 PyQt5 工具包。PyQt5 是基于 python 的一个针对 Qt5 应用的一套框架。Qt5 是以功能强大而著称的一个 GUI 库。它拥有 600 多个类、6000 多种方法和功能，也是一个多平台的工具包。

PyQt5 的类被分成了几个模块，包括 QtCore, QtGui, QtTest 等。QtCore 模块包括了非 GUI 的核心代码，主要用于处理时间，文件和目录，各种数据类型，流，URL，mime 类型，线程或进程等。QtGui 是一个图形化用户界面设计的模块。

PyQt5 的优势主要有：跨平台的效果好，它可以兼顾多种平台，包括 Windows, Linux, Max；易上手，它是基于面向对象思路设计的，命名、继承等都很规范；功能强大，可以实现 GUI 的绝大部分功能；开源免费，维护简单；文档丰富，参考 Qt 文档，编写代码较容易。

4.2 可视化结果分析

可视化方面，本文采用 pyqt5 制作 GUI，使用 Python3.8 语言开发 PyQt5,利用 PyQt5 提供的 button, label 等接口。在新建窗口时候，采用 Widget 进行开发；按钮选用 pushButton，并设置其大小、文字等属性；情感分析输入文本框采用 label，并调整字体颜色、大小；在最后的输出结果时候，根据输出结果进行三分类 ，分为“正面评价”、“负面评价”、和“中性评价”，并设计相应图片显示出来。

如图 4.1 所示，首页由标题“中文情感分析”，四个功能按钮“情感预测”、“正向情感词云”、“负向情感词云”、“关于”，和一个背景图片构成。标题的汉字通过 WindowsTitle 功能创建，并修改其字体、颜色、大小；四个功能按钮通过 pushButton 创建，并链接到之后的页面里；图片则通过 QtGui 加载位于项目内的一个图片。

如图 4.2 所示，点击首页“情感预测”按钮，跳转到“中文情感预测系统”界面，含有两个文本框、一个“点击预测”按钮、和一个情感结果显示的图片。文本框利用 textEdit 创建，并修改其大小、文字大小、位置、并有提示文字“请输入文本”和“返回结果”；“点击预测”按钮通过 pushButton 创建，并修改背景颜色、字体、字号、位置；“返回结果”文本框链接到 biLSTM 模型，将“输入文本”的文字输入进模型，并从模型的 Sigmoid 函数返回结果到后端，后端将与结果对应的输出文本返回在文本框，并修改其字体、字号等。右侧的图片也根据返回的结果输出对应的图片。当结果做出负面评价、正面评价、中性评价时的反馈结果分别如图 4.3、图 4.4、图 4.5 所示。

“正向情感词云”和“负向情感词云”均输出通过训练文本找到的最高频出现的词语，利用 WordCloud 实现。图 4.6 和图 4.7 分别为正向情感词云和负向情感词云的展示。

图 4.1 首页界面

图 4.2 中文情感预测系统界面

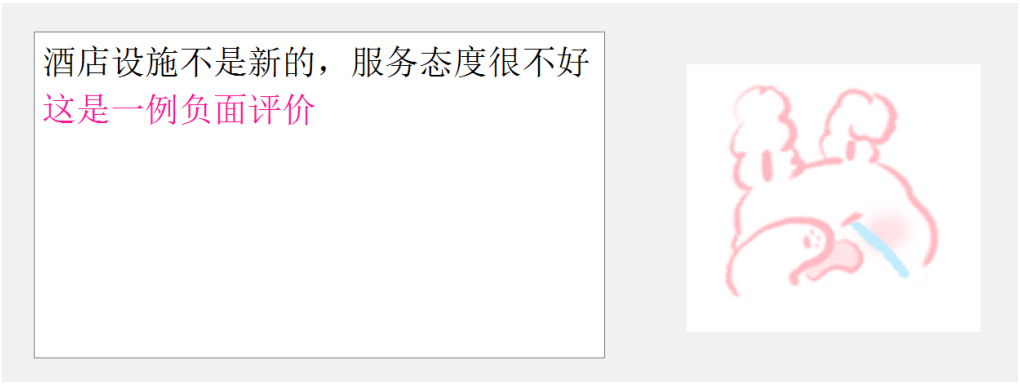

图 4.3：负面评价结果返回演示界面

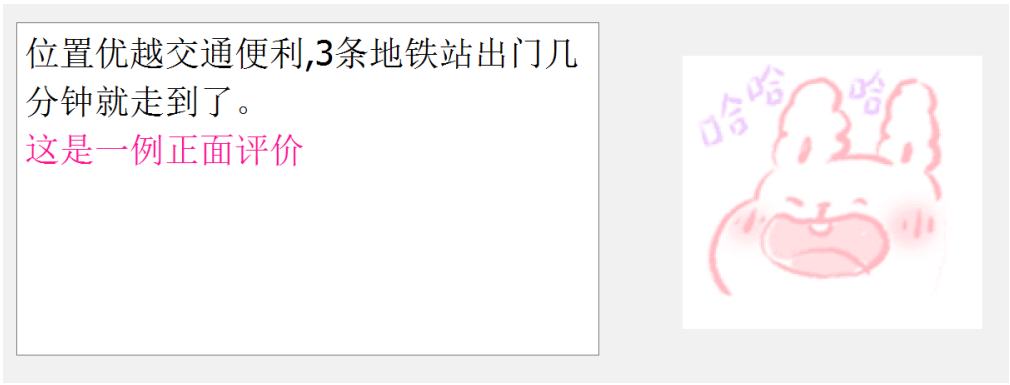

图 4.4：正面评价结果返回演示

图 4.5：中性评价结果返回演示

图 4.6：正向情感词云展示

图 4.7：负向情感词云展示

4.3 主要代码介绍

主要代码如下：

```
class Stats(QtWidgets.QMainWindow):
def __init__(self):
super(Stats,self).__init__()
self.setupUi(self)
self.retranslateUi(self)

def setupUi(self, MainWindow):
        MainWindow.setObjectName("MainWindow")
        MainWindow.resize(1000,700)
```

```
# MainWindow.setStyleSheet()
        self.centralwidget = QtWidgets.QWidget(MainWindow)
self.centralwidget.setObjectName("centralwidget")
self.pushButton_1 = QtWidgets.QPushButton(self.centralwidget)
self.pushButton_1.setGeometry(QtCore.QRect(250, 180, 180, 60))
self.pushButton_1.setStyleSheet("font: 12pt, Microsoft Yahei;\n""")
self.pushButton_1.setObjectName("pushButton_1")

self.pushButton_2 = QtWidgets.QPushButton(self.centralwidget)
self.pushButton_2.setGeometry(QtCore.QRect(550, 180, 220, 60))
self.pushButton_2.setStyleSheet("font: 12pt \"Microsoft Yahei\";\n""")
self.pushButton_2.setObjectName("pushButton_2")

self.pushButton_3 = QtWidgets.QPushButton(self.centralwidget)
self.pushButton_3.setGeometry(QtCore.QRect(550, 300, 220, 60))
self.pushButton_3.setStyleSheet("font: 12pt \"Microsoft Yahei\";\n""")
self.pushButton_3.setObjectName("pushButton_3")

self.pushButton_4 = QtWidgets.QPushButton(self.centralwidget)
self.pushButton_4.setGeometry(QtCore.QRect(250, 300, 180, 60))
self.pushButton_4.setStyleSheet("font: 12pt \"Microsoft Yahei\";\n""")
self.pushButton_4.setObjectName("pushButton_4")

self.label = QtWidgets.QLabel(self.centralwidget)
self.label.setGeometry(QtCore.QRect(400, 50, 351, 111))
self.label.setObjectName("label")

        w = QtWidgets.QWidget(self.centralwidget)
        w.setGeometry(250, 400, 700, 800)
```

```
            png = QtGui.QPixmap('./pics/main_pic2.png')
            l1 = QtWidgets.QLabel(w)
            l1.setPixmap(png)
            l1.move(100, 20)
            w.show()

            MainWindow.setCentralWidget(self.centralwidget)
    self.retranslateUi(MainWindow)
            QtCore.QMetaObject.connectSlotsByName(MainWindow)

    self.pushButton_1.clicked.connect(self.prediction)
    self.pushButton_2.clicked.connect(self.pos_wordcloud)
    self.pushButton_3.clicked.connect(self.neg_wordcloud)
    self.pushButton_4.clicked.connect(self.about)

    def prediction(self):
            ui_hello_1.show()
    # MainWindow.close()
```

4.4 本章小结

本章主要介绍本文运用到的可视化工具 PyQt5，以及可视化结果的分析和主要代码展示。并且利用相应方法实现多个窗口之间的相互切换，以及在窗口内调用中文情感分析项目的运行和结果返回。

本可视化界面分为五个部分，主界面、情感预测、正向情感词云、负向情感词云、关于。在主界面中点击相应按钮进入相应界面；在情感预测中输入文本点击“情感预测”，返回“正面评价”、“负面评价”、和“中性评价”，并有相应图片显示出来。正向情感词云和负向情感词云均分别显示一张对应正向情感语料和负向情感语料的统计词云。关于界面显示作者的基本信息。

第 5 章 总结与展望

5.1 论文总结

以互联网为载体的网络模式诞生出了许多的社交平台和购物平台，也产生了大量的留言、评论文本，这类文本往往带有用户的主观评价和情绪表现。此类数据的数量十分庞大，如果仅仅通过安排员工来进行阅读与分析这些文字，这需要消耗大量人力物力和时间，而且准确率并不能得到很好的保障，如何通过自然语言处理的方法利用计算机有效地在海量文本数据中获取和分析舆情具有很大的意义。

本文里提出了一种利用 biLSTM 模型来处理文本情感分析的方法，用于中文文本情感分析。具体来说，采用多层神经网络对每一层的信息进行过滤，最后，将模型在公共数据集上进行测试并得到准确率的反馈，验证了该模型的优点与可行性。

本文的训练和测试数据选用的是谭松波酒店评论语料，详细阐述了中文文本情感分析与可视化技术和对评论语句的情感预测与打分。获取了原始数据之后，对 4000 条原始数据进行预处理，包括分词、去停用词、清理重复评论等后，对剩下的数据进行文本挖掘与情感分析和可视化。工作流程为：文本预处理、文本索引化、训练样本切割、准备词向量矩阵、搭建 biLSTM 模型、情感分析与预测、可视化等。

通过训练的结果分析，可以发现，在情感分析与预测时，得到的 output 预测结果往往是趋向于 0 或者 1 的，没有太多处于 0.5 左右，因此可以认为训练的效果比较好，结果是有信心的。训练出来并应用在测试集上的数据，得到的准确率 Accuracy 为 88.50%，预测准确率也是比较高的，结果具有可信度。

尽管这种方法具有很高的竞争力，但是很难准确地预测具有深层隐藏情感或疑问情感的文本情感。在今后的工作中，将进一步探讨如何准确判断模棱两可的句子的情感特征。

本论文同时对情感分析模型制作了一个可视化工具，让使用者可以自订测试的文本，点击预测按钮即可进行情感预测，简化了使用难度。可视化界面分为五个部分，主界面、情感预测、正向情感词云、负向情感词云、关于。可视化工具利用相应方法实现多个窗口之间的相互切换，以及在窗口内调用中文情感分析项目的运行和结果返回。

5.2 展望

基于 biLSTM 模型的中文文本情感分析与可视化涉及了多方面的知识、技术、方法。本文涉及的项目依然有一些新的问题需要解决，在以下几个方面，还可以有更进一步的持续研究：

1、本文只突出使用了酒店评论语料的数据用于训练模型，此类数据在情感方面的确定性较强。但对于一些较为中立和多重情感混合在一起的语料，本模型的准确性还有待提高。

2、随着文本语料数据量增加，可视化界面可以添加一些统计的图表，增加可视化的强度。

3、将数据不局限于酒店评论语料，可以在多个领域都使用这个模型。

参考文献

[1] 罗正军，柯铭菘，周德群. 基于改进型 LSTM 的文本情感分析模型研究.南京航空航天大学，江苏，南京.210016.计算机技术与发展，第 30 卷，第 12 期.

[2] 李丹. 南华大学. 天猫博物馆旗舰店文创类热销产品的客户评价分析.

[3]王洪伟,刘勰,尹裴. Web 文本情感分类研究综述[J]. 情报学报,2010,29(5):931-938.

[4] 鹿鹏. 研究园地.文本情感分析的研究现状与展望,20170718.

[5] Jun Zhang. 2020. Sentiment Analysis of Movie Reviews in Chinese. Uppsala Universitet (Sweden), Advisor(s) Pettersson, Eva. Order Number: AAI28078251.

[6] Syed Muzamil Basha and Dharmendra Singh Rajput. Evaluating the Impact of Feature Selection on Overall Performance of Sentiment Analysis. In Proceedings of the 2017 International Conference on Information Technology (ICIT 2017). Association for Computing Machinery, New York, NY, USA, 2017, 96–102.

[7] G. Xu, Y. Meng, X. Qiu, Z. Yu and X. Wu. Sentiment Analysis of Comment Texts Based on BiLSTM. in IEEE Access, vol. 7, pp. 51522-51532, 2019, doi: 10.1109/ACCESS.2019.2909919.

[8] 刘文秀，李艳梅，罗建，李薇，付顺兵. 基于 BERT 与 BiLSTM 的中文短文本情感分析. 西华师范大学,南充,四川,637009,太原师范学院学报（自然科学版）第 19 卷，第 4 期.

[9] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao,Peng Zhou, Bo Xu. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. 2017.

[10] 曾诚，温超东，孙瑜敏，潘列，何鹏. 基于 ALBERT-CRNN 的弹幕文本情感分析[J/OL]. 郑州大学学报(理学版).

[11] 贵向泉，高祯，李立. 融合 TCN 与 BiLSTM+Attention 模型的疫情期间文本情感分析. 西安理工大学学报.

致谢

写下这段话的时候，也代表我与四年的大学生涯就此告别。人生若只如初见，何事秋风悲画扇，回首过往四年，我只有感激与感谢，感谢一路相伴的老师、同学，以及每一位陪我走过大学四年的人。

首先我要感谢我的父母，这四年的大学生活里，我能够取得今天让我十分满意的成绩并且成功申请出国留学，与父母物质、精神上的支持密不可分，也正是因为他们，才有今天独立、自信的我。

此次毕业论文的设计里，我的论文指导老师马长林老师给予了我极大的帮助和支持，从项目设计框架到数据分析、功能完善，和论文撰写的规范等等，每个阶段都给予了我指导。这次毕业设计的机会，是我人生里十分宝贵的一次学习经历，感谢马老师让我学会了如何独立设计并完成一个机器学习项目，完成一项极富挑战性的科研。

接着，我要感谢与我并肩走过这四年的计算机学院和 1702 班的同学们，感谢有你们的支持和帮助。在 1702 班当团支书这几年，因为有你们的陪伴和鼓励，我才能够有如今的自信、不怯场，以及有你们这么多朋友，很幸运。

另外，我还要感谢我的室友们：罗皓天、任祎铭、吕英豪。大学四年，我们相处很融洽，经常一起学习、一起约饭，考试周相互帮助，晚上交流人生理想。这四年的学习里，有你们，才有今天的我。

最后还要感谢每一位陪我走过这四年里一段或整段的朋友们，从校史馆到马基课、科研小队、开黑小组、旅游小队们，这四年里的欢笑和泪水，快乐与忧愁，能够分享与你们，是我的精神支柱和食粮。

敲完这段话，我的大学四年也画上了一个句号上，人生路添了一个逗号，我向我的母校华中师范大学致以崇高的敬意，也向我人生路上的每一位老师、朋友表达感谢与祝愿，就祝各位恭喜发财吧。人生路漫漫，新的研究生生活也就是新的起点，希望自己能够后进有所宗，绝学得所继。

当我写下“致谢”，也代表我已经走远。