学号：201772330　　密级：

长江大学

硕士研究生学位论文

基于深度神经网络的网络短文本情感分类研究

专业领域：　电子与通信工程

研究方向：　深度学习

研 究 生：　王 锭

指导教师：　杜 红 教授

论文起止日期：2017 年 9 月至 2020 年 6 月

学号：201772330　　密级：

长江大学

硕士研究生学位论文

基于深度神经网络的网络短文本情感分类研究

专业领域：　电子与通信工程

研究方向：　深度学习

研 究 生：　王 锭

指导教师：　杜 红 教授

论文起止日期：2017 年 9 月至 2020 年 6 月

Research on emotion classification of network short texts based on deep neural network

Field： Electronics and Communication Engineering

Direction of Study：Deep Learning

Graduate Student： Wang Ding

Supervisor：Prof. Du Hong、

School of Computer Science

Yangtze University

September,2019 to April,2020

Research on emotion classification of network short texts based on deep neural network

Field：Electronics and Communication Engineering

Direction of Study：Deep Learning

Graduate Student： Wang Ding

Supervisor： Du Hong

School of Electronic Information

Yangtze University

April,2019 to May,2020

摘 要

随着信息化时代的来临，互联网技术的发展愈加成熟，各类网络媒体也应运而生。从最开始的 QQ 聊天，到后来的豆瓣影视、新浪微博等，人们可以随时随地的发表自己的想法和意见。这种便捷快速的交互信息方式背后带来的则是日益增长的数据量，这些数据中包含着人们对于事物或者事件的看法和意见，利用自然语言处理技术对这些数据进行分析并发现其中所包含的情感倾向，对于我们进行舆情监测、商品营销、金融分析等实际应用有着重要影响。

文本情感分类又称之为为文本倾向性分析，是近些年来自然语言处理领域的研究热点之一，吸引了很多研究学者的关注。其中基于深度神经网络的情感分类算法鉴于其优异的特征提取能力，已经慢慢成为了解决文本情感分类问题的主流方法之一。本文基于深度神经网络对网络短文本进行文本情感分类研究，首先借助网络爬虫获取豆瓣电影的影评，随后对爬取的数据按照相应的准则进行预处理和情感标注，从而获取了文本情感分类数据集。

为了充分利用文本情感资源，使用 word2vec 工具将词性特征以及词汇特征分别向量化，随后进行向量拼接并以此作为卷积神经网络的输入。在此基础上，考虑到传统的最大池化方式容易丢失特征信息，使用 k-max 池化代替最大池化，提出 KMCNN 模型。实验证明，基于 KMCNN 的文本情感分类相较于其它分类模型有更好的分类性能。

考虑到对文本序列特征的学习需要结合上下文联系，在 KMCNN 的基础上提出 KMCNN-GRU 模型，在特征提取层将 KMCNN 提取的局部重要特征和 GRU 提取的序列特征进行融合，以特征融合的方式加强模型对于文本情感的捕获能力，并进行多组对比实验。实验结果表明，基于 KMCNN-GRU 的文本情感分类模型可以学习到更多的语义特征，相较于其它分类模型有更好的稳定性以及分类精度。

关键词： 深度神经网络，情感分类，word2vec，卷积神经网络，GRU

Abstract

With the advent of the information age, the development of Internet technology is becoming more and more mature, all kinds of network media emerge at the historic moment, from the beginning of tencent QQ chat, to the later douban film, sina weibo.People can express their ideas and opinions anytime and anywhere, and this convenient and rapid way of exchanging information brings about an increasing amount of data.These data contain people's views and opinions on things or events. The use of natural language processing technology to analyze these data and find the emotional tendency contained in them has an important impact on our practical applications such as public opinion monitoring, commodity marketing and financial analysis.

Text emotion classification, also known as text orientation analysis, is one of the research hotspots in the field of natural language processing in recent years, attracting the attention of many researchers, and many practical and effective classification algorithms have been proposed.Based on the deep neural network model, the emotion classification algorithm has gradually become one of the mainstream methods to solve the text emotion classification problem due to its excellent feature extraction ability and the advantages of hardware computational support.Based on the deep neural network, this paper conducts text emotion classification research on network short texts. Firstly, it obtains the film review of douban movie with the help of network crawler, and then preprocesses and emotion labeling the crawled data according to the corresponding criteria and standards, so as to obtain the text emotion classification data set.

In order to make full use of the text emotional information, word2vec tool is used to vectomize the part of speech features and lexical features respectively, and then the vector splicing is carried out and used as the input of the convolutional neural network.On this basis, considering that the traditional maximum pooling method is likely to lose feature information, a KMCNN model is proposed to replace the maximum pooling with k-max pooling.Experiments show that KMCNN based text emotion classification has better classification performance than other models.

Considering that the learning of text sequence features needs to be combined with context, a kmcnn-gru model is proposed on the basis of KMCNN, and the locally important features extracted by KMCNN and the sequence features extracted by GRU are fused in the feature extraction layer to enhance the ability of the model to capture text emotions in the way of feature fusion.The experimental results show that the text emotion classification model based on kmcnn-gru can learn more semantic features and has better classification accuracy than other classification models.

Key words: text sentiment classification, word2vec, convolutional neural networks, GRU

目　录

第 1 章 绪论

1.1 研究背景及意义

随着无限通信技术的不断发展，网络的传播不在局限于有限连接，人们对于网络的认知也随着移动通信端的不断发展而有了很大转变。人们每天都在和数据信息进行交互，不知不觉间已经和数据捆绑在了一起，将我们原本在现实生活中进行的活动慢慢转移向了网路，如购物、聊天以及观影等等。这种生活状态的转变同样带动了互联网经济的快速发展，同时也改变着人们的消费观点。在这样一种趋势下，手机等移动端已经成为了人们进行互联网活动的主要途径。而随着移动端软件便利度的提升，人们越来越热衷于将自己对于现实生活中的一些意见或者想法通过简短的文字（即网络短文本）分享到电商平台或社交媒体上，例如微博热点评论、美团外卖评论以及豆瓣影视评论等等，并且也可以通过他人发布的信息中获取自己需要的内容。这样一种新型的信息交互方式产生了大量带有用户主观情感倾向的文本数据，而单纯以人工的方式去处理并分析这些数据是不现实的，因此文本情感分类应运而生。

电影起源于十九世纪的欧洲，近些年来得到了迅猛发展。人们日常工作之余，观影成为了主要的休闲娱乐方式之一。而随着信息化时代的到来，人们的观影模式也发生了改变，使用移动端设备观看电影成为了趋势。而这种转变的到来，也同样促使着视频网站的发展。除了提供常用的影片播放功能，还提供给用户在线评论的功能。用户可以在观看影片之余对影片进行评价，这些评价中包含用户自身对于影片的一些情感态度，包括对影片的人物、剧情或者音效背景等的看法和意见。这些看法和意见可以帮助其他用户在观影前进行影片筛选，更快的找到适合自己的影片。

文本情感分类也可以称之为意见挖掘。在信息化时代，面对日益增长的数据量，通过使用计算机代替人工的方式去识别文本中所包含的情感倾向，利用归纳总结的方式对文本数据进行分析的过程称之为文本情感分类。情感倾向主要是指人们的情绪的好坏或者对于一件事物的认可程度，比如高兴或者悲伤，满意或者不满意等等。早期文本情感分类主要采用机器学习的方法，例如对垃圾邮件的筛选等，在节省了人力资源消耗的同时也提高了工作效率。而随着科技的不断发展以及深度学习的出现，特别是随着人工智能领域相继取得重要突破，包括在人脸识别，语音识别等，使得文本情感分类也得到越来越多专家学者的关注。目前，文本情感分类领域已经出现了很多高效实用的算法，在实际应用场景中发挥着巨大的作用。例如，消费者在电商平台购买商品时首先要做的是查看商品信息以及商品的评价数据，这些评价数据在一定程度上能够帮助消费者更加明确自己的购买意图；而商家也可以通过分析建模对这些评价数据进行归纳分类，如哪些商品更受欢迎以及哪些

商品具有怎么的缺点等等，使得商家可以更加清晰的了解市场的行情以及走向，制定更加合理的市场策略。不仅如此，文本情感分类还可以用于推进很多实际应用的研究。比如在舆情监测方面，可以通过情感分类技术阻止恶意谣言的传播，保证网络环境的整洁以及安全性；在金融分析方面，可以给金融分析人员提供关于股票走势的参考。因此，从日益增长的网络短文本数据中分析出文本的情感倾向具有非常高的研究意义。

多种多样的文本信息表现形式以及复杂多变的结构，使得文本情感分类的研究非常具有挑战性。本文基于以上研究背景，对文本情感分类领域进行研究学习。

1.2 国内外研究现状

文本情感分类是很多研究应用的基础，吸引着众多的专家学者的关注。可以将分类任务大致分为三个研究层次：篇章级、句子级、词语级。基于篇章级的方法主要是对于一个文章整体的一个情感倾向的输出，即总体的正面意见或者负面意见；基于句子级别的方法主要依赖于情感数据库，通过标注句子中的情感类别强度来实现对句子的情感分类；基于词语级的情感分类的分析对象为分词后的词汇项，就词汇的情感倾向分析文本的情感极性。就分类结果而言，一般只包含正面情感倾向以及负面情感倾向，也存在有多分类的情况，类似于喜欢，一般，不喜欢等更加细粒度的情感分类。而网络短文本主要以句子的形式存在，因此本文的研究主要是基于句子级别的文本情感二分类研究。

通过研究学习发现文本情感分类研究主要经历三个发展时期：（1）构建情感字典对文本进行分析；（2）采用机器学习算法对文本数据进行学习训练；（3）使用深度学习的方法自动发现文本内部特征。其中前两种方法相对于深度学习的方法更加完备，基于其长时间研究成果的积淀以及研究资源的积累使得其在一定的时期取得了不错的研究结果。而随着互联网的快速发展，文本数据集数量的大幅度递增，使得传统的文本分类方法很难在保持分类精度的同时还能快速处理大量数据。而深度学习的出现使得文本情感分类领域焕然一新，其不依赖于人工经验的特征提取以及硬件算力的支持，使得其成为了近些年来的研究热点。相比于通过提取特征进行分类研究的机器学习法以及深度学习法，字典法不需要考虑文本的特征结构以及语义表征形式，其重点是在于情感字典的造构，并约定相应的计算规则来判定文本的情感极性。该方法虽然简单易于理解，但是对于日益增长的数据量以及网络新词的不断涌现，很难再去构造完备的情感字典，因此字典法的适用范围相对较小。

在使用特征提取的方法中，机器学习法依赖于人工经验对于文本数据的理解，基于概率论或统计学的方式对文本进行特征提取，然后训练相应的分类器进行文本分类。该方法可以将无序的数据转换成相对有用的特征信息，并且可以通过训练一定数量的数据样本去学习数据内部的浅层统计结构，因此取得了一定的研究成

果。但是由于该方法过度依赖人工的主观经验，很难提取到数据内部深层次的特征，特别是对于文本序列等非线性数据，因此该方法只能适用于小数量级的文本分析。基于深度学习的方法是以机器学习为基础，模块化的神经网络结构对底层特征进行依次提取并组合成更抽象的分布式表征形式，从而以简单的神经元实现复杂现实问题的表示，相对于机器学习法能取得更好的分类效果。

1.2.1 基于情感字典的文本情感分类

基于情感字典的方法是早期的文本分类的雏形，当时是通过简单的包含关系来分析文本的情感极性。该方法最为关键的是对于情感词典的构建，情感字典需要包含语言学中通用的情感词，以此来获得更加全面的极性分析。目前比较有代表性的词典包括英文领域的 WordNet[1]，中文领域的 HowNet[2]以及《情感词汇本体》[3]等。

除了借助外部工具情感词典，同时还需要约定相应的计算规则来判别文本的情感极性。Kamps 等人[4]提出一种通过计算形容词之间的距离的方式来判断形容词的情感极性。Turney 和 Littman 以 WordNet 词典为基础，计算词典中形容词与 good 和 bad 的距离来分辨相似程度，以此来分析文本的情感倾向[5]。在基础的情感字典构建完成之后，为了使字典更加完备，通常是在其之上进行词典的扩充。词典的扩充方式一般为寻找同义词或者将不同的词典进行合并消重。Yang Min 等人[6]在基础情感词典的基础上，在特定的训练语料中提取主题词来扩增特定领域的词典。Qiu 等人[7]在原有词典的基础上，通过组合叠加的方式对情感词与观点词进行组合，构建出新的情感词典。除此之外，还可以结合实际研究背景进行词典的扩充。Li 等人[8]基于迁移学习的方式，考虑到某些特定领域数据集不足的情况，结合其它相近领域的标注数据充实词典。陈晓东[9]等基于网络爬虫提出基于中文微博的网络情感词典。杨飞等人[10]结合中文酒店评论，对高频词赋予权重并扩展基础的情感字典，计算情感加权值进行情感分析，精准度相对于基础情感字典提高了 10%左右。

基于情感字典的情感分类是早期使用较多的情感分类方法，模型相对简单，不需要考虑文本的特征结构以及语义表征形式，其重点是在于情感字典的构造，并约定相应的计算规则来判定文本的情感极性。相比于通过提取特征进行分类研究的机器学习法以及深度学习法，其方法简单并且计算量小，在情感分类任务中取得了一定的研究成果。但是随着互联网文本数量呈爆炸性增长，越来越多的“网络新词”层出不穷，构建完备的情感字典太过于耗时耗力。因此，为了减少资源消耗以及提高精准度，基于机器学习的方法受到了国内外专家学者的更多关注。

1.2.2 基于机器学习的文本情感分类

不同于字典法，机器学习法不需要借助外部工具，更多的注重于文本特征的提

取，特征提取的好坏在很大程度上决定分类模型的优劣。根据提取的特征属性以生成模型（朴素贝叶斯，隐形马尔科夫模型）或判别模型（SVM、logistic 回归）对文本的类别进行判决。

生成模型是通过从数据中学习到文本类别的联合概率分布，然后根据条件概率进行分类判别。而判别模型主要是直接学习数据内部的决策函数，以文本类别之间的最优分割超平面来对不同的文本类别的差异。例如，对于正类和负类两种文本类别的判定，生成模型会首先根据正类的特征在数据中学习出一个正类模型$P_{(w_1|x)}$，根据负类的特征在数据中学习出一个负类模型$P_{(w_2|x)}$，然后提取出待分类文本的特征x_0代入$P_{(w_1|x)}$以及$P_{(w_2|x)}$中进行比较，概率更大的为输出类别。而判别模型则是从训练的历史数据中学习决策函数，对于未分类的文本则根据学习到的决策函数来输出类别。可以看出，生成模型可以还原出数据的联合概率分布，但学习和计算过程比较复杂；判别模型则是直接进行预测，准确度高，但是不能反映数据自身的特性。

Pang 等人[11]首次使用机器学习进行文本分类任务，并进行多组对比实验验证了模型的有效性。Ni 等人使用不同的特征提取方式 [12]，并分别结合朴素贝叶斯和 SVM 算法训练文本情感分类模型，验证了使用 SVM 实现情感分类可以得到更好的效果 。冯成刚等人[13]使用不同的特征权重结合 SVM 进行微博文本分类，发现使用信息增益作为特征选择时准确率最高。徐健锋等人[14]结合机器学习的方式来优化语义理解，验证了在不同领域下具有稳定性和有效性。王大伟等人[15]提出一种 PCA-SVM 算法，对文本词向量使用 PCA 进行降维，在减少计算量的同时最大程度的保存原始数据特征。Agarwal 等人[16]基于 Twitter 数据，在朴素贝叶斯以及 EM 算法中采用多种词性特征的组合结构进行对比实验，实验表明朴素贝叶斯以及 EM 算法能在文本情感分类任务取得较好成绩。

基于机器学习的方法相对成熟，可以将无序的数据转换成相对有用的特征信息，并且不需要依赖于词典资源。但是由于该方法过度依赖人工对于数据的理解，很难去学习到数据内部深层次的特征信息，特别是对于文本序列等非线性数据需要结合上下文语义进行研究，因此该方法只能适用于小数量级的文本分析。而随着深度学习的兴起以及硬件加速计算的突破，越来越多的 AI 应用也随之落地。其中人脸识别等图像领域的应用更是走进了人们的生活中，而文本情感分类作为 NLP 研究的基础，也吸引了越来越多的国内外研究学者的关注，并逐步取得了一系列的进展。

1.2.3 基于深度学习的文本情感分类

基于深度学习的方法是通过构建深层的神经网络结构，从大量的数据样本中自动学习其内部的统计结构，以此来进行特征提取，因而能够发现文本数据内部更

深层次的语义特征，从而获得更高的分类精度。

模块化的神经网络结构以层层叠进的方式对数据内部的统计结构进行学习，层级越高神经元所包含的底层特征信息越少，更多的是关于目标信息（文本类别），从而以简单的神经元实现复杂现实问题的表示。2006 年，由 Hinton 等通过在神经网络中引入层次化结构[17]，使得神经网络不仅具有快速特征的提取能力，而且还解决了深度神经网络难以训练的问题。Kim 等人[18]用卷积神经网络（Convolutional Neural Network，CNN）进行文本分类任务，取得了相对于传统机器学习算法更好的效果，打开了使用 CNN 解决自然语言处理问题的大门。梁军等人通过构建递归神经网络模型用于微博情感分类任务中情感特征挖掘[19]。谢博等人提出一种半监督卷积神经网络情感分析模型，使用不同的特征信息对不同的信息通道进行文本情感信息的学习，并验证了其有效性[20]。在算法模型方面，Tomas Mikolov 等人[21]提出了 FastText 模型,能够在一定程度上提高模型对数据的处理速度。而循环神经网络（Recurrent Neural Network，RNN）基于其可以对历史信息学习的特点，在文本分类领域也取得了非常好的成绩。 Zhang 和 Liu 等提出一种树型 LSTM 神经网络模型结构，并取得不错的分类成绩[22]。张翠等通过融合特征的方式结合 CNN 以及 RNN，提出一种特征融合的深度学习算法，并通过实验的出该算法与传统的 CNN 以及 LSTM 相比准确率分别提高了 2.56 和 1.87 个百分点[23]。

基于深度学习的方法应用广泛，并且随着硬件加速的突破以及越来越多框架的开源，使得深度学习成为 21 世纪最受关注的研究领域之一。基于其优异的数据表征能力，在图像、语言、音频和文本领域取得了非常多的研究成果。目前，虽然基于深度学习的方法取得了一定的研究成果，但依旧存在着很多问题。本文以深度神经网络为基础，结合词性特征丰富文本表达，使用基于 k-max 池化的卷积神经网络进行文本情感分类研究。同时，为了有效增强文本特征的提取能力，在特征提取层面通过向量拼接的方式将 CNN 以及 GRU 两种特征模型提取的文本向量特征进行融合，从而提高模型的性能。

1.3 研究内容

传统的文本情感分类算法包括词典法以及机器学习算法，具备成熟的理论支持以及丰富的研究资源。词典法主要借助于情感词典等外部工具，并事先约定好相关计算规则，从而得出文本的情感极性，比较知名的情感字典包括 WordNet、HowNet 等。机器学习法利用人工经验提取特征，结合相应算法进行模型训练学习，从而达到分类效果，常用的算法为朴素贝叶斯以及 SVM 等算法。深度学习的方法是近几年的热点，并在图像、语音等领域都取得了非常不错的效果，同时也吸引着越来越多的研究学者致力于其中。

本文主要研究内容如下：

（1）数据集的构建。首先利用网络爬虫爬取豆瓣影视评论作为训练语料，并对文本数据进行清洗去噪等预处理工作。随后对数据集进行标注，选取三次标注结果一致的文本数据作为最终的数据集。

（2）基于 CNN 的文本情感分类。为了更有效的利用情感资源，使用词性标注的方式，将词性表达与词汇向量相结合作为文本表示。在此基础上，考虑到传统的最大池化方式容易丢失特征信息，使用 k-max 池化方式代替最大池化，提出模型 KMCNN。实验结果表明，KMCNN 在文本情感分类任务上具有良好的分类性能，能在去除一定噪声影响的同时保留局部重要特征。

（3）基于特征融合的文本情感分类。考虑到单纯的基于词语级粒度进行特征提取可能无法涵盖更多的文本表征信息，在对不同的深度学习算法进行研究分析后，提出融合特征的 KMCNN-GRU 模型，在特征提取层面通过向量拼接的方式将 CNN 以及 GRU 两种特征模型提取的文本特征向量进行融合。使用 keras 进行模型搭建，并构建多组实验来验证模型的有效性。

1.4 本文的组织架构

本文一共包括五个章节：

第一章 绪论。介绍本文的研究背景和意义，并对国内外研究现状进行说明。

第二章 相关理论和技术。简要阐述关于文本数据预处理相关技术以及深度学习相关方法和理论。

第三章 基于 CNN 的文本情感分类。为了有效利用文本情感资源，在文本表示层添加词性特征，并结合 k-max 池化提出 KMCNN 模型。

第四章 基于特征融合的文本情感分类。为了充分提取文本特征信息，在 KMCNN 的基础上，采用特征融合的方式将 CNN 提取的局部重要特征以及 GRU 提取的长序列特征进行融合，提出模型 KMCNN-GRU。

第五章 总结和展望。对于全文所涉及的文本情感分类进行梳理总结，并对一些不足的地方进行分析以及展望。

第 2 章 相关理论和技术

本章主要介绍研究中所涉及的相关知识。早期的文本分类方法主要包括字典法以及机器学习法，依赖于情感字典的方法简单易懂，不需要考虑文本的特征结构以及语义表征形式，其重点是在于情感字典的造构，并约定相应的计算规则来判定文本的情感极性。基于机器学习的方法相对成熟，其依靠于人工主观经验对于文本数据的理解，基于概率论或统计学的方式对文本进行特征提取，然后训练相应的分类器进行文本分类。基于深度学习的方法是以机器学习为基础，以模块化的神经网络结构为基础，对底层特征进行层层提取组合成更抽象的分布式表征形式，相对于机器学习法能取得更好的分类效果。

2.1 文本处理以及文本表示

随着信息化时代的不断进步，文本的表现形式也多种多样，因此在工程任务之前，需要对文本数据进行预处理。图 2-1 为文本预处理主要流程：

图 2-1 文本处理流程图

Figure 2-1 Flow chart of text processing

2.1.1 数据获取

步入 21 世纪初，互联网已经融入了每家每户的生活之中。随着智能终端的不断推广，各种观影网站也如雨后春笋般出现，人们可以随时随地的观影，不再拘泥于场地的限制。而这种休闲方式的转变也促进了线上视频网站的发展，线上观影逐渐成为了人们娱乐休闲的主要方式之一。不仅如此，在万物互联的背景下，影评的存在使得人们不仅可以从别人发布的影评中知悉电影内容，还可以在观影后将自己对于电影的看法或者意见悉数表达于评论中。这样一种新型的信息交互方式带来了数目递增的评论数据，并且这些数据中包含了大量用户个人的情感倾向信息。因此本文选择国内专业的电影网站--豆瓣电影作为数据获取源，并对获取的数据进行相应的预处理工作，选取其中有效的部分作为模型的训练数据集。

豆瓣电影是国内知名的影视网站，网站内包含了很多国内外的影视资源，基于其简洁高效的服务态度以及高清可视的视频资源吸引了大量用户的关注，同时也带来数目递增的影评数据。影视评论数据主要分为两种，其中长文本通常字数没有限制，字数一般在 20 字以上，而短文本则是几个字到十几个字不等，部分评论示例如表 2-1 所示。

网络爬虫是一种能够按照事先编写好的规则去自动爬取网络上的数据[24]。本文基于 python 语言编程实现爬虫功能，利用 python 的第三方库 urllib 抓取目标网页内容，结合相关 xml 知识进行网页解析并获取我们实验所需数据，爬虫流程图如图 2-2 所示。

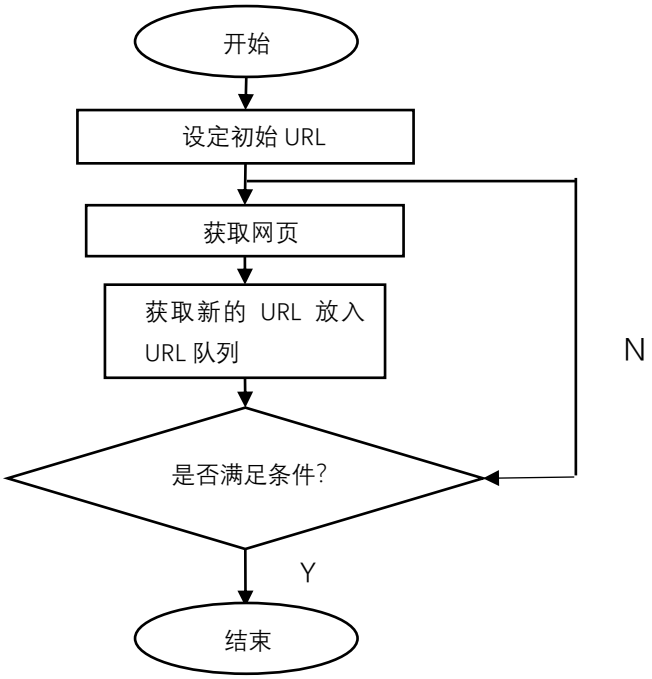


图 2-2 网络爬虫流程图

Figure 2-2 Flow chart of web spider

表 2-1 部分影评内容

Table 2-1 Part of the reviews

情感极性	电影评分	影评内容
正面情感评价	5	喜欢张译，不愧是影帝。
	4	有些幸福无关爱情喜欢这个故事
	5	我强烈推荐这部电影及原声。
负情感评价	3	给人印象深刻比阿甘正传好一些。
	2	不看影评几乎不知道说什么。
	1	要多不困才能看完。

2.1.2 数据清洗

通过网页爬虫采集的文本数据中会包含很多无用的部分，这些内容不包含情

感倾向，但是会占用大量内存，并且在训练时会影响分类精度。因此我们需要对分词后的数据进行清洗。具体如下：

（1）去除无效字符。采集的数据主要是来源于网络爬虫，一般会包含一些 html 代码或者 css 标签等，删除这些无用内容可以节约内存资源。

（2）去除非中文语料。本文研究主要是针对中文网络短文本，因此需要对采集的数据中包含的一些非中文单词进行清理，如英文单词 “like”、“hate”等。 通过删除这些语料有助于数据的规范化。

（3）去除繁体字语料。中文文字书写方式主要包括繁体和简体两种[25]。考虑到影评中的大部分数据都是中文简体，所以删除数据集中包含的繁体字的文本数据使得数据更加规范。

表 2-2 去停用词示例

Table 2-2 An example of Stop-words

数据版本	分词项
原始数据	真为吴京的演技尴尬，总是摆出一副大义凌然的样子
分词	真为/吴京/的/演技/尴尬，总是/摆/出/一副/大义/凌然/的/样子
去停用词	真为/吴京/演技/尴尬 /摆/一副/大义/凌然/样子

2.1.3 文本分词

汉语的文字表达不同于英语等其他语种，其继承于古汉语的传统，词语之间没有空格。本文是基于中文语境下的网络短文本情感分类研究，因此要对获取的数据进行分词操作。

词语作为最小的情感单元[26]，蕴涵着丰富的情感意义，相同的一句话经过不同的分词方式输出的意义也会产生很大差异。因此，分词方式的好坏在一定程度上影响着文本分类的准确性。

常见分词算法如下：

(1) 基于统计的分词：统计语料库中所有词语信息，基于概率的方式去计算待切分字符串与语料库中所有词的邻接次数，邻接概率越大，则成词的概率越大。

(2) 基于字符串匹配的分词[27]：需要借助于词典等外部工具，通过扫描匹配的方式判断字符串的子串是否在词典中，如果在则进行切分。

(3) 基于知识理解的分词：以传统算法为基础将文本进行分词，随后结合语法和语义的分析，能够在一定程度上避免词语的二义性问题。该方法虽然相较于传统

算法具有一定优势，但需要大量语言知识信息作为支撑。

在文本的表达中，一些句子的情感倾向也会通过词语反映出来。而词性作为词语的属性之一，也包含着比较明显的情感倾向，和词汇的语义表达相辅相成。常见的词性有名词、动词以及形容词等，如“讨厌”、“难受”、“愤怒”等动词表达就具有一定的负面情感倾向，如“兴奋”、“愉悦”、“亲切”等形容词则带有一定的正面情感倾向。由此可以看出，文本的词性属性也可以作为文本情感极性的表达之一，因此本文将词性标注任务加入了文本预处理任务中。

基于词汇级别的词性标注是通过结合词汇所在句子的整体语义表达，以及词汇在句子中的位置进行标注。词性标注的方法有两种 ：规则法以及统计法。基于规则的方法需要借助于外部工具，通过规则库与词语进行比较；基于统计的方法则是采用概率统计的方式计算词语以及对应的词性的概率[28]。目前常用的中文分词工具有中科院的分词器 NLPIR[29]、哈工大的 Hanlp[30]等。考虑到分词器的功能性以及易用性，本文采取 Hanlp 作为分词以及词性标注的工具。

随着互联网的日新月异，越来越多的人使用网络进行交互，由此产生了很多网络新词。如“外貌协会”、“三无青年”以及“高富帅”等词语，这些词语言简意赅，但却包含着丰富的情感信息。不仅如此，这些新词以其新颖的表达方式吸引着人们的追捧，经常被用于日常的口语交流中。但是其不同于常见的词汇结构，以传统的分词方式进行分词会破坏这些词语原本的语义，如“外貌/协会”，原意是指女生只注重男生外表，而分词后则表示成了“外貌”以及“协会”两个词语，完全失去了其本身所想表达的意义。因此，需要将这些网络新词进行处理。本文以人工搜索的方式将热门的网络新词进行整理归纳，并加入到 Hanlp 工具中的用户自定义词典中。

2.1.4 文本表示模型

文本数据表现形式多种多样，不能直接与计算机进行交互，因此要将其转换成数字向量的形式。

（1）布尔模型：布尔模型[31]通过将二值化与集合论的结合的方式进行模型检索，使用“0”和“1”两种状态代表特征项是否在分词后的词项集合中出现，经过模型转换后的向量是一串只包含 0 和 1 的数字。模型虽然易于实现，但是无法精准的判定特征词的重要性。

（2）向量空间模型：向量空间模型由 Saltaon 等人[32]提出，其不同于二值化的布尔模型，而是通过对每一个特征项赋予具体的权重信息，权重大小的增减则是依据于某个特征项与文本之间的相关性。相关性越强相应的特征权重就会增大，反之则会减小。虽然向量空间模型相较于二值化的布尔模型表达能力有所提升，但是其构建的特征是无序的，所以难以有效的表示文本的上下文关系。

（3）概率主体模型：概率主题模型是一种统计模型[33]，其主要是从文本中提取抽象的主题信息进行文本表示，主要适用于长文本数据。

2.1.5 基于 word2vec 的文本表示

对短文本实现向量化的方式主要有两种，一种是基于泛化的方式，如布尔模型，将所有文本特征都看作相同权重，结构虽然简单，但是却缺乏文本表示能力；另一种是基于向量空间的方法，虽然区分了不同特征的权重大小，但是其无序性并不能很清晰的表达出文本的上下文联系。并且对于语义相近的词语，很难将其精确映射到对应维度中，如“土豆”、“马铃薯”等。此外，还有基于语言概率模型的文本表示，如 n-gram。

n-gram 属于生成模型，n 的取值一般为 2 或 3。预测词语出现的概率需要学习其前 n-1 个词汇项的联合概率分布。：

$$P(w_1, w_2, \dots w_K) = \prod_{k=1}^{K} P(w_k | w_1, , w_2 \dots w_{k-1}) \qquad (2-1)$$

虽然实现简单，但是对于长文本序列会产生巨大的计算量，因此只适用于短序列文本。并且由于模型参数的概率分布是基于离散型变量，同样没有办法解决近似语义问题。随着神经网络的出现，研究人员尝试使用神经网络语言模型，并且取得了不错的研究成果。

最早的神经网络语言模型由 Bengio[34]在 1986 年提出，使用四层神经网络对词汇项的 n-1 个前项进行映射，通过计算每个词汇项的概率分布将其向量化。而 Mikolov 等人在其基础上进行完善，提出 word2vec 工具。

word2vec 的出现使得词向量的训练更加高效，以分布式表示结合神经网络结构，将文本序列转换成数字向量的形式。主要包括两种模型：连续词袋模型（CBOW）以及 Skip-gram 模型。CBOW 是在词袋模型的基础上进行改进，通过当前文本的上下文来推断出当前文本。对于词袋模型的通俗理解是，句子或者文章的文字都可以放在一个“袋子”中，不需要考虑词的顺序和句子的语法，还可以将单个文字出现的频率作为文本分析的特征。

Skip-gram 模型是在 n-gram 的基础上的改进，采用跳跃取词的方式。不同于 CBOW 模型，该方法通过当前词项推断其前后 n-1 个词项出现的概率。使用跳跃取词的方式在一定程度上使得模型增强对文本的长距离信息的学习能力，而随着跳跃距离的增加，模型训练的复杂度也会同时增加。

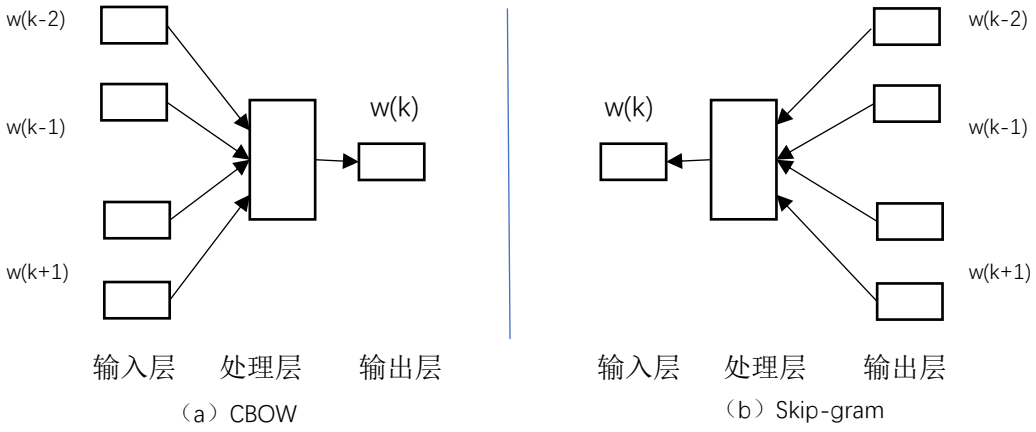


图 2-3 Skip-Gram 与 CBOW 结构示意图

Figure 2-3 Skip-Gram and CBOW Structure Map

基于 word2vec 的方式是 word embedding 的表现形式之一，以分布式词向量表示结合神经网络，在 k 维向量空间中对词向量进行表征，并利用训练完成的词向量来度量词项之间的相似程度。相较于传统方法带来的高维稀疏以及相似词项难以区分的问题，词嵌入方法不仅能够从大量文本数据中学习到词向量表征形式，而且还可以在不损失特征信息的条件下将词语向量进行降维，减少计算复杂度。

2.1.6 特征提取

在中文的语义中，对于特征的理解是事物异于其它事物的点，也可以解释为事物与其它事物不相同的地方，寻找这些异常点可以被用来更好的区分事物，因此特征提取方式的优劣对于分类任务来说显得尤为重要。词汇单元作为最小的情感单元，常通过特征选择方法计算分词后的词项与文本类别的关联程度，常用的有以下几种方法：

（1）基于词频（DF）

基于词频（Term frequency,简称 TF）的方式是通过人工的主观经验所得，认为特征的重要性与词汇在文本中出现的次数有关，对于出现次数较少的词汇特征则认为是不重要的特征。但对于文本数据中所包含的词汇出现次数相近时，很难提取出有效特征，因此该方法具有一定的局限性。

$$tf(x,y)=\frac{n_{x,y}}{\max\{n_{x',y}:x'\in y\}} \qquad (2-2)$$

其中$n_{x,y}$表示词汇项x在文本y中出现的次数，分母$\max\{n_{x',y}:x'\in y\}$表示在当前文本y中出现频次第二的词汇项x'。

（2）TF-IDF(词频-逆文件频率)

TF-IDF 是一种基于统计学的特征提取方法[35]，对于特征的提取主要依赖于两点：词频以及逆文件频率（IDF）。IDF（逆文件频率）是在 TF 的基础上，结合文本类别来衡量词汇项对于文本类别的区分力度。

具体公式如下：

$$idf\left(x, Y = \log\frac{N}{|\{y \in Y; x \in y\}| + 1}\right) \quad (2-3)$$

其中 N 代表总的文本数目，$\{y \in Y: x \in y\}$代表文本中包含词汇项x的文本数目，同时为了防止分母为 0，通常会在分母处加上数字 1。

TF-IDF 具体公式如下：

$$tf - idf_{(x,y,Y)} = tf_{(x,y)} \times idf_{(x,Y)} \quad (2-4)$$

TF-IDF 更倾向于选择与类别信息更相关的词汇项，相较于 TF 方法适用性更广。

（3）互信息

互信息(mutual information，简称 MI)源自于信息论中，主要是衡量两个变量间相互关系[36]。

可由以下公式计算：

$$mi(x, Y) = \sum_{i=1}^{z} p(Y_i) log\frac{p(x|Y_i)}{p(x)} \quad (2-5)$$

其中$p(x|y)$ 表示类别Y_i的文本总数占总文本数的比例，z 表示文本的总类别，$p(x|Y_i)$表示类别Y_i的所有文本中包含词汇项x比例，$p(x)$表示所有文本中包含词汇项x的文本数量占据总文本数的比例。互信息是对词汇项与文本类别的相关性进行度量，相关性越大，即词汇项对于文本类别的区分就更重要，反之则表示当前词汇项对于文本类别的区分作用不明显。互信息的取值有三种状态：当值为 0 时，则表示词汇项与文本类别毫不相关；当值为正数时，则表示词汇项与文本类别正相关；当值为负数时，表示负相关。基于互信息的方法不通过词项与文本类别的关系进行假设，所以更适用于文本分类。

2.2 深度学习

在传统的文本情感分析任务中，主要是通过不同的数据表示方式来对文本进行分析，而这些数据表示的方式基本依赖于人工经验所得，所以带有一定程度的人类主观认知，并不能很好的挖掘出数据的深层次特征。所以，怎样挑选出文本数据中的合适特征至关重要。基于深度学习的方法不同于传统的分类方法，其能从文本中自动地发掘复杂的特征，学习数据内部的统计结构，而不采用人为的主观因素来选择特征。并且随着硬件突破带来的算力支撑，使得采用深度学习进行大规模数据处理成为了可能。深度神经网络的概念源于人工神经网络[38]，早期研究人员通过模拟动物神经网络，以线性函数结合非线性表达的神经元进行分布式并行特征提取。这种特征提取的方式依赖于训练数据的规模大小，数据规模越大，模型就能学习到更为深层的特征，同时也能避免使用人工特征提取的过程。

2.2.1 卷积神经网络

不同于人工神经网络，CNN 对于特征的提取方式更加注重于局部重要特征的提取，每一层的权重和偏置都相同（每个通道之间权重不共享，即引入了先验知识），使得不同位置的同一特征能够被识别。基于其空间共享以及稀疏卷积带来的优势，使得 CNN 在图像领域中取得了非常优异的成果[37]。虽然文本数据不同于图像结构，但以词向量表示的文本数据也可以使用 CNN 进行训练学习。最早基于 CNN 的文本情感分类模型是 Kim 等人于 2014 年首次提出的，并且取得了不错的研究成果，打开了使用 CNN 进行文本分析的大门。

在 CNN 结构中，输入层是以二维矩阵的形式送入隐藏层中，其模型框架如图 2-4 所示：

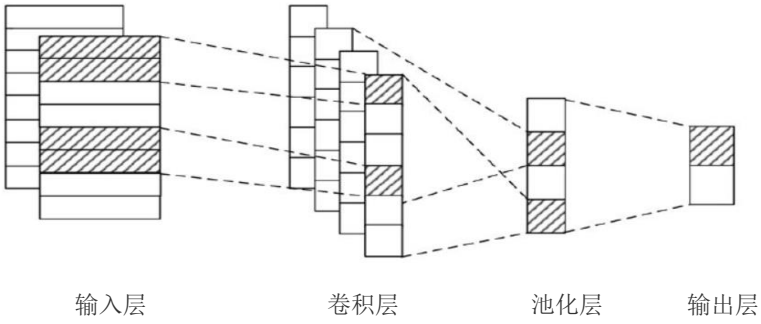


图 2-4 卷积神经网络结构图

Figure 2-4 Structure diagram of CNN network model

由图 2-4 可知，CNN 的每一层之间都是输出作为输入，直接交互，并以堆叠的方式对底层特征进行学习。不同于人工神经网络，CNN 引入稀疏连接的结构，代替传统的密集连接方式。稀疏连接的大小取决于滑动窗口的大小，其一般取值为奇数，主要是为了在卷积的过程中锚点在中间，使得在模板匹配的过程中避免位置发生迁移。

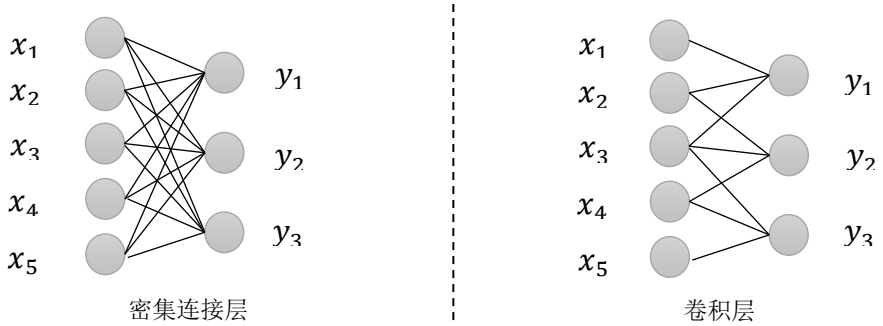


图 2-5 连接方式结构图

Figure 2-5 Structure diagram of connections

对于密集连接层：

$$y_1 = W_{11}x_1 + W_{12}x_2 + W_{13}x_3 + W_{14}x_4 + W_{15}x_5 \quad (2-6)$$

$$y_2 = W_{21}x_1 + W_{22}x_2 + W_{23}x_3 + W_{24}x_4 + W_{25}x_5 \quad (2-7)$$

$$y_3 = W_{31}x_1 + W_{32}x_2 + W_{33}x_3 + W_{34}x_4 + W_{35}x_5 \quad (2-8)$$

对于卷积层：

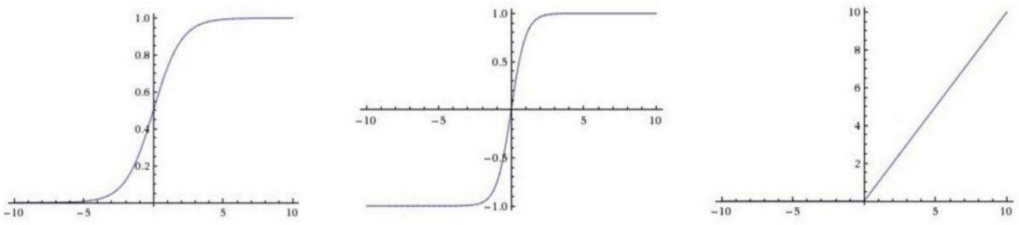
$$y_1 = W_{11}x_1 + W_{12}x_2 + W_{13}x_3 \quad (2-9)$$

$$y_2 = W_{21}x_2 + W_{22}x_3 + W_{23}x_4 \quad (2-10)$$

$$y_3 = W_{31}x_3 + W_{32}x_4 + W_{33}x_5 \quad (2-11)$$

其中$W_{i,\ j}$为权重矩阵。由上公式 2-6 到 2-11 可以看出，基于密集连接层的特征提取是对所有输入数据进行提取，而基于卷积的方式则是对于局部区域进行连接，以滑动窗口的方式进行特征提取。因此相对于密集连接的方式而言，能够极大的减少在模型训练过中所需的参数数量。

在现实生活中，大部分的事物都是以非线性的状态存在，单纯的依赖于线性表达很难对特征进行精确描述。因此，为了强化神经网络模型的学习能力，在卷积层中，每个卷积核除了线性表达部分外，还包括一个非线性的单元。常见的激活函数包括 tanh（曲线正切）、sigmoid（S 型函数）以及 Relu(修正线性单元)等。



sigmoid　　tanh　　Relu

图 2-6 激活函数示意图

Figure 2-6 Structure diagram of activation function

（1）sigmoid 激活函数

sigmoid 常用在二分类任务中，适用于前向传播。但是也存在一个很严重的缺点，就是输出不是以零为中心（零均值），即函数的输出范围为[0,1]。模型在训练过程中，如果输入的神经元的数值总是为正数或者负数的时候，那么权重参数 W 在训练时也就全部为正数或者负数的输出，即其下降趋势呈 Z 字型下降，会导致模型收敛曲线陷入波动的状态，无法快速的收敛。同时由图 2-6 可以看出，sigmoid 函数两端都有一段平缓部分，当数据分布在模型训练的过程中逐渐偏移到两端的时候，会造成梯度消失现象。因此，当神经网络结构较小时，sigmoid 表现更好。

具体公式为：

$$f(x)=\frac{1}{1+e^{-z}} \qquad (2-12)$$

（2）tanh 激活函数

tanh 不同于 sigmoid 的结构，其输出的取值范围为[-1,1]，因此不存在零均值问题。但由图 2-6 可以看出，函数两端依然存在平缓区域，也就是存在梯度消失问题。其公式如下：

$$f(x)=\frac{e^{x}-e^{-x}}{e^{-x}+e^{x}} \qquad (3-13)$$

（3）Relu 激活函数

由图 2-6 可知，Relu 函数的取值范围为[0,+∞]，其曲线的形状为直线状态，因此该函数的导数始终为常数，能够避免模型训练过程中产生梯度消失的问题。此外，由于 Relu 函数的求导不涉及到浮点运算，所以在反向传播的过程中会有加速运算的效果。但该方法对参数初始化以及学习率的要求较高，即当梯度更新到[-∞,0]的状态时，这个状态下的神经元将无法再次被其它数据单元再次激活（此时的梯度为 0），在一定程度上会丢失数据的多样化。其公式表达如下：

$$f(x)=\max(0,x) \qquad (2-14)$$

经过卷积层提取后的特征图存在一些无用以及相似部分，因此卷积层后面一般会接有一个池化层，用来去噪以及减少训练参数的数量。池化操作主要是通过滑动窗口的方式对从卷积层提取的特征进行局部区域内的特征提取（具有代表性的特征），选取的特征一般为局部区域的平均值或最大值作为输出，在保证特征原有强度的同时去除部分无用信息。

图 2-7 为最大池化方式：

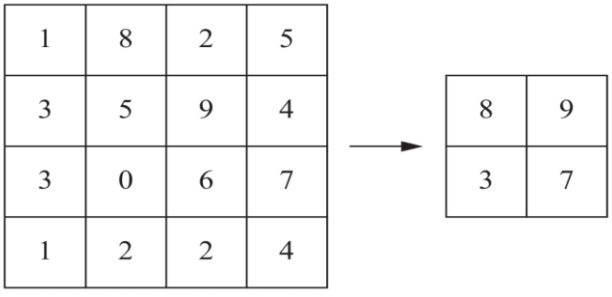


图 2-7 最大池化结构图

Figure 2-7 Structure diagram of Max pooling

上图的滑动窗口大小为 2*2，对 4*4 大小的特征图进行依次扫描，通过最大池化的方式对每一个大小为 2*2 大小的局部区域选取其中最大值作为输出，使得特征图缩减至 2*2。由此可以看出，池化层的作用主要表现为对数据的压缩降维，并

尽量保存特征图中的重要特征。池化方式主要包括两种，其中最大池化是通过选取领域中最大值作为整体领域的输出值，适应于特征差异较大的情况。均匀池化则是计算领域之中所有单元的平均值作为领域的输出，适用于特征间差距相对较小的情况。除此之外，池化层还具有平移不变性，当局部区域有少量偏移的时候，并不影响最终的特征输出。

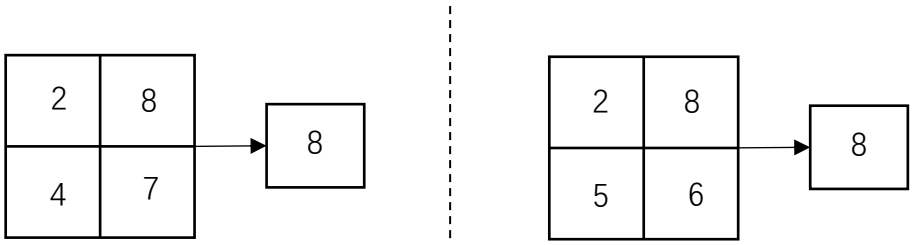


图 2-8 平移不变性图示

Figure 2-8 Illustration of translation invariance

在模型的输出层，一般是以 softmax 函数计算各个类别的概率分布情况，选取概率最大的作为最终的类别输出。具体公式如 2-15，其中e^{x_i}表示向量 T 中的第 i 个预测结果，t 表示分类的类别数。

$$P_i = \frac{e^{x_i}}{\sum_{j=1}^{t} \mathrm{e}^{x_j}} \qquad (2-15)$$

2.2.2 循环神经网络

（1）传统的循环神经网络

对文本数据等非线性序列特征的学习，除了对文本特征本身的提取，还需要结合上下文语义去发掘更深的文本表征形式。与早期的人工神经网络不同，RNN 具有记忆性，通过当前的输入信息结合之前节点的历史信息进行输出。这种模型结构能够很好的结合上下文联系，因此能以很高的效率对文本序列的非线性特征进行学习。

主体结构如图 2-9 所示：

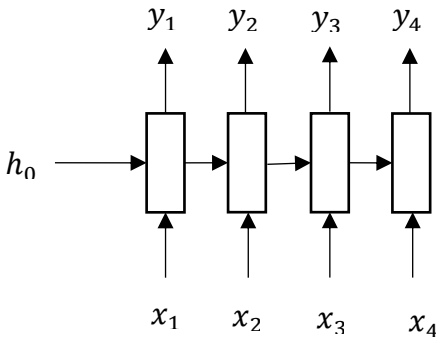


图 2-9 循环神经网络结构示意图

Figure 2-9 Structure diagram of RNN network model

文本序列的表达一般基于词汇成句的方式，单纯的词汇只能包含很少的语义信息，需要结合上下语义进行语义表征。图 2-9 为常见的循环网络结构的多对多结构，即输入数目与输出数目相同，此外还有一对多、多对一等等。其中x_t表示在 t 时刻的输入，y_t表示在 t 时刻的输出，同样作为 t+1 时刻的输入。当前节点的输出y_t除了与当前节点的输入x_t有关外，还与上一节点输出的隐状态y_{t-1}有关，计算公式如下：

$$y_t = f(W_1 y_{t-1} + W_2 x_t) \tag{2-16}$$

$$\sigma_t = softmax(U y_t) \tag{2-17}$$

其中W_1、W_2以及 U 表示参数矩阵，f代表激活函数，其主要作用是增加模型的非线性表达，使得模型能够更加清晰的刻画出文本特征。通过输入的x_t以及上一个节点的隐状态可以计算出当前的输出以及隐状态y_t，经过 softmax 函数进行概率输出，得到预测的类别，并与真实的类别进行比较得到误差，最终使用反向传播算法进行参数更新（包括参数矩阵W_1、W_2以及 U），使得模型可以更好的拟合文本数据。

RNN 的模型结构特性使得其在处理文本等非线性序列的时候有着很大的优势。虽然其优异性显著，但是也同时存在着缺点，其中最大的一个缺点就是随着 RNN 网络深度的加深，容易产生梯度消失，并且由于其每次进行计算都需要根据前 n 个单词进行上下文联系，因此计算速度也是一个非常大的问题。因此有研究人员对 RNN 进行优化和改进，提出了 LSTM 以及 GRU。

（2）LSTM（Long Short Term Memory networks）

RNN 擅长处理文本等非线性序列，其主要优势体现在能够处理文本数据的“长依赖“问题，即能够持久存储特征信息。这种性质使得模型在反向传播时会随着“依赖距离”的增加而产生梯度弥散。虽然传统 RNN 在理论上是可以学习到数据特征的长依赖信息，但是实际却因为梯度消失问题而难以学到数据的长距离特征信息。

针对以上问题，Hochreiter 等人提出了 LSTM 模型[38]，在一定程度上让 RNN 能够真正应用于实际。和传统的 RNN 结构一样，LSTM 同样采用链式连接方式，不同的是其单元内部存在三种门控结构。

LSTM 模型结构如下：

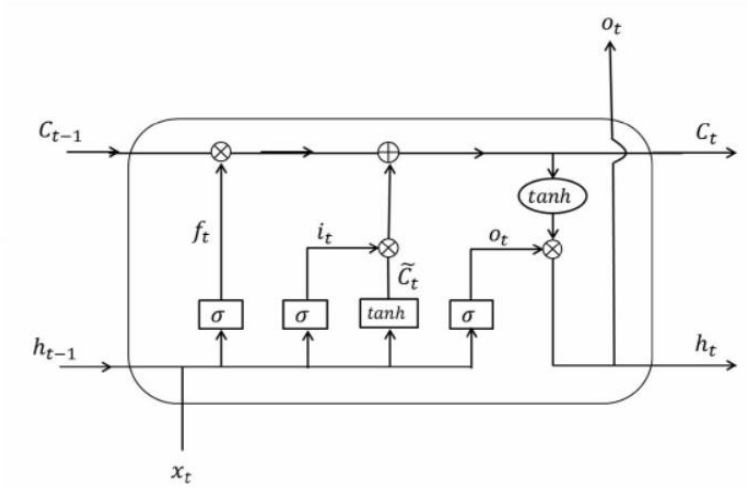


图 2-10 LSTM 网络结构示意图

Figure 2-10 Structure diagram of LSTM network model

由图 2-10 可知，三个门控结构为输入门、遗忘门以及输出门。通过三个门控单元使得 LSTM 可以选择性对于结构中的每个时刻的状态进行控制，其中C_{t-1}是上一个节点单元更新后传递的隐状态，C_t表示当前节点单元更新后的隐状态。门控结构是由一个激活函数为 sigmoid 的神经网络以及一个数乘操作组成，其工作方式为：输入的数据经过神经网络输出一个大小为 0 到 1 的数值，这个数值的大小是控制当前的输入有多少可以通过门控单元继续传递。三种门控单元的具体实现公式如下。

输入门：

$$Input_{gate_t} = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{2-18}$$

$$\widetilde{C_t} = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{2-19}$$

遗忘门：

$$Forget_{gate_t} = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{2-20}$$

$$C_t = Forget_{gate_t} * C_{t-1} + Input_{gate_t} * \widetilde{C_t} \tag{2-21}$$

输出门：

$$Output_{gate_t} = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{2-22}$$

$$h_t = Output_{gate_t} * \tanh(C_t) \tag{2-23}$$

其中，$\widetilde{C_t}$表示当前时刻的隐状态，W_i、W_c、W_f、W_o以及b_i、b_c、b_f、b_o皆为参数矩阵（权重矩阵以及偏置矩阵）。三种门控单元中，输入门控单元主要是根据当前时刻的输入以及前一时刻的输出来控制是否需要改变当前节点的状态，如果不需要则直接通过当前节点继续传递到下一节点；遗忘门是对上一时刻的节点

所存储的历史信息进行控制；输出门则是控制当前时刻节点是否应该对下一节点产生影响。可以看出，LSTM 通过三种门控结构的形式传递以及更新文本特征的历史信息，能够减少在传播过程中的产生的梯度消失或爆炸问题，使得模型能够真正的学习到非线性序列的长距离依赖信息。

2.3 优化算法

在深度神经网络的训练过程中，参数空间中存在着很多鞍点或局部最小值点，使得损失函数在训练过程很难更新至最优状态，因此需要对反向传播算法进行优化。早期的优化方式是采用控制训练批次大小来加速模型的训练速度。但是训练批次的大小难以选择，特别是对于凹凸不平的参数空间，如果以相同的学习率进行参数更新就很容易使得损失函数波动不收敛。因此，通过控制全局学习率或者局部学习率的优化算法成为了人们研究的重点。

2.3.1 梯度下降

梯度下降算法（Gradient Descent，GD）是最基础的反向传播算法之一，其学习率不会在训练过程中发生改变，一次对全部样本的梯度进行参数更新，因此在模型训练时训练速度很慢，时常导致模型陷入局部最小点。

具体更新公式如下：

$$\theta = \theta - \mu \cdot \Delta J_{(\theta)} \tag{2-24}$$

其中μ表示学习率。GD 算法简单易理解，但其参数更新规则过于简单，使得其很难在复杂的参数空间中帮助参数更新到最优状态，因此研究人员在其基础上提出了几种改进的算法。

2.3.2 随机梯度下降

随机梯度下降（Stochastic Gradient Descent，SGD）的更新方式不同于梯度下降，对于参数更新的数量其每次只会随机选择样本中的一个，因此更新速度相对较快。

具体更新公式如下：

$$\theta = \theta - \mu \cdot \Delta J_{(\theta;x_i,y_i)} \tag{2-25}$$

其中x_i、y_i表示随机选取的第i个样本和其相应的标签。随机梯度下降算法在速度上要优于梯度下降算法，但是对于样本数据量过大的情况，单个的样本很难代表样本的特征选择，会导致损失函数随机波动，难以快速收敛。因此其不适用于特征差距较大的数据集。

2.3.3 小批量随机梯度下降

小批量随机梯度下降（Mini-Batch Gradient Descent，MBGD）对于每次训练批次的选择取值进行了优化。不考虑学习率的情况下，反向传播的快慢取决于每次更

新参数的数量。更新参数数量过大时，训练速度就会减慢；而更新参数数量较小时，损失函数容易受单个样本的影响而波动。

具体更新公式如下：

$$\theta = \theta - \mu \cdot \frac{1}{n}\sum_{i=1}^{n} \Delta J_{(\theta;x_i,y_i)} \qquad (2-26)$$

MBGD 可以在一定程度上使用了局部特征的信息进行梯度传播，不会因为单个样本的偏差而导致损失函数波动，因此收敛过程相对稳定一些，并且可以使用矩阵向量化的方式加速每个批次数据的运算。其缺点是批次大小的选择不当也会带来一些问题，如批次过大就造成收敛速度过慢，批次过小就会导致损失函数波动，同样难以快速收敛。

2.3.4 Adagrad

反向传播算法中，除了对训练批次的选择，还要考虑更新过程中对于学习率的改变。学习率主要是控制参数在更新过程中的更新速度，对于凹凸不平的参数空间，如果使用一层不变的学习率，损失函数很快就会陷入局部最小值区间。因此需要在训练过程中根据梯度状态对学习率进行更改，使得其在参数空间比较平缓的区域可以加速通过。具体更新公式如下：

$$\theta_{i+1,j} = \theta_{i,j} - \mu \cdot \frac{1}{\sqrt{G_i + \epsilon}} \cdot g_{i,j} \qquad (2-27)$$

其中，$\frac{1}{\sqrt{G_i+\epsilon}}$表示历史梯度和，相当于一个约束项，$g_{i,j}$表示当前批次的梯度。我们一般是希望模型在刚开始训练的时候，反向传播的速度加快，能够快速到达最优状态附近；而当在训练快要结束的时候，我们一般会希望传播速度减慢，使得损失函数逐渐逼近最优点，防止梯度动能太大跳过最优区间。由式 2-27 可以看出，前期训练时$g_{i,j}$较小，约束项较大时，能够加速梯度传播；后期训练时$g_{i,j}$较小，历史梯度较大使得约束项变小，因此能减小梯度传播速度。

2.4 本章小结

本章介绍本文涉及到的相关技术，主要分为两个部分：数据部分以及模型部分。数据部分主要包括对文本数据的采集、清洗去噪、分词以及文本表示方式，并且介绍了文本特征提取的相关方法。模型部分主要介绍了 CNN 以及 RNN。卷积神经网络主要介绍卷积层以及池化层相关特性，并对比分析稀疏连接的优势和池化所带来的平移不变性。循环神经网络主要介绍传统的 RNN 以及 LSTM，并分析常用的 RNN 结构无法对序列的“长距离依赖”的原因，以及 LSTM 如何进行优化。通过对相应理论技术进行分析学习后，为后续的模型的提出以及性能优化提供支撑。

第 3 章 基于 CNN 的文本情感分类

随着网络数据的爆炸式增长，使得文本的表现形式出现多样化。同时也出现了很多问题，例如主题不明确、语言不规范、新词较多等。对于这些问题，使用人工处理的成本较高，而传统的机器学习模型无法得到文本数据中深层次的语义信息。基于以上分析，本章基于卷积神经网络进行文本情感分类研究。首先对模型的输入使用 word2vec 工具进行词向量转换，随后将其作为 CNN 网络的输入并对模型进行训练。在模型训练的过程中，考虑到不同维度的词向量可能会对模型的分类精度产生一定影响，分别生成不同维度的词向量进行对比实验。在此基础之上，提出模型 KMCNN，使用 k-max 池化方式代替传统的池化方式。实验结果表明，KMCNN 在文本情感分类任务上具有良好的分类性能，能在去除一定噪声影响的同时，保留局部重要特征。

3.1 实验设计

3.1.1 实验数据集

本文通过网络爬虫爬取豆瓣影评（url=http://movie.douban.com）作为研究数据，一共爬取 173584 条电影评论，其中长文本 58496 条，短文本 115088 条。本文主要研究的基于网络短文本的情感分类研究，因此需要去除长文本数据，并经过数据清洗处理之后得到 82368 条数据。

数据清洗完成之后，需要对数据进行标注。本文对采集的数据集先后进行三次数据标注，为了尽量防止人工错标的情况，选取三次结果相同的数据进行二次复标。最终得到数据集包括正向数据 6580 条，负向数据为 7860 条。考虑到样本的平衡性，选取正负样本数据中的各 6000 条作为最终的实验数据，同时抽取正负样本数据中各 25%作为测试集。

详情见表 3-1：

表 3-1 数据分布图示

Table 3-1 Illustration of data distribution

训练集	
正项情感倾向	4500
负面情感倾向	4500
测试集	
正项情感倾向	1500
负面情感倾向	1500

3.1.2 实验环境

基本系统：win10 系统，内存为 16GB，显卡为 GeForce GTX 1060；

编程语言：python；

编程软件：vscode；

深度学习框架：keras。

3.2 算法设计

3.2.1 实验流程设计

文本序列语义多变，其语义表达不依赖于某个词语，需要结合文本的上下文信息进行分析。如：“这个水果的味道我非常喜欢”，“这个水果的味道非常符合我的胃口”，两段文本中虽然组成的词语不相同，但是所表达的情感极性是一样的。同样，相同的词语在不同的语义环境下所带给整体文本的情感极性也会不同。如：”这款冰箱的性价比真的很高”,“这款冰箱的价格真的很高”,对于相同的情感词“很高”，但是第一种表达为褒义，第二个表达却为贬义，因此需要学习更深层次的文本特征进行文本情感分类。

卷积神经网络的特征提取方式是以模块化的神经网络结构为基础，以层层叠进的方式对特征进行学习。随着神经网络层数的增加，每一层所提取的特征会变得越来越抽象，层数越高的神经元中包含着越来越少的输入信息，而关于文本类别的信息会逐渐增多，最终形成具有更加抽象的语义信息的特征。对于提取的抽象语义信息，使用分布式的概率分布进行特征表征，从而以简单的神经元实现复杂现实问题的表示，相对于机器学习的分类方法能取得更好的分类效果。

具体模型结构如图 3-1 所示：

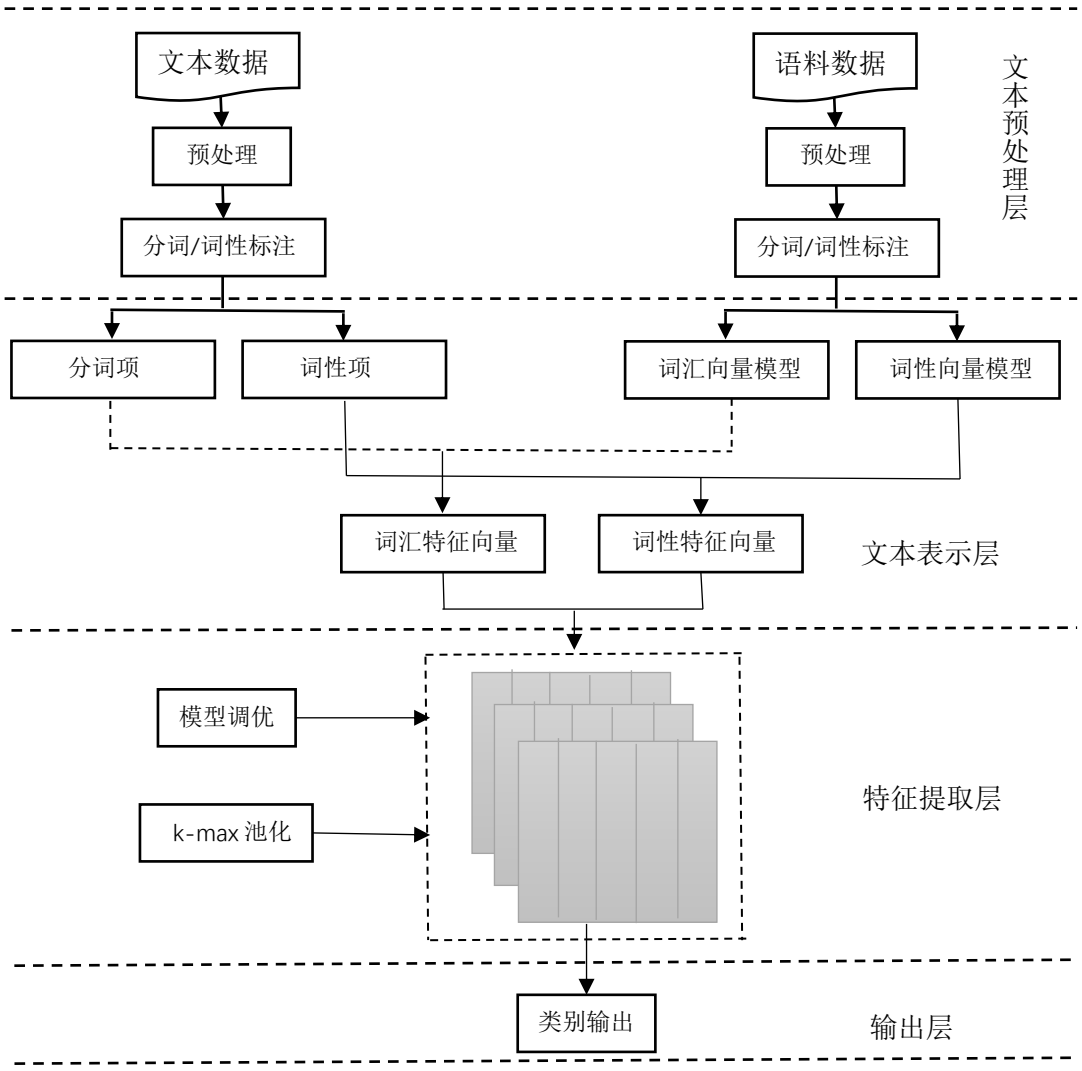


图 3-1 实验流程图

Figure 3-1 Flow chart of the experiment

为了统一输入，将文本向量转换为词向量矩阵的形式。并对原始采集的文本进行相应的预处理，随后对词性进行标注，将分词后的文本数据以及词性数据通过 word2vec 工具进行训练得到不同维度的向量矩阵。

由图 3-1 可知，KMCNN 主要包括四个部分：文本预处理层、文本表示层、模型训练层以及输出层。

（1）文本预处理层

本文的数据主要来源于网络短文本，通过网页爬虫采集的文本数据中会包含很多无用的部分，这些内容不包含情感倾向，但是会占用大量内存，并且在训练时会影响分类精度。因此我们需要对分词后的数据进行处理。数据的清洗主要包括去除网页字符、非中文语料以及繁体字等，主要是为了保证数据的规范性，使计算机能够更加清楚的了解数据，在一定程度上也能提高模型的分类精度。

（2）文本表示层

文本数据是非结构化数据，其表现形式多种多样。卷积神经网络擅长处理像图像这样可以通过像素表示成二维矩阵的数据，无法直接处理文本数据，因此需要通过词向量的方式将文本数据转换为数字向量。此外，考虑到对文本数据的特征提取是属于对非线性特征序列的学习，需要结合上下文之间的联系来加强文本特征的表征，因此在文本表示层除了常见的词汇向量矩阵，还结合词性标注的方式来强化文本表征能力。

词性标注经常应用于文本预处理任务中，是给词汇指定一个词类或词汇类别标记的过程[39]。在自然语言处理领域有着诸多应用，如文档分类、关键词检索以及机器翻译等。词性主要分为有具体意义的实词以及没有具体意义的虚词，实词主要包括名词、动词以及形容词等，而虚词则包括介词、连词以及助词等[40]。一句完整的文本表达通常会混合使用实词和虚词。文本的组合形式与词性有着很大的关联，常见的结构有动词+副词、副词+形容词等，如“我/喜欢/看/电影”对应的词性表达就是[r，v，v，n]，其中 r，v，n 分别表示代词、动词和名词。由此可以看出，词性也可以作为文本特征的一种表现形式，在一定程度上体现出词汇特征与上下文语义之间的潜在关系。由于词性特征无法被计算机直接识别，因此同样参照词汇向量的生成方式生成词性向量矩阵，并按照拼接的方式将两个特征向量矩阵进行融合，具体公式表示如下：

$$h_i = w_i \oplus c_i \tag{3-1}$$

其中，符号⊕表示行方向的连接，h_i表示由第 i 行的词汇向量以及词性向量拼接而成的向量。

拼接样式如图 3-2 所示：

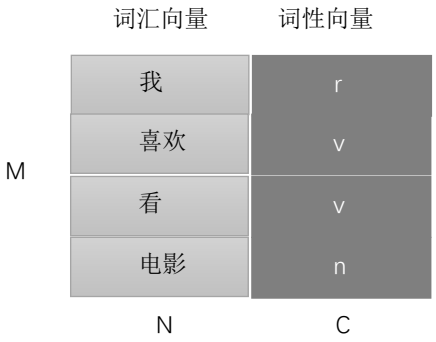


图 3-2 拼接向量

Figure 3-2 Stitch Vector

由图 3-2 可知，词汇向量矩阵维度大小为 M×N，词性向量矩阵的维度大小为 M×C。其中 M 表示文本数据集中词汇数目，N 表示词汇向量维度，C 为词性向量维度。在输入层将词性特征于词汇特征拼接组成新的向量矩阵，维度大小为 M×

(N+C)。

本文首先使用开源分词工具 Hanlp 对采集的数据集进行相应的预处理，包括分词以及词性标注任务等。同时考虑到日益发展的网络传播所带来的信息口语化的改变，如“辣眼睛”、“盘它”以及“狗带”等等，这些网络新词不仅蕴涵丰富的情感色彩，还受到众多网络用户的追捧，成为一种新时代网络文化。为了不破坏这类词的语义信息，本文也将这些词语以及它们的词性加入到分词工具的词典中进行扩充，最后采用基于 skip-gram 方法的 word2vec 训练词向量以及词性向量。

其中部分 word2vec 模型的训练参数如表 3-2 所示：

表 3-2 word2vec 模型参数部分示例

Table 3-2 Example of The Parameters of the Word2vec Model

超参数	参数值
窗口大小	3*3
迭代次数	5
模型依据	skip-gram
词向量维度	50/100/150/200
采样值	1e-3

考虑到词向量的维度大小会对分类结果造成影响，因此分别生成了四种不同维度大小的词向量矩阵并进行对比实验。经过分词后的单条文本最多的分词项为 76 项，对于分词数小于 76 项的则通过添加 0 项的方式使得分词长度一致。

实验结果如图 3-3 所示：

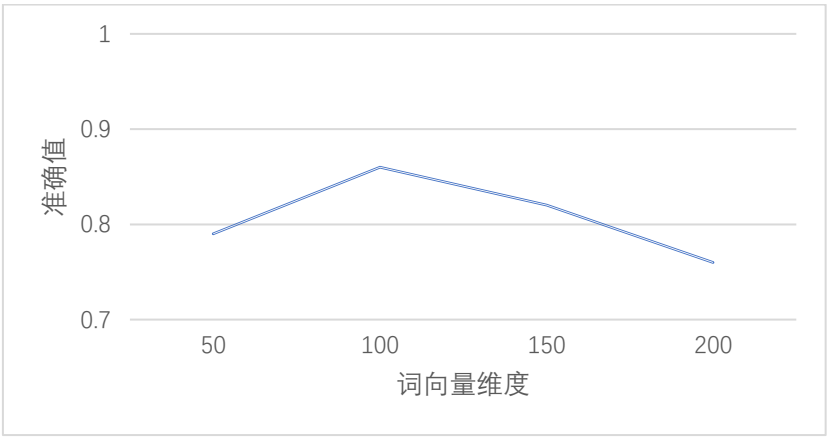


图 3-3 词向量维度实验对比结果

Figure 3-3 The experimental results of the word vector dimension

词向量维度作为 CNN 模型的输入属性之一，不同大小的维度会对模型的分类

精度产生一定影响。由图 3-3 可知，当维度大小为 100 维时，模型的准确性最好，表明此时的模型对于输入数据的拟合程度最好。

（3）特征提取层

特征提取层主要由 CNN 组成，CNN 是目前深度学习领域非常具有代表性的神经网络之一，目前已经被大范围的应用到图像的各个领域中。而随着 Kim 等首次采用 CNN 进行文本分类任务，取得了相对于传统机器学习算法更好的效果，打开了使用 CNN 解决自然语言处理问题的大门。相较于其它的神经网络结构，CNN 的优势主要体现在两个方面：

（1）卷积层结合池化层的方式使得 CNN 对于局部特征有着优异的提取能力，并且提升了模型的鲁棒性。

（2）基于其连接的稀疏性以及参数共享的模式，使得 CNN 所需的参数远小于其它神经网络。不仅能够简化模型的复杂性，还在一定程度上使得模型具有更好的泛化性。

文本情感分类算法根据不同的粒度大小大致可分为基于词语级粒度、基于篇章级粒度以及基于句子级粒度[41]。基于 CNN 的文本情感分类一般基于词语级粒度，借助于中文分词工具将文本数据进行分词处理，然后结合 word2vec 将分词后的词汇特征以及词性特征转换成可以作为神经网络输入的向量形式。本文所应用的 CNN 模型主要包含四个主体层级：卷积层、池化层以及全连接层。

具体结构如图 3-4 所示：

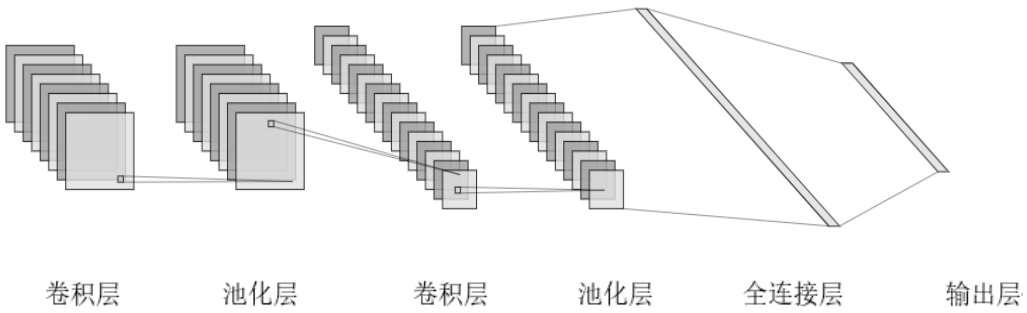


图 3-4 基于 CNN 的文本情感分类模型结构

Figure 3-4 Text emotional classification model structure of CNN

①卷积层

卷积层以滑动窗口的方式进行特征提取，如图 3-5 所示。

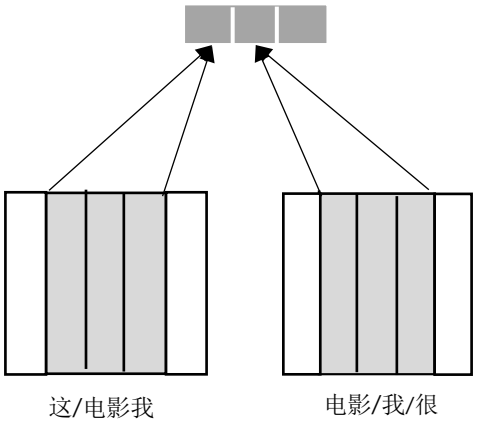


图 3-5 卷积运算

Figure 3-5 Operation of convolution

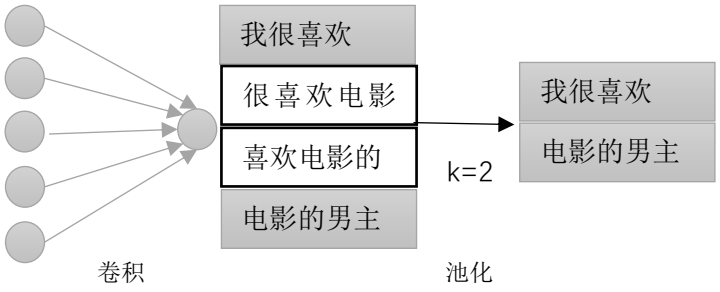
在图 3-5 中，长度为 3 的窗口作为一个整体，从左往右以步长为 1 的方式进行卷积运算。例如，对于“这/电影/我/很/喜欢”这句话，以词向量的形式输入卷积神经网络，经过一次卷积后得到以下短语特征：{这/电影/我/；电影/我/很；我/很/喜欢}。滑动窗口的大小代表着感受野的大小，滑动窗口越大，带来的更大的感受野，但同时也会带来计算复杂度的增加。因此常见的滑动窗口大小一般为 3。

② k-max 池化

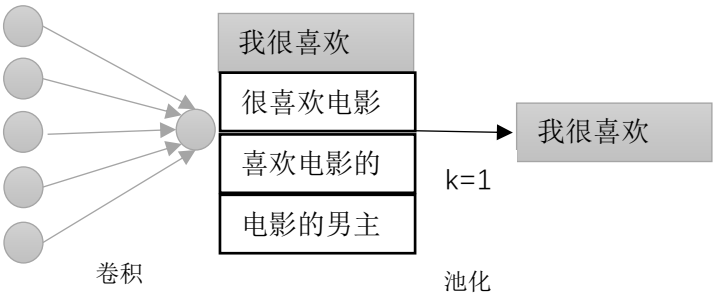
池化层是卷积神经网络中另外一个不可或缺的部分，一般是存在于在连续的卷积层中间，主要是用于压缩数据和参数的量，以此来减小过拟合。池化方式主要包括两种：最大池化以及均匀池化。最大池化是通过选取领域中最大值作为局部区域的输出值，适用于特征相差较大的情况。均匀池化则是计算领域之中所有单元的平均值作为局部区域的输出值，适用于特征相差较小的情况。

最大池化能够在一定程度上突出局部重要特征，但是由于方式太过于简单粗暴，容易造成信息的丢失。本文使用 k-max 池化方式代替传统的池化方式，主要区别在于 k-max 池化认为每一个池化区域并不是只有一处重要特征，而是排序在前 k 个的特征都很重要。

具体结构示意图如图 3-6 所示：



（a）k-max 池化



（b）最大池化

图 3-6 池化方式对比图

Figure 3 comparison of pooling mode

由图 3-6 可知，相较于传统的最大池化，k-max 池化能够保留前 k(经过实验验证，最终选择 k=2)个文本特征，在一定程度上丰富了语义信息，可以表达同一类特征出现多次的情形，即可以表达某类特征的强度。并且相应的保留了部分位置信息，但是这种位置信息只是特征间的相对顺序，而非绝对位置信息。

③全连接层与输出层

在全连接层中，将特征h利用非线性函数映射成维度大小为 t 的向量（本文的情感倾向只有两种，所以 t 为 2），具体公式如下：

$$H = Relu\left(W_h \cdot x + b_h\right) \tag{3-2}$$

其中W_h和b_h为全连接层的权重矩阵以及偏置矩阵。利用 softmax 函数得到每一个类别的概率分布，将概率最大的类别作为最终的类别输出。

$$P_i = \frac{e^{x_i}}{\sum_{j=1}^{t} e^{x_j}} \tag{3-3}$$

其中e^{x_i}表示文本向量 T 中第 i 个元素。

3.2.2 模型训练

（1）偏差-方差

偏差主要是表示模型对数据的拟合程度，而方差则是表示数据扰动所带来的影响，如噪声等。通过对方差以及偏差的分析，能够快速高效的反映出模型的泛化能力。

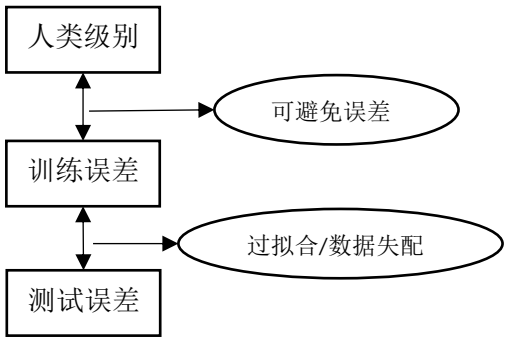


图 3-7 误差分析图

Figure 3-7 Analysis chart of error

三种常见的拟合状态包括欠拟合、正常以及过拟合。欠拟合一般指模型在模型没有很好的拟合数据；而过拟合则是模型对于训练数据的匹配过高，在使用新的数据集进行测试时会产生很大的方差。

常用的减少过拟合问题主要包括正则化、dropout、early stoping 以及数据扩增等。而对于欠拟合的控制，一般有两种：①挖掘更多的特征，增加输入特征空间维度。②选择更复杂的模型（增加网络的深度、神经元个数以及基数等）

（2）参数选择

深度学习模型在训练时除了需要大量的数据，还会涉及到很多超参数的选择以及训练技巧的应用，我们不仅需要考虑最终的准确率的高低，还需要结合数据特点来综合考量模型的优劣。本节主要涉及相关超参数的选择以及模型优化所做的选择。

对于超参数的选择主要包括以下几个方面：

①卷积核的选择；

②激活函数的选择；

③批量处理大小；

④Dropout 大小

常见的超参数选择方式主要包括网格搜索、随机搜索以及贝叶斯优化等。考虑到模型的量级以及超参数的数量，对于部分超参数的选择采取网格搜素的方式。其中包括：卷积核的大小采用（1，3，5，7）四组方案进行搜索；卷积核的数量采用（16，32，64，128）四组方案进行搜索；批量处理大小（Batch_Size）采用（32，16，8，4）四组方案进行搜索。经过多次实验之后，模型的最终参数选择如表 3-3 所示：

表 3-3 部分模型参数示例

Table 3-3 Partial model parameter examples

	卷积核数目	卷积核大小	其他参数设置	
卷积层-1	16	5*5	Dropout	0.5
池化层-1	16	2*2	Batch_Size	8
卷积层-2	32	5*5	激活函数	ReLU
池化层-2	32	2*2	词向量维度	100

在模型训练过程中，常常会因为模型的复杂度以及数据不匹配等问题造成过拟合问题，Dropout 通过对神经元进行随机失活的方式来降低特征之间的关联性[42]。对于 Dropout 大小的选择本节也采取网格搜索的方式，分别选择大小为（0.1，0.3，0.5，0.7）四组方案进行实验，最终选定数值大小为 0.5 作为 Dropout 的数值，其实验结果如图 3-8 所示。

由图 3-8 可知，随着 Dropout 数值的依次递增，模型的准确率也随之在不断变化着。当数值由 0.1 到 0.5 的过程中，模型的准确率有着很明显的增长，主要是因为模型在训练之后产生了过拟合问题。通过随机失活部分神经元，使得模型的泛化性增强，进而使得分类准确率上升，而随着失活的神经元数量达到 70%的时候，模型已经无法很准确的拟合数据，处于欠拟合状态，因此准确率开始下降。

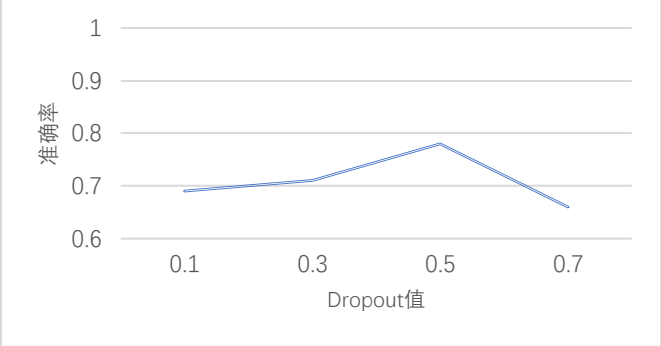


图 3-8 Dropout 实验结果

Figure 3-8 Results of Dropout experiments

（3）模型优化算法

①正则化

正则化的作用主要是防止过拟合，常见的正则方式包括 L1 正则、L2 正则等[43]。本文选择 L2 正则进行正则化，具体公式如：

$$y_i = w_i \cdot x + b_i + \frac{1}{2}\lambda w_i^{2\cdot} \qquad (3-4)$$

其中w_i、b_i为权重矩阵以及偏置矩阵，$\frac{1}{2}\lambda w_i^{2\cdot}$则为惩罚项，通常不对偏置$b_i$进

行正则化，主要是因为每一个偏置仅控制一个单变量，不对其正则化也不会产生太大的方差。加入 L2 正则之后，使模型更倾向于使用所有特征，而不是严重依赖于输入特征中的小部分特征，适用于无关联的特征。正则化相当于是模型中的先验信息，会引导着损失函数的最小值朝着约束的方向迭代，在一定程度上能够优化迭代速度并且减少波动。

②Adam(adaptive moment estimation，自适应矩估计)

传统的 SGD（梯度随机下降）算法比较单一，在模型训练过程中对学习率的状态不会进行改变。具体更新公式如下：

$$J_{(x)} = \frac{1}{n}\sum_{i=1} J_{(x_i)} \tag{3-5}$$

$J_{(x)}$为目标函数，$J_{(x_i)}$表示第x_i个样本的目标函数，那么目标函数在x处的梯度为：

$$\nabla J_{(x)} = \nabla \frac{1}{z}\sum_{i=1} J_{(x_i)} \tag{3-6}$$

梯度下降就是每次更新迭代的过程中需要 z 个样本的梯度，而 SGD 就是随机从所有样本中随机选择一个样本$J_{(x_i)}$进行参数的更新。在反向传播过程中，当遇到参数空间较为平缓的方向，容易陷入局部最小值或者鞍点因此我们需要动态改变学习率的状态，当历史梯度平方和比较小时，通过调整学习率（全局学习率）来加速迭代过程。本文选择 Adam 作为优化算法进行模型训练。

①梯度下降的缺点就是每次更新时只与当前位置的梯度方向进行，Adam 中通过加入动量变量v_t来控制不同方向的梯度，使得各个方向的梯度移动方向一致，减小传播过程中所产生的震荡。

$$v_t \coloneqq \beta_1 V_{t-1} + (1-\beta_1) g_t \tag{3-7}$$

同时还加入累加状态变量S_t进行指数加权平均，以及对小批量梯度进行指数加权平均。

$$s_t \coloneqq \beta_2 V_{t-1} + (1-\beta_2) g_t \odot g_t \tag{3-8}$$

采用偏差修正之后的更形式如下：

$$\hat{v}_t \coloneqq \frac{v_t}{1-\beta_1^t} \tag{3-9}$$

$$\hat{S}_t \coloneqq \frac{s_t}{1-\beta_2^t} \tag{3-10}$$

根据动量变量$\hat{v}_t$以及累加状态变量$\hat{S}_t$可以对小批量状态进行更新：

$$g_t' = \alpha \frac{\hat{v}_t}{\sqrt{\hat{s}_t + \varepsilon}} \tag{3-11}$$

$$x_t \coloneqq x_{t-1} - g_t' \tag{3-12}$$

其中α表示学习率，β_1，β_2，ε为超参数，通常β_1，β_2一般取值为 0.99，ε一般为 1e-8。由上面的公式可知，Adam 通过动态的适应学习率[44]，相较于梯度下降算法的单一学习率，在非凸优化问题上取得了很大的优势。

3.2.3 评价标准

考虑到本文是基于二分类问题的研究，因此评价标准主要采用准确率(Precision)、召回率(Recall)以及 F 值（F-measure）。其中，将正面情感预测为正面情感用 TP 表示，将正面情感预测为负面情感表示为 FN，将负面情感预测为正面情感用 FP 表示，将负面情感预测为负面情感用 TN 表示。

（1）准确率

准确率是表示在所有被分类的文本数据中，TP 以及 TN 所占的比例。其计算公式如下：

$$Precision = \frac{TP}{TP + FP} \tag{3-13}$$

（2）召回率

召回率表示在所有被分类的正例样本中，TP 所占的比例，其计算公式如下：

$$Recall = \frac{TP}{TP + FN} \tag{3-14}$$

（3）F 值

F 值是通过度量准确率以及召回率的值来综合考量不同算法的优劣，其计算公式如下：

$$F = 2 \cdot \frac{Precison * Recall}{Precision + Recall} \tag{3-15}$$

3.2.4 对比实验设计

（1）SVM：采用逆文档频率（TF-IDF）作为特征提取，使用 SVM 进行训练并输出相应的文本情感倾向。

（2）CNN：机器学习法对于文本特征的提取在很大程度上受到人为经验的影响，而基于深度学习的方法则是通过学习数据内部的统计结构，使其能从文本中自动地发掘复杂的特征，因此能够达到很好的分类效果。本节主要基于 CNN 进行文本分类，分为两组对比模型：CNN 和 Static-CNN，CNN 以及 TCNN。其中 CNN 模型的词向量参数矩阵初始化为随机生成，而 Static-CNN 则是基于迁移学习的方式；使用中文维基百科中预训练的词向量参数矩阵；TCNN 则是在 CNN 的基础上使用结合词性特征的文本表示作为输入。

（3）KMCNN ：KMCNN 主要是在基础的 CNN 模型上进行了改进，结合词性特征丰富文本表示。在此之上，使用 k-max 池化策略代替传统的最大池化策略。

3.3 实验结果与分析

3.3.1 实验结果

本节采用上一节提到的评价标准，分别计算出每个模型的准确率（P）、召回率（R）以及 F 值，实验结果如表 3-4 所示。

表 3-4 对比实验结果示意图

Table 3-4 schematic diagram of comparative experimental results

模型	准确率	召回率	F 值（%）
SVM	77.29	77.98	77.63
Static-CNN	80.15	80.56	80.35
TCNN	79.42	79.51	79.46
CNN	79.03	79.18	79.10
KMCNN	81.92	80.86	81.48

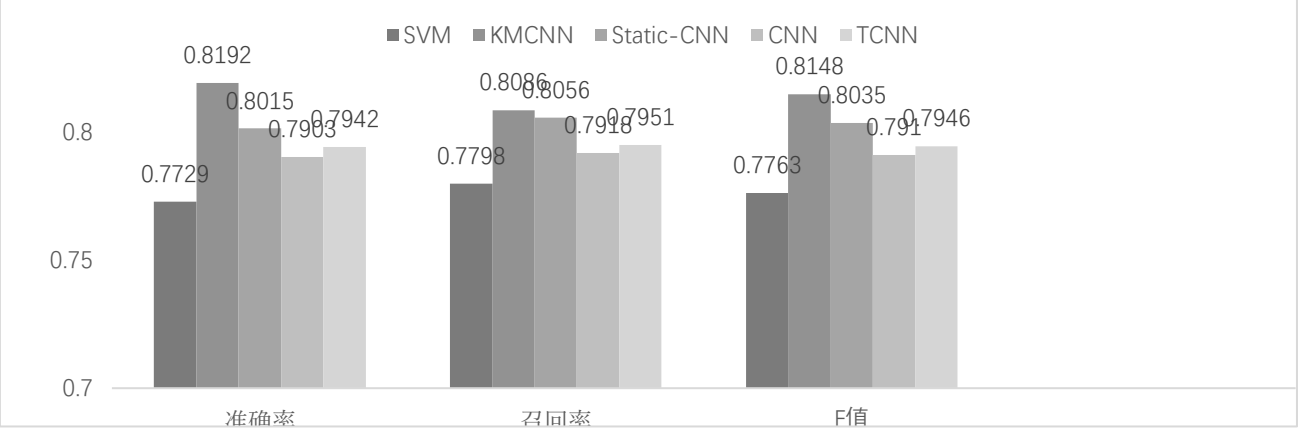


图 3-9 实验结果对比图

Figure 3-9 Comparison of experimental results

3.3.2 实验结果分析

通过图 3-9 的对比分析发现，基于 CNN 的文本情感分类方法在一定程度上要优于 SVM。原因推断：传统的机器学习方法采用的特征提取方式很大程度依赖于人为经验，虽然也能取得一定的成果，但是不能发掘更深层次的文本特征。而基于卷积神经网络的分类方法则是通过学习数据内部的统计结构，能够自动的提取文本数据特征，发掘更深层次的文本表达，因此有更好的分类性能。

通过对比 CNN 模型以及 Static-CNN 的实验结果可知，相对于词向量参数矩阵初始化的无规则生成，经过迁移学习的 word2vec 得到的词向量能够取得更好的分类性能。原因推断：基于 Skip-gram 方法的 word2vec 算法能够从大量的语料数据中学习到文本的表征形式，并且其本身也是一种神经网络模型，通过训练模型可以

学习词语间的位置关系进一步反映词汇之间的语义关系。因此使用预训练的参数矩阵能够更好表征这种映射关系以及更加精准的刻画词语特征，分类性能也就更好。

由图 3-9 的实验结果可知，KMCNN 在结合词性表征以及 k-max 池化两种优化方式之后，模型的性能有了一定的提升。原因推断：①对文本数据的特征提取是属于对非线性特征序列的学习，需要结合上下文之间的联系来加强文本特征的表征信息，因此在文本表示层结合词性标注的方式来强化文本表征能力。②传统的最大池化对于特征图的压缩过于简单，虽然能够在一定程度上突显出局部重要特征，但是容易丢失其它重要特征。而 k-max 池化认为在池化区域存在多个重要区域，通过选择前 k 个重要特征获取更多的文本特征。这种方式在一定程度上抽取到包含更加完整的语义特征，从而达到提升词语级粒度特征的丰富性。

3.4 本章小结

本章主要研究基于 CNN 的文本情感分类研究，并基于在传统的 CNN 模型基础上提出 KMCNN。为了更有效的利用情感资源，使用词性标注的方式，在传统的词向量表达的基础上，将词性表达与词汇向量相结合作为文本表示，实验证明结合词性特征的文本表达在一定程度上能够提升分类精度。在此基础上，考虑到传统的最大池化方式容易丢失特征信息，使用 k-max 池化方式代替最大池化，对特征进行有效提取。实验证明该模型相较于传统的 CNN 模型有更好的性能表现。此外，组织多组对比实验来探究词向量维度以及基于预训练词向量对模型分类精度的影响。实验结果表明，在其它条件一致时，对于本文的数据集而言，使用维度大小为 100 时的文本分类模型性能最好。

第 4 章 基于特征融合的 KMCNN-GRU

4.1 引言

由第三章可知，CNN 在文本情感分类任务上有不错的性能表现，但是单纯的依赖于词语级粒度进行特征提取可能无法涵盖更多的文本特征信息。针对以上问题，本章基于特征融合的方式提出模型 KMCNN-GRU，利用并行通道的方式，分别使用 KMCNN(去除最后的全连接层)以及 GRU 提取文本数据的特征，并将拼接融合后的文本特征送入 softmax 层，输出文本类别。

4.2 实验数据集与实验环境

4.2.1 实验数据集

本章采用的实验数据集与第三章所采用的数据集相同，都是来源于豆瓣影视的文本数据评论。对文本进行相应的预处理工作，数据清洗完成之后对数据进行标注。最终选取正负数据中的 6000 条作为最终的实验数据，同样抽取各自 25%的数据作为测试集。

4.2.2 实验环境

基本系统：基于 win10 系统，内存为 16GB，显卡为 GeForce GTX 1060；

编程语言：python；

编程软件：vscode；

深度学习框架：keras。

4.3 算法设计

4.3.1 实验设计流程

CNN 虽然在文本情感分类任务上取得了不错的成绩，但是对于文本数据等非线性特征的学习需要更多的联系上下文语义。卷积方式的优势在于能在一定程度上突出文本局部重要特征，而 RNN 则更擅长对长序列特征的学习。基于以上分析，本章基于 KMCNN 以及 GRU 对文本特征提取方式的不同，提出模型 KMCNN-GRU，将 KMCNN 以及 GRU 提取的文本特征进行结合，以获取更多的文本特征。这种方式在一定程度上抽取到包含更加完整的语义特征，从而提升词语级粒度特征的丰富性。

具体模型结构图如图 4.3 所示：

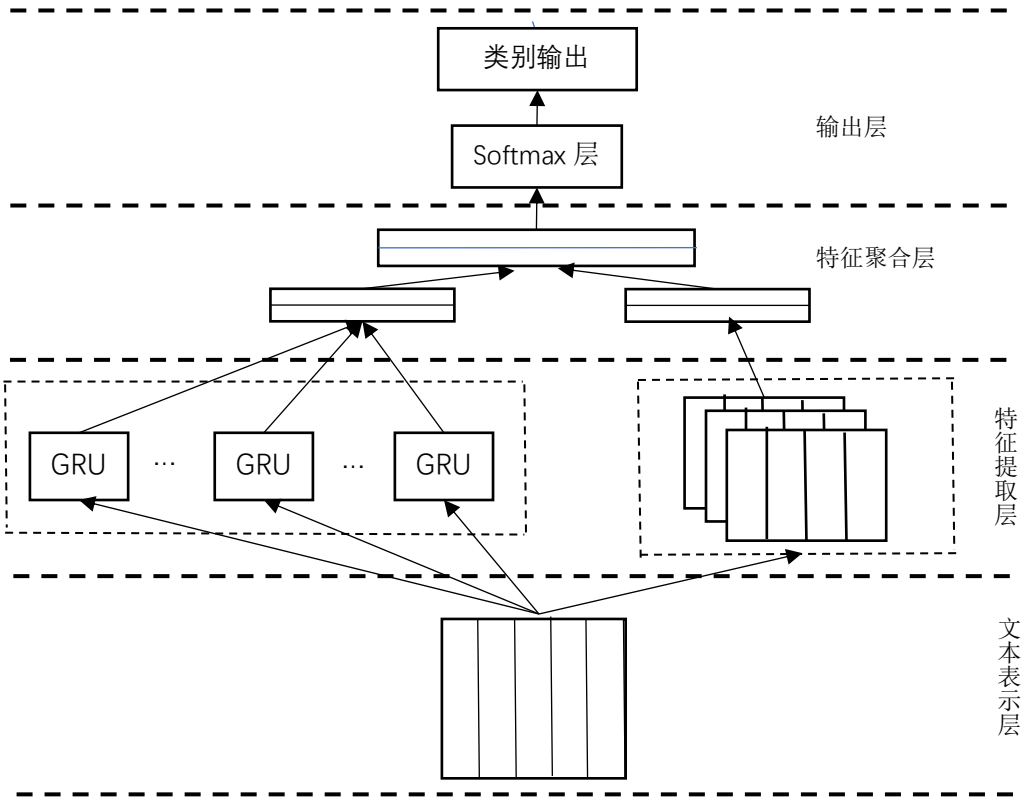


图 4.1 实验流程图

Figure 4.1 Flow chart of the experiment

如图 4.3 所示，KMCNN-GRU 结构主要包括四个部分：文本表示层、特征提取层、特征聚合层、输出层。

（1）文本表示层

本章的文本表示与第三章相同，对于原始的文本序列 T=$\{x_1, x_2, \ldots, x_n\}$，首先进行清洗去噪，然后进行分词处理，最后使用基于 skip-gram 模式的 word2vec 将文本数据以及词性标注数据进行聚合，并表示成向量矩阵。

（2）特征提取层

特征提取层主要包括两个部分，分别为基于 KMCNN 情感分类特征提取以及基于 GRU 情感分类特征提取。

①基于 KMCNN 情感分类特征提取

基于 KMCNN 情感分类特征提取与第三章模型结构基本相同，去掉最后的全连接层不输出分类结果，将经过池化层的文本特征$\boldsymbol{h}_{CNN}$作为特征聚合层的输入。

②基于 GRU 情感分类特征提取

RNN 具有记忆性，因此善于处理文本序列等非线性数据。而传统的 RNN 在对序列进行编码时经常会面临长距离依赖消失的问题。有鉴于此，Hochreiter 等人在 1997 年便针对该问题提出了 LSTM，LSTM 在传统 RNN 的基础上添加了三个门

控结构以及节点之间的信息传递结构，每个节点可以自动的控制是否使用上一个节点传递来的历史信息，同样也可以控制是否将自身的信息传递给下一个节点。该方法可以在一定程度上减少传统 RNN 出现的梯度弥散现象，并且能够实际应用于现实问题中。

Cho 等人于 2014 年提出的门限循环单元(Gated Recurrent Unit, GRU)[45]，GRU 相对于 LSTM 减少了一个门控结构，因此使得模型在训练过程中避免了一部分计算量，同时其性能近似于 LSTM，因此使用相对广泛。

GRU 的输入和输出与普通的 RNN 结构是一样的。首先假设当前的输入为x_t，和上一个节点传递下来的隐状态h_{t-1},这个状态包含了之前节点的相关信息。GRU 通过结合x_t与h_{t-1}可以得到当前节点的输出y_t以及传给下一个节点的隐状态h_t。

GRU 输入输出结构如图 4.1 所示。

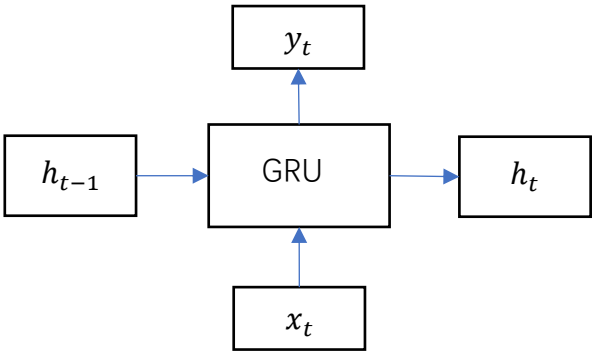


图 4.2 GRU 结构示意图

Figure 4.2 Structure diagram of GRU network model

不同于 LSTM 的三门结构，GRU 只有更新门以及重置门。但是同样可以通过控制这两个门控单元决定最终的输出信息，相比于 LSTM 其优势在于保存序列中的历史信息的同时还能加速计算，因此对于研究人员而言更具选择性。

GRU 结构图如 4.2 所示。

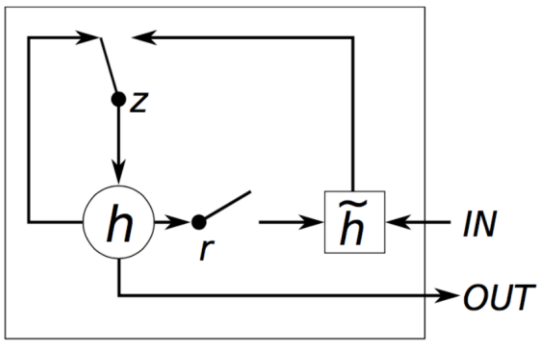


图 4.3 GRU 输入输出示意图

Figure 4.3 Structure diagram of GRU input and output

具体更新过程如下：

首先，需要根据上一个节点传输的隐状态h_{t-1}与当前节点的输入x_t来获取更新门以及重置门的状态。其中 r 为控制重置的单元，z 为控制更新的单元，σ表示函数 sigmoid，主要是将数据变换为 0 到 1 的数值，以此来充当门控信号。计算公式如下：

$$r_t = \sigma(W^r x_t + U^r h_{t-1}) \tag{4-1}$$

$$z_t = \sigma(W^z x_t + U^z h_{t-1}) \tag{4-2}$$

通过门控信号r_t以及z_t，可以计算出当前单元的隐状态h_t。首先需要计算出当前节点的隐藏信息$\tilde{h}$，计算公式如 4.3 所示。其中r_t主要是用来控制需要保存多少之前的记忆，当r_t为 0 时，则$\tilde{h}$只包含当前单元的信息。计算公式如下：

$$\tilde{h} = \tanh(W x_t + r_t U h_{t-1}) \tag{4-3}$$

门控信号z_t主要作用于更新门，通过控制上一个节点的隐状态h_{t-1}以及当前节点信息$\tilde{h}$，得到最终的隐状态h_t。计算公式如下：

$$h_t = z_t \circ h_{t-1} + \tilde{h} \circ (1 - z_t) \tag{4-4}$$

（1） $z_t \circ h_{t-1}$：表示对上一节点的隐状态进行控制，对其中一些不重要的信息进行遗忘，其中z_t的取值范围为 0~1。当门控信号越接近于 1 时，表示保留的数据越多，反之则表示遗忘的越多。

（2）$\tilde{h} \circ (1 - z_t)$：表示对当前节点信息的$\tilde{h}$进行控制，（1-$z_t$）表示对当前节点的信息选择的程度进行度量。保留当前节点的信息越多，则（1-z_t）的数值越大，反之则越小

（3）$h_t = z_t \circ h_{t-1} + \tilde{h} \circ (1 - z_t)$：通过$1 - z_t$与$z_t$的联动，对于传输进来的数据信息以权重为$z_t$进行控制，同时使用权重为$1 - z_t$对当前节点的信息进行弥补。

基于 GRU 的文本特征提取使用 4.2.1 节中提到的 GRU 模型，得到模型的输出向量$h_t (t = 1,2 \ldots n)$，将得到的$h_1 h_2 \ldots h_n$送入池化层得到h_{GRU}。

其结构如图 4.4 所示：

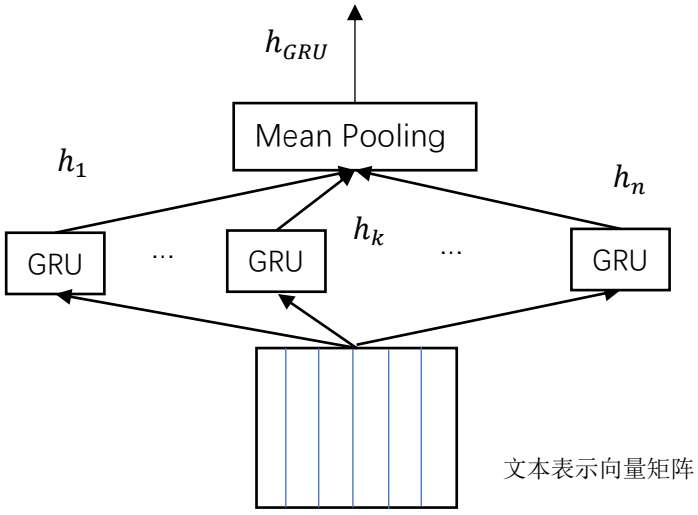


图 4.4 GRU 模型设计图

Figure 4.4 Design of GRU model

③特征聚合层

由特征提取层得到文本特征h_{CNN}以及h_{GRU}。其中h_{CNN}表示 CNN 提取的局部重要文本特征，h_{GRU}则表示循环神经网络 GRU 对于文本全局特征的提取。通过向量拼接的方式对提取的两种文本特征进行聚合，具体表示如下（[:]表示向量拼接）：

$$h = [h_{CNN}:h_{GRU}] \tag{4-5}$$

④输出层

从特征聚合层得到聚合后的文本特征h后，与传统的模型结构一致，将特征h接入全连接层，使用激活函数进行非线性映射成维度大小为 t 的向量（本文的情感倾向只有两种，所以 t 为 2），具体公式如下：

$$H = \tanh\left(w_h \cdot \mathrm{x} + b_h\right) \tag{4-6}$$

其中W_h和b_h为参数矩阵（权重矩阵以及偏置矩阵）。之后使用 softmax 函数计算每一个类别的概率分布，将概率最大的类别作为最终的类别输出：

$$P_i = \frac{e_i}{\sum_{j=1}^{t} \mathrm{e}^j} \tag{4-7}$$

其中e_i表示文本向量 T 中第 i 个元素。

4.3.2 模型训练

本章的所提出的 KMCNN-GRU 模型与第三章的 CNN 模型结构相似，只是特征提取层面略有差异。这里主要阐述一下 GRU 部分超参数的选择以及模型优化算法的使用。

（1）超参数的选择 0

与第三章相同，本章的超参数选择方式也是通过网格搜索进行选择，通过固定

参数多次实验进行参数的调节。分别考虑以下超参数对于实验的影响：①Dropout，取值范围为{0.1，0.3，0.5，0.7}；②训练批次大小，取值范围{2，8，16，32}；③GRU 隐藏神经元个数，取值范围为{32，64，128，256}。

部分超参数选择结果如下：

表 4-1 超参数选择

Table 4-1 Superparameter selections

超参数	大小
Dropout	0.5
训练批次	16
学习率	0.01
GRU 隐藏单元数	128

（2）模型优化

为了避免模型在训练中太早的结束学习，在 GRU 部分采用基于梯度的 RMSprop 算法进行算法优化。不同于 SGD，RMSprop 维护了一个累加状态变量s_t，并对梯度g_t做指数加权平均。具体更新公式如下：

$$s_t \coloneqq \rho s_{t-1} + (1-\rho) g_t \odot g_t \tag{4-8}$$

利用累加状态s_t以及梯度g_t将目标函数自变量中的每个元素按元素运算重新调整，然后更新自变量：

$$x_t \coloneqq x_{t-1} - \frac{\alpha}{\sqrt{s_t + \epsilon}} \odot g_t \tag{4-9}$$

其中ρ为超参数，α为学习率，$g_t \odot g_t$表示指数加权平均。由上式可以看出，当梯度更新至参数空间较为平稳的地方时，当历史梯度平方和较小时 RMSprop 通过调整学习率（全局学习率）使得模型的学习不会过早的进入停止状态。

4.3.3 评价标准

本章评价标准与第三章相同，同样采用准确率（Precision）、召回率（Recall）以及 F 值作为评价标准。

4.3.4 对比实验设计

为了验证 KMCNN-GRU 的有效性，设置以下对比实验：

（1）SVM：采用逆文档频率（TF-IDF）作为特征提取，并将提取的特征送入 SVM 中进行训练并输出相应的文本情感倾向。

（2）Static-CNN：同第三章提到的 CNN 模型一致，使用词汇结合词性向量作为文本表示，维度大小为 100。

（3）GRU：模型结构如图 4.4 所示，使用维度大小为 100 的词向量输入层，

后面接入 GRU 层，隐层神经元个数设置为 128 个。

（4）KMCMM：在基础的 CNN 上进行了改进，结合词性特征丰富文本表示。在此之上，使用 k-max 池化策略代替传统的最大池化策略。

（5） KMCNN-GRU：基于 KMCNN 以及 GRU 对文本特征提取方式的不同，将 KMCNN 以及 GRU 提取的文本特征进行结合，以达到式获取更多的文本特征。

4.4 实验结果与分析

4.4.1 实验结果

实验结果如表 4-3 所示，主要对比 SVM、Static-CNN、KMCNN、GRU 以及 KMCNN-GRU 五种模型在网络短文本中的分类性能。

表 4-3 实验结果

Table 4-3 Results of Experiments

模型	准确率	召回率	F 值 (%)
SVM	77.29	77.98	77.63
Static-CNN	80.15	80.56	80.35
GRU	81.32	79.24	80.26
KMCNN	81.92	80.86	81.48
KMCNN-GRU	82.51	82.74	82.62

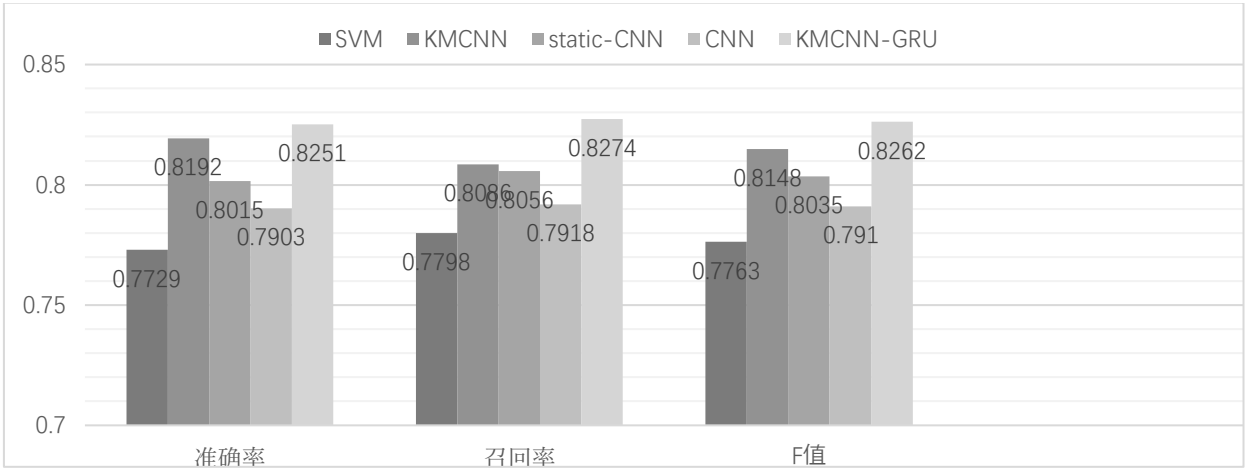


图 4.5 实验结果对比图

Figure 4.5 Comparison of experimental results

4.4.2 实验结果分析

由实验结果可知，基于深度神经网络的分类模型在性能上要优于传统的 SVM，原因推断：CNN 以及 RNN 都是结构化模型，以层次交叠的方式对底层特征进行以此提取，不需要依赖人类的主观意识，能够更好的发掘深层次的特征。卷积神经

网络的特征提取方式优势在于能够捕获句子的局部重要特征，而对于文本序列这类非线性特征的学习需要依赖于长距离的上下文联系，因此 GRU 的分类精度要略优于 CNN。

KMCNN-GRU 相较于传统的循环神经网络模型性能有了一定的提升，虽然在准确率上没有明显增加，但是召回率以及 F 值显然相对于其它模型有了一定的提升，表明使用特征融合的方式能够提升模型的特征提取能力，提高模型的鲁棒性。

4.5 本章小结

CNN 以及 RNN 都是以模块化的结构对底层特征进行层层提取，以此学习到更为抽象的表征形式。不过对于文本序列等非线性结构，需要联系当前词项的上下文语境进行分析，单纯的基于 CNN 进行特征提取很难完成对文本复杂结构进行表示。因此本章在第三章提出的卷积神经网络模型的基础上，提出 KMCNN-GRU 模型。在特征提取层面通过向量拼接的方式将 CNN 以及 GRU 两种特征模型提取的文本向量特征进行融合。同时为了验证该模型的有效性，进行了多组对比试验并分别就不同的评价标准进行分析对比，实验结果验证了该模型的有效性。

第五章 总结与展望

5.1 工作总结

文本情感分类是自然语言处理领域中重要的分支之一，旨在通过建模归类分析出文本数据中所蕴涵的情感倾向，是近些年来研究的热点。早期的文本分类由于数据量较小并且硬件算力的限制，使得人们的研究方向主要集中于字典法与机器学习法，并且取得了一定的成绩。随着互联网的快速发展，文本数据集数量的大幅度递增，使得传统的文本分类方法很难在保持分类精度的同时还能快速处理大量数据。基于深度学习的方法不依赖于人工经验的特征提取，并且随着硬件算力的不断突破，对于大规模的数据处理成为了可能。

通过研究学习发现：字典法虽然简单易于理解，但是对于日益增长的数据量以及网络新词的不断涌现，很难再去构造完备的情感字典，因此字典法的适用范围相对较小。对于机器学习法，其依赖于人工经验对于文本数据的理解，基于概率论或统计学的方式对文本进行特征提取，然后训练相应的分类器进行文本分类。该方法可以将无序的数据转换成相对有用的特征信息，并且可以通过训练一定的数据量去学习到数据内部的浅层统计结构，因此取得了一定的研究成果。但是由于该方法受人为主观意识影响，对数据集的优劣有很强的敏感性，因此在文本分类的过程中鲁棒性不高，很难去学习到数据内部深层次的特征信息，因此该方法只能适用于小数量级的文本分析。基于深度学习的方法是以机器学习为基础，以模块化的神经网络结构层层叠进的学习数据内部的统计结构，以底层特征进行再次提取组合成更抽象的分布式表征形式，从而以简单的神经元实现复杂现实问题的表示，相对于机器学习法能取得更好的分类效果。但是，在现阶段的深度神经网络分类模型中常常缺乏对情感资源的有效利用以及对情感特征的有效提取。

针对以上分析研究，本文在对文本情感分类相关技术以及深度学习相关知识进行学习研究的基础上，提出两种模型 KMCNN 以及 KMCNN-GRU，使用豆瓣影评数据进行多次实验。同时为了验证模型的有效性，设立多个对比实验。

具体工作如下：

（1）利用网络爬虫爬取豆瓣影视评论作为训练语料，同时对文本数据进行清洗去噪，保证语料的整洁性。由于本文是基于监督学习下的文本分类研究，所以需要对文本数据进行标注。并且为了保证标注的准确性，分别进行不同时间的三次标注，最终选取三次结果相同的文本数据作为最终的训练语料。

（2）为了更有效的利用情感资源，使用词性标注的方式，将词性表达与词汇向量相结合作为文本表示。在此基础上，考虑到传统的最大池化方式容易丢失特征信息，使用 k-max 池化方式代替最大池化，提出基于 CNN 的文本情感分类模型 KMCNN。实验证明该模型相较于传统的 CNN 模型有更好的性能表现。此外，本

文还组织多组对比实验来探究词向量维度以及基于预训练词向量对模型分类性能的影响。实验结果表明，在其它条件一致时，使用维度大小为 100 时的文本分类模型性能最好。

（3）考虑到单纯的基于词语级粒度进行特征提取可能无法涵盖更多的文本表征信息，在对不同的深度学习算法模型进行研究分析后，提出融合特征的 KMCNN-GRU 模型，在特征提取层通过向量拼接的方式将 CNN 以及 GRU 两种模型提取的文本特征拼接融合，对特征进行更有效的提取，实验证明该模型在一定程度上能提高模型的鲁棒性。

5.2 未来展望

随着大数据时代的到来，网络短文本数量的暴增，使得文本情感分类领域也吸引了越来越多研究学者的关注。本文以深度神经网络为基础，分别基于卷积神经网络以及循环神经网络的进行文本情感分类研究，基本完成了研究任务并对提出的模型进行了验证。但是仍然还有许多方面值得去探究和优化，主要体现在以下几个方面：

（1）如何更加充分的提取网络短文本的情感特征信息。基于词汇粒度级别的文本特征提取往往难以涵盖有效的情感信息，本文虽然在文本表示层结合了词性特征，并且使用不同的特征提取方式，但是仅仅是处于词汇级别的研究，特征提取方式还有很多的优化空间。

（2）实验中使用的文本数据较少。基于深度神经网络的文本情感分类需要大量的数据来学习其内部的统计结构，因此需要收集更多的数据。此外，模型的分类精度还取决于分词的结果，对于分词方式的优化以及网络新词的覆盖也是后续研究的重点。

（3）实验环境以及算力的支持。在深度神经网络的训练过程中，除了算法优化以及数据的选择之外，还需要硬件算力的支持。模型的容量越大，相对来说更能精确的刻画出数据的表征信息。在以后的学习研究中，可以通过增加网络层数、神经元个数以及训练迭代次数等方式优化模型结构。

致　谢

时光如水，转瞬间三年又过去了，仿佛昨天还是初入校园的少年。人的一生中，总是充满了离别和相遇。即将要离开校园，去往一片新的天地，就仿佛三年前第一次来这里一样。母校带给我太多的成长和回忆，突然要离开，心中还是充满了不舍。想想自己的读书生涯中，从小学到大学，从大学到研究生，一路上走过了风风雨雨，所幸的是，一路上都有老师与同学们的陪伴。如今离别之际，我要感谢在这段人生旅途中一直陪伴我、呵护我的人们。

首先要感谢的是我的导师杜红教授，无论从学习还是生活一直给予我支持和帮助，本篇论文也是在杜老师的细心指导下完成的。杜老师知识渊博，视野开阔，并且有着丰富的项目经验，在论文的完成过程中认真细心的给我提出很多宝贵的意见，鼓励我在研究过程中大胆想象，勇于创新和实践。在杜老师的悉心教导下，我的毕业论文才能有条不紊的顺利完成。

与此同时，还要感谢我的学长康杜，感谢他在学习和生活中对我的帮助。还要感谢我的同学们，他们分别是金亚楠、熊艳湫、余田椿、杨少波、刘辉，在论文的写作过程中，我们一起探讨学习，收到了他们的许多宝贵意见。在研究生的三年生涯中，我们一起学习奋进，在科研的道路上共同成长。

最后还要感谢我的父母，正是有了他们的坚定支持，我才可以顺利完成学业。感谢他们在我成长过程中所付出的努力和心血，总是无条件的支持我、鼓励我，唯有奋勇向前才能不辜负父母的期盼。

在我成长的道路上，感谢一路有你！谢谢！

参 考 文 献

[1] WordNet :https://wordnet.princeton.edu.

[2] HowNet ：http://www.keenage.com/html/c_bulletin_2007.htm.

[3] 情感词汇本体：http: / /ir.dlut.edu.cn /EmotionOntologyDownload.aspx.

[4] Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure. SemanticOrientations of Adjectives[C]// LREC. 2004, 4: 1115-1118.

[5] Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association [J]. Acm Transactions on Information Systems, 2003, 21(4):315-346.

[6] Yang Min,Peng Baolin,Chen Zheng. A Topic Model for Building Fine-Grained Domain-Specific Emotion Lexicon[C]. Association forComputational Linguistics（ACL）,2014:421-426.

[7] Qiu G, Liu B, Bu J, et al. Opinion word expansion and target extraction through doublepropagation[J]. Computational Linguistics, 2011, 37(1):9-27.

[8] 周咏梅, 阳爱民, 杨佳能. 一种新闻评论情感词典的构建方法[J]. 计算机科学, 2014, 41(8):67-69.

[9] 陈晓东. 基于情感词典的中文微博情感倾向分析研究［D］.华中科技大学，2012.

[10] 杨飞,吴颖丹,王鑫颖.基于基础词典扩展的中文酒店评论情感分析[J].湖北工业大学学报,2019,34(01):107-110.

[11] Bo Pang,Lillian Lee,Shivakumar Vaithyanathan. Thumbs up: Sentiment Classification Using Machine Learning Techniques[C]. ACL-02Conference on Empirical Methods in Natural Language Processing,2002:79-86.

[12] Ni X, Xue G R, Ling X, et al. Exploring in the weblog space by detecting informative and affective articles[C]// International World Wide Web Conference. 2007:281-290.

[13] 冯成刚,田大钢.基于机器学习的微博情感分类研究[J].软件导刊,2018,17(06):58-61+66.

[14] 徐健锋,许园,许元辰,张远健,刘清.基于语义理解和机器学习的混合的中文文本情感分类算法框架[J].计算机科学,2015,42(06):61-66.

[15] 王大伟,周志玮,曹红根.基于 PCA-SVM 算法的酒店评论文本情感分析研究[J].现代计算机,2019(21):13-17+49.

[16] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of Twitter data[C].The Workshop on Languages in Social Media. Association for Computational Linguistics, 2013:30-38.
[17] Hinton G E, Osindero S, Teh Y W. A Fast Learning Algorithm for Deep Belief Nets [J]. Neural Computation, 2014, 18(7):1527-1554.
[18] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
[19] 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）：155-161.
[20] 谢博,叶颖雅,陈振彬,黎树俊,陈珂.基于半监督卷积神经网络的文本情感分类[J].广东石油化工学院学报,2018,28(06):31-35.
[21] Tomas Mikolov, Quoc V. Le, Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation[I]. arXiv:1309.4168v1, 2013.
[22] Zhang L, Liu B. Sentiment Analysis and Opinion Mining [J]. Synthesis Lectures on Human Language Technologies, 2016, 30(1):152-153.
[23] 张翠,周茂杰.一种基于 CNN 与双向 LSTM 融合的文本情感分类方法[J].计算机时代,2019(12):38-41.
[24] 华云彬,匡芳君.基于 Scrapy 框架的分布式网络爬虫的研究与实现[J].智能计算机与应用,2018,8(05):46-50.
[25] 毛伟. 互联网资源标识和寻址技术研究[D].中国科学院研究生院（计算技术研究所）,2006.
[26] 王山雨. 面向产品领域的细粒度情感分析技术[D].哈尔滨工业大学,2011.
[27] 苏其龙. 微博新词发现研究[D].哈尔滨工业大学,2013.
[28] 李华栋,贾真,尹红风,杨燕.基于规则的汉语兼类词标注方法[J].计算机应用,2014,34(08):2197-2201.
[29] http://ictclas.nlpir.org/
[30] http://hanlp.linrunsoft.com/
[31] Manning C D, Raghavan P. Introduction to Information Retrieval [M]. 人民邮电出版社, 2010.
[32] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the Acm, 1975, 18(11):613--620.
[33] 徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8):1423-1436.

[34] Bengio Y, Ducharme Réjean, Vincent Pascal. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research. 2000, 3(6):1137-1155.
[35] Wu H C, Luk R W P, Wong K F, et al. Interpreting TF-IDF term weights as making relevance decisions [J]. Acm Transactions on Information Systems, 2008, 26(3):55-59.
[36] Oren Etzioni, Michael Cafarella, Doug Downey, etc. Web-Scale Information Extraction in KnowItAll. In Proceedings of the 13th international conference on World Wide Web, 2004, 100-110.
[37] Choi Y, Kim N, Hwang S, et al. Thermal Image Enhancement using Convolutional Neural Network[C]// Ieee/rsj International Conference on Intelligent Robots and Systems. IEEE, 2016.
[38] Hochreiter S. LSTM can solve hard long time lag problems[C]// International Conference on Neural Information Processing Systems. MIT Press, 1996:473-479.
[39] 吕颖. 基于半监督学习的文本情感分类平台的设计与实现[D].山西大学,2016.
[40] 吴晓芳.张斌语法研究特点回溯[J].文化学刊,2017(02):48-54.
[41] 王文凯. 基于深度神经网络的文本表示及情感分析研究[D].郑州大学,2018.
[42] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4):págs. 212-223.
[43] 刘且根. 基于滤子函数的正则化方法的研究[D].上海交通大学,2009.
[44] 李思明. 基于复合高斯模型的杂波统计分析与建模[D].哈尔滨工业大学,2015.
[45] Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machinetranslation: Encoder-decoder approaches[J]. arXiv preprint arXiv:1409.1259, 2014.

个 人 简 介

王锭，男，汉族，1995 年 5 月 19 日出生于安徽省安庆市。2018 年 6 月通过六级考试;2017 年 6 月毕业于安徽新华学院电子信息工程学院，获工学硕士学位；2017 年 9 月进入长江大学电子信息学院，攻读电子与通信工程专业硕士学位，修完 33 个学分，完成了培养方案上规定的理论知识学习，成绩良好。

2017 年至 2019 年连续三年获长江大学二等学业奖学金。

2018 年 8 月以第一作者的身份在《电脑与信息技术》上发表题名为“基基于深度神经网络的电影评论情感分类研究”论文一篇。

学号：_____

研究生学位论文原创性声明和版权使用授权书

原创性声明

我以诚信声明：本人呈交的硕士学位论文是在 XXX 教授指导下开展研究工作所取得的研究成果。文中关于“……”的结论和关于“……”的结果系本人独立研究得出，关于“……”的结论和关于“……”的结果系本人独立研究得出（或与 XXX 合作研究得出），不包含他人研究成果。所引用他人之思路、方法、观点、认识均已在参考文献中明确标注，所引用他人之数据、图件、资料均已征得所有者同意，并且也有明确标注，对论文的完成提供过帮助的有关人员也已在文中说明并致以谢意。

学位论文作者（签字）：

签字日期：　　年　　月　　日

版权使用授权书

本人呈交的硕士学位论文是本人在长江大学攻读硕士学位期间在导师指导下完成的硕士学位论文，本论文的研究成果归长江大学所有。本人完全了解长江大学关于收集、保存、使用学位论文的规定，即学校有权保留并向国家有关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。本人授权长江大学可以将本学位论文的全部或部分内容编入有关数据库，可以采用影印、缩印、数字化或其它复制手段保存论文，学校也可以公布论文的全部或部分内容。（保密论文在解密后遵守本授权书）

学位论文作者（签字）：

签字日期：　　年　　月　　日