**大连海事大学**

**毕 业 论 文**

┊┊┊┊┊┊┊装┊┊┊┊┊┊┊订┊┊┊┊┊┊┊线┊┊┊┊┊┊┊

**二○一九年六月**

网络文本数据的爬取与分析

专业班级： 软件工程2班

姓 名： 王 冶

指导教师： 李 楠

信息科学技术学院

**摘 要**

本文针对互联网中的网络评论数据（以豆瓣影评为例），在机器学习的算法的基础上，分析其夹杂的个人情感，从而实现对于评论文本积极、消极情感的判断，以及灌水评论、垃圾评论的识别和剔除，达到舆情分析的初步效果。

第一，使用python对豆瓣网站的电影评论数据进行爬取，对于初步处理的数据使用MonGoDB进行持久化操作；第二，提取数据库中的影评文本和对应的评论得分，为每一条评论标注标签；第三，方便向量化操作，使用“结巴”工具做分词处理 ，对每一条文本数据进行分词处理、并生成TFIDF特征向量矩阵；第四，使用机器学习中的算法（朴素贝叶斯、支持向量机）以及深度学习中的卷积神经网络进行文本的情感分析。

本次实验基于Linux操作系统，使用python，使用VsCode编辑器编写程序，借助于机器学习工具包Scikit-learn和深度学习框架Tensorflow，以多种方法对文本的情感进行的训练分析，得到了对应的效果。

**关键词：网络爬虫；文本分类；情感分析；机器学习；深度学习**

**ABSTRACT**

Based on the network comment data in the Internet (taking douban film review as an example), this paper USES machine learning algorithm to analyze the personal emotions contained in it, so as to judge the positive and negative emotions of the comment text, as well as to identify and eliminate comments and garbage comments, and achieve the initial effect of public opinion analysis.

First, use python to crawl the movie review data of douban website, conduct preliminary data processing and store it in MonGoDB database. Second,

Extract the movie review text and corresponding comment score in the database, label each comment; Thirdly, using the Chinese word segmentation tool of jieba, word segmentation was performed on each piece of text data and the TFIDF eigenvector matrix was generated; Fourthly, the model in machine learning (naive bayes, support vector machine) and the convolutional neural network in deep learning are used for the emotional analysis of text.

This experiment is based on the Linux operating system. Python is used as the development language, and the VsCode editor is used to write the program. With the help of the Sklearn machine learning kit, the training and analysis of the text's emotion are carried out in various ways, and the ideal effect is obtained.

**Keywords: Web crawlers; Text classification; Emotional analysis; Machine learning; Deep learning**

**目 录**

[第1章 绪论 5](#_Toc9857034)

[1.1 课题研究的背景及意义 5](#_Toc9857035)

[1.2 文本情感分析的研究现状 5](#_Toc9857036)

[1.2.1 基于情感词典的文本情感分析方法 5](#_Toc9857037)

[1.2.2基于机器学习的文本情感分析方法 6](#_Toc9857038)

[1.2.3基于深度学习的文本情感分析方法 6](#_Toc9857039)

[1.3 论文组织结构 6](#_Toc9857040)

[1.4 本章小结 7](#_Toc9857041)

[第2章 基础知识及相关技术 8](#_Toc9857042)

[2.1 python网络爬虫技术 8](#_Toc9857043)

[2.2 文本预处理技术 8](#_Toc9857044)

[2.3 机器学习算法、深度学习算法 8](#_Toc9857045)

[2.3.1朴素贝叶斯 8](#_Toc9857046)

[2.3.2 支持向量机（SVM） 9](#_Toc9857047)

[2.3.3 卷积神经网络（CNN） 13](#_Toc9857048)

[2.4 本章小结 14](#_Toc9857049)

[第3章 网页爬取和数据预处理 15](#_Toc9857050)

[3.1 网页爬取 15](#_Toc9857051)

[3.2 数据预处理 16](#_Toc9857052)

[3.2.1去除脏数据 16](#_Toc9857053)

[3.2.2中文分词处理 17](#_Toc9857054)

[3.2.3停用词处理 17](#_Toc9857055)

[3.2.4平凡词、独特词处理 17](#_Toc9857056)

[第4章 朴素贝叶斯算法在中文文本情感分析中的应用 19](#_Toc9857057)

[4.1 贝叶斯定理与朴素贝叶斯 19](#_Toc9857058)

[4.1.1 贝叶斯公式 19](#_Toc9857059)

[4.1.2 朴素贝叶斯在文本情感分析中的具体应用 19](#_Toc9857060)

[4.2 朴素贝叶斯中文文本情感分析处理流程 20](#_Toc9857061)

[4.2.1 中文文本预处理 20](#_Toc9857062)

[4.2.2 停用词处理 20](#_Toc9857063)

[4.2.3 中文分词 20](#_Toc9857064)

[4.2.4 自定义词典 21](#_Toc9857065)

[4.2.5 TF-IDF特征提取 21](#_Toc9857066)

[4.2.6 中文词性标注 22](#_Toc9857067)

[4.2.7 词频统计 22](#_Toc9857068)

[4.2.8 训练数据、验证数据 22](#_Toc9857069)

[4.3 本章小结 26](#_Toc9857070)

[第5章 支持向量机算法在中文文本情感分析中的应用 27](#_Toc9857071)

[5.1 支持向量机·核函数 27](#_Toc9857072)

[5.2 不同核函数的SVM文本情感分析对比 28](#_Toc9857073)

[5.2.1 线性核函数实验效果： 28](#_Toc9857074)

[5.2.2多项式核函数实验效果 29](#_Toc9857075)

[5.2.3 高斯核函数实验效果 29](#_Toc9857076)

[5.3 本章小结 30](#_Toc9857077)

[第6章 卷积神经网络算法在中文文本情感分析中的应用 31](#_Toc9857078)

[6.1 文本预处理 31](#_Toc9857079)

[6.1.1 基于字粒度的文本预处理 31](#_Toc9857080)

[6.1.2 基于词粒度的文本预处理 31](#_Toc9857081)

[6.2 中文文本情感分析时卷积神经网络的结构 32](#_Toc9857082)

[6.3 卷积神经网络训练、验证、测试结果 33](#_Toc9857083)

[6.4 本章小结 34](#_Toc9857084)

[第7章 实验结果对比与不足之处 35](#_Toc9857085)

[7.1 实验结果对比 35](#_Toc9857086)

[7.1.1 对比准确率、召回率、F-Score 35](#_Toc9857087)

[7.1.2 数据量对比 36](#_Toc9857088)

[7.2 问题与不足 36](#_Toc9857089)

[参考文献 37](#_Toc9857090)

[结论 39](#_Toc9857091)

[致谢 40](#_Toc9857092)

**网络文本数据的爬取与分析**

# 第1章 绪论

## 1.1 课题研究的背景及意义

当今社会，在信息化的影响下，每个人都拥有着丰富多彩的网络生活，伴随着大流量app、网站的出现与流行，数以万计的网络数据进入人们视野，而垃圾评论、恶意灌水评论的大量存在混淆了人们的视听，错误引导舆论，甚至严重至歪曲事实真相。

基于这样的现实背景，分析网络文本评论数据、过滤垃圾评论、以及进行舆情把控便具有重大的意义。本文在这样的现实背景下，借助于机器学习的相关算法，对数据集中的数据进行了相关处理，包括分词、提取文本的特征、向量化、预测分析等，从而初步实现对于网络文本数据的正负向情感分析、以及垃圾评论的过滤。

本文立足于上述的现实背景和需求，通过研究机器学习算法（朴素贝叶斯、支持向量机） 、深度学习算法（卷积神经网络）并将其应用到中文文本分类这一问题，而这样的做法对于推动网络舆情把控，以及对于特定网络文本的分析有着重要意义。

## 1.2 文本情感分析的研究现状

现如今，能够对中文文本进行情感分析，并且得到广泛应用和认可的方法如下所示：

1. 构建情感词典，对中文文本的情感倾向进行分析
2. 使用机器学习算法对文本的情感进行分析和预测
3. 那么随着进年来GPU大幅度的提升了运算速度。深度学习的相关算法也可以被应用到这一领域，主要包括卷积神经网络，循环神经网络

在接下来的章节中，本文将详细的对于以上三种方法的特点和现阶段发展的状况进行介绍，同时也会列举各自的不足之处。

### 1.2.1 基于情感词典的文本情感分析方法

使用人工构建的情感词典，进行分析，通常是以人工搭建的情感词典为基础，，根据文本中出现的具有感情色彩的词语，和情感词典进行比较，从而对文本的情感做出判断。就研究现状而言，日常生活中常用的搜索引擎便将这一方法应用了进来。不过基于情感词典的文本情感分析方法需要投入大量的人力，将非常耗费人力成本。

### 1.2.2基于机器学习的文本情感分析方法

机器学习算法主要分为两类，有监督和无监督。有监督就是需要对实验数据集进行人工的预处理，而无监督并不需要这一处理。

有无监督的区别在于：是否使用人工标注的数据作为数据集进行训练分析，近年来，包括使用朴素贝叶斯、SVM支持向量机、决策树等机器学习算法在文本情感分析问题上面都取得了非常不错的效果，朴素贝叶斯基于贝叶斯公式，将“已知特征求解类别”的问题转化成“已知类别求解特征”和“类别概率”的乘积问题；SVM基本思想可以简单用两个部分来解释。第一，需要找到能够将不同类别实验数据合理分隔开来的媒介，这个媒介称其为分隔超平面。第二，需要使分隔超平面之间几何距离最大。本文将使用sklearn机器学习库实现这两种机器学习算法，不断调参优化，最终实现中文文本情感分析。

### 1.2.3基于深度学习的文本情感分析方法

主要使用的方法有：CNN（卷积神经网络）、RNN（循环神经网络）。卷积神经网络通过卷积核的卷积运算，获取上层输入数据的重要元素或特征，在经过池化层的进一步处理，池化层有两种主要的方法，一个是Maxpolling，另外一个是平均化池化。在经过池化层的进一步特征提取，最终在全连接层对分类问题进行回答，因此卷积神经网络的上述流程完美契合分类任务的要求。而循环神经网络由于自身与卷积神经网络不同的特点，同时属于前馈神经网络，导致其通道之间具有了前后时间上的联系。所以循环神经网络对于分类的任务并不是很适应。目前国内外关于文本情感分析的研究还远远没有达到饱和，由于模型的限制，以及中文数据的不确定性，这也导致了关于中文的情感分析还没有达到英文情感分析的高度。本文使用卷积神经网络这种方法，通过使用Tensorflow搭建网络，从而满足实验和数据的需求。

## 1.3 论文组织结构

本文的内容组织结构如下：

第一章：介绍当前网络生活中的大流量应用所带来的垃圾评论对人们生活的影响，讨论网络生活中的信息是否具有分析的价值。最后讲述网络文本情感分析主要的研究方法和现阶段的研究现状。

第二章：介绍本实验过程中所用的相关技术、软件工具以及必备的基础知识。最后，本章着重介绍所使用的机器学习、深度学习算法即朴素贝叶斯、支持向量机、卷积神经网络。

第三章：介绍使用python网络爬虫技术对豆瓣网站进行爬取，同时介绍本文爬虫的具体逻辑；除此之外，还将介绍如何使用正则表达式过滤脏数据、冗余数据；除此之外，也将对然语言处理中的文本预处理内容——分词、去除停用词部分进行详细阐述。

第四章：详细介绍机器学习中的分类算法——朴素贝叶斯，以及它在文本情感分类中的相关处理过程，同时详细介绍在使用该算法进行文本的情感分类时，文本预处理的步骤，以及如何使用机器学习工具包sklearn实现这种算法，并讨论影响文本情感分析的具体因素和遗留的具体问题。

第五章：详细介绍机器学习中的算法——支持向量机，包括支持向量机的数学公式推导、拉格朗日对偶问题的解决，以及SMO算法的数学原理。除此之外还将介绍如何使用sklearn工具包实现支持向量机并应用到中文文本情感分析中。

第六章：详细讨论卷积神经网络(CNN)的结构层次，包括卷积层、池化层、Relu函数、全连接层等。以及如何使用Tensorflow对网络进行搭建，如何预处理文本来满足神经网络的输入要求，展示使用cnn进行文本的情感分析的具体实验结果。

第七章：主要对比朴素贝叶斯、支持向量机、卷积神经网络的实验效果，以及如何调参优化试验模型，使实验效果达到最优。

## 1.4 本章小结

本章主要介绍了文本情感分析的现实背景、以及应用意义。同时也阐述了实现文本情感分析的三种主流方法的原理以及相应的应用现状，主要包括基于情感词典、基于机器学习、基于深度学习的文本情感分析方法。除此之外，还列举了各个章节的内容安排。本文将着重使用机器学习、深度学习这两种方法实现网络文本的情感分析，同时还会涉及中文分词，生成词向量、文本预处理等相关内容。

# 第2章 基础知识及相关技术

网络文本数据的爬取与数据分析涉及到python网络爬虫、机器学习、深度学习、自然语言处理等相关内容，本文将对网络爬虫技术、自然语言处理中的中文分词，生成词向量、特征提取，机器学习中的朴素贝叶斯、支持向量机和深度学习中的卷积神经网络进行详细介绍。

## 2.1 python网络爬虫技术

网络爬虫作为重要的网络编程技术，被成功应用到搜索引擎技术中，能够代替人快速阅读网络数据，并进行下载。Python语言由于其工具库的存在，能够帮助我们非常方便的从网络上爬取自己需要的数据。

## 2.2 文本预处理技术

文本与处理技术的使用目的是为了得到信息量更多，信息更有用的数据，同时也是本实验中非常重要的一项处理流程。而文本预处理技术主要包括去除脏数据、中文分词处理、停用词处理、平凡词独特词处理等。通过这些方法，能够从海量数据中提取自己需要有价值的数据，同时也能够降低数据的规模，这对于本次实验都有着非常重要的作用。

## 2.3 机器学习算法、深度学习算法

### 2.3.1朴素贝叶斯

贝叶斯决策论是在概率与统计学的背景下，对数据集实施决策的基本方法，，当所有相关概率在已知的理想情况下，贝叶斯决策理论将根据这些相关概率和误判损失进行最优分类，此时分类问题便可以顺利得到解决。

而基于贝叶斯公式如公式（2.1），我们知道。

 （2.1）

——类别的先验概率

——样本A对于类别B的条件概率

——归一化证据因子

——后验概率

这其中，我们把P(B)成为“类别”的先验概率；而p(A|B)是样本A对于类别B的条件概率；P(A)称为用于归一化的证据因子，那么对于给定的样本A，p(A)与类别标记没有任何关系，那么问题：“在已知特征样本A的情况下，求解该样本为B的概率？”就将转换成如何根据训练数据估计先验概率P(B)、条件概率P(A|B)的乘积问题。

类先验概率P(B)表示的是各类样本占总样本的比例，反映了子样本的数量情况。根据大数定律可知，当训练集的样本数量充足时，各个样本满足独立同分布时，P(B)即可用各类样本出现的频率来表示。

类条件概率P(A|B)，表示的是关于所有特征样本A的联合概率，那么当各个特征相互独立的时候，类条件概率便可转换为各个子特征属性的条件概率乘积

此时，当我们假设各个特征属性发生的可能性互独立，那么每个特征属性都将对分类的结果产生影响，这时我们将贝叶斯分类器称为朴素贝叶斯分类器。“朴素”即代表特征之间独立假设的成立。

那么由上述条件可知，朴素贝叶斯的数学公式应用到特征分类领域如公式（2.2）所示。

 （2.2）

由于各个特征属性之间相互独立，那么条件概率便可以写成各个子特征的条件概率之积，而先验概率仍然保持不变。那么朴素贝叶斯分类器的训练流程即是，基于训练集来计算类别的先验概率P(B)，从而估计每一个特征属性的条件概率。

朴素贝叶斯分类器共有三种模型，分别是多项式模型、伯努利模型、高斯模型。三个模型各有特点，也又所区别。其中多项式模型的特征为单词、特征值为该类单词出现的词频占百分比，并且在多项式特点的朴素贝叶斯分类器中，特征向量多为离散型向量，应用于文本分类；伯努利模型以文本为特征，特征值为布尔型数据，标为0或者1；高斯模型中，特征向量是连续性变量，并且假定所有特征的取值是符合高斯分布的。高斯模型适用于连续性变量预测。

### 2.3.2 支持向量机（SVM）

支持向量机算法模型在1995年被提出之后，得到了迅速发展，在分类识别、回归预测等不同方面，均取得非常不错的效果。支持向量机根据其使用的核函数可分为：线性、拉普拉斯、多项式、高斯、Sigmoid类型的SVM。本文就情感分析问题上主要使用的是线性支持向量机，故此下文对线性支持向量机以及其数学推导过程进行详细的总结。

支持向量机算法的使用需要对数据集进行人工干预。在二维平面上，散落着很多数据点，假设数据点仅有两类，那么我们可以找到一条直线对其进行分割，这样便可以使异类的数据点分割开来。同理在三维的空间中我们仍然可以找到一个面，将数据点分割开来，继而将维度上升至n维，那么也必定能够找到n-1的对象将n维中的数据分为不同的类别。这个n-1维的对象称为分隔超平面。在分割的过程中，离分隔超平面最近的点叫作支持向量。在实际应用中，最优的分隔超平面往往会具有最好的分类效果，所谓最优，就是指分隔超平面两侧的支持向量间的距离最大，当满足这个条件时，我们把它称为最大分类间隔超平面。

通过数学建模可知，在二维的情况下，分隔超平面的线性方程如公式（2.3）所示。

 （2.3）

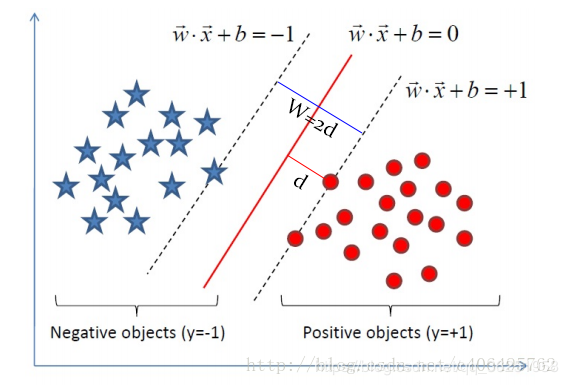
——分隔超平面的法向量

——分隔超平面方程的截距量

那么计算支持向量到分隔超平面的距离如公式（2.4）所示。

 (2.4)

同时，如图2.1所示。



**图2.1支持向量机图解**

对于支持向量来说，其满足的线性方程如公式（2.5）所示。

 (2.5)

因此计算异类支持向量之间的距离如公式（2.6）所示。

 (2.6)

——分隔超平面法向量的绝对值

那么求解最大分类间隔的超平面即为求解两个异类支持向量之间的最大距离，也就是求解W的最小值。为了方便研究W的最值问题以及方便求导，将求解W的最小值转化成½|W2|的最小值，即MIN(½|W2|)。同时对于分隔超平面每侧的数据点，也就是支持向量。均满足下面的关系式如公式（2.7）所示。

 (2.7)

那么给出支持向量机的基本数学关系如公式（2.8）所示。

 (2.8)

在约束条件如公式（2.9）所示。

 (2.9)

在后续的求解过程中，由于问题本身就是一个凸二次规划问题，故此可以使用拉格朗日对偶问题的求解思路进行解决。首先设置拉格朗日α因子，并规定α大于等于0。那么得到的拉格朗日函数如公式（2.10）所示。

 (2.10)

根据拉格朗日函数式，对W、α求偏导如公式（2.11）（2.12）所示。

 (2.11)

 (2.12)

进而推导拉格朗日函数式如公式（2.13）所示。

 (2.13)

同时问题转换成如公式（2.14）所示。

 (2.14)

其应该满足的条件如公式（2.15）所示。

 (2.15)

则问题由原来的拉格朗日函数式问题转换成求解合适α，使关系式取得最大值，那么求解α也就是smo算法问题，smo算法的数学推导如下：

1. 计算误差如公式（2.16）。

 (2.16)

1. 计算上下界如公式（2.17）。

 (2.17)

1. 计算η如公式（2.18）。

 (2.18)

1. 更新αj如公式（2.19）。

 (2.19)

1. 修正αj 如公式（2.10）。

 (2.20)

1. 更新αi如公式（2.21）。

 (2.21)

1. 更新b1和b2如公式（2.22）。

 (2.22)

1. 根据b1、b2更新如公式（2.23）。

 (2.23)

### 2.3.3 卷积神经网络（CNN）

#### 2.3.3.1 传统的神经网络的基本结构、优缺点

传统的神经网络是根据生物的神经网络系统的特点，模仿着生物通过神经网络对现实世界的判断，从而实现算法对于现实问题的交互反应。对于生物神经网络系统而言，神经元与神经元之间的联系依靠神经元分泌的化学物质，上一个神经元分泌的化学物质刺激下一个神经元，并最终激活该神经元，实现神经元之间的信息互通。

传统神经网络的基本框架是：输入层，隐含层（通常包含多层），输出层，而每一层都是由若干的神经元组成的，不同的神经元所携带的信息也不尽相同。输入层作为接收数据的基础，接收数据之后传递至下一层神经元。在传递的过程中，神经元之间经过激活函数变换，传递信息即权重，从而激活下一层的神经元，并这样一直传递下去，最终在输出层得到结果，回归或者分类等等。但是对于传统的神经网络，由于每层的神经元都需要进行传递信息和运算，将导致运算规模较大、运算速度较慢，不满足本次实验的要求。

卷积神经网络（CNN）针对传统神经网络的缺点很好的做了弥补和优化，例如如果我们去识别一张“狗”图片，我们不需要分析每一个像素点之后，才能得出这是一只狗的结论，如果查看眼睛、耳朵等部位也可以得出同样的结论。那么这样无论速度、还是准确率都将获得提升。卷积神经网络的正是通过卷积运算，将输入的数据重要特证提取出来，从而实现通过局部特征进行分析，得出结论。

#### 2.3.3.2 卷积神经网络基本结构：

卷积神经网络的基本构成包括卷积层、池化层、全连接层，dropout层。下面将分别进行介绍：

1. 卷积层的作用就是类似于特征选择器，通过卷积核与原特征矩阵进行卷积运算，从而提取原特征矩阵中的关键特征元素。这样也就解决了传统神经网络的参数、特征太多问题。同时卷积运算的公式（2.24）如下：

 （2.24）

其中n为原矩阵维度，f为卷积核大小，p为填充大小，s为步长。在这里p即填充大小，填充的目的在于保证经过卷积运算之后得到的矩阵大小仍然和原输入矩阵大小相同。S作为步长，规定了卷积运算时横向、纵向移动的长度。

1. 池化层可以理解为在卷积层经过卷积运算的基础上，将运算后得到的数据输入到池化层，而池化层通过既定的运算规则，进一步的提取重要的特征，从而降低运算的复杂性，将准确率提高。池化层进行池化的方式具体有两种，一种是取最大值MaxPooling对局部数据进行代替，另一种是取平均值Avgpooling代替局部数据。前者是根据步长，去相应矩阵中的最大值进行合并，后者是根据矩阵元素的平均值，作为新的代表元素进行合并。
2. Dropout层是研究人员为了避免在训练模型、测试数据的时候呈现“过拟合”现象而选择使用的技巧，通俗来讲就是通过设置概率阈值，每次训练的过程中按照一定的概率丢弃相应的神经网络，最终使训练的网络减少，通过实验验证，这种做法对于降低过拟合很有帮助。
3. 全连接层作为卷积神经网络的最后一层，也就是输出层，对于问题的解决具有重要意义，在这一层上，之前经过卷积、池化后的特征矩阵将在全连接层上做最后的运算，最终得到我们需要的结果。以上便是卷积神经网络的基本构成，在后文中，将会介绍在实验的过程中，如何搭建具体的卷积神经网络，以及怎样使用神经网络进行情感分析。

## 2.4 本章小结

在本章中，第一部分首先讲述了python网络爬虫技术，以及本文爬取数据的具体逻辑，之后对于爬到的电影影评数据进行了预处理，去掉了其中的英文、字符、表情、标点，这样整个数据集中便只剩下了由纯中文构成的中文影评文本。那么为了文本向量化，本章又讲述了使用jieba中文分词工具对预处理过后的文本进行分词，从而得到分词后的评论文本。而对于分词后的评论文本，还讲述了如何使用去除停用词，也就是对于感情色彩并不重要的词或字。从而得到最终的数据集。第二部分本章具体的介绍了机器学习算法、深度学习算法的特点。同时也对这些算法中的数学推导部分做了详细的阐述。

# 第3章 网页爬取和数据预处理

本节将对本次实验的网络爬虫部分和数据与处理部分进行详细的阐述，同时也会对爬虫的逻辑流程通过流程图来展示，除此之外，具体的数据预处理方法也会被详细的介绍到。

## 3.1 网页爬取

本次实验选取带有感情色采的电影评论语句作为数据集，通过使用python网络爬虫技术爬取了数据共四万余条。下面将详细的介绍具体的爬虫逻辑。

本文使用python-requests模块对豆瓣电影网站进行了爬取，共爬到电影评论数据4万条左右，数据内容主要包括电影的文本评论、对应评论的官方打分等相关内容。

本文需要情感丰富的文本评论数据作为数据集进行训练分析，而评论文本又不宜过长，故此选择了豆瓣电影的电影短评数据作为本次实验的训练集数据，网站数据如下图3.1所示。



**图3.1** **豆瓣影评页面**

矩形框内便是需要爬去的电影短评数据，箭头指向的便是这条评论的得分数据，而豆瓣影评的文本数据也是依据这一数据进行标注。通过图2.1可以看出，对每一部电影豆瓣都存在着分为1、2、3、4、5星的电影评论，这些评论的星数也恰恰说明了这些电影短评的感情倾向，本次实验以3星为中介，对1、2星的评论数据标记为消极的情感倾向，而对4、5星的电影短评数据标记为积极的情感倾向。那么这样对于整个爬虫数据的分类别已经完成。下文将展示本次实验的爬虫的逻辑。

具体的爬虫逻辑如图3.2所示。



**图3.2 爬虫逻辑**

对于上文中获得的大量豆瓣电影评论数据，要进行特殊字符以及英文字符的处理，本文使用的工具包括正则表达式、BeautifulSoup等工具，同时还要根据评论的得分情况，对其进行人工标注，1,2标注为0，表示为消极情感。4和5分标注为1，用来表示为正向情感。

## 3.2 数据预处理

### 3.2.1去除脏数据

由于爬虫爬取下来的网络数据包含大量的HTML标签、表情、符号、以及包括很多英文字符，这些字符的存在将大大降低情感分析的精准度以及效率。本文使用Python中自带的Re模块即正则表达式模块对英文字符、表情、符号进行剔除，同时借助于BeautifulSoup过滤HTML标签。

### 3.2.2中文分词处理

英文较中文来讲，由于空格的存在，天然不需要分词这一处理。而中文则不同。目前中文分词的工具包包括jieba、THULAC以及北大开源工具包pkuseg。目前分词工具实现的原理基本是基于规则、统计、语义、理解这四种方式实现。而本文使用jieba以统计为特点，共有三种模式：精确模式，全模式和搜索引擎模式。而这三种模式都能够对文本进行分割，精确模式能够将句子精确地切分开，特别适合做文本分析。全模式虽然能够很快的扫描句子词语，但也会出现句子、词汇歧义问题。搜索引擎模式，常常应用于搜索引擎分词。本文的分词采取的是“精确模式”，这样利于文本的情感分析。

### 3.2.3停用词处理

停用词是一些不会影响当前句子感情的词或字，在做自然语言处理的时候，更多的采取将这些词过滤掉，从而提高后续工作的效率。目前对于停用词的处理，更多的都是通过查找的办法进行，通过遍历的方式去除停用词。本文选用了哈工大停用词表，对豆瓣影评分词后的数据进行了停用词过滤。本文以5000条影评数据为例，对比了在未去除停用词和去除停用词之后的特征数，如图3.3所示。



**图3.3 停用词处理效果图**

可以看到再去除停用词后，特征数由26748下降到26517，一旦实验遇到大规模的数据，去除停用词便显得尤为重要。

### 3.2.4平凡词、独特词处理

平凡词指的是在众多文本数据中均出现的词语，例如“电影”，由于出现的频数过高，其所代表的情感意义也就不是很突出，过于平凡。独特词指的是，在众多文本数据中某个词语仅出现很少的次数，也就是仅在少数文本数据中存在，那么它所代表也只是少数的特点，并不具备说服力。为了降低特征矩阵的特征数，改善实验的准确率，本文采用统计词语在文件中出现的次数frequence、词语的文件占比率rate两项数据作为参考指标，设置阈值，将词语的文件占比率rate超过0.8，词语在文件中出现的次数frequence少于10次的词语全部剔除，极大的降低了特征数。如图3.4所示。



**图3.4 平凡词、独特词处理效果图**

可见，在设置阈值，剔除平凡词和独特词之后，特征数成功下降到原来的十分之一。

# 第4章 朴素贝叶斯算法在中文文本情感分析中的应用

## 4.1 贝叶斯定理与朴素贝叶斯

根据统计学相关知识可知，过去由于数据量过于庞大，而计算量过于繁重，贝叶斯定理在当时并没有引起人们的注意，如今，计算机的出现不但加快了数据的计算速度，同时也扩大了数据的运算规模，而贝叶斯定理也深入的应用到各个领域，在机器学习领域，朴素贝叶斯算法也成功的应用到文本分类、垃圾评论过滤、以及预测分析的问题上。

朴素贝叶斯算法，是在贝叶斯定理的基础上，提出相关假说，并将贝叶斯公式成功进行了应用。在原贝叶斯定理的基础上，如果假设问题的各个特征相互独立且发生的概率互不影响，此时，贝叶斯定理也就精确为朴素贝叶斯。

### 4.1.1 贝叶斯公式

根据统计数的相关知识，可知贝叶斯公式如公式（4.1）所示。

 （4.1）

P(A|B)、P(B|A)均表示条件概率，而对应的含义分别为：

P(A|B)：在事件B发生的条件下，事件A发生的概率

P(B|A)：在事件A发生的条件下，事件B发生的概率

其中，P(A|B)称作后验概率，P(A)称为先验概率，P(B|A)/P(B)称为可能性函数，又称做影响因子。那么由以上含义，条件概率便有了新的阐释方式：

条件概率 = 可能性函数（影响因子） \* 先验概率

### 4.1.2 朴素贝叶斯在文本情感分析中的具体应用

而在具体的文本情感分析、垃圾邮件分类等的问题中，假设问题的各个特征之间互不影响、相互独立，那么此时贝叶斯公式便发生适当转化，也就是朴素贝叶斯公式。

在文本情感分析的过程中，将分词后的句子的每个词作为该句子的特征，并且这些特征出现的可能性均相互独立。那么此时问题便转化成：

“在已知特征的情况下，求解对应情感类别的概率”，公式表达如下：

P(类别|特征)

那么根据贝叶斯公式可知：已经提到特征是不止一个的，那么对于一句中文文本如何使用朴素贝叶斯公式求解呢？

并且已经假设各个特征之间相互独立，且互不影响，那么此时由朴素贝叶斯公式可得到公式（4.2）。

 （4.2）

可见，在特征特别多的文本情感分类问题上，朴素贝叶斯仍然适用。

## 4.2 朴素贝叶斯中文文本情感分析处理流程

基于机器学习算法朴素贝叶斯的中文文本情感分析处理流程主要包括：中文文本预处理、停用词处理、自定义词典、中文分词、TFIDF特征提取、中文词性标注、词频统计、训练数据、验证数据。下面将详细介绍。

### 4.2.1 中文文本预处理

由于数据集在爬取的过程中，难免出现夹杂冗余数据的情况，而冗余数据主要包括：标点、符号、表情、英文等。这些词的存在，不仅加大了算法处理数据的负担，同时也大大降低了文本情感剖析的准确率，所以在进行数据训练之前必须过滤掉这些脏数据。本文使用的方法是利用python的正则模块re，匹配英文、标点符号、表情，并去除掉这些数据。

### 4.2.2 停用词处理

什么是停用词，停用词是自然语言数据（或文本）处理时对一句话或一短文本并未实质意义的词或字，这些字或词即被称为Stop Words（停用词）。可见，停用词在文本的情感分析上面，并不具备很好的情感倾向，故此完全可以去除掉停用词，这样也就降低了这些词对于情感分析的影响，从而提高本次实验的准确性。以下面的句子为例：

“电影情节复杂，汇聚了不同的元素，这部电影说实话很好看。导演很了不起”

在这句中，停用词包括“的”、“和”、“啊”可见这些词或字并没有对这句文本的情感正负性产生任何影响，因此还是应该去掉，从而降低算法训练数据的压力。

### 4.2.3 中文分词

由于中文使用标点符号，分隔句子，这种特点便加大了词与词之间的联络，为了分析中文文本的情感，必须将句子分隔成不同的词语，这样每个词语作为该文本的特征，而整个句子也被分割成了多个词语，这样一个中文句子便被分割成了若干个词语。例如该句：

“结局真的出乎意料韩寒导演他给我们青春的回忆啊”

进过中文分词处理后：

“结局 真的 出乎意料 韩寒 导演 我们 青春”

只有经过分词处理后，一个中文文本的句子才可以转换成向量矩阵进行运算，从而对中文的情感剖析做下一步的处理。

### 4.2.4 自定义词典

自定义词典的原因主要是：无论多么好的分词工具，也不可能将自己的数据集中的中文文本全部准确无误的分隔成词，比如从上文我们可知，“韩寒”作为具体的人称被分词工具成功的发现，但是“沈腾”却并没有被发现并提取出来，这样也就造成了分词的误差，而太多的分词误差便会导致整个文本情感分析的准确性，因此我们必须自己根据数据集的情况手动添加词语进入自定义词典，并在分词工具分词前加载进去。这样便能够使分词按照我们的意愿进行，也防止了一些词汇被误分隔，一些词汇没有被发现的问题的出现。具体例子如下：

未定义词典：

”打败 自己 的 只有 自己 韩寒 沈 腾 给 我们 青春 的 回忆”

自定义词典后：

“打败 自己 的 只有 自己 韩寒 沈腾 给 我们 青春 的 回忆”

可见，沈腾作为名词被识别出来了。

### 4.2.5 TF-IDF特征提取

当中文文本被分词处理后，得到了若干词组成的向量，，而对于确定文本的情感倾向，需要进一步提取文本的特征。分词后将得到很多词汇，这些词汇有的对于句子的情感有作用，有的则无足轻重，那么在大文本数据量的背景下，特征维度过大，导致算法的运算时间过长，对实验的结果将会产生很大的影响。而如果只考虑词频，按照词频的高低选取固定数目的词作为该句文本的特征，那么便会出现如下情况：

“我今天非常不开心，因为我的书包丢了”这句话经过分词后，得到如下：

“我 今天 非常 不 开心 ， 因为 我 的 书包 丢 了”，可知“我”这个字出现了两次，是整个句子出现频率最高的词，不过如果只选择我作为该句话的特征，也就是仅考虑词频高低的情况下，实验的效果也会受到影响。

TF-IDF算法是我们常说的“词频-逆文本频率”。TF-IDF的主要思想也就是，如果一个词在当前文本中出现的次数很高，即TF很高，且其在所有文本中出现的次数很低，也就是IDF很高，那么我们认为这个词具备很好的分类能力，故此将其提取出来。

本文使用机器学习工具包sklearn中的TFIDF模块实现文本的向量化，并且根据每个词的TFIDF特征值，提取相关词作为当前语句的特征，从而实现降低词向量维度，同时提高准确率。

### 4.2.6 中文词性标注

词性标注主要是对词汇进行词性的查找及标注，也就是在分词的之后，根据词性表将分好的词语标注词性，目前比较权威的汉语词性表是ICTCLAS汉语词性表，它主要将词语的词性归类为：

\* 实词：名词、动词、形容词、状态词、区别词、数词、量词、代词

\* 虚词：介词、副词、、助词、连词、叹词、拟声词

本文在ICTCLAS汉语词性表的基础上，使用中文分词工具包jieba分词工具，在分词的同时，也对词语的词性进行了标注，更好的区分了动词、副词、形容词等等。具体词性标注效果如下(仅列举部分)：

表演：v 台词： n 没： v 刻意： v 营造: v 笑点: n 燃点 : n

### 4.2.7 词频统计

对于词频的统计，在TFIDF特征值处理的时候，已经计算过了，此处不再赘述。

### 4.2.8 训练数据、验证数据

完成上述步骤之后，本文采用了朴素贝叶斯算法，也属于机器学习算法中的一种，利用算法模型训练数据，通过大量数据的训练，让算法模型的学习能力增强，并使用测试集检验模型的效果，最终得到相关的指标。本文采用两种方式实现朴素贝叶斯训练数据。下文将详细介绍，与此同时本文还选择了准确率、召回率、以及F-score、学习曲线作为模型的评判标准。由混淆矩阵可知：

True Positive(真正，TP)：将正类预测为正类数

True Negative(真负，TN)：将负类预测为负类数

False Positive(假正，FP)：将负类预测为正类数误报 (Type I error)

False Negative(假负，FN)：将正类预测为负类数→漏报 (Type II error)

准确率指得是经过模型训练、分析、预测得到的分类正确的信息数与总信息数的比值，准确率越高，说明该分类器的分类效果就越好，分类能力越强。准确率计算如公式（4.3）所示。

 （4.3）

召回率指的是经过模型训练、分析、预测得到的正确的分类数量与总共的正确数量的比值。召回率计算公式如公式（4.4）所示。

1.  （4.4）

score是指当准确率、召回率发生冲突时，对二者进行综合的考虑得到的指标。F-score的具体计算公式如公式（4.5）所示。

 （4.5）

其中α为调和参数

本文先后采用了自己完成的朴素贝叶斯算法进行分类训练，预测分析和使用机器学习工具库sklearn朴素贝叶斯分类模块进行分类，具体的分类效果以及评判标准如下：

1. 使用python编写实现的朴素贝叶斯算法，并在训练集上面训练模型，进行测试验证。本文以9893条豆瓣影评进行训练，2000条豆瓣影评进行测试，得到的最终结果如图4.1所示。

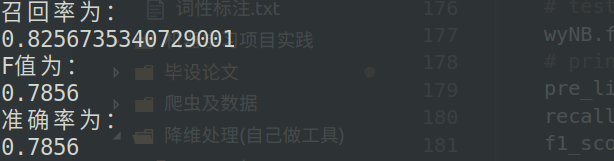


**图4.1 朴素贝叶斯**

可以看到，训练完的模型在测试集上面的准确率在87%左右。不过自实现的算法在大数据量的情况下，算法运行的时间将会大大增加，而且实验最终的准确率也大幅度降低。这体现了自实现朴素贝叶斯算法的不完备性。

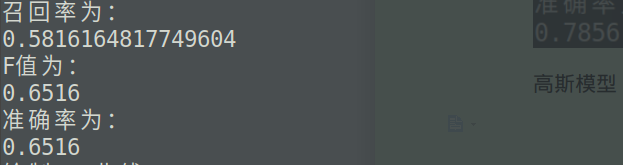
1. 使用sklearn实现朴素贝叶斯算法，并根据现有数据集情况应用算法。作为Python的强力的机器学习库，sklearn对于数据的加载,数据的预处理都做了很好地封装，同时非常方便就可以完成数据的训练与测试。在上文的论述中，我们了解到朴素贝叶斯共有三种模型，分别是多项式模型，高斯模型，伯努利模型，而这三种模型都有着不同的应用领域。多项式模型，更多的用于特征是离散的情况；高斯模型更多的处理连续特征的情况；与多项xi式模型类似，Bernoulli 模型适用于特征是离散的情况，与其去别的是，Bernoulli 模型中每个特征的取值只能是1和0。而对于中文情感分析问题，本文分别使用了三种模型对其分析效果进行了对比。主要对比如下：
2. 准确率、召回率、F-Score对比：

多项式模型结果如图4.2所示。



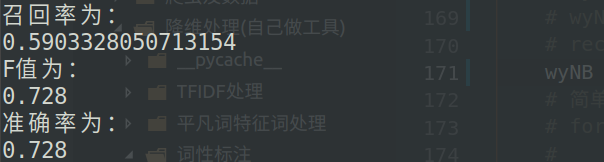
**图4.2 多项式模型结果**

高斯模型结果如图4.3所示。



**图4.3 高斯模型结果**

伯努利模型结果如图4.4所示。

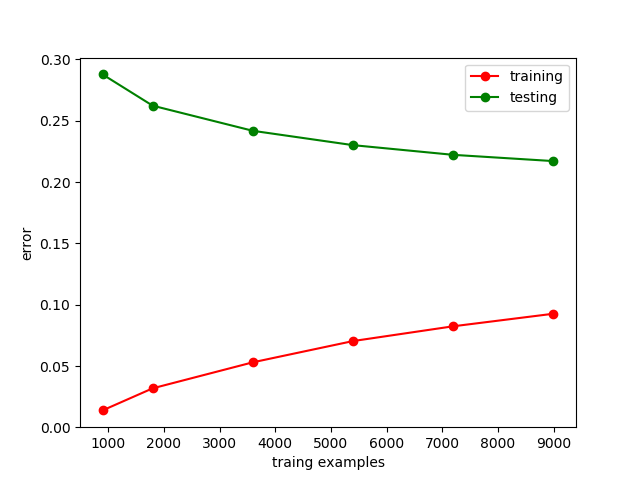
****

**图4.4 伯努利模型**

通过对比可以看到，无论是准确率、还是召回率、亦或是F-Score指标，都是多项式模型更适合做情感分析。

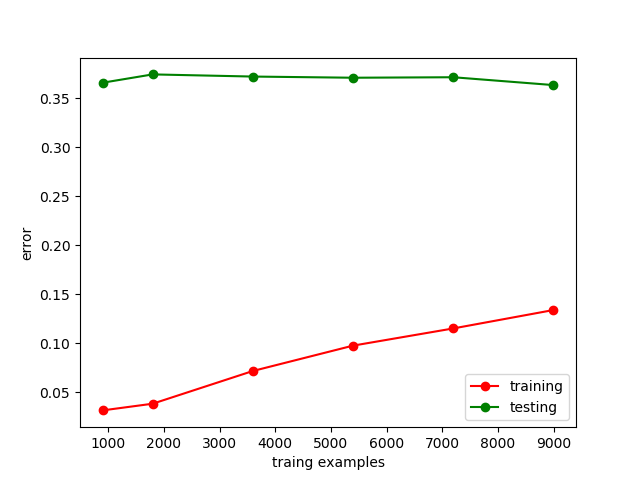
1. 通过学习曲线，对比拟合程度

多项式模型学习曲线如图4.5所示。



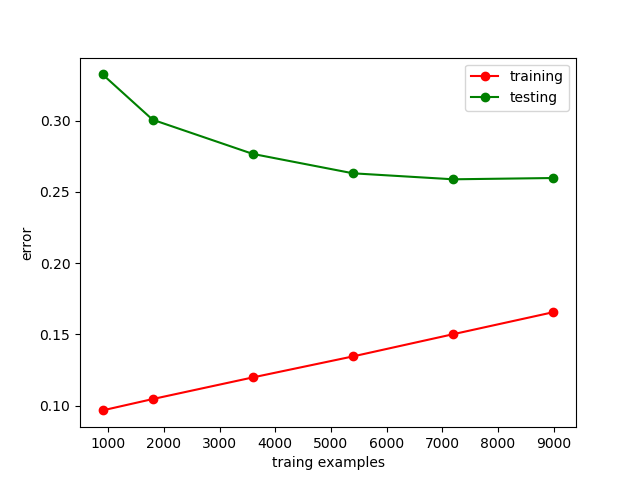
**图4.5 多项式学习曲线**

高斯模型学习曲线如图4.6所示。



**图4.6 高斯模型学习曲线**

伯努利模型学习曲线如图4.7所示。



**图4.7 伯努利模型学习曲线**

通过图4.1、图4.2、图4.3、图4.4、图4.5、图4.6、图4.7分析可知：

高斯模型在训练集逐渐增大的情况下，训练集的出错概率也在逐渐增大，而测试集准确率也并没有随着数据的增大有所改善。伯努利模型和多项式模型，测试集、训练集的学习曲线均有着相同的趋势。二者均随着数据量的增加而不断收敛至一个标准值，且由图可以判断二者均出现了欠拟合问题，不过就准确率、召回率的高低比较而言，多项式模型具有更好的实验效果。

## 4.3 本章小结

在本章中，具体介绍了朴素贝叶斯算法如何应用到中文文本的情感分析中来，同时在4.2小节中，对朴素贝叶斯情感分析的处理流程本章进行了详细的阐述。主要包括文本预处理、特征提取、训练数据、测试数据等。并在最后分别展示了朴素贝叶斯的三种模型高斯模型、伯努利模型、多项式模式分别用于情感分析的效果对比，主要包括召回率、准确率、F-Score等。除此之外，还分别绘制了学习曲线从而查看不同的模型的拟合程度。同时本章也对比了不同数量的数据集对于实验效果的影响，可以看到在增大数据集之后明显实验效果会改善。最后本章介绍了sklearn工具包，以及使用其完成朴素贝叶斯算法的编写和应用等。

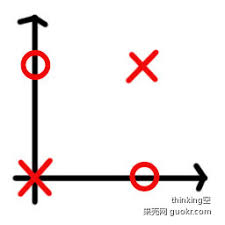
# 第5章 支持向量机算法在中文文本情感分析中的应用

机器学习算法支持向量机在中文文本情感分析中的主要流程为：文本预处理、中文分词、向量化、特征提取、svm训练数据、测试数据等。而文本预处理、中文分词、向量化、特征提取这些部分的内容均与朴素贝叶斯中文文本分类的步骤相同，在这里便不再重复赘述。本文使用机器学习工具包sklearn实现svm分类模型的训练，并在测试集上进行了测试，也取得不错的效果。

## 5.1 支持向量机·核函数

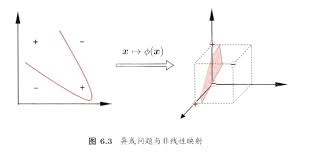
对于支持向量机的基本定义、最大间隔、支持向量以及拉格朗日对偶问题的使用等已经在上文详细介绍过，在此不做过多赘述。而对于支持向量机而言，所使用的核函数不同，得到的分类效果也不尽相同。

在之前讨论的支持向量、以及分隔超平面的相关内容，我们首先是在假设数据点线性可分的基础上进行推导的。那么如果数据点并不是线性可分的呢，例如图5.1异或问题。



**图5.1 异或问题**

对于这种问题我们不可能找到一条超平面或者直线将数据点分割成两个类别，那么为了解决这类问题，我们引入映射函数将原样本空间中数据点映射至更高维的特征空间中，这样做的目的是使数据点在更加高维的特征空间中依然是可分的。例如图5.2，异或问题便得到了解决。



**图5.2 核函数解决异或问题**

而在上文中使用的映射函数被称为核函数，可想而知在使用svm支持向量机算法做文本的情感分析的时候，我们最终的目的是能够通过算法找到分隔超平面从而将数据点分为两类，这样的做法也就相当于找到优质的特征空间，而特征空间的选择与核函数息息相关，故此适当的选择核函数对于文本的情感分析至关重要。SVM支持向量机的常用核函数有线性核，多项式核，高斯核等。下面详细介绍这些核函数：

1. 线性核函数：

线性核的表达式如公式（5.1）所示。

 （5.1）

处理线性可分的数据点样本时更多地采取线性核函数，如果输入的数据点是线性可分的。那么输入线性核函数后形成的特征空间与原样本空间是完全一致的。

1. 多项式核函数：

多项式核的表达式如公式（5.2）所示。

 （5.2）

类似于解决异或问题，为了将原样本空间映射成高维的特征空间，此时多项式核函数非常合适。从而实现原样本点不可分转换成在特征空间内可以分割的情况，对于多项式核函数来说，运算较线性核来说更加复杂，运算的规模更大，速度更慢。

1. 高斯核函数：

高斯核函数的表达式如公式（5.3）所示。

 (5.3)

高斯核函数也可以将原样本空间的数据点，通过高斯函数映射到高维的特征空间，不过与多项式核函数相比，高斯核函数的参数更少，并且实用性更强，无论大样本数据还是小样本数据都可以取得非常不错的效果。

## 5.2 不同核函数的SVM文本情感分析对比

本文共使用了三种核函数对svm中文情感分析进行了实验，并取得了不同的效果。

### 5.2.1 线性核函数实验效果：

**注：以下实验均使用豆瓣影评为数据集，其中训练集5000条，测试集1200条**

使用sklearn，设置参数,kernel = 'linear'，即为线性核函数。使用线性核函数时的情感分类效果如图5.3所示。



**图5.3 线性核函数实验结果**

可以看到，分类的准确率达到了74.08%，召回率达到了71%左右。

### 5.2.2多项式核函数实验效果

实验结果如图5.4所示。



**图5.4 多项式核函数实验结果**

在同样的数据集下如果和函数采用多项式核函数的话，准确率下降到了0.48，召回率达到了100%，这并不能体现分类的结果达到了很好的水平，因为召回率反应了该分类模型对正向情感分类的效果非常好，但是准确率不足一半又恰恰说明了该分类模型对负向也就是消极的情感并不具备很好的辨识能力。可见多项式核函数并不适合此次中文文本情感分类。

### 5.2.3 高斯核函数实验效果

实验结果如图5.5所示。



**图5.5 高斯核函数实验结果**

由图5.5所示，高斯模型在同样的数据集下的分类效果和多项式核函数模型相同，故此可以推断，在5000训练集，1200测试集的情况下，线性核函数的分类效果更加优越。当然，在后文的调参优化、增大数据集操作之后，高斯、多项式核函数也会展现出改善的效果。

## 5.3 本章小结

本章主要介绍了支持向量机（SVM）基本结构，在5.1中本章主要介绍了支持向量机的核函数以及核函数的用途，同时讲述了svm的三种主要核函数。在下一节中，本文使用sklearn实现不同核函数的支持向量机的中文文本的感情分析，最终得到不同的效果，可以从上文的图片得知，只有线性核达到了本次实验的基本要求。

# 第6章 卷积神经网络算法在中文文本情感分析中的应用

在第三,第四章的内容中,主要讲述的是在机器学习的基本算法基础上，从而对中文文本的情感进行分析,主要的算法包括朴素贝叶斯,支持向量机。本章将使用非传统方法来对中文文本的进行分析，也就是使用卷积神经网络对现有数据集——豆瓣电影评论进行分析。下面将详细介绍具体流程。

## 6.1 文本预处理

前文中无论是使用朴素贝叶斯、还是使用支持向量机算法，都需要对数据集进行文本预处理，而使用的深度学习中的卷积神经网络进行分析时，同样是需要进行文本预处理的，不过此处的文本预处理与前文并不一致，在本章中，将使用两种方式实现卷积神经网络，一种是基于字粒度，而另外一种是基于词粒度。当然实现的方法不同，对应的文本预处理也就不一致。

### 6.1.1 基于字粒度的文本预处理

之所以是基于字粒度，原因是不对影评数据集做分词处理，而是将训练集中所有评论的数据逐字进行统计，提取出不重复的所有字，作为字符表，用于后期生成每个影评数据的向量。具体的做法如下：

1. 读取MongoDB数据库中的影评数据，同时剔除重复数据，不能将重复的数据作为训练集，这样将导致数据测试的准确率下降。利用正则表达式、BeautifulSoup等工具剔除文本中掺杂的标点符号、英文数字。
2. 将处理好的文本数据按照相应比例分为训练集、测试集、验证集写入不同的文件，用于后期的训练验证和测试。
3. 为了统计每个字出现的频率，需要遍历训练集数据，使用Counter类过滤重复汉字，并按照频率高低进行排序，生成字典的形式，也就是key为汉字，value为频率。写入文件中，形成全部的字符表。为后期的训练时生成向量做准备。

以上便是基于字粒度的文本预处理流程，最主要的目的就通过预处理得到全部的不重复字符表，以作为后期向量生成的标准和参考。

### 6.1.2 基于词粒度的文本预处理

基于词粒度的文本预处理，便是不在以字符作为输入单位，而是将影评数据进行分词，得到由不同词汇组成的文本，然后统计全部的不重复词作为词汇表。

1. 读取MongoDB数据库中的影评数据，同时剔除重复数据，不能将重复的数据信息当做训练集，这样将导致数据测试的准确率下降。利用正则表达式、BeautifulSoup等工具剔除文本中掺杂的标点符号、英文数字。
2. 将处理好的文本数据按照相应比例分为训练集、测试集、验证集写入不同的文件，用于后期的训练验证和测试。
3. 遍历训练集数据，使用结巴工具分词，统计分词后的词语的频率，使用Counter类过滤重复词语，并按照频率高低进行排序，生成字典的形式，也就是key为词语，value为频率。写入文件中，形成全部的词汇表。为后期的训练时生成向量做准备。

词粒度的做法更能很好的体现评论语句与词汇的联系性，分词后得到的词汇

表的词汇量要远远大于基于字符粒度的总量。也就是说在使用分词后的训练数据时，神经网络的维度要变得更大。

## 6.2 中文文本情感分析时卷积神经网络的结构

在第二章时，本文讲述了卷积神经网络的框架构成，包括输入层、卷积层、池化层、全连接层、输出层。那么在本章中，将会详细的介绍，在本次实验的流程下，如何搭建符合数据集情况的网络框架。

1. **词向量嵌入层**

词向量嵌入层作为是本文自定义的网络的第一层，其作用便是将通过文本预处理得到的词汇表，将每句影评文本通过建立的已知词汇表，映射成相应维度的词向量矩阵。由于影评文本的长度不一，那么得到的词向量的形状便不一样，这样的话不利于之后的数据批处理，故此可以采用设置阈值，同时使用keras中的特殊标记<PAD>，自动的将文本的长度扩大到制定阈值，这样也就保证了每个文本的映射的词向量长度都是固定的，对数据的批处理阶段会更方便。

1. **卷积层、池化层**

在上文中，我们已经提到过卷积层的作用相当于特征提取，而卷积层所做的运算主要是卷积运算。为了选取更能代表数据样本的特征，卷积运算十分重要。在本文中，设定卷积核数目为256个，同时指定卷积核大小为5。同时为了进一步的提取特征，降低噪声的影响，避免过拟合，池化层的设置也必不可少。

1. **全连接层、Dropout层**

上文也记录了全连接层、Dropout层的详细作用，分别是最后的运算和丢弃部分神经元防止过拟合现象。本章中我们设定全连接层神经元的个数为128个，设定分类器，并在参数上赋值为积极、消极两种类别。同时设置dropout比例，最后链接relu激活函数。主要代码如下：

1. **定义损失函数、设置准确率函数**

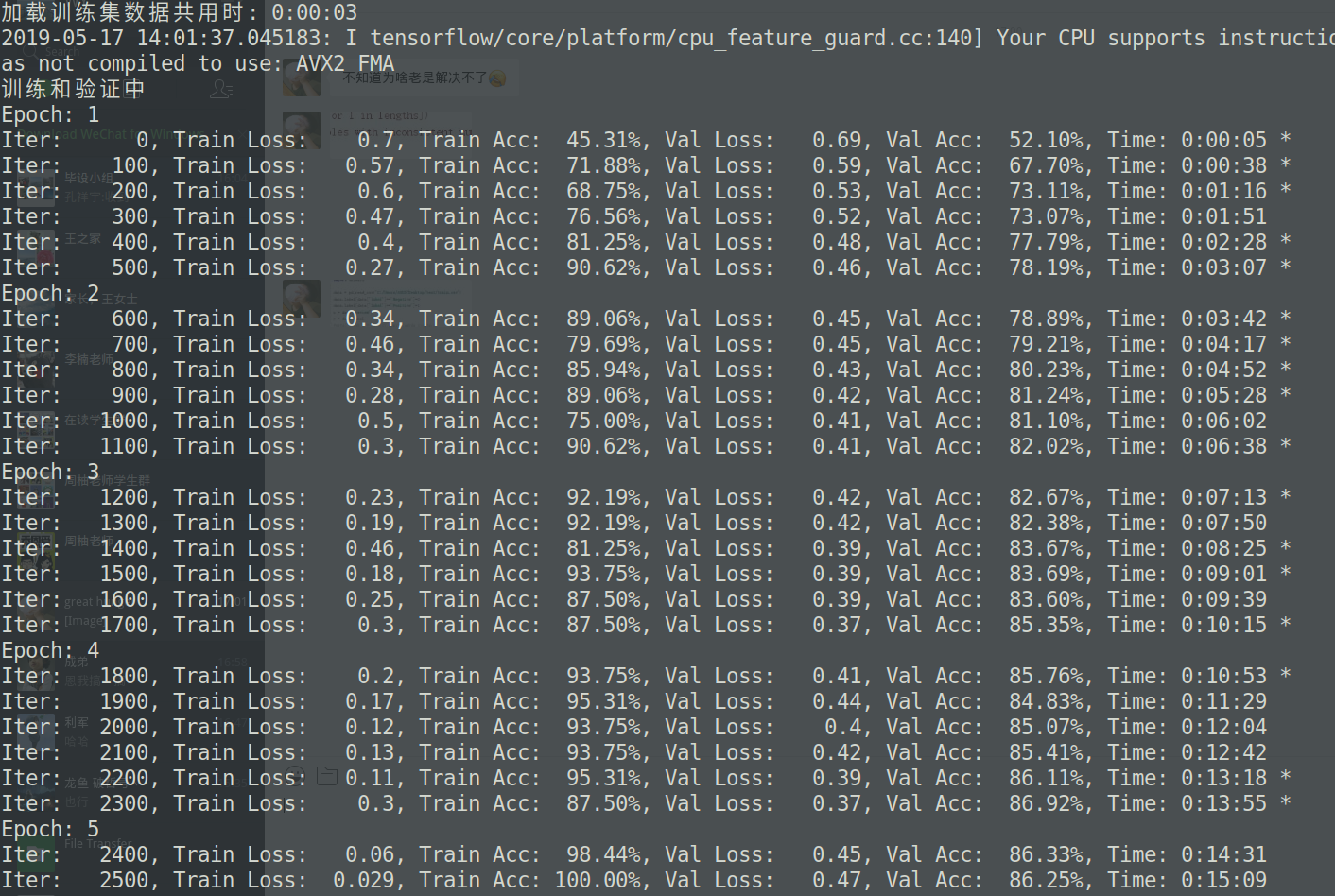
损失函数是对当前训练的模型所造成误差的一种度量方式，而度量值越小说明模型的误差越小，也体现模型对于当前分类问题的适应性。为了度量误差的大小，我们使用交叉熵损失函数来对模型的误差程度进行评估。定义了准确率函数，目标是在训练阶段、测试阶段记录模型的性能情况。

## 6.3 卷积神经网络训练、验证、测试结果

在上一节中，详细讲述了怎么搭建卷积神经网络。本节在已有网络的基础上，对训练集、验征集数据进行了训练和交叉验证，之后在测试集进行了测试，得到的具体分类指标如下：

1. **数据训练阶段**

训练情况如图6.1所示。

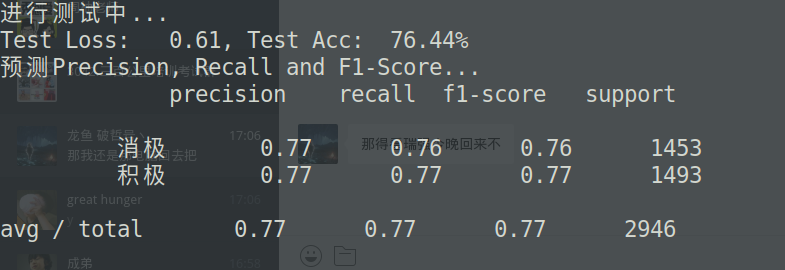


**图6.1 神经网络训练数据**

从图中可以得出结论，随着不断地迭代训练，训练集的损失不断减小，训练集的准确率越来越高，可见模型在训练集上的学习能力逐渐增强；但可以看到，验征集的准确率虽然也在一直提高，但与训练集相比还要差很多，除此之外，验征集的损失也一直在0.45左右。

1. **测试阶段**

测试的结果如图6.2所示。

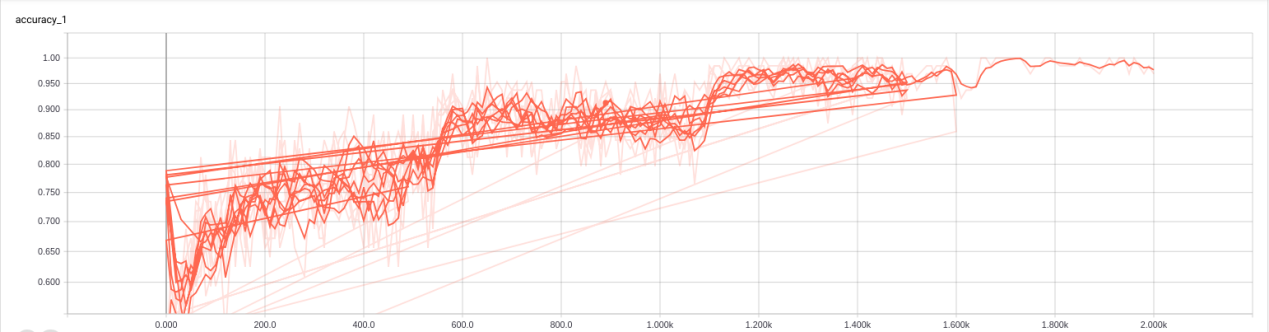


**图6.2 神经网络测试数据**

从图6.3.2能够看到，模型在测试集上的准确率为76.44%，并且损失率也达到了0.61，结合验征集的验证指标，可以得出结论：经过训练集训练的模型并不被具备很好的泛化能力，这说明模型在训练的过程中学习的太好，将噪声当做数据学习了，最终可能会出现过拟合问题。

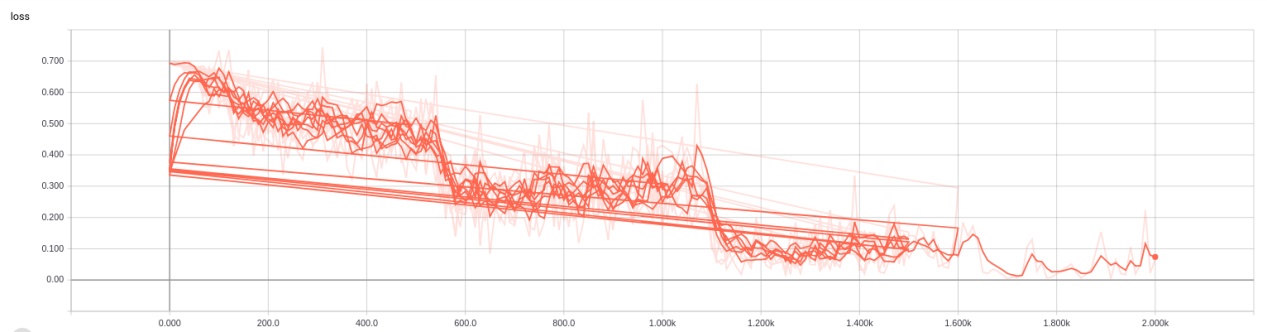
1. **Tensorboard绘制的曲线**

Accuracy曲线如图6.3所示。



**图6.3 准确率函数**

Loss曲线如图6.4所示。



**图6.4 损失函数**

虽然模型的训练准确率总体趋势为逐渐增大，损失率总体趋势为逐渐减少，不过准确率的波动很大，说明模型的稳定性和泛华性并不是很好。

## 6.4 本章小结

在本章中，讲述了如何在已有数据集的基础上，对卷积神经网络的网络模型进行设计和实现，以及设置那些超参数会对实验效果产生影响。同时通过使用tensorflow成功搭建词向量嵌入层、卷积层、池化层、全连接层，并对4万余条数据进行了训练、验证、测试，得到了相应的实验结果。同时根据tensorboard训练准确率、损失率图表，可以得出当前神经网络模型还存在待优化的可能，也出现了过拟合的问题有待解决。

# 实验结果对比与不足之处

## 7.1 实验结果对比

本次毕业设计，共使用了三种方法对中文文本的情感进行了分析和预测。对于这三种方法，在实践的过程中都出现了不同的问题，最终也都呈现出不尽相似的实验结果，那么下文将从几个不同的方面进行分析和比对。同时，由于数据集、模型等存在的问题，整个实验也出现了若干问题，最后本文也将会对这些问题进行总结。

### 7.1.1 对比准确率、召回率、F-Score

具体实验结果和实验数据将会以图表的形式在下文中展示出来。

1. 基于机器学习算法的实验结果：

在上文的介绍中，可以看到在此实验中多项式模型的朴素贝叶斯算法得出了更好的效果，从准确率、召回率、还有综合考虑的F-Score,MultinomialNB模型根据优越性。

同样在上文中，当核函数指定为线性核函数时，达到了最好的分类效果。二者的具体情况表7.1所示。

表7.1 实验结果对比

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | 朴素贝叶斯实验 | 支持向量机实验 | 卷积神经实验 |
| 准确率 | 0.793 | 0.748 | 0.77 |
| 召回率 | 0.825 | 0.711 | 0.77 |
| F-Score | 0.793 | 0.740 | 0.77 |
|  |  |  |  |

通过对比可知，无论是哪个指标，在同等数据集下朴素贝叶斯的分析效果都要更优越于支持向量机。

1. 基于深度学习算法的实验结果：

在上文中，也给出了卷积神经网络在中文文本情感分析的实验结果，由此表可知，深度学习中的卷积神经网络模型相对于机器学习的两个算法，并没有在三个指标上呈现出更好的结果。

### 7.1.2 数据量对比

虽然在以上的对比中，机器学习的算法要稍稍优于深度学习的卷积神经网络，但这并不足以说明问题。将本次实验从训练数据量、训练规模的角度考虑，卷积神经网络的表现更好。如下描述：

1. 朴素贝叶斯算法当数据量达到30000条以上时，无论是算法的运算速度、准确率都会出现下降问题，同时根据学习曲线可知，模型的学习能力已经出现下滑。
2. 支持向量机算法的数据运算规模最高不能超过5000条（仅限于本实验），那么除此之外，当数据量增加时，会出现内存溢出，当然这也跟硬件设备有直接关系。
3. 卷积神经网络在实验中表现顽强，由于卷积核运算的特点，当数据量增加至40000条时，模型的学习能力依然在上升，同时硬件内存条件良好。而且逐步增大数据量，神经网络的性能仍然在增加。

## 7.2 问题与不足

本次实验遗留的问题依旧很多，下面将一一列举：

1. 数据集不规范：

本次实验的需求是对中文文本进行情感分析，根据这一需求，本文选择豆瓣电影的短评作为数据集，但是由于其自有的特点，当一条评价的情感是负面的，也就是批评这部电影的时候，由于评价的内容量，价值量更大，这样的评价豆瓣也将其标记为正向情感。这就导致了数据集错误，从而可能导致实验结果偏低。

1. 卷积神经网络模型鲁棒性差：

本文搭建的卷积神经网络，是基于Google的开源框架TensorFlow，由于其语法的复杂性，导致神经网络的搭建并不理想，未能在最后的评价分析上使用L2正则化来降低过拟合问题对实验结果的影响，除此之外，本次试验中由于网络模型健壮性差，数据集不规范的问题，导致神经网络在训练和学习的过程中，出现了过拟合问题，对实验结果造成了很大的影响。

# 参考文献

[1]张尧政,邓少灵.基于文本情感分析的企业网络舆情应对策略比较研究[J].电子商务,2019(05):32-35.

[2]李坚孝,吴家隐,李云锦,范振锷,谢永辉.基于互联网+机器学习的心理健康预警技术[J].电子技术与软件工程,2019(08):148.

[3]刘书齐,王以松,陈攀峰.基于CNN-ATTBiLSTM的文本情感分析[J/OL].贵州大学学报(自然科学版),2019(02)[2019-05-26].https://doi.org/10.15958/j.cnki.gdxbzrb.2019.02.17.

[4]洪巍,李敏.文本情感分析方法研究综述[J].计算机工程与科学,2019,41(04):750-757.

[5]王义,戴月明.基于混合互信息算法的文本情感分析[J/OL].计算机应用研究:1-6[2019-05-26].https://doi.org/10.19734/j.issn.1001-3695.2018.08.0537.

[6]陈洁,邵志清,张欢欢,费佳慧.基于并行混合神经网络模型的短文本情感分析[J/OL].计算机应用:1-6[2019-05-26].http://kns.cnki.net/kcms/detail/51.1307.TP.20190329.1643.008.html.

[7]苏灵松,应捷,杨海马,肖昊琪.双通道卷积记忆神经网络文本情感分析[J/OL].软件导刊:1-5[2019-05-26].http://kns.cnki.net/kcms/detail/42.1671.TP.20190326.0923.014.html.

[8]王娟,张志勋.基于改进主题模型的微博短文本情感分析研究[J].信息与电脑(理论版),2019(06):134-135+141.

[9]龚泽阳,徐华韫,何正杰,崔金真,吴珂.基于深度学习的中英文混合短文本感分析[J].信息与电脑(理论版),2019(05):55-57.

[10]王宁,李世林,刘堂亮,赵伟.基于注意力机制的BiGRU判决结果倾向性分析[J].计算机系统应用,2019,28(03):191-195.

[11]刘毓,赵云阁.基于教学评价的中文短文本情感分析[J].现代电子技术,2019,42(06):30-33+37.

[12]姬晨,郭延哺,金宸,段云浩,李维华.一种基于卷积神经网络的跨领域情感分析[J].云南大学学报(自然科学版),2019,41(02):253-258.

[13]邓存彬,虞慧群,范贵生,朱虎超.燃气客服热线的中文文本情感分析[J].华东理工大学学报(自然科学版),2019,45(01):140-147.

[14]苏秀芝,左国才,张珏.基于深度学习框架的短文本情感分析方法研究[J].数字技术与应用,2019,37(02):80+82.

[15]余传明,原赛,王峰,安璐.大数据环境下文本情感分析算法的规模适配研究:以Twitter为数据源[J].图书情报工作,2019,63(04):101-111.

[16]余传明,原赛,王峰,安璐.大数据环境下文本情感分析算法的规模适配研究:以Twitter为数据源[J].图书情报工作:1-10.

[17]周梁,方兴龙.基于商品评论主题模型的隐含狄利克雷分布研究[J].安徽工程大学学报,2019,34(01):78-84.

[18]李青松.文本情感分析研究[J].现代计算机(专业版),2019(04):21-25.

[19]冒小栋,范涛.基于文本情感分析的共享单车用户满意度研究[J].计算机系统应用,2019,28(01):222-227.

[20]王荣波,沈卓奇,黄孝喜,谌志群.面向中文短文本情感分析的改进特征选择算法[J].杭州电子科技大学学报(自然科学版),2019,39(01):45-50.

# 结论

本文通过对比三种方法对于中文文本的情感进行分析，得出了以下结论：

1. 机器学习算法，深度学习算法都可以用于分析文本的情感，不过从数据的规模，模型的稳定性来说，深度学习的卷积神经网络更具有优势。
2. 朴素贝叶斯算法中，对于文本的向量化操作，利用TF-IDF算法提取特征，构建词向量，得到的是实验结果更好。
3. 不同模型的朴素贝叶斯算法对于实验结果有很大的影响，要根据实验数据集选择合适的模型。
4. 支持向量机算法可以通过调节惩罚系数，调整模型达到最优。
5. 卷积神经网络的搭建过程中，卷积层起到了关键的作用，但是并不是卷积层越多，效果就越好，要反复调整，达到最优。

# 致谢

**致 谢**

行文至此，预示着我的大学生涯也即将结束。回首四年的大学时光，往事依然历历在目，这四年里有过大一的无知、大二的贪玩、也有大三大四的努力与沉淀。可无论怎样，这四年都是属于自己不平凡的四年，都是若干年后回味起来仍然津津有味的四年。

若要用一个词进行总结，那必定是“感谢”一词。

感谢这四年陪伴自己的好友。忘不了与你们一同走过的那些时光，因为它是我人生最宝贵的经历。

感谢于我传道受业解惑的老师们，您们的谆谆教诲我将铭记于心，学生后悔未能珍惜课堂时光，深知为时已晚，却也无力弥补，只能在此祝福老师们身体健康，桃李天下。

感谢我的父母，是你们给了我充足的空间去成长，给了我足够的权利去选择。作为您们的孩子，四年时光，伴您左右少之又少，无以回报，但父亲您对我的要求，母亲您对我的教诲，漫漫人生，孩子绝不会忘。

感谢19岁到23岁的自己，是你的年少无知、懵懵懂懂让我明白了原来吃亏是福，是你的敢于尝试，不愿低人一等的倔强给了我如今的自信，感谢你！

挥别过去，放眼未来，更加艰巨的任务还需要自己去完成，希望自己仍然保有19岁、20岁的满腔热血，21岁、22岁的毅力与坚持，学好本事，放低身姿，以梦为马，不负韶华，成为更好的自己。