

摘要

自 2019 年 12 月起，新冠肺炎疫情开始爆发并逐渐蔓延，它带来了不计其数的人民生命财产的损失。由此出现了各种程度的害怕、担忧、恐慌等负面情绪，在此背景下舆论引导和维护社会稳定成为了社会治理中的热点话题。基于情感分析技术的舆情发展研究在把控舆情走向的过程中有利于维护社会稳定以及增进国民信心。

本文基于从微博中与疫情相关的 230 个话题中采集的 10 万条微博评论文本数据集对情感分类问题进行研究和探索。在数据预处理阶段本文使用了多种针对评论文本的数据清洗方法对数据进行去噪处理。并通过 jieba 分词包来实现文本语句的分词，接着使用哈工大停用词表对不含有分类信息的文字进行剔除。

在实验部分本文利用 Pytorch 框架基于 Word2Vec、Bert 以及 ERNIE 预训练模型搭建了多种神经网络，结果表明模型中基于双向长短期记忆神经网络搭建的模型表现会更好，并且 ERNIE 针对中文文本的词嵌入效果最好。

基于以上结论本文提出的 ERNIE-RCNN (CLS) 模型以 ERNIE 作为词嵌入模型接入 Bi-LSTM 模型并通过池化层对 Bi-LSTM 的所有隐层状态以及词嵌入向量进行降维，最终通过将池化层的结果与 CLS 对映向量进行拼接后接入全连接层实现情感分类。本文的创新之处在于保留了 CLS 对映向量并与池化层的结果进行拼接。结果表明 ERNIE-RCNN (CLS) 模型能够实现从更多数据信息中提取更多的有效信息，本模型在微博评论文本数据集上的情感分类任务中准确率达到了 74.95%，Macro-F1 值为 0.7163，优于本文中其他模型对比实验的结果。

关键词：自然语言处理、情感分类、新冠疫情、ERNIE 预训练模型、RCNN 神经网络模型

abstract

Since December 2019, the new type of coronal pneumonia epidemic has started to break out and gradually spread, which has caused countless losses of people's lives and property. As a result, negative emotions such as fear, worry and panic have emerged in various degrees. In this context, public opinion guidance and maintenance of social stability have become hot topics in social governance. The research on the development of public opinion based on emotion analysis technology is conducive to maintaining social stability and enhancing national confidence in the process of controlling the trend of public opinion.

This paper studies and explores the issue of emotion classification based on the data set of 100000 weibo comment texts collected from 230 topics related to the epidemic in weibo. In the data preprocessing stage, this paper uses a variety of data cleaning methods for comment text to de-noise the data. And jieba word segmentation package is used to realize the word segmentation of text sentences, and then the words that do not contain classification information are eliminated using the Harbin Institute of Technology's stopword list.

In the experimental part, this paper uses the Pytorch framework to build a variety of neural networks based on Word2Vec, Bert and ERNIE pre-training models. The results show that the model based on Bi-LSTM in the model will perform better, and ERNIE's word embedding effect for Chinese text is the best.

Based on the above conclusions, the ERNIE-RCNN (CLS) model proposed in this paper uses ERNIE as a word embedding model to access the Bi-LSTM model, and reduces the dimension of all hidden layer states and word embedding vectors of Bi-LSTM through pooling layer. Finally, the emotion classification is achieved by splicing the results of pooling layer and CLS mapping vector and accessing the full connection layer. The innovation of this paper is that the CLS mapping vector is retained and spliced with the results of the pooling layer. The results show that the ERNIE-RCNN (CLS) model can extract more effective information from more

data information. The accuracy of this model in the emotion classification task on the Weibo comment text dataset is 74.95%, and the Macro-F1 value is 0.7163, which is better than the results of other models in this paper.

Keywords: natural language processing, emotion classification, COVID-19 epidemic, ERNIE pre training model, RCNN neural network model

目录

1、绪论.....	1
1.1 研究背景和意义.....	1
1.1.1 研究背景.....	1
1.1.2 研究意义.....	3
1.2 研究现状.....	3
1.2.1 基于情感词典的情感分类研究现状.....	4
1.2.2 基于机器学习的情感分类研究现状.....	5
1.2.3 基于深度学习的情感分类研究现状.....	6
1.3 论文主要研究内容.....	8
2、相关理论与技术研究.....	10
2.1 文本预处理.....	10
2.1.1 文本数据清洗.....	10
2.1.2 中文分词.....	10
2.2 词向量转化模型.....	11
2.2.1 Word2Vec 词嵌入模型.....	12
2.2.2 Bert 词向量预训练模型.....	13
2.2.4 ERNIE 中文预训练模型.....	18
2.3 基于深度学习的文本分类模型.....	20
2.3.1 TextCNN 卷积神经网络.....	20
2.3.2 双向长短期记忆神经网络模型.....	22
2.3.3 RCNN 神经网络模型.....	24
3、数据预处理及评价指标.....	26
3.1 数据说明.....	26
3.2 数据预处理.....	27
3.3 评价指标.....	29
4、模型设计与实验.....	31
4.1 ERNIE+RCNN (CLS) 模型结构.....	31
4.1.1 词嵌入层.....	33
4.1.2 RCNN 神经网络层.....	34
4.1.3 输出层.....	34
4.2 模型参数设置.....	35
4.2.1 词嵌入层.....	35
4.2.3 神经网络层.....	36
4.3 模型训练.....	37
4.3.1 基于 Word2Vec 的模型对比试验.....	37
4.3.2 基于 Bert 以及 ERNIE 的模型对比试验.....	38
4.3.3 消融实验.....	40

5、总结与展望.....43

 5.1 研究工作总结.....43

 5.2 不足与展望.....44

参考文献.....45

1.绪论

1.1 研究背景和意义

1.1.1 研究背景

自我国接入互联网以来我国网民数量逐年爆发式增长，根据 2022 年 8 月中国互联网网络信息中心（CNNIC）所发布的《第 50 次中国互联网络发展状况统计报告》显示截至 2022 年 6 月的网民画像数据我国网民数量已达到 10.51 亿人，互联网普及率已达到 74.4%，网民人均每周上网时长为 29.5 个小时，众多网民中使用手机上网的用户比例达 99.6%。这表明民众提供和获取信息的速度更加迅速。一方面由此所带来的便利包括我们可以轻而易举地得到实时信息，我们所传递的观点、情感也可以迅速在网络上扩散。另一方面由此产生的舆论所带来的影响是空前的，如今的互联网中媒体进入门槛低，信息的发布和传递速度非常迅速，负面的舆论谣言等如果任由其在网络上扩散和发酵会对民众产生非常严重的误导和煽动。可以说互联网已经成为人们思想文化的枢纽以及舆情扩散的加速器。

舆情的传统意义是作为民意的基本反映。而伴随着网络的普及，又赋予了舆情新的含义，相较于人们口口相传的议论和流言，网络舆情更易于调查，可以通过直接调取的方式获取原始信息资源，我们作为用户也可以通过浏览、爬虫等方式获取信息。而人们在街头的议论却只能通过调查的方式获取，这里面存在成本高，信息准确率低的缺点。在开放的互联网媒体平台上非理性信息以及具有煽动性的谣言等极易引发公众的对立情绪，由此可能导致一系列的突发事件以及社会矛盾的激化。在我国政府有很多大数据信息部门会对网络舆论情况及导向进行监控和引导。除此之外，对于一些企业而言深度掌握网络舆情方向能够更好地把握市场走向，有利于企业制定发展战略。

在众多网民中 99.6%的用户使用手机上网，并且手机在当今社会作为生活必需品其中的社交平台已经逐渐成为人们生活中所必须的。社交平台作为一种基于互联网的服务，用户们可以在其中建立公共性质的个人页面，用户们通过主页建立连接。根据德本咨询与《互联网周刊》的调查显示 2021 年社区社交平台排行前五分别为：微信、QQ、微博、百度贴吧、知乎。新浪微博自 2009 年 8 月 14 日开始内测起发展至今已作为最受欢迎的社交媒体平台。新浪微博不同于百度等基于门户、搜索的互联网入口，它基于公开的平台架构让用户能够公开发表信息动态，继而通过裂变式传播建立用户之间传播联系的桥梁从而实现信息的实施分享。根据 2022 年 3 月 3 日晚微博发布的 2021 年报表示 2021 年应收达到了 22.2 亿美元，较去年增长了 34%。归母净利润为 4.28 亿美元，事实上新浪微博通过利用用户发布和传导的信息针对用户进行广告投放来实现大部分的营收。在这个流量即资产的时代微博平台占据了很大的优势，根据 2022 年 11 月 17 日发布的 2022 年第三季度报显示微博 9 月的月活跃用户数量已经达到了 5.84 亿，手机端用户占其中的 95%。而全球知名社交平台 Twitter 月活跃用户数量仅接近 3 亿。据此，我们可以认为微博用户分享和传递的信息基本可以代表我国大多数的民意舆情。

自 2019 年 12 月起，新冠肺炎（Corona Virus Disease 2019, COVID-19）疫情开始爆发并逐渐蔓延，它带来了不计其数的人民生命财产的损失。为世界卫生安全带来了前所未有的挑战。由于大量人员伤亡和经济财产损失，广大民众的身心健康受到严重威胁，出现了各种程度的害怕、担忧、恐慌等负面情绪，在此背景下舆论引导和维护社会稳定成为了社会治理中的热点话题。伴随着新冠肺炎疫情在各地的此起彼伏，这引发了不同程度的网络舆论。人们通过社交平台表达惊恐、担忧、乐观等不同倾向的信息。依照文本数据处理以及文本数据的情感分析方法通过各大社交媒体平台的数据可以实现以下研究：溯源舆情发展路径以及现状，探究不同地区在不同程度的疫情下舆情差异，研究舆情发展规律。

情感分析任务作为自然语言处理的子任务在社交媒体以及舆情监测等领域有着非常广泛的应用。当前情感分析任务主要包含词级别情感分析、句子/篇章级情感分析以及目标级情感分析。其中词级别情感分析任务的代表就是通过给词赋予情感信息来构建情感词典。而句子/篇章级情感分析任务主要是

针对文本整体给予正负向的预测。目标级情感分类则将评价目标上升为文本中具体的评价对象，针对其进行多属性的情感分析。

1.1.2 研究意义

习近平总书记指出，“网民来自老百姓，老百姓上了网，民意也就上了网。群众在哪儿，我们的领导干部就要到哪儿去，不然怎么联系群众呢？”“善于运用网络了解民意、开展工作，是新形势下领导干部做好工作的基本功。”在这个疫情完全放开的大环境下存在民众心理恐慌、过量抢购药品等情况。正确引导网络舆情走向从而稳定民众情绪已经成为当下热点话题，当面对重大突发事件时网络部门需要在最短的时间内披露事实从而以正视听进而让谣言传播的范围压缩到最小。做好网络舆情控制便是完善新时代意识形态工作，同时也是完善走好网上群众路线工作，更是凝聚发展共识的途径。

针对社交媒体平台上获得的新冠肺炎疫情过程中的舆情文本信息分析民众的情感倾向，可以帮助相关平台和部门了解民众的实时情绪，进而制定相关政策从而达到稳定民众情绪、维护社会稳定的目的。

由于网络评论的随意性、自由性，微博评论文本数据长度差异较大会产生传统表现较好的自然语言处理深度学习模型效果表现较差，用词新颖可能会导致传统的分词程序效率降低，表情符号、标点符号表情、重复符号的应用为数据带来了更多噪声。基于以上特点本文通过在大多数主流自然语言处理模型上进行实验来选出最适合此类任务的模型。能够为以后处理多噪声评论文本提供参考。

1.2 研究现状

情感分类技术作为一类很特殊的文本分类任务在很多领域拥有很大的应用需求和一定的商业价值，已经转变为自然语言处理领域的热点问题。从最初的通过人阅读打标进行分类，到通过构建情感词典以及基于文本中的关键字和词语进行匹配而后通过一定的计算规则来进行分类，再到后来机器学习方法的出现，通过一些简单的转化将文本转化为数值型数据，进而借助机

器学习模型进行分类。近年来深度学习技术的提出和不断发展伴随着计算机计算能力的不断提升, 诞生出诸多各式各样的优秀的模型, 为传统的文本分类问题提供新思路 and 方向。按照情感分类的粒度不同, 情感分类任务又有二元、三元以及多元情感分类之分。其中二元只考虑正向和负向的情感极性, 三元主要分为积极、中性以及消极的情感极性。事实上较粗的情感粒度在很多情感分类场景中应用价值较低, 在很多情况下会有更细的情感粒度划分, 其中最经典的是 Ekman 于 1972 年提出的人的基本情感, 主要包括愤怒、厌恶、恐惧、快乐、悲伤、惊讶六种人类情绪。

1.2.1 基于情感词典的情感分类研究现状

通过构建情感词典实现情感分类是早期情感分类领域较为常用的方法。基于情感词典的情感分类问题先检索情感词典并利用适应的计算规则得出目标文本的情感极性数值, 然后依照一定的情感阈值来实现情感分类。该方法的核心在于搭建完备详尽的情感词典和制定与目标任务相适应的情感值计算规则。英文情感词典中比较主流的是 Word Net, Hu^[1]等人在 Word Net 的基础之上, 利用类似单词之间的相似关系即同义词和反义词关系对已有的情感词典进行了补充, 使样本中能提取的信息更加详细。Turney^[2]提出了点间互信息法, 基于《Word Net》词典中形容词的正向和负向的相似度来进行评论分类, 主要是将一个切分的文本单位的情感倾向计算为该短语与正向词之间的相似度减去这个文本单位与负向字之间的相似度, 最后通过文本的平均情感倾向来对文本进行分类。Kamps 等人^[3]基于 Turney 的方法, 通过计算文本语料中词语与情感词典中基准词之间的相似度, 完成文本的情感分类。

国内文本情感分类研究任务中情感词典方法开始的时间相对较晚, 但经过国内学者不断的探索下迅速发展。林鸿飞教授等^[4]构建了第一个中文文本的情感词语本体库, 为中文文本基于情感词典的情感分类的快速发展奠定了基础。至今比较热门的中文文本情感词典包括知网的情感分析用词语集 Hownet、台湾大学的简体中文情感极性词典 NTUSD、清华中文褒贬义词典以及大连理工大学开发的中文情感本体库 DLLG 等。孙艳等^[5]针对一些标注样本不易获取的问题通过将 LDA 模型与情感模型相结合的形式即无监督 NTSU 模型, 利用采样标

签的形式基于主题情感词进行分类。栗雨晴等人^[6]（2016）则建立双语多类情感词典来处理文本中同时存在中文和英文的情况通过双语多类的极性分析实现文本分类。迄今为止情感词典方法经过了广泛的应用和发展，但伴随着文本数据的多样化它的弊端越来越明显。

第一，所构建的情感词典的质量对情感分类结果影响很大；其次，随着文本多样化，很多新词语的涌现、旧词语出现频率的降低以及旧词新义的出现会导致情感词典会有很高的维护成本；最后，词汇的多种情绪语义倾向，也会降低情感词典的效率。

1.2.2 基于机器学习的情感分类研究现状

使用机器学习模型进行文本情感分类的方法是通过将文本切分为最小单元，再通过提取特征的方式对单元进行向量化，最后利用机器学习算法进行情感分类。这相对于情感词典方法的优势在于它以算法的形式将文本上下文信息加以利用来大幅提升准确率并降低了人工成本。最常见的词的特征表示方法为词袋模型（Bag Of Words）它假设词语是独立存在的，它只考虑词出现的频率而忽略了上下文的联系信息，基于 one-hot 编码形式的词向量又常常会出现维度灾难、数据过于稀疏导致训练不收敛的问题。常见的机器学习算法包括朴素贝叶斯、最大熵法、支持向量机、条件随机场、EM 算法等。Pang 和 Li 等人^[7]使用词袋特征转化对电影评论文本数据进行转化并使用最大熵模型、朴素贝叶斯、支持向量机等机器学习算法进行情感分类，发现一元词特征下 SVM 分类模型效果最好。Wang 等人^[8]提出结合了朴素贝叶斯和 SVM 的优势的 NBSVM 模型，通过朴素贝叶斯来提取文本特征，再将文本特征向量通过 SVM 模型实现情感分类。姬东鸿等人^[9]通过将多种词性的特征相组合来完成特征提取接着使用支持向量机和条件随机场进行分类，结果发现不同的模型对不同特征的适配度也不同。事实上在模型相对成熟的情况下整个模型在应用中的性能会很大程度上依赖特征提取的效率。故接下来对于如何提取文本信息中的有效信息而非噪声成为了文本分类问题中研究的焦点。Kim 等人^[10]对评论文本进行分类时引入了位置特征，结合评价词等语义特征，结果发现一元词特征即 unigram 只在样本集较小时会有优势。刘康等人^[11]为解决过量标记

导致的分类准确率低的问题通过层叠式 CRFs 模型通过多个 CRF 模型从粗到细分布式判断极性方向及强度并探究句子文本二元分类的类别以及分类强度的之间的层级冗余关系。以上这些监督学习都会存在一定的缺陷，人们面对海量的文本数据时需要花费很大的人工成本来打标和维护；而由于人们的主观性，会导致打标质量的良莠不齐，这些都会影响最终的实验结果。为降低人工成本、提升分类效率，半监督学习和无监督学习逐渐受到研究者们的关注。李宇峰^[12]通过半监督多标记支持向量机验证了半监督学习可以通过少量有标注数据以及大量无标注数据显著提升分类器的泛化能力。

机器学习方法摆脱了对情感词典的依赖，它的特点在于模型结构简单、运算快、可解释性强。随着计算机硬件系统不断升级，计算机算力的不断提升催生了很多能够产生更高准确率的复杂模型，深度学习应运而生。

1.2.3 基于深度学习的情感分类研究现状

基于深度学习实现文本分类问题主要分为词嵌入即词的向量化和神经网络分类器来实现文本分类。词嵌入机制往往是通过计算词语在多个语义维度上的概率而组成的向量，这里的向量往往是低维稠密的，这使得 one-hot 编码导致的维数爆炸问题得到了解决，在模型的训练中会更容易收敛。比较常见的词嵌入方法主要包括 Word2vec 里面的 CBOW 以及 Skip-gram 以及 Global Vectors 等，随着深度学习的发展，强大的 Bert 和更适用于中文的 Ernie 预训练模型应运而生。Word2vec^[13]是由 Google 公司于 2013 年提出的，它通过双层神经网络结构来实现词语的向量化，并且可以实现词向量之间的距离可以反映词语之间的语义距离。Word2vec 一经提出便得到了研究者的广泛应用。2014 年 Pennington^[14]进一步提出 GloVe，他利用了文本语料数据集中的全局信息，并通过最小二乘法设定了新的目标函数，但是它还是没有解决 Word2vec 中一词多义的问题。2018 年 Peters 等人^[15]提出了 ELMo (Embedding from language models)，它使用双向 LSTM 进行特征提取，它采用了预训练和微调相结合的形式。它首先通过预训练对每个词语生成三个词嵌入特征，接着通过微调来进一步训练，最后通过调整这三个词嵌入特征的权重来实现解决一词多义的问题。而 BERT^[16]的问世使得词嵌入部分性能得到了前所未有的

的提升, Bert 也采用了预训练和微调相结合的形式。在预训练时通过大量无标签的文本数据对 Bert 进行预训练, 之后又通过少量特定任务场景中的有标签数据进行微调。在实际模型使用的过程中使用预训练的模型结构和模型参数, 这样就大大降低了对特定任务场景中的有标签数据量的需求。在 Bert 大放异彩的时期各种优秀的变体模型不断涌现, 其中百度于 2019 年发布的 Ernie 预训练模型在中文文本预训练任务中效率卓越, 它受到 Masking 方法的启发, 通过运用知识掩盖策略来对文本中的信息(汉字、短语以及命名实体等)进行随机掩盖。

在文本分类任务中使用最为广泛的神经网络主要是卷积神经网络(Convolutional Neural Networks)、长短时记忆网络(Long Short-term Memory Networks)以及 Transformer 等。卷积神经网络(CNN)是一种包括输入层、卷积层、池化层、全连接层以及输出层的神经网络。Kim^[17]通过在英文文本分类问题上使用卷积神经网络, 它通过在一维上横向滑动卷积核来提取局部特征信息, 他发现卷积神经网络的分类效率要高于其他神经网络。在此基础上 Zhang 等人^[18]对句子结构进行了改进, 他们发现以句子矩阵形式进行排列的信息分类效果更好。何炎样等人^[19]通过将文本中表情符号信息的情感空间映射与 MCNN 相结合构建了 EMCNN 模型, 增强了模型捕捉语义的能力。而 DCNN^[20]则是通过迭代多层池化层和卷积层, 通过池化和卷积操作实现降维, 使得句子中相距很远的词语能够产生交互信息。循环神经网络(Recurrent Neural Network)能够通过定向循环对信息进行保留和传递, 从而捕捉前文的信息, 故更加适合去处理序列信息结构。Hochreiter^[21]发现循环神经网络在处理长序列信息时在训练的过程中会存在梯度消失和梯度爆炸的问题。而 LSTM 和 GRU 则是在 RNN 的基础上添加了门控机制, 其中 LSTM 是通过三个门控单元来实现信息的舍弃、保留和传递, 从而克服了 RNN 里梯度消失和梯度爆炸的问题。而门控循环单元(Gate Recurrent Unit)则是在保留 LSTM 的基本结构和功能的基础上进行了简化, 使得模型只有两个门控单元。事实上无论是 LSTM 还是 RNN 都只能读取和保留一个方向上的信息, 为使得对信息特征的充分提取, Schuster 等人^[22]提出了双向循环神经网络(Bidirectional Recurrent Neural Network)来提取前向信息和后向信息。Huang 等人^[23]则首次通过拼接双向 LSTM 与条件随机场大幅改善了命名实体识别的效果。Zhang

等人^[24]则通过双向 LSTM 实现了产品评价的中文文本分类,事实证明双向 LSTM 在情感分类的效果要远高于单向 LSTM 模型。Rajabi 等人^[25]则通过拼接双向 LSTM 和 CNN 组成了多通道的 Bi-LSTM-CNN 模型在进行多标签文本分类任务中取得了不错的分类效果。事实上这样的复合拼接模型的分类效果普遍会强于原始的 LSTM 与 CNN,但是这些模型都不能自动学习到文本信息中对分类起决定作用的关键信息,换句话说更为理想的模型可以赋予关键信息更高的权重。因此,越来越多的文本分类任务中加入了注意力机制。吴小华等人^[26]通过对短文本字向量化后在双向 LSTM 模型中融入了自注意力机制,通过动态调整特征权重的方式实现了情感分类准确率的提升。

基于深度学习的情感分类一直备受研究者的关注,目前文本分类任务中表现最好的模型均是各种各样的深度学习模型。深度学习模型通过模型参数内部的架构实现了文本特征的自动提取,研究者们也在不断探索和研究更为优秀的深度学习算法。事实上深度学习模型的可解释性和可视化一直是研究者们尚未解决的问题,在这其中还有很多值得研究的地方。随着深度学习神经网络复杂度的提升,对计算能力更高的计算机依赖程度也越来越大。

1.3 论文主要研究内容

本文以新冠疫情爆发之初(2020.1.1至2020.2.20)网民们在新浪微博上疫情相关词条下的评论文本为研究对象,以文本情感倾向性分类为目标进行探究,本文主要通过以下两个方面进行探究:

(1)本文通过借助多种深度学习模型进行微博评论文本的情感极性分类。首先,通过 Word2vec 预训练模型或者随机初始化对文本字符进行向量化,针对词级切分以及字符级切分分别进行词嵌入,从而探究最适合微博评论文本的情感分类任务的词嵌入方式。接着通过 TextCNN、双向 LSTM 模型进行初步分类预测。通过多次实验探究得出适应评论短文本的分类效果最佳的模型结构。接着通过构建 DCNN 解决普通卷积神经网络无法对距离较远的词中间产生联系的缺点。构建双向 LSTM+CNN 的拼接模型加大对文本特征提取效率,通过使用池化层进行降维解决普通双向 LSTM 对较长的文本时无法学习关键信息的问题。

(2) 本文通过使用 Bert 以及 Ernie 作为预训练模型实现词嵌入来提升微博评论文本的情感极性分类精度。通过构建 BERT+FC、BERT+Bi-LSTM 、BERT+TextCNN、BERT+RCNN、以及 BERT+DPCNN 进行初步训练，接着通过 ERNIE+FC、ERNIE+Bi-LSTM 、ERNIE+TextCNN、ERNIE+RCNN、以及 ERNIE+DPCNN 探究结合知识掩盖策略的预训练模型在高噪声的评论短文本分类中效果提升是否显著。最终将 ERNIE 与 RCNN 以及 CLS 向量信息相结合构建 ERNIE+ RCNN (CLS) 模型。本文通过大量实验对比不同模型对微博评论短文本的情感分类任务的分类效果，探究最适合此类任务的模型结构，并通过消融实验来验证模型每部分的效果。

2.相关理论与技术研究

2.1 文本预处理

文本表示主要分为文本预处理以及文本向量化两个部分。文本预处理主要包括数据清洗和分词两个部分。由于原始文本中存在很多噪声，去除这些干扰性信息有助于提升输入模型数据的质量从而提升模型的分类效果。

2.1.1 文本数据清洗

事实上根据数据类型以及数据不同数据清洗的方法也不尽相同。针对微博评论的文本数据主要存在以下几种文本数据清洗的方法：微博评论文本中通常会存在“@+用户名”、“#+词条名”、“网页链接”以及 URL 等对文本分类没有帮助的噪声信息。在清洗过程中我们要对这一类信息进行剔除。微博评论文本中通常会存在 emoji 表情等富含信息但不属于传统文本范畴的表情符号，针对这一类表情符号需要通过对应的表情-文本库进行替换。除此之外微博评论数据中还存在一定的繁体字、英文评论，针对这一类文本信息会通过翻译转化为简体字文本。文本数据在文本切分后也存在一些不含有分类信息的文字，例如“嗯”、“哦”、“噢”等语气词，还有“因此”、“于是”等这样的连词，针对此类信息通常会使用中文停用词表、哈工大停用词表、百度停用词表、机器智能实验室停用词库等中文文本常用的停用词表来进行剔除。

2.1.2 中文分词

在自然语言处理中为了能使得文本数据合理地被转化为向量，对文本进行切分是必不可少的。针对文本的切分主要分为字级切分以及词级切分两种方

式。字级分词就是针对每个字符进行切分，这个的优点是操作简单，但是会丧失很多词语级别的语义信息。而词级切分就涉及到中文分词的内容，事实上中文分词本质上是一种划分词边界的方法，主流的中文分词方法有 n 元语法分词、最短路径分词、循环神经网络分词、由字构词分词、Transformer 分词等。本文主要对 jieba 分词工具进行了尝试和运用。Jieba 分词工具是通过由字构词的方法使用基于隐马尔可夫模型的维特比算法。由字构词的主要思想是每一个字在词中都会有一个特定的构词位。维特比算法主要用来解决动态规划中的寻找最优情况的问题，主要是基于局部最优寻找全局最优，通过条件概率计算寻找所有路径中具有最大概率的路径。除此之外，隐马尔可夫模型可以识别未出现过的新词，从而解决 OOV 问题。在使用过程中 Jieba 是基于 python 开发的并已经开源，在调取和使用上都非常方便。

2.2 词向量转化模型

在文本预处理后往往需要进行词向量的转化任务。训练词向量转化模型的过程实际上也是实现文本信息转化的过程。词向量表示主要分为离散表示和分布式表示。其中离散表示方法主要有 one-hot 编码、词袋模型等。One-hot 编码又称为独热编码，它所生成的词向量维度与词典大小相同。大小为 n 的词表中第 i 个词通过此模型生成的向量为长度为 n 并且在第 i 位为 1 其他位均为 0。这样的词向量转化方式简单，但缺点在于它只包含了词频信息，并不能涵盖语义信息，并且由于词向量维数过高会产生维度灾难的问题。

词的分布式表示方法的理论基础是对于上下文语义环境相似的词语它们的语义也要相近。事实上分布式表示对词进行向量转化的本质在于词的语义由它的上下文决定。故所有文本的分布式表示方法主要区别在于表述上下文环境的方法不同以及描述上下文环境与目标词之间的关系的模型不同。根据建模方法的不同词的分布式表示方法主要分为基于聚类的分布表示、基于神经网络的分布表示以及基于矩阵的分布表示这三种。其中比较常用的是基于神经网络的分布表示，也称为词嵌入。词嵌入通过用稠密低维的向量对词语进行语义表示，既解决了独热编码维度灾难的问题又包含了语义信息。比较常见的词嵌入方法包括 Word2Vec、Elmo、GloVe 以及 Bert 等。

2.2.1 Word2Vec 词嵌入模型

Word2Vec 是 2013 年 Google 团队提出的一种词嵌入模型并开源了模型训练工具包。它的主要思想是根据文本词语来预测该词的上下文环境以及根据目标词的上下文环境来预测目标词。所以词嵌入向量只是模型的副产物之一。通过模型训练可以得到一个由词转化为向量的矩阵，每个词都有其对应的向量。词义相近的词语被映射到向量空间中的位置距离也相近，而反义词等词语关系被反映到向量空间中的位置关系也会较远。所以 Word2Vec 模型可以很好地反映词语与词语之间的相似关系。Word2Vec 模型主要包含了两种模型结构：连续词袋（CBOW）模型以及跳字（skip-gram）模型。Word2Vec 模型的训练方式有两种优化方法，分别是 softmax 以及 negative sampling。

连续词袋（CBOW）模型的本质在于根据词语的上下文环境来预测这个目标词语。连续词袋模型主要包含三层神经网络结构，分别是输入层、隐层以及输出层。它的输入为上下文词语在词典中的独热编码，这里假设词典大小为 V ，上下文的单词个数为 C ，故输入为 C 个大小为 V 维的稀疏向量。假设隐层大小为 N 维，于是输入层到隐层的权重矩阵为 $W_{V \times N}$ 。将 C 个大小为 V 维的稀疏向量分别与共享的权重矩阵 $W_{V \times N}$ 相乘得到的 C 个 N 维向量相加求平均得到隐层向量。由于要对目标词进行预测，故输出层的形式应为一个 V 维向量。由隐层转化为输出层的权重矩阵为 $W'_{N \times V}$ 。将隐层向量与权重矩阵为 $W'_{N \times V}$ 相乘并通过激活函数得到输出层 y 。 y 中最大的数字所处的位置对应的单词为模型对目标词所作出的预测。通过对 y 与真实目标词的独热编码进行比较得到的损失函数作为模型的损失函数，这里的损失函数常用交叉熵函数。最后在训练模型的过程中通过梯度下降的方式对两个权重矩阵 $W_{V \times N}$ 与 $W'_{N \times V}$ 进行更新。在训练结束后通过将每个单词的独热编码向量与训练得到的权重矩阵 $W_{V \times N}$ 相乘就能得到每个词语的词嵌入向量。

跳字（skip-gram）模型的本质在于根据目标词语来预测这个词语的上下文环境。跳字模型主要包含三层神经网络结构，分别是输入层、隐层以及输出层。它会根据参数设置选取目标词前后固定数量的词语作为目标词的上下文环境。它的输入为目标词在词典中的独热编码，这里假设词典大小为 V ，故输入为大小为 V 维的向量。假设隐层大小为 N 维，于是输入层到隐层的权重矩阵

为 $W_{V \times N}$ 。将 V 维的输入向量与权重矩阵 $W_{V \times N}$ 相乘得到大小为 N 维的隐层向量。由于要对上下文环境进行预测，故输出层的形式应为一个 V 维向量。由隐层转化为输出层的共享的权重矩阵为 $W'_{N \times V}$ 。将隐层向量与共享权重矩阵 $W'_{N \times V}$ 相乘并通过激活函数得到输出层 y 。 y 中数字代表这个数字所处的位置对应的单词为模型对上下文环境中的词的概率。通过对 y 与真实上下文词语的独热编码进行比较得到的损失函数作为模型的损失函数，这里的损失函数常用多个上下文词语交叉熵函数求和的结果。最后在训练模型的过程中通过梯度下降的方式对两个权重矩阵 $W_{V \times N}$ 与 $W'_{N \times V}$ 进行更新。在训练结束后通过将每个单词的独热编码向量与训练得到的权重矩阵 $W_{V \times N}$ 相乘就能得到每个词语的词嵌入向量。

2.2.2 Bert 词向量预训练模型

Bert 作为 Google 团队于 2018 年发布的自然语言处理领域最具影响力的论文之一。它作为一个动态文本预训练模型主要由 Transformer 的编码器部分组成。相对于 Word2Vec 的连续词袋模型以及跳字模型这样的静态词嵌入预训练模型，Bert 通过使用多层的双向 Transformer 编码结构在并行化训练的情况下能够更完善地解决一词多义的问题，在进行词嵌入时对语义的表达更加完善。事实上 Transformer 的编码器部分主要由自注意力机制和前馈型神经网络组成。Bert 模型结构如图 2-1 所示，E 层为输入层，T 层为输出层，编码器网络为多个 Transformer 编码结构堆叠而成。其中 E 层输入 $\{E_1, E_2, E_3, \dots\}$ 表示文本的输入 E_i 代表第 i 个位置上的文本单元，这里的文本单元为单个字以及一些额外的标识符。T 层输出 $\{T_1, T_2, T_3, \dots\}$ 表示每个输入文本单元对应的词向量。原始的 Bert 一共有两个版本，分别是 Bert-Large 和 Bert-base，本文的预训练模型使用了 Bert-base 架构，下文对模型结构的介绍会以 Bert-base 为准。Bert-base 的编码器部分主要包含了 12 层 Transformer 编码结构。在使用上我们可以基于 Bert 预训练模型构建网络，针对我们的任务进行训练时由于 Bert 预训练模型的参数只是通过预训练参数文件进行了初始化，我们会使 Bert 预训练模型的参数也参与训练进行一定程度上的微调。

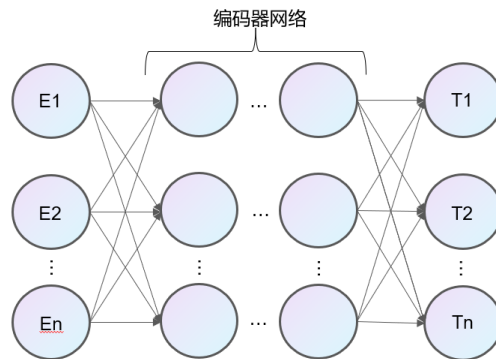


图 2-1 Bert 模型结构图

Bert 的输入层为词向量编码层，如图 2-2 所示，不同于 Word2Vec 中使用独热编码作为模型的输入，Bert 通过将三种包含不同信息的 Embedding 进行相加得到最终的输入。第一类 Embedding 为包含词义的 token embedding，这里不仅包含了输入文本，还额外加入了<CLS>、<SEP>标识符。<CLS>标识符会放在每个样本的第一位，<CLS>标识符在最后一层的输出向量包含了整个样本的语义信息，通常可以接入下游的文本分类等利用整个样本语义信息的模块。不同于文本中别的词所对应的输出向量，<CLS>标识符作为无语义信息的标识符可以无倾向地融合其他文本中的语义信息。在同一个样本中句子与句子间会加入<SEP>标识符用来区分句子。常见的输入的形式为“<CLS>标识+句子 A+<SEP>标识+句子 B+<SEP>标识”。第二类 Embedding 为 segment embedding，它主要包含了不同句子之间的位置信息。第三类 Embedding 为 position embedding，它主要包含了每个字之间的位置信息，可以解决多头注意力机制不能提取位置顺序信息的问题。

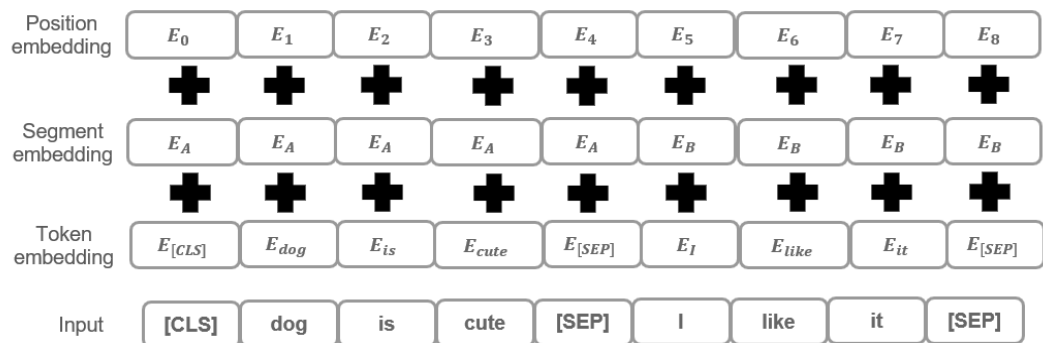


图 2-2 Bert 词向量编码层

Bert 的编码结构部分与 Transformer 编码结构大致相同。Transformer 编码结构主要由两部分组成，分别是多头注意力机制以及前馈型神经网络，如图 2-3 所示。这里的核心是自注意力机制，它的本质是通过某种运算直接计算编码过程中每个位置的注意权值，然后以权和的形式计算整个句子的隐向量表示。所谓多头注意力机制，实际上是对原始输入序列多组的自注意力结果进行加工。然后，每一组自注意力结果又被拼接在一起，得到最终多头注意力的输出，Bert-base 中将并行自注意力数目设置为 12。具体为针对同一个文本生成的 V 、 K 、 Q 自注意力矩阵通过线性变换成 12 份 v 、 k 、 q 矩阵，再独立地对其进行多个自注意力机制的运算，对得到的多个输出矩阵进行拼接，得到的拼接矩阵与一个权重矩阵 W 相乘得到最终的多头注意力模块的结果。运算步骤图如图 2-4 所示。通过矩阵运算为： $Multihead(Q, K, V) = (b_1, b_2, b_3 \dots b_{12})W$ 其中 $b_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ 。

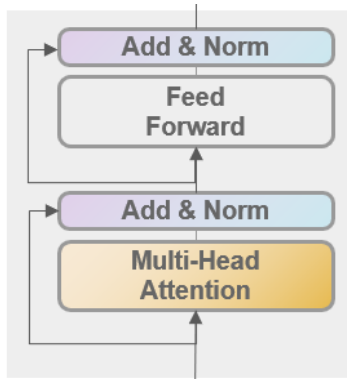


图 2-3 Bert 编码结构图

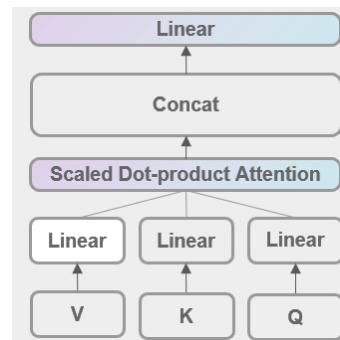


图 2-4 多头自注意力机制运算步骤

Bert 编码结构中的前馈型神经网络为两层线性层，中间包含了一个 Relu 激活函数以及一层 dropout 机制，这里前馈型神经网络的作用是加深我们的网络结构。在多头自注意力机制以及前馈型神经网络后均应用了层标准化、残差连接以及随机失活。其中层标准化是在词嵌入的维度上进行标准化处理，这是为了防止当底层网络在训练过程中分布发生变化时高层网络会由于训练而产生适应性的移位，由此会减缓模型收敛的速度。而残差连接是将本模块的输入和输出相加作为下一个模块的输入，它的思想是当我们的网络要远远深于最优网络结构时会多出很多层多余的神经网络结构，残差连接有助于促

进这些多余的层结构成为恒等映射。在 Bert 预训练模型中存在 12 层多头注意力机制的堆叠，很容易由于网络过于复杂而产生模型退化的问题，对于神经网络来说学习 $F(x) = x$ 的难度要远高于学习 $F(x) = 0$ ，故残差连接可以帮助避免这部分问题。在训练过程中深层次网络由于参数初始化往往接近于 0 会使得在训练过程中在进行梯度传递时通过链式相乘得到的梯度接近于 0，会在训练初期发生梯度消失的问题，而残差连接对于下个模块的输入为 $F(x) + x$ ，会使得每个模块的梯度都加一，从而避免了梯度消失的问题。而随机失活主要是针对模型参数过多而样本较少导致的模型过拟合的问题。过拟合是指往往参数较多的模型在训练后期会出现在训练集上表现过好但在测试集上表现很差。在训练前我们需要指定 dropout 的超参数 P ， P 的取值范围在 0 到 1 之间，dropout 会使得模型参数在概率为 P 情况下设为 0，使得神经元在 P 概率下停止工作。Dropout 函数如下所示：
$$y = \begin{cases} \frac{x}{1-p} & r > P \\ 0 & r \leq P \end{cases}$$
 r 服从 0 到 1 的均匀分布， y 表示模块输出值。一般 dropout 只在训练中使用。

由于 Bert 具有无监督学习的优点，这在预训练阶段大幅节省了训练样本的成本。Bert 在预训练任务中定义了两类任务从而实现模型对上下文语义的准确表述，分别是 Masked Language Model (MLM) 掩码语言模型以及 Next Sentence Prediction (NSP) 邻句检索模型。

掩码语言模型的本质在于它会随机掩盖或替换输入文本中 15% 的字然后模型会通过剩下的文本信息去预测被掩盖或替换的部分。这样会产生一个问题，由于有些字可能出现的频率比较少又被随机掩盖了，在后期微调过程中模型可能会遇到没有见过的词，为解决此类问题，针对这 15% 的字进行了以下处理：其中 80% 会使用 <MASK> 标识进行替代，10% 会将目标字用随机的字来替换，10% 保持不变。这样的处理也可以防止模型仅仅学习到 <MASK> 标识的分布而非上下文语义信息，与此同时保持不变的目标字也保证了模型不会学习到错误的语义信息。10% 的随机替换成其他字会对样本添加不超过 2% 的噪声，从而提升模型的性能。通过掩码语言模型的预训练，Bert 词嵌入能够捕捉到一词多义的语义信息。

邻句检索模型主要是判断两个句子是否为邻句，通过使模型训练对句子关系的判断来提升模型对语义的理解以及对整句话的分类表征。模型会选取句

子对进行判断，通常设置 50% 为原本正确的邻句，剩下 50% 为从所有样本中随机选取的非目标句相邻的句子。由于只需要判断句子与句子之间的关系，所以这里的信息需求仅限于句子的全局信息。在预训练时会利用<CLS>标识在最后一层的输出向量对邻句检索模型进行训练。通过邻句检索模型的预训练，Bert 对于<CLS>标识的输出向量能够更好得融合整个句子的信息。事实上这是模型对于句子的相似度进行了学习和训练，所以 Bert 预训练模型在后续微调时可以在自然语言推理等任务中有着优异的表现。

在文本分类中我们往往需要整个文本的全局信息，而传统的神经网络递归神经网络在进行文本分类任务时会出现一些弊端。如使用基于 LSTM 或者 RNN 的神经网络在进行文本分类时通常会基于最后时刻的输出，这样的序列输出会存在一定的顺序性。模型对序列的感知对于文本中的感知也会因位置的先后而不同，继而无法实现平行化感知。由此我们需要应用注意力机制来平衡不同时间步文本单元的权重，通过训练获得更适合分类的权重。具体将注意力机制实施在 LSTM 网络中我们会在每个时间步后的输出进行提取并接入权重矩阵，而不是单纯地使用最后时刻的输出。

近年来自注意力(self-attention)机制被广泛使用于神经网络中。Bert 内部编码器网络结构中通过将多个自注意力模块拼接组成了多头注意力机制实现了优异的词嵌入效果。自注意力机制首先将文本中的字向量 x_i 与一个权重矩阵 W 相乘得到输入向量 ai 。接着将 ai 与三个权重矩阵 Q 、 K 、 V 相乘得到这个字的查询向量 q 、键向量 k 以及值向量 v 。其中查询向量 q (query)主要是用来与其他字对应向量进行匹配，键向量 k (key)主要是被用来与目标字对应向量进行匹配，值向量 v (value)则代表每个字包含的语义信息。接着计算这个字的查询向量与其他字的键向量进行点积运算计算注意力值，接着对这些点积都除以 $\sqrt{d_k}$ (d_k 是序列向量的维数)进行缩放，并通过 softmax 函数进行归一化得到注意力权重 $\alpha_{ij} = \text{softmax}(\frac{q_i k_j}{\sqrt{d_k}})$ 。最后根据计算得到的权重对值向量进行加权，最终得到目标字在通过自注意力机制后的向量。经过自注意力模块运算后得到的向量会融合目标字与其上下文环境中的其他字的相关度，使得模型对字的理解会结合上下文的语义。具体运算过程如图 2-5 所示：

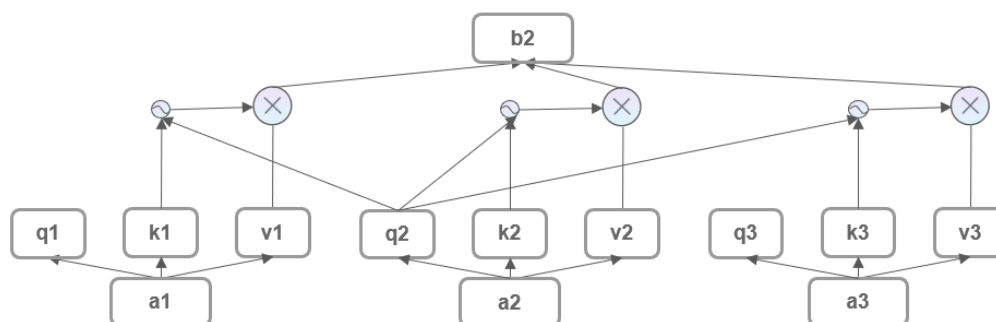


图 2-5 自注意力机制运算过程

2.2.4 ERNIE 中文预训练模型

百度公司于 2019 年提出了 ERNIE(Enhanced language Representation with informative Entities)知识增强的语义理解框架。它在中文文本自然语言处理任务中的表现要优于 Google 的 Bert 以及 XLNet。ERNIE 在预训练过程中利用了更多的文本语义知识也丰富了训练任务，使得词嵌入变得更加高效。在基于中文文本的分辨语义相似度、文本分类、命名实体识别等任务中均优于 Bert 预训练模型。

GLUE 是由华盛顿大学、纽约大学和 Deepmind 工作者提出的通用语言理解评估平台，它可以用来评估模型在九种已有自然语言理解任务中的性能，涵盖了自然语言推理任务、单句任务、释义任务等。GLUE 基于模型在这九个任务上的平均得分进行排名。其中 ERNIE 以 9.1 分的得分排到了第五名的位置。尽管评估语言是英文，不是 ERNIE 擅长的中文文本自然语言处理领域，仍然可以反映出 ERNIE 预训练模型强大的性能。

ERNIE 在模型结构上与 Bert 基本相同，ERNIE 在随机遮盖字的机制上做出了改进。ERNIE 在随机掩盖文本样本时将文本语言知识划分为三类，分别是单词级掩盖、词组级掩盖以及命名实体级掩盖。单词级掩盖就是和 Bert 一样对样本中的每个字进行随机掩盖。词组级掩盖是对样本中的词组进行随机掩盖，然后让模型对这些词组进行预测，通过训练模型模型能够学习到词语

与词组之间的关联以及词组的语义信息。命名实体级掩盖是对样本中的命名实体进行随机掩盖，然后让模型对这些命名实体进行预测。通过将样本切分成一个个命名实体并对这些命名实体间的关系进行学习，模型能够更准确地理解实体与实体间的关系，有利于理解文本知识间的依赖来使得模型更具泛化性。

除此之外 ERNIE 基于 Bert 还对预训练任务进行了改进，它在三个层次的语言知识上丰富了预训练任务，使得模型能够捕捉到更多语义知识。在单词级别的预训练任务中除了掩码语言模型还有预测首字母大写任务 (Capitalization Prediction Task) 以及段落词语预测任务 (Token-Document Relation Prediction Task)。预测首字母大写任务的创建是由于单词中往往首字母大写的词语会有特别的含义，模型通过对首字母是否大写来学习这部分语义信息。段落词语预测任务是预测一个文本片段中的词是否会出现别的片段，这有助于学习一些文本语义信息对关键词的依赖。

在语法级别以及语义级别的预训练任务不仅拓展了邻句检索模型还增加了片段顺序复原模型(sentence deshuffling pre-training task)、语义关系预测 (Discourse Relation Task)，拓展后的邻句检索模型是句子距离预测模型(sentence distance pre-training task)。片段顺序复原模型主要是将一个段落切分成 n 个片段，然后对片段在原始段落中的正确顺序进行预测，这是一个多分类的问题，有助于学习语言语法间的依赖关系。句子距离预测模型则是一个三分类问题，它对两个句子进行了以下三种预测：两个句子是相邻的、两个句子不相邻但处于同一个段落中、两个句子处于两个不同的段落。句子距离预测模型能学习到语义以及语法双重语言信息。语义关系预测则是对两个句子间的修辞关系进行预测来提升模型对句间语义的学习。

在训练形式上 Bert 是两个预训练任务同时训练，ERNIE 由于预训练任务繁多而采用了融合多预训练任务的持续学习。如果每种任务单独依次训练可能会导致模型遗忘早期训练任务中包含的语义信息，如果多种任务同时训练会导致如果有新任务的加入就会使得模型重新训练。而持续多任务学习在保留原有任务的基础上引入新任务加入训练，在保留模型在过去学到的知识的同时能根据新任务加深对文本的学习。具体任务训练图如下图所示：

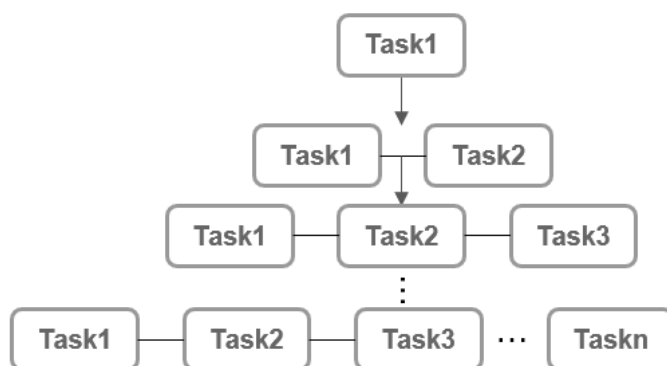


图 2-6 ERNIE 持续多任务学习流程图

2.3 基于深度学习的文本分类模型

深度学习作为机器学习的延伸,主要基于深层次的神经网络结构针对数据进行表征学习从而实现模型能像人脑一样解释数据。研究者在不断构建更深层次更复杂的网络结构来达成对数据理解更加深入的目的。深层神经网络(DNN)、深度信念网络(DBN)、卷积神经网络(CNN)以及循环神经网络(RNN)等富有代表性的神经网络结构在很多领域均有着优异的表现。得益于深层次的神经网络结构,深度学习在自然语言处理中的分词、机器翻译、文本分类以及信息抽取等任务中的表现要优于传统机器学习方法。事实上在模型流程中深度学习分类模型往往是在词嵌入层模型之后,通过词嵌入层模型输出的词向量来进行深度模型的训练任务。基于上文的词嵌入模型,本文使用了 TextCNN、Bi-LSTM、DPCNN 以及 RCNN 深度神经网络结构,下文会对这些模型的模型结构以及原理进行介绍。

2.3.1 TextCNN 卷积神经网络

卷积神经网络作为最常用的深度学习网络结构之一通过多个卷积核在数据矩阵中滑动的过程中进行卷积计算实现模型对数据的感知。卷积神经网络由于其独特的深层网络结构最初应用于计算机视觉领域。至 2014 年 Kim 等人提出了应用于自然语言处理领域文本分类任务的 TextCNN,并取得了优异的

分类效果。首先 TextCNN 通过将词向量拼接构成数据矩阵；类似于 CNN，在数据矩阵通过卷积层时 TextCNN 会利用多个不同大小的卷积核通过卷积运算进行特征提取。在数据通过卷积层后会生成特征矩阵，接着让这些特征矩阵通过池化层提取其中最大数值或者平均数值来实现数据降维以及特征提取。本文中的池化层选用最大池化层。最大池化层的特点在于提取其中最大特征，并丢弃其他表现较弱的特征，从而达到降低噪音的效果。最后将结果进行拼接后输入全连接层并通过 softmax 进行归一化处理。最终输出一个长度为类别数目的向量，将向量中最大值所在位置对应类别作为模型对文本类别的预测。TextCNN 的优势在于它拥有独特的卷积运算方式能够捕捉局部信息，并且通过共享卷积核的方法来优化模型的感知能力，因此 TextCNN 在特征提取中通常有着比较稳定的效果。具体模型结构如图 2-7 所示：

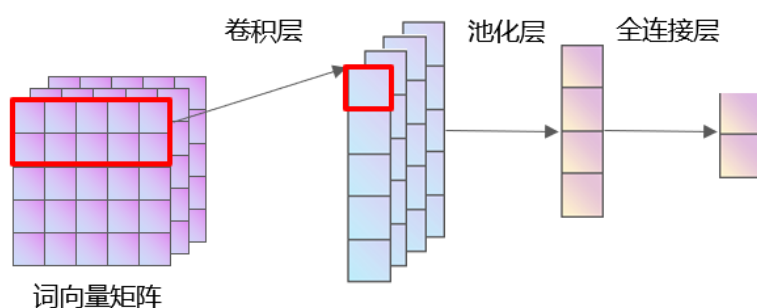


图 2-7 TextCNN 模型结构图

TextCNN 不同于传统 CNN 模型，TextCNN 只有一层卷积层并且卷积核通常是 2 维的，在使用多种词嵌入模型时可能会采用 3 维卷积核。除此之外 TextCNN 的卷积运算以及池化运算只会沿着词向量矩阵的长进行移动，这缘于文本数据的结构特点。为了能够对每个词的完整信息进行感知，TextCNN 的卷积核的宽会与词向量的长度相同。基于语法特征，文本中用于分类任务的信息特征通常是稀疏的，故卷积核长度通常会比较短。又由于能够反映同一情感倾向的文本特征是多样的，为了能够捕捉尽量多的特征信息，模型会对于同一尺寸设置多个卷积核。只有设置足够多的卷积核，卷积核之间才能互补并针对文本提取更全面的信息，从而产生更好的特征提取效果。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/827144125013006036>