诚 信 声 明

我声明，所呈交的毕业论文是本人在老师指导下进行的研究工作及取得的研究成果。据我查证，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得其他教育机构的学位或证书而使用过的材料。我承诺，论文中的所有内容均真实、可信。

毕业论文作者签名： 签名日期： 年 月 日

基于tensorflow的论坛帖子情绪分析

**[摘 要]** 随着互联网的快速发展和普及，人们倾向于在各种社交论坛上发表自己的观点和评论。这些论坛帖子中包含着大量丰富的情感信息，使用计算机处理和理解这些信息就要用到文本情绪分析技术。近年来随着机器学习的发展和成熟，其中很多方法被用于文本情绪分析问题。本文首先梳理了文本情绪分析问题的技术流程。然后利用tensorflow机器学习计算框架，搭建了BP神经网络模型和卷积神经网络模型用于文本情绪分析，并通过实验进行最优数据处理方法选择和参数调优，使得模型体现较好的性能。最后，将两种模型集成为具有实用性的文本情绪分析软件，使非专业人员能够方便地使用软件进行各类论坛帖子数据情绪分析。

**[关键词]** 文本情绪分析；机器学习；特征向量

The sentiment analysis of forum posts based on tensorflow

**Abstract：**With the rapid development and popularization of the Internet, people tend to express their opinions and comments on various social forums. These forum posts contain a wealth of emotional information, it requires sentiment analysis for computer to process and understand these information. In recent years, with the development and of machine learning, many methods of machine learning are used in sentiment analysis. This paper first combs the technical flow of sentiment analysis. Then use the tensorflow framework to build the BP neural network model and the convolution neural network model for sentiment analysis. Next, optimal data processing method and parameter optimization are carried out through experiments, which make the model better performance. Finally, the two models are integrated into practical sentiment analysis software, so that non professionals can easily use software to analyze the emotional analysis of various forum post data.

**Keywords：**sentiment analysis；machine learning ；feature vector

目 录

[1 绪论 1](#_Toc514060821)

[1.1 选题背景及意义 1](#_Toc514060822)

[1.2 文献综述 2](#_Toc514060823)

[1.3 研究框架 4](#_Toc514060824)

[2 相关理论知识及技术 5](#_Toc514060825)

[2.1 数据预处理 5](#_Toc514060826)

[2.2 特征向量表示 6](#_Toc514060827)

[2.2.1 文本特征向量表示 6](#_Toc514060828)

[2.2.2 词向量表示方法 7](#_Toc514060829)

[2.3 特征选择 8](#_Toc514060830)

[3 基于BP神经网络的情绪分析 10](#_Toc514060831)

[3.1 模型构建及实现 10](#_Toc514060832)

[3.1.1 文本预处理 10](#_Toc514060833)

[3.1.2 特征向量表示和特征选择 11](#_Toc514060834)

[3.1.3 BP神经网络模型搭建 11](#_Toc514060835)

[3.2 实验过程及分析 13](#_Toc514060836)

[3.2.1 特征向量表示和特征选择方法 13](#_Toc514060837)

[3.2.2 神经网络模型参数调优 14](#_Toc514060838)

[4 基于卷积神经网络的情绪分析 17](#_Toc514060839)

[4.1 模型构建及实现 17](#_Toc514060840)

[4.1.1 文本预处理及词向量表示 17](#_Toc514060841)

[4.1.2 卷积神经网络搭建 18](#_Toc514060842)

[4.2 实验过程及结论 20](#_Toc514060843)

[5 实用情绪分析软件开发 22](#_Toc514060844)

[5.1 软件设计与实现 22](#_Toc514060845)

[5.2 软件效果展示 23](#_Toc514060846)

[结论 26](#_Toc514060847)

[致谢 27](#_Toc514060848)

[参考文献 28](#_Toc514060849)

# 绪论

## 选题背景及意义

随着计算机和互联网技术的飞速发展，网络的普及率越来越高，网民规模日益庞大，“全民上网”正逐渐成为现实。中国互联网络信息中心（CNNIC）发布的第41次《中国互联网络发展状况统计报告》中指出：截至2017年12月，我国网民规模达7.72亿，普及率达到55.8%[[1]](#endnote-2)[1]。

互联网技术的革新与发展并不仅仅体现在网民规模的增长上，不断创新的互联网模式为用户提供多样化的、覆盖面更广的线上线下服务，使得人们的生活越来越便捷。移动支付比例持续提高并向农村渗透，手机付款已经逐渐成为了一种生活习惯；共享单车和网约车服务提高了人们的出行效率；线上政务服务使人们能够随时随地缴纳水电气费；除了基础生活服务外，网络娱乐的形式也更加丰富。随着越来越多的人活跃在网络上，人们积极地在各种网络平台上发表观点和看法，产生了大量的文本数据。由于其庞大的数据量，采用人工方法进行处理是非常困难的，利用计算机技术对用户评论、社交数据进行分类、处理、应用有着重要意义，文本情绪分析就是其中一个研究方向。

文本情绪分析是指将文本按照所包含的主观情绪进行分类，一般实现正向和负向的二级分类，或正向、负向、中性的三级分类。网络论坛中的评论文本，包括对商品的评价、对事件、政策的态度等。对这些文本进行情绪分类，对于网络舆情检测、商业市场分析等有着重大意义[[2]](#endnote-3)[2]，并为进一步的文本分析、提取、过滤、自动文摘建立基础。

## 文献综述

一般认为，有关情绪分析的研究最早由Bo Pang[[3]](#endnote-4)[3]于2002年提出，之后的时间中吸引了很多研究人员的关注。随着计算机技术特别是机器学习与深度学习的逐渐发展，越来越多的方法与技术被用于文本情绪分析，很多方法在此领域都取得了不错的成果。现有文本情绪分析的方法主要可以分为两大类，即无监督学习和监督学习。

文本情绪分类中基于无监督学习的方法一般为基于词典和规则的分析方法，这种方法依赖于完善的情感词库和准确的语法规则分析。早在文本情绪分析概念提出前，就有很多有关词语情感的研究可以用于情感词典的构建。Hatzivassiloglou 和McKeown[[4]](#endnote-5)[4]在1997年基于一种语言现象“相同情绪的词常以and连接；相反情绪的词常以but连接”提出预测词语情绪倾向的方法，先选取部分初始种子集利用and, but进行词语拓展，根据词语相似度进行聚类构建情感词典。Turney[[5]](#endnote-6)[5]等人在2002年提出利用基准词（excellent,poor）与目标词之间的点互信息(PMI)去度量两个词的相关程度从而进行词语倾向性分析。Hu和Liu[[6]](#endnote-7)[6]在2004年利用WordNet中同义词与反义词的关系信息来预测词语极性，扩充情感词典。在国内，朱嫣岚[[7]](#endnote-8)[7]等人在2005年利用HowNet中语义相似度和语义相关场的计算来分析中文词语倾向，拓展中文情感词典。另外也有针对专门领域情绪分析的研究，2012年陈晓东[[8]](#endnote-9)[8]等人利用点互信息对新浪微博数据分析，建立了微博情感词典，以此为基础进行情绪分析。在word2vec工具出现后，2016年黄仁[[9]](#endnote-10)[9]等人利用word2vec计算出的词语相似性信息建立情感词典，对商品评论进行情绪分析。均取得了不错的成果。

基于监督学习的方法以机器学习方法为主，是近年来研究的热门。采用机器学习方法时，往往将文本情绪分析问题看做一种文本分类问题，将文本表示为特征向量，输入分类器进行分类。BO Pang[3]等在2002年首先在情绪分析问题中使用机器学习方法，用词袋模型表示特征向量，选用最大熵模型、朴素贝叶斯、支持向量机分类器进行准确率比较，最终实验表明当使用支持向量机分类器和unigrams能够得到更好的效果。之后的很多研究多集中在特征向量表示和特征提取方法的改进，但由于句子语义的复杂性，文本向量化始终存在很多问题。随着深度学习的提出及发展，为特征提取找到了新的方向。卷积神经网络模型由于可以进行局部特征值的提取，被广泛应用在语音识别、图像识别等多个领域。Yonn Kim[[10]](#endnote-11)[10]首次将卷积神经网络模型应用在文本分类中，取得了不错的效果。在国内，中国科学院唐慧丰[[11]](#endnote-12)[11]等将机器学习各种分类方法和特征表示、特征选择方法用于中文情感分类中并比较其准确性，结果表明采用Bigrams特征表示方法、信息增益特征选择方法、支持向量机分类器能取得较优的结果。刘龙飞[[12]](#endnote-13)[12]等将卷积神经网络方法用于微博情绪分类研究，分别对词向量级和字向量级原始特征进行操作，均取得了比较好的性能。

虽然机器学习的方法将文本情绪分析问题看做分类问题，但情绪分析相对于其他文本分类问题，由于情感表达的复杂和歧义、对领域依赖性较高等特点，更难取得较好的准确率。中文分类相较于英文分类，由于中文词词性词义复杂，对其情绪分析的难度更高。

除了以上提及的研究外，还有很多专家和研究人员为文本情绪分析研究投入了很多精力，并在各自方向取得了不错的成果。自然语言处理以其独特的魅力吸引着越来越多的研究人员。

## 研究框架

本文主要研究论坛帖子评论文本的情绪分析问题，基于google于2013年开源的tensorflow机器学习计算框架，分别搭建BP神经网络和卷积神经网络，并采用不同特征提取和选择方法、不同神经网络参数进行比较分析、提出改进方案。最终得到较为准确的情绪分析方法，并组织成一个实用的情绪分析系统。

第一章主要介绍了选题的背景及其意义，国内外研究人员在此问题研究上已经取得的进展，并简要介绍了中文文本情绪分析的难点所在。

第二章介绍有关情绪分析问题的相关理论知识，主要包括文本预处理、中文分词、特征向量表示和特征选择的有关方法。

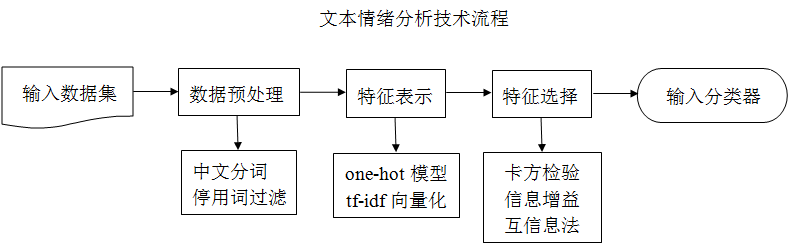
第三章设计并实现利用BP神经网络进行文本情绪分类的模型，并针对不同特征向量表示方法及模型参数进行实验，给出实验数据和相关结论。

第四章设计并实现使用卷积神经网络进行文本情绪分类的模型，利用卷积层进行特征提取，给出模型调优对比实验数据及相关结论。

第五章集成上文中构建的两种机器学习模型，设计并实现实用性文本情绪分析软件，使非专业人士能够方便的进行文本情绪分析。

# 相关理论知识及技术

随着自然语言处理逐渐成为近年来计算机研究和机器学习的热门领域，应用于文本情绪分析的技术也越来越成熟。本章主要对文本情绪分析中相关理论知识和通用技术做简要介绍，为下文进一步的研究奠定基础。



图‑1 文本情绪分析技术流程图

## 数据预处理

在自然语言处理工作中，往往将词语作为处理的基本单位。在英文语言中有空格作为自然明确的词语分割符，但在中文中词与词之间没有分隔符，我们在处理中文文本时就多了一步分词的步骤。不仅如此，由于汉语的词汇和语义复杂性，同一个句子有不同的分词方式，且它们可能表示不同的含义。这使得中文分词成为了一项颇有难度的工作，并且分词的质量会在很大程度上影响后续处理的效果。

常用的分词方法有基于词典和规则的方法、基于统计的方法。基于词典和规则的方法是根据已有词典，利用一定规则进行匹配。常见方法有：1）正向最大匹配：将文本从左向右取词典中最长词条长度作为待匹配字段，若待匹配字段在词典中，则将该词从文本中切下来，若不在，则去掉该字段最后一个词继续匹配。2）逆向最大匹配与正向最大匹配方式相同，区别在与匹配时从右向左进行。3）双向最大匹配将正向与逆向匹配到的词同时收录。基于统计的方法利用机器学习和概率模型进行分词，基本思想是根据字与字同时出现的概率和组合频率来判断是否能组成一个词。目前常用方法有隐马尔可夫模型、最大熵模型、N元文法模型等。

随着中文分词需求量越来越大，出现了很多分词工具供研究人员使用。比较常用的有：中科院计算所NLPIR、哈工大LTP、Hanlp分词器，基于python的jieba分词等。

对文本进行分词后，往往会发现有很多常用词出现频率很高，但对于文本分析基本没有帮助，如“的”“这”“那”“了”等。此时一般会选择使用停用词词典进行停用词过滤，从而节省存储空间，提升处理和计算效率。

## 特征向量表示

### 文本特征向量表示

中文文本用汉字表示，但计算机无法理解自然语言。因此需要将文本表示成计算机能够识别的形式，一般是向量形式。同时还需要保证向量能够尽量表达文本的特征。

常用的特征表示基本方法为词袋模型，也称one-hot模型。该模型以词表长度作为向量长度，向量的每一个位置代表一个词，用布尔值来简单表示该词在或不在所表示的文本里。

例如词典为：{我 小花 喜欢 阳光 需要}

“我喜欢小花”可以表示为：[1 1 1 0 0]

“小花需要阳光”可以表示为：[0 1 0 1 1]

One-hot的方法只表示文本中是否有某个词出现，而没有考虑他们的出现频率、重要程度。

为了表征这些信息，可以使用tf-idf模型来表示文本向量，这种模型考虑了词语出现的频率及其辨别能力。基本思路是，如果一个词在某个文本中出现次数较多，而在其他文本中出现较少，那么这个词可以更大程度的表征该句子的特征。其中tf指词频，为词条在文本中出现的频率。idf被称为“逆文本频率指数”，计算方法为文本总数目除以包含词语x的文本数目，将得到的商取对数。if与idf的乘积就是词语的在文本中tf-idf值。进行文本特征向量表示时，计算词表中每一个词在文本中的tf-idf值，将所组成的权值向量作为特征向量[[13]](#endnote-14)[13]。

### 词向量表示方法

随着深度学习的发展及其在语音识别、图像识别等方面的成功应用，深度学习在自然语言处理领域也逐渐有了更多的应用。利用深度学习的方法进行情绪分析需要将文本中的词表示为向量形式。传统的词向量表示方法为one-hot Encoder方法，与文本特征向量表示的方法类似，在与词表长度相同的向量中，将词在词表中出现的位置设为1。这种方法维度过高，且不能表示词与词之间的关联。word2vec的词向量表示方法推出后，很快成为了深度学习在自然语言处理领域的必要工具。

word2vec方法原理基于Tomas Mikolov[[14]](#endnote-15)[14][[15]](#endnote-16)[15]的论文，利用神经网络的方法将词语表示为实数向量，word2vec由谷歌在2013年开源。word2vec包含两种模型：CBOW和Skip-gram。CBOW模型是由一个词的上下文来预测该词；Skip-gram则相反，是通过一个已知词来预测上下文。在使用过程中可以通过指定窗口值来确定需要考虑的上下文窗口。

在word2vec出现之前，已经有利用神经网络来训练词向量的研究，传统的CBOW模式是使用带一个隐层的神经网络。网络的输入是one-hot表示的目标词上下文词向量，输出是所有词的最大softmax概率，也就是目标词的one-hot词向量。所求的词向量结果为神经网络的参数。Skip-gram与CBOW模型相反，输入为目标词one-hot词向量，输出为softmax概率最高的上下文词向量。由于词汇表很大，使得这种模型的计算量非常大。

Word2vec也使用了CBOW和Skip-gram两种模式，但改进了训练所用的神经网络模型。给出了两种优化方式，Hierarchical Softmax方式使用哈夫曼数来代替从隐藏层到输出层的映射；Negative Sampling方式使用负采样方法来简化概率值的求解。

使用word2vec训练出的词向量能够较好的表示词语特征及词与词之间的关系。

## 特征选择

将文本表示为特征向量后，特征向量维度通常非常大，这是由于特征词数量过大造成的。过大的维度为后续的计算造成了很大的困难，并且某些特征项的区分度较低，实际上对后续的分析处理没有太大的影响。使用特征选择的方法，选择对文本分类影响力大的特征项，能够有效的降低特征维度，并可以从一定程度上提升分类的准确率。

常用的特征选择方法有卡方检验、信息增益、互信息等。

本文中选用卡方检验作为特征选择方法，卡方检验方法的基本思想是通过实际值和理论值的偏差大小判断理论的正确性，偏差较小时接受原假设，偏差较大时判断原假设不成立。应用在特征选择中时，通常将特征词x与类别t不相关作为原始假设，被衡量的值应该为x在t中出现的频率和x在所有文档中出现的频率。此时实际值可以被看做是x在t中的频率，理论值可以看做x在全类别中的频率。利用卡方检验中的偏差衡量计算公式来计算特征词（也就是特征向量中的特征项）与类别t的卡方值。卡方值越大说明特征项与类别越相关，实际运用时按照需要选择前k大的特征。

经过分词、停用词过滤、特征向量表示、特征选择之后将文本转化成了计算机可以理解的维度适宜的特征向量。

# 基于BP神经网络的情绪分析

BP神经网络是一种利用信号正向传递、误差反向传递的多层前反馈神经网络。利用对目标误差函数最小值的不断逼近，能够进行多维函数的拟合且性能良好，当前应用非常广泛。本章尝试将全连接BP神经网络应用在文本情绪分析中，并利用多次实验进行参数调优，最终实现准确率较好的分析算法。

## 模型构建及实现

### 文本预处理

现有的可以用于情绪分析的中文标注语料集较少，且没有统一的评测数据来评测模型准确率。本文在选取数据集时，二分类数据集使用中科大谭松波等人提供的酒店评论语料，正负向各2000条，且使用该数据集作为模型构建实验评测标准。需使用到三分类（正向、负向、中性）数据集的地方，利用爬虫爬取豆瓣影评，以豆瓣评星作为基础标注，一、二星预标注为负向，三星预标注为中性，五星预标注为正向，又进行了手动筛选，最终得到三种分类各约3000条数据。

本文选择python的jieba分词库作为分词工具，选择适合文本分析的精确模式即尽可能精准的对文本进行分词，并使用自定义词典添加了部分自定义词，如“很好”“非常好”之类对情绪分析很有帮助且正常分词可能会被切分的词。分词工具还可以进行词性标注，用于后续特征表示。分词及词性标注结果示例如图所示：



图表 ‑1 分词及词性标注结果示例

分词之后进行停用词过滤，在通用的停用词词典中删去一些对文本情绪分析有所帮助的程度副词，最终形成包含1893个字符的停用词词表，对分词后文本进行停用词筛选。

### 特征向量表示和特征选择

对分词后的文本集，使用python的sklearn库选取出现次数大于5次的词形成词表。

在选取特征向量表示方法时，对多种方法进行了实现，并实验评测其准确度以便选取较优的方法。

首先使用one-hot模型进行特征向量表示。之后利用python的sklearn库实现了使用tf-idf方法表示的特征向量。考虑到文本情绪分析相较于其他分类问题的特殊性，提出了one-hot模型的变形实现方案，按照我们的普遍认知，副词和形容词对情绪表达的贡献度最高，动词次之，名词及其他词表达情感较小。因此将one-hot模型中对每个词出现与否的布尔值表示改为按词性设置为不同权重。具体实验及结果见下文。

本文使用卡方检验作为特征选择方法，取卡方检验值前k大作为选择后的特征。最终每个文本形成一个k维向量作为后续神经网络的输入。

### BP神经网络模型搭建

本文中使用的BP神经网络模型为带隐层的全连接神经网络结构，隐层的数目和结点数由后续实验决定。本文使用tensorflow框架进行神经网络的搭建与训练验证，只需进行计算图的构建，由框架实现计算过程。

本文所使用的BP神经网络模型如下图所示：

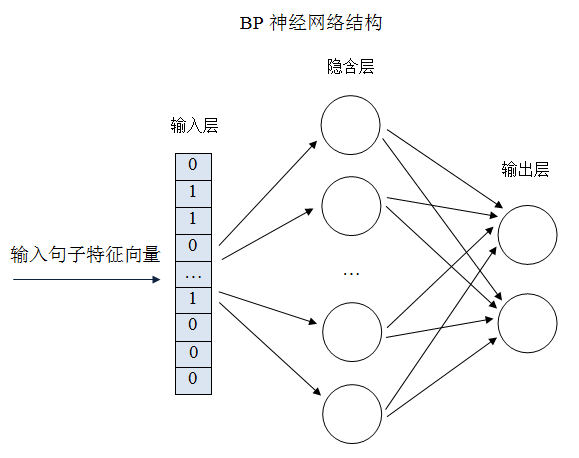


图 ‑2 BP神经网络结构

神经网络输入为文本的特征向量，输入层结点数为特征向量维度。网络输出为分类结果，输出结点数为类别数，例如：正向表示为[1,0]，负向表示为[0,1]。

神经网络训练过程中，首先将参数设为随机值，定义交叉熵代价函数为损失函数。利用优化算法向着损失函数取极小值的方向不断调整参数。

在优化算法选取过程中，首先选用了随机梯度下降算法。实验过程中发现，在调整隐层结点数目时需随之调整算法的学习率，否则由于学习率过大或过小有时网络的训练会出现问题。因此将优化算法改为自适应梯度下降算法（adam），算法能够自行为不同参数计算使用自适应性学习率。同时为了防止过拟合现象，在损失函数中加入了l2正则化系数。

计算准确率时，在数据集中随机选取10%数据作为验证集，其余数据作为训练集，在训练集中进行多次迭代训练网络参数，最后在验证集上计算准确率。

通过合理的参数设置，该网络能够取得较好的分析准确率，具体实验过程及准确率见下文。

## 实验过程及分析

### 特征向量表示和特征选择方法

初始进行实验时，选用带两个隐层的神经网络，隐层结点数设置为100，特征向量维度即输入结点数设为200，在二分类数据集上进行实验。

首先对三种文本特征向量表示方法进行实验：分别是one-hot模型，考虑词性的one-hot模型，tf-idf模型。分别进行五次实验，对比实验数据：

表格 ‑1特征向量表示方法准确率比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 特征向量表示 | 1 | 2 | 3 | 4 | 5 | 平均值 |
| one-hot | 0.836 | 0.838 | 0.833 | 0.815 | 0.818 | 0.828 |
| one-hot 词性 | 0.775 | 0.797 | 0.823 | 0.801 | 0.813 | 0.8018 |
| tf-idf | 0.792 | 0.754 | 0.764 | 0.803 | 0.754 | 0.7734 |

通过实验数据可以看出在这三种方法中朴素的one-hot模型对情绪分析而言效果更好。针对词性区分权重的方法并没有取得更好的效果，这也从某种程度上符合文献[3]中的实验结论，相比于只采用形容词作为特征，采用所有词作为特征可以更好的达到情绪分析的效果。Tf-idf权重向量方法在本实验中也没有取得更好的效果。

本实验使用卡方检验方法作为特征选择方法，为了验证特征选择方法的有效性，将使用了特征选择与未使用特征选择的方法作为对比实验：

表格 ‑2使用特征选择准确率比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 特征选择 | 1(准确率) | 2 | 3 | 4 | 5 | 平均值 |
| 卡方检验 | 0.836 | 0.838 | 0.833 | 0.815 | 0.818 | 0.828 |
| 不使用 | 0.775 | 0.797 | 0.823 | 0.801 | 0.813 | 0.8018 |

不使用特征选择时，特征向量维度为词表长度，本实验中为3190，使用卡方检验方法选择200维向量。使用卡方检验后模型平均准确率比不使用特征选择提高了2%，实验表明特征选择方法不仅能够有效降维降低计算复杂度，也一定程度上增加了分类准确率。

### 神经网络模型参数调优

以双隐层神经网络为基础，输入特征向量维度为200，调整隐层结点数为50、100、200，进行对比实验：

表格 ‑3不同隐层结点数准确率比较（输入向量维度为200）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 隐层结点数 | 1(准确率) | 2 | 3 | 4 | 5 | 平均值 |
| 50 | 0.813 | 0.811 | 0.831 | 0.772 | 0.846 | 0.8146 |
| 100 | 0.836 | 0.838 | 0.833 | 0.815 | 0.818 | 0.828 |
| 200 | 0.841 | 0.854 | 0.816 | 0.831 | 0.831 | 0.8346 |

实验表明，随着结点数的增加，分类的准确率有小幅度提升（隐层结点数不能超过输入结点数）。每增加一倍结点数，准确率平均值上升了1%左右。

上述实验中，将特征选择后向量维度定200，神经网络输入结点数自然也为200。为了明确输入结点数对准确率的影响。以及再次验证分类准确率随隐层结点数的变化。将特征选择后向量维度调整为400，改变隐层结点数，再次进行对比实验：

表格 ‑4 不同隐层结点数准确率比较（输入向量维度400）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 隐层结点数 | 1(准确率) | 2 | 3 | 4 | 5 | 平均值 |
| 100 | 0.828 | 0.826 | 0.816 | 0.819 | 0.816 | 0.821 |
| 200 | 0.824 | 0.843 | 0.851 | 0.856 | 0.805 | 0.8358 |
| 400 | 0.82 | 0.89 | 0.854 | 0.844 | 0.849 | 0.8514 |

比较隐层结点数可以看出分类的准确率会随着结点数的增加而有小幅度提升，再次验证了上文实验中的结论。比较输入向量维度与准确率的关系可以看出，到目前为止分类的准确率会随着向量维度的增加而提升。增加向量维度到800，设置隐层结点数为800，再次进行实验，将不同向量维度实验数据列表进行比较：

表格 ‑5 不同输入向量维度准确率比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 输入向量维度 | 1(准确率) | 2 | 3 | 4 | 5 | 平均值 |
| 200 | 0.841 | 0.854 | 0.816 | 0.831 | 0.831 | 0.8346 |
| 400 | 0.820 | 0.890 | 0.854 | 0.844 | 0.849 | 0.8514 |
| 800 | 0.854 | 0.852 | 0.836 | 0.818 | 0.834 | 0.8388 |

当将向量维度增加到800时，准确率并没有进一步提升反而有所下降，可能是由于出现了冗余特征项造成，过大的向量维度也对计算造成了困难。特征向量维度为400时准确最高，在之后的实验及应用中将统一采用400维向量。

为了比较隐层数量对分类结果的影响，以特征向量维度为400，隐层结点数为400，对单隐层、双隐层、三隐层网络进行对比实验：

表格 ‑6 不同隐层数目准确率比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 隐层数 | 1(准确率) | 2 | 3 | 4 | 5 | 平均值 |
| 单隐层 | 0.849 | 0.892 | 0.857 | 0.829 | 0.854 | 0.8562 |
| 双隐层 | 0.820 | 0.890 | 0.854 | 0.844 | 0.849 | 0.8514 |
| 三隐层 | 0.844 | 0.829 | 0.861 | 0.826 | 0.854 | 0.8428 |

随着网络层数的增加，分类的准确率并没有得到提升，因此可以选取单隐层网络作为训练模型。

经过多次实验后，最终得到准确率最高的基于BP神经网络的训练方法为：使用one-hot模型进行文本特征向量化，卡方检验作为特征选择方法，最终得到维度为400的特征向量，输入单隐层网络进行训练，隐层结点数为400。

# 基于卷积神经网络的情绪分析

卷积神经网络最早由Hubel与Wiesel[[16]](#endnote-17)[16]提出，多用于语音和图像识别。近年来也逐渐用于自然语言处理领域，并取得了很好的效果。卷积神经网络的特征就是使用卷积层与池化层的混合，可以提取文本的高级特征。本章将构建卷积神经网络模型用于文本情绪分析问题，利用tensorflow计算框架实现网络模型，并通过实验比较模型准确率。

## 模型构建及实现

### 文本预处理及词向量表示

此处使用与上一章节相同的中文分词和停用词过滤方法。

卷积神经网络的输入应为文本词向量所构成的矩阵，矩阵的长为词向量维度，矩阵的宽为文本长度。为了使不同长度的文本组成相同大小的矩阵，应以最大文本长度为矩阵宽，不足长度的文本在构成矩阵时进行向后补零操作。

输入的词向量可以有两种选择，随机向量化和经过训练后的词向量。

使用随机向量化的方式，为每个词在词表中建立编号，利用高斯分布对每个词的词向量进行随机初始化，完全由卷积层去进行特征学习，设置词向量维度为400。

使用经过训练后的词向量，相当于为卷积层提供了一部分先验知识，在此基础上再进行学习，往往会得到更好的效果。本文中使用word2vec作为词向量训练方法，word2vec语言模型使用由维基百科训练出的模型，词向量维度也为400。

### 卷积神经网络搭建

本文所使用的卷积神经网络模型结构如图所示：

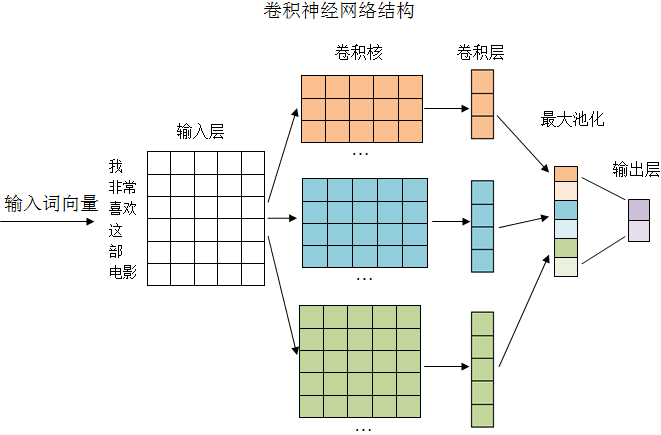


图 ‑1 卷积神经网络结构

图中模型输入由词向量堆叠成的句子向量矩阵，使用宽度为3、4、5的卷积核（卷积核可以有多个）对特征向量卷积。再使用最大池化层提取卷积得到的向量中最重要的特征。使用全连接方法连到输出层。

输入层中输入的为词向量所组成的二维矩阵，若文本最大长度为n，词向量维度为400，矩阵大小就为n\*400。

卷积层的作用是进行局部特征的学习，相当于用过滤器（卷积核）来过滤矩阵的局部区域，根据卷积核的参数不同得到矩阵的局部区域特征。卷积核计算的实质是将被卷积区域与卷积核中值相乘求和，卷积核工作原理示例：

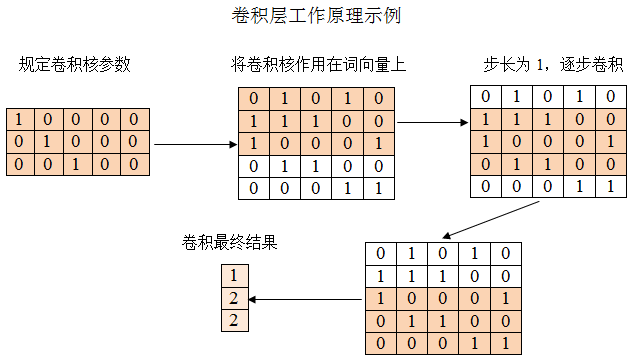


图 ‑2 卷积层工作原理示例

在用于文本处理时，卷积层其实就是学习n-gram特征，也就是词语的上下文特征。在对图像进行卷积操作时，使用的卷积核往往是n\*n大小，而在处理文本时，卷积是整行整行进行的，卷积核长度为词向量长度，这样学习到的就是以词为基本单位的特征而不会将词拆分成几部分。在卷积层对卷积窗口的步长设置时，为了学习每个词的特征，要将步长设置为1，即卷积核每次移动长度为1格。为了能够学习到不同窗口大小的n-gram特征，可以定义多种不同的卷积核宽度，本实验中定义卷积核宽度为3,4,5。并为定义每种卷积核数量为64个。经过卷积操作后，得到3\*64个一维特征图。

最大池化（max-pooling）层用于提取卷积层所得到的特征图中的最重要信号。该层操作时从每个一维的特征图中提取最大的特征项。最终输出为所有特征图的最大特征项，在本实验中为一个长度为192的一维向量。

之后是全连接层，利用softmax方式将得到的长度为192的一维向量连接到输出层，输出层为文本分类结果。

该卷积神经网络模型中，优化器依然使用自适应梯度下降算法（adam）。在全连接层中，为了防止过拟合，使用了l2正则化系数进行参数约束。同时还使用了dropout技术，在训练过程中随机让网络结构中某些结点权重暂时不更新。本实验在训练过程中将dopout系数设置为0.5，即随机让一半结点暂时不更新权重。

准确率计算过程与上一章相同，随机选取数据集的10%作为验证集，在训练集迭代训练网络完成后，使用验证集进行准确率计算。

## 实验过程及结论

对两种输入词向量表示方法，随机词向量生成和word2vec向量生成进行比较实验：

表格 ‑1 不同词向量表示方法准确率比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 词向量表示 | 1(准确率) | 2 | 3 | 4 | 5 | 平均值 |
| 随机生成 | 0.845 | 0.7875 | 0.815 | 0.806 | 0.825 | 0.8157 |
| Word2vec | 0.867 | 0.865 | 0.861 | 0.86 | 0.865 | 0.8636 |

通过实验可以看出，利用随机生成的词向量取得了81%的准确率，从某种程度上验证了卷积的效果。而提供了先验知识的word2vec词向量表示方法对分类的准确率有明显提升。这也是正是word2vec在深度学习的自然语言处理领域广泛使用的原因。

对卷积神经网络的最优结果与上一章中所用的BP神经网络最优结果进行对比，结果如下表：

表格 ‑2 不同神经网络算法准确率比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络模型 | 1(准确率) | 2 | 3 | 4 | 5 | 平均值 |
| CNN | 0.867 | 0.865 | 0.861 | 0.86 | 0.865 | 0.8636 |
| BP | 0.849 | 0.892 | 0.857 | 0.829 | 0.854 | 0.8562 |

通过对比可以看出，卷积神经网络模型所得到的平均准确率要略高于BP神经网络，且卷积神经网络所得到的准确率较稳定。说明利用卷积神经网络的方法能更好的学习文本的语义特征。

# 实用情绪分析软件开发

## 软件设计与实现

在论文的最后，利用上文中经过多次实验验证的两种用于文本情绪分析的算法，设计并实现一个实用的文本情绪分析软件。使得用户不需理解技术细节，只需提供必要的数据，就能够利用软件方便的进行文本情绪分析。

根据上文实验结论，使用卷积神经网络的方法相较与BP神经网络能够更加准确而稳定的进行情绪分类。但由于卷积神经网络的结构较为复杂，因此相比而言运行速度更慢。因此软件提供两种运行模式，标准模式对应卷积神经网络算法，快速模式则对应BP神经网络算法。用户可以根据自身需要，综合考虑速度和准确性选择运行模式。

软件为用户提供两种模型训练和使用方法，自定义标注数据集方法，即需要用户自行提供已经标注好的训练数据集，软件使用用户提供的数据集进行模型训练，再将训练出的模型应用在待分类数据集上。使用默认模型方法，即使用内置的训练好的模型（目前默认模型由上文中实验用数据集训练得到）。由于文本情绪分析问题具有较强的领域依赖，不同领域的数据相同的情绪特征较少，比如将用酒店评论数据集训练出的模型作用在影评数据集上，所得到的分析准确率可能较差。因此，推荐用户使用自定义数据集训练的方法来进行模型训练和使用。

软件支持二分类（正向和负向）、三分类（正向、负向和中性）两种分类类别数。其中三分类的准确率低于二分类，随着分类越来越细致而导致的准确率越来越低也是文本情绪分析领域的一个普遍现象。大于三种的分类由于难以保证准确率，因此不提供支持。

用户选择运行模式、模型训练方法、分类类别数之后，需提供必要的数据集文件和结果存储地址。软件会根据用户输入，调用对应的算法实现文本情绪分析并将分析结果写入目的地址。

软件使用python实现，神经网络部分使用tensorflow计算框架，另外使用了sklearn、numpy等第三方库。

## 软件效果展示

软件目前只实现了命令行效果，以选择快速模式、自定义数据集方法、二分类为例，软件运行时输入选择效果如下图：



图 ‑1软件输入效果展示

软件要求目标数据文件为txt格式，每行对应一个待分析的文本。自定义数据集中标注数据文件也为txt格式，不同类别文本在不同文件中，同样每行对应一个已经分类标注好的文本数据。利用自定义数据集，模型训练迭代20次，每次迭代输出损失函数值，运行结束时，输出分析结果文件地址，运行效果如下图：

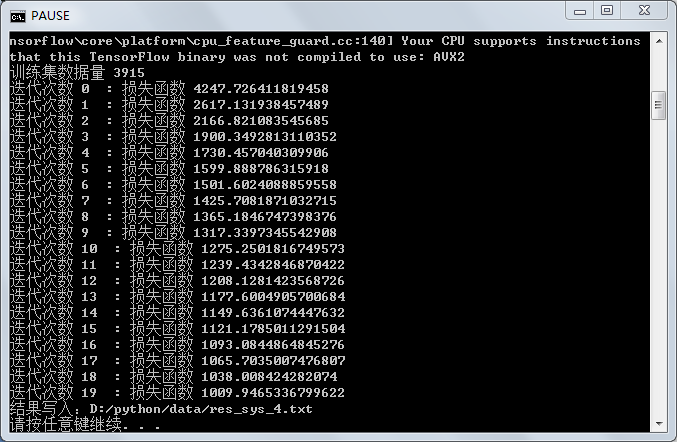


图 ‑2 软件运行过程展示

打开软件写入的分析结果文件，可以看到分析效果（为使效果方便分析，图中文本正确分类结果都应为负向）：

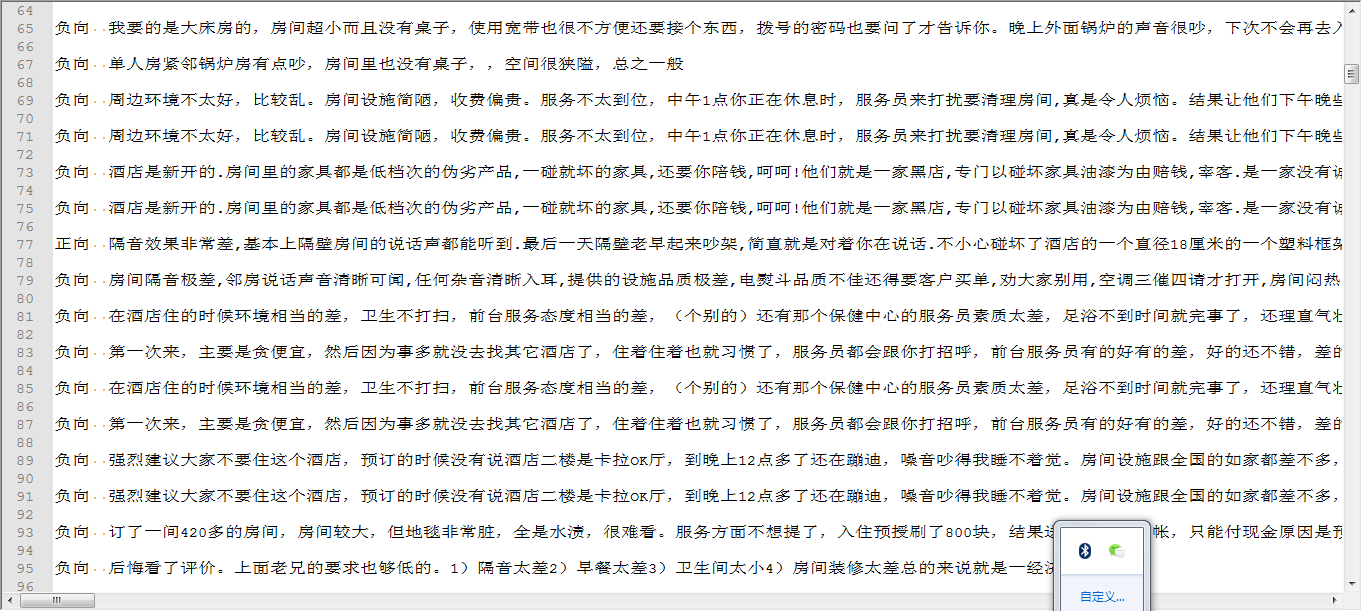


图 ‑3 软件分析结果展示

可以看到分类结果形式为每行代表一个文本，文本前标注得到的分类类别，由以上示例可以看出，软件所得到的分类结果基本正确，但也有一定的误差，上述行标77的文本就被错误的标注成了正向。一般来说，文本中有情感转折的数据和表述不常见的数据更容易判断失误。但总体来说，软件运行结果误差率是可以接受的。

# 结论

在互联网时代的背景下，大量的文本数据不断涌现在网络中，对于自然语言处理的研究变得越来越热门，文本情绪分析作为其中一个方法有着重大意义。本文使用tensorflow计算框架，将两种机器学习算法应用到对论坛帖子的情绪分析上，主要成果有：

（1）搭建BP神经网络模型进行文本情绪分析。通过实验得到较优处理方法：one-hot模型进行特征表示、卡方检验方法用于特征选择，构建单隐层网络，最终得到平均准确率约为85.6%。

（2）搭建卷积神经网络模型进行文本情绪分析。使用维基百科训练出word2vec模型，将其用于词向量表示，输入卷积神经网络，最终得到平均准确率为86.4%。得到一种稳定且准确的文本情绪分析方法。

（3）设计并实现具有使用性的文本情绪分析软件，集成两种文本情绪分析方法、支持情绪二分类与三分类，支持自定义数据集训练网络。使得非专业人士也能够方便地进行文本情绪分析。

同时，论文有一定的不足之处，情绪分析准确性某种程度上与数据集大小成正比，由于优秀标注数据集的获取困难，没能使用更大量的数据进行网络训练。由于实验用计算机性能有限，没能对网络模型中参数进行更全面的实验。另外最终实现的软件只有命令行界面，若有时间，应对软件进行进一步设计和改善。

# 致谢

感谢我的导师陈双平老师，谢谢他在毕业设计期间给予我的指导和点拨，引导了我毕设的方向，使我少走了很多弯路。其次，还要感谢在这四年的学习中教过我的所有老师们，他们不仅传授给我了知识和技能，也在我的学习生活中给了很多帮助和建议。另外，感谢我的室友和同学们，在四年的大学生活中给我带来了很多欢乐和温暖。

参考文献

1. [1]第41次《中国互联网发展状态统计报告》[OL].中国互联网络信息中心.

   http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P02018030540

   9870339136.pdf .2018.01 [↑](#endnote-ref-2)
2. [2]相若晨．中文文本情感分析研究[J]：硕士学位论文．扬州：扬州大学.

   2016 [↑](#endnote-ref-3)
3. [3] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?

   Sentiment Classification using Machine Learning Techniques[C]. Proceedings of the Conference on Empirical Methods in Natural

   Language Processing . PWiladelphia.2002.79—86. [↑](#endnote-ref-4)
4. [4] Vasileios Hatzivassiloglou and Kathleen R. McKeown[C]. 1997. Predicting the Semantic Orientation of Adjectives[C]. Association for Computational Linguistics -1997. 174–181 [↑](#endnote-ref-5)
5. [5] Turney R Thumbs up or thumbs down? Semantic orientation applied unsupervised classification of reviews[C]. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA, 2002. 417-424. [↑](#endnote-ref-6)
6. [6] Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions

   on the Web[C].In WWW ’05: Proceedings of the 14th international conferenceon World Wide Web. 2005:342--351. [↑](#endnote-ref-7)
7. [7]朱嫣岚,闵锦,周雅倩,黄萱菁,吴立德.基于HowNet的词汇语义倾向计算 [J]. 中文信息学报. 2006(01):14-20 [↑](#endnote-ref-8)
8. [8]陈晓东. 基于情感词典的中文微博情感倾向分析研究[J]：硕士学位论 文．武汉：华中科技大学，2012 [↑](#endnote-ref-9)
9. [9]黄仁,张卫．基于word2vec的互联网商品评论情感倾向研究[A]．计算 机科学. 2016. 43(6A)：387-389 [↑](#endnote-ref-10)
10. [10] Y. Kim. Convolutional Neural Networks for Sentence Classification[C]. Proceedings of the 2014Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha. Qatar. 2014:1746-1751. [↑](#endnote-ref-11)
11. [11]唐慧丰,谭松波,程学旗.基于监督学习的中文情感分类技术比较研巧[J]. 中文信息学报. 2007.21(6): 88-94. [↑](#endnote-ref-12)
12. [12]刘龙飞,杨 亮,张绍武,林鸿飞. 基于卷积神经网络的微博情感倾向性 分析[A].中文信息学报.2015.29[6]: 195-165. [↑](#endnote-ref-13)
13. [13]张英. 基于深度神经网络的微博短文本情感分析研究[J].硕士学位论文. 郑州.中原工学院. [↑](#endnote-ref-14)
14. [14] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word

    representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013. [↑](#endnote-ref-15)
15. [15] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119. [↑](#endnote-ref-16)
16. [16] Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. Journal of Physiology. 1962;160(1):106-154. [↑](#endnote-ref-17)