**中文情感分类算法**

**实验报告**

组员：蒋浩浩 卢丽华 聂卉 薛景

**摘要：**20世纪初以来，文本的情感分析在自然语言处理领域成为了研究的热点，吸引了众多学者越来越多的关注。对于中文文本的情感倾向性研究在这样一大环境下也得到了显著的发展。本文主要是基于机器学习方法的中文文本情感分类，主要包括：使用开源的Markup处理程序对XML文件进行分析处理、中科院计算所开源的中文分词处理程序ICTCLAS对文本进行分词处理、去除停用词等文本预处理，在基于向量空间模型VSM的文本表示，使用卡方统计量CHI的进行特征选择，TFIDF权值计算方法进行的特征权值计算，最后使用支持向量机SVM进行中文文本情感的分类。在实验过程中，对比了特征数量的大小对情感分类结果的影响、布尔权值、TF权值和TFIDF权值三种不同计算方法对情感分类结果的影响以及SVM分类器对于不同类型文本数据的分类效果。从整个实验结果来看，TFIDF权值计算相较于其他两种更有利于文本的情感分类。SVM分类器作为文本情感分类器对于不同类型的文本数据，其分类效果不同，但总体上取得了较好的效果。

**关键词：中文文本情感分类 SVM分类器 特征选择**

目录

[1绪论 1](#_Toc408826742)

[1.1 国内外研究现状 1](#_Toc408826743)

[1.2 存在的问题和挑战 2](#_Toc408826744)

[2 基于机器学习的文本分类方法 2](#_Toc408826745)

[2.1 文本预处理 3](#_Toc408826746)

[2.2 文本表示模型 4](#_Toc408826747)

[2.3 特征选择 4](#_Toc408826748)

[2.4 特征加权 5](#_Toc408826749)

[2.5 分类方法 7](#_Toc408826750)

[2.6 评价标准 7](#_Toc408826751)

[3 实验方法与步骤 8](#_Toc408826752)

[4 实验结果及分析 10](#_Toc408826753)

[4.1 实验数据与程序说明 10](#_Toc408826754)

[4.2 特征数量的大小对情感分类影响实验 10](#_Toc408826755)

[4.3 不同权值计算方法对情感分类结果影响实验 11](#_Toc408826756)

[4.4 SVM分类器对于不同类型文本数据分类效果实验 11](#_Toc408826757)

[4.5测试文本数据分类实验 12](#_Toc408826758)

[5总结 12](#_Toc408826759)

[6参考文献 13](#_Toc408826760)

[附件 人员分工 14](#_Toc408826761)

## 

## 1绪论

### 1.1 国内外研究现状

文本情感分类是文本分类中的一个重要分支，也称之为意见挖掘。简而言之，文本的情感分类就是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程[1]。情感分类中，按照处理文本的粒度不同，可分为词语短语级、句子级、篇章级等几个研究层次[2]。在此，对词语短语级的情感分类进行详细阐述。在情感分析中可以认为构成篇章的基本单位包括词、短语、和固定搭配，对于它们的褒贬程度的度量是判别文本情感倾向的基础。国外许多学者做了大量研究，其主要研究方法分为基于字典和基于语料库两种。

基于语料库的词语短语级的情感判别主要是根据它们的语法特性，对大规模语料库进行信息挖掘，从而得到统计数据并对其极性做出判断。在研究早期，学者发现由某些连词连接的形容词具有相同或相反的极性，Hatzivassilolou和Mckeown[4]利用大规模语料库华尔街日报中的连接词信息来自动识别形容词的情感倾向，利用聚类算法将它们归属于褒义或贬义的类别集合。Turney和Littman[5]提出了点互信息的方法判别词的褒贬倾向。Dave等[6]从语料中抽取特征集合，通过分析此特征集合和己标记文本的关系来判定词汇的语义倾向。

基于词典的词语短语的情感判别，主要是根据词典WordNet或HowNet中词语间的关联来判别词语的极性。sista等[7]将GI(General Inquirer)和WordNet中的褒义和贬义词作为种子词，得到一个扩展后的较大规模情感词集合，并以此作为分类特征，利用机器学习方法对文本褒贬义进行了自动分类。Faye Baron和Graeme Hirst[8]从文档中抽取倾向性强的搭配作为种子词汇，取得了较好的分类效果。相对于英语，中文的词汇、短语的情感分析研究起步较晚。中科院自动化所的王根等[9]提出了词语倾向性的极坐标方式，并采用均衡化的互信息方法计算了词语倾向性。北京理工大学的李钝博士[10]把短语归结为一个非递归的基本词汇及依存关系的集合，提出了一种基于短语中心词之间依存概率统计分析方法，并将其应用于对短语的倾向性和倾向强度的计算。

## 1.2 存在的问题和挑战

情感分类的应用十分广泛，因此近年来的发展迅速，取得的不小的进步，同时我们也该看到，由于情感分类问题较复杂，不仅仅是单一的文本分类或文本挖掘任务，因而在研究过程中还存在很多问题与挑战。

**1）情感语义的机器理解问题**

人类的自然语言情感表达十分复杂，特别是网络评论的形式更加灵活多变，要使机器精确的理解文本中的情感内容，不能简单的提取词语作为特征，还必须结合语言学方面的知识，借助于文本上下文和领域相关性对情感语义进行分析处理。

**2)特征提取问题**

文本分类中一般采用词袋法表示文本的特征，然而由于情感表达中有许多诸如隐喻、反话等复杂的语言形式，且上下相关，因此简单采用词袋法提取特征并进行分析的效果极其有限，如何提取对情感分析具有更大价值的特征依然是一个有待完善的课题。

**3)领域依赖**

由于情感表达在不同的领域差别较大，所以无论是在有监督的学习方法还是无监督学习方法，情感分类都面临着领域依赖问题。

**4)语料库建设问题**

情感分类领域，许多研究者自己通过互联网获取语料，并进行人工标注，即使相同领域语料，但语料内容相差较大且标注标本不统一，造成实验结果很难进行比较。特别是在中文的情感研究领域，国内的公开语料库较少，资源匮乏。

## 2 基于机器学习的文本分类方法

基于机器学习的情感分类问题，它的处理过程大致可以分为两个部分，一部分是学习过程，另一部分是情感分类过程。其中，学习过程包括训练过程和测试过程，训练过程中对训练集进行训练得到分类器，用其对测试集进行情感分类，将测试的结果反馈给分类器，进一步改进训练方法，生成新的分类器，最后利用最终生成的分类器对新的文本进行情感分类，其基本流程如图1所示。

图1 基于机器学习的文本情感分类流程图

## 2.1 文本预处理

文本的预处理是进行文本情感分类的第一步，预处理结果的好坏直接影响到今后的分析处理能否顺利进行。文本预处理的目的是从文本语料库中规范地提取出主要内容，去除与文本情感分类不相关的信息。对于中文的预处理，其主要操作包括规范编码，过滤非法字符，分词处理，去除停用词等步骤。

**1）文件规范编码处理**

从网上下载的语料库存储格式可能千差万别，对实验带来很大困扰。所以第一步一般都是对语料库进行数据格式的归一化处理。

**2）中文分词处理**

中文文本单词之间没有天然的分隔符，因此在提取特征之前，首先要对中文文本进行分词。分词处理能够将连续的汉字序列按照一定的规则重新切分为词或词组。切分好的词或词组将会作为文本的特征用于情感分类分析过程，因此能否高效、正确的对中文进行分词成为中文情感分析的重要任务。中国科学院计算技术研究所专门开发了汉语词法分析系统ICTCLAS(Institute of Computing Technology，Chinese Lexical Analysis System)。ICTCLAS的主要功能包括中文分词、词性标注、新词识别、命名实体识别等功能，它的分词性能和分词精度都较高，是目前最受好评的汉语分词开源系统。

**3）停用词去除**

文本中包含许多助词、虚词等词性的单词以及在文本中经常出现的高频词汇但其本身对情感分类意义不大，这些词汇我们将它们统称为停用词。停用词表的构造一般有两种方式，人工方式或机器自动统计。停用词的存在不但会增加存储空间，而且很可能形成噪声，影响情感分类的精度，因此需要过滤文本中的停用词。

## 2.2 文本表示模型

文本是一种非结构化的数据，由大量字符构成，计算机无法直接处理字符类型的数据，因此需要将普通文本的内容转变为计算机能够读懂的数据形式，即将文本进行形式化表示。本文采用向量空间模型来表示文本。

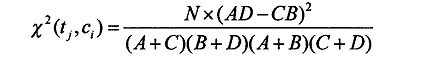
向量空间模型(VSM)[11]是由salton等人在1975年提出的一个基于统计的文本表示模型，并成功的应用在著名的SMART系统中。向量空间模型可以形式化的表述为，对于给定的文本D=D{t1，w1;t2，w2;…tn，wn}其中ti表示第i个特征项(最小不可分割的语言单位，如字、词、短语),wi表示特征项ti所拥有的权值，n为特征项的总数。每一个文本可以表示为一个向量(w1,w2…wn)，wi的计算可以使用布尔权重法，词频权重和TFIDF权重。

向量空间模型对文本的表示效果较好，可以将文档表示成空间向量进行运算，且具有较强的可计算性和可操作性。向量空间模型是文本分类中应用最为广泛的文本形式化模型。向量空间模型的缺点也十分明显，忽略了特征的次序和位置关系，且不考虑文本的长度，不考虑语义联系，认为所有的特征项都是独立的，只考虑文本所属类别的文档中出现特征项的频率，在情感分类的应用中存在一定的局限性。

## 2.3 特征选择

在对文本进行预处理并对其进行形式化表示后，得到了一个高维稀疏的特征空间，特征的数量可以达到几万维甚至是几十万维，不仅使得运算时间变长，而且会在很大程度上降低分类的准确度。特征选择(Feature Selection)就是从原始的高维特征集合中选择一小部分特征作为分类器的分类特征。特征选择过程需要通过构造好的评估函数对每个特征进行评分，然后按照评分的大小对特征向量进行降序排序，最后选择一定数量的特征作为分类特征集合。目前常用的特征选择方法有文档频率(DF)、信息增益(IG)、统计量(CHI)、期望交叉熵(ECE)和互信息(MI)等，本文采用卡方统计量来进行特征选择，并选取了不同数量的特征进行了对比测试。

卡方统计方法[12]用来衡量特征tj和文档类别Cj之间的统计相关强度，并假设tj和Cj之间符合具有一阶自由度的X2分布。特征相对某类的X2统计值越高，则其含有的信息量越多，与该类的相关性越大。CHI的计算公式如下：



其中，tj表示特征项，Ci表示类别，N表示文本总数，A表示包含特征tj，且属于类别Ci的文档数，B表示包含特征tj，但不属于类别Ci的文档数，C表示属于类别Ci但不包含特征tj，的文档数，D表示不属于类别Ci且不包含特征tj的文档数。当有多个类别时，分别计算特征tj对每个类别的CHI，然后取最大值作为特征的CHI值。CHI特征选择方法可以在降低85％以上特征维度的情况下取得较好的分类结果。

## 2.4 特征加权

特征选择过程中选择了最能代表文本内容的特征向量，但是这些特征对文本分类的影响不尽相同，因此，有必要对经过选择的特征进行加权，对表征能力强的特征赋予较大权重，对具有较弱类别区分能力的特征赋予较小的权重，这样可以有效抑制噪声。特征加权(Feature Weighting)就是对特征集合中每个特征根据其对分类的贡献程度赋予一定权值的过程。常用的特征加权方法有布尔权重，词频权重，TFIDF权重。本文分别使用了这三种权重计算方法，并对三种方法得到的分类结果进行了对比分析。

设文本类别集合D={d1,d2,…,dn}，dk∈D(k=1,2,…,n)，n为文本总数；类别集合C={c1,c2,…,cm}，ci∈C(i=1，2，．．．，m)，特征集合T={t1,t2,…,tn}，tj∈T(j=1，2，．．．，n)，m为特征总数；wij表示特征tj在文本dk中的权重。

**1)布尔权重(Boolean Weighting)**

布尔权重是最简单的权重计算方法，其计算公式如下：



若特征tj，在文本di中出现，则其权重wij为1；若特征tj在文本di中不出现，则其权重wij为0。

**2)词频权重(Term Frequency，TF)**

词频权重是指用特征在文本中的出现次数作为权重。不同类别的文本集合，特征出现的频率有较大差异，因此特征频率信息可以作为文本分类的重要参考之一。一般情况下，对于一个类别，出现频率较高的特征应该包含更多的类别信息，说明此特征对分类具有重要的意义，以词频作为参考赋予较大权重，反之应该赋予较小权重。TF权重计算公式如下：



其中tfij表示特征tj在类别ci中出现的次数。

词频权重相对于布尔权重，不再是均值权重，而是根据文本集合中特征出现的次数对特征进行加权，从权值函数构造来看有了一定的进步，然而词频权值对高频特征过度依赖，忽略了一些带有大量类别信息的低频特征。

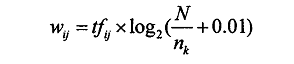
**3)TFIDF权重**

将特征词频和逆文档频率相结合用于特征权重计算，在实验中取得了较好的效果，因此TFIDF也成为了最经典也是最广泛使用的一种特征权重计算方法。逆文档频率(Inverse Document Frequency，IDF)是以包含特征的文档数为参数构造特征权重函数。其核心思想是：在大多数文档中出现的特征所带有的类别信息相比于在少量文本中出现的特征较少，也就是说，若一个特征同时出现在多个文档中，它所携带的类别信息较少，它的重要程度较低。逆文档频率在计算时常采用对数形式，其计算公式如下：



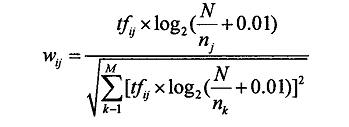
其中，ni为文本集合中包含特征tj的文本数。

集合次品权重函数就形成了TFIDF权重，计算公式如下：



其中，tfij表示特征tj在文本ci中出现的次数，ni为文本集合中包含特征ti的文本数。

为了消除文本长度对特征权重的影响，要对特征的权重进行归一化处理，归一化后的TFIDF特征权重函数如下：



## 2.5 分类方法

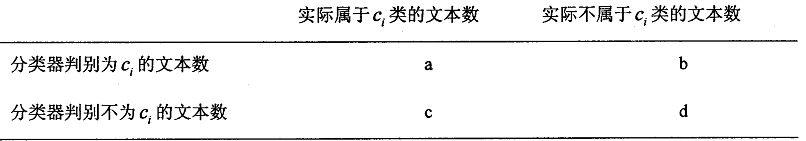
分类器是文本分类问题中的核心部分，在进行文本分类过程中常用的分类器有朴素贝叶斯分类器(NB)，支持向量机(SVM)，最大熵分类器(ME)，K近邻分类器(KNN)等，本文中主要使用支持向量机来进行分类，用现有的LibSVM完成实际操作，并对book类别、dvd类别和music类别的训练集分别进行了测试，对比测试结果。

支持向量机(Support Vector Machine)是Cortes和Vapnik[13]等人首先提出，它在解决小样本、非线性及高维模式识别中表现出许多优势，并能够推广应用到函数拟合等其他机器学习问题中。其基本思想是首先通过非线性变换将输入空间映射到一个高维特征空间，然后根据核函数在这个新空间中求取最优线性分类平面。将支持向量机应用于文本分类表现尤为突出，其分类的精确率和召回率都较高，且具有较好的稳定性，虽然它在大数据集上的训练收敛速度较慢，需要大量的存储资源和较高的计算能力，但其分隔面模式有效地克服了特征冗余、样本分布以及过拟合等因素的影响，具有较好的泛化能力。

## 2.6 评价标准

在进行文本分类实验时，需要将文本集合划分为训练集合和测试集合，在进行划分时一般采用K折交叉验证法，将其得到的实验结果的平均值作为实验的最终结果。对结果的评价已经有了许多标准，常用的有查准率、查全率等。

表1 文本类别归属判别表



查准率是指分类器判别为ci类别的文本数与实际属于ci类别的文本数的比值，其计算公式如下：



查全率是指实际属于ci类别的文本数与分类器判别为ci类别的文本数的比值，其计算公式如下：



查准率和查全率分别反映分类器不同方面的性能：查准率反映分类器的准确性，查全率反映分类器的完备性。采用准确率还是召回率作为分类器的评价性能取决于实验者所侧重的目标，两个评价标准是互补的，单纯提高其中一个评价标准会导致另一个评价标准的降低。对于一个优良的分类器应该同时具备较高的查准率和查全率。

## 3 实验方法与步骤

为了更好的理解情感分类的相关算法与步骤，结合相关论文方法与语料库进行了一系列的实验。实验过程框架图如图2所示，实验步骤如下：

1. 将训练文本数据文件作为XML文件处理，提取评论文本数据信息；
2. 对评论文本数据进行中文分词处理与停用词去除处理，并统计词频；
3. 结合预处理结果、词频信息以及情感标签，使用CHI方法进行特征选择；
4. 对评论文本进行特征权值计算并归一化权值，使用VSM模型表示文本；
5. 结合归一化后的特征权值向量，利用支持向量机分类器进行分类器训练；
6. 对测试文本进行情感分类的学习预测。

图2 中文情感分类实验框架图

结合实验框架图的基础上，针对分类过程中的各个部分，通过大量的实验进行验证。主要包括以下几个方面：

1. 特征数量的大小对情感分类结果的影响。按照理论情况，选取全部的特

征必然会得到最好的分类效果，但是会导致特征维数的急剧增大，从而导致“维数灾难”，以至后续的分类将会很难进行。故通过实验验证特征数量对分类效果的影响，力求做到能够在不大幅度降低分类效果的基础上，选取一个合适并且便于计算的特征数量。

1. 布尔权值、TF权值和TFIDF权值三种不同计算方法对情感分类结果的影

响 。从大量的论文中分析得出，TFIDF权值计算方法在情感分类中效果最好，这里通过实验对三种不同的权值计算方法进行验证。

1. SVM分类器对于不同类型文本数据的分类效果。不同类型的数据语料由

于其内容以及情感变化的不同，在同时使用SVM分类器的过程中其分类效果可能不同。故采取实验的方法验证不同类型文本数据在使用SVM分类器进行情感分类的效果区别。

最后，结合上述三个实验验证结果，对给定的测试语料进行情感分类实验。

## 4 实验结果及分析

## 4.1 实验数据与程序说明

整个实验的相关算法利用C++在Windows平台下完成，使用ICTCLAS中文分词开源程序、LibSVM以及开源的XML文件处理程序Markup完成。 具体实验代码详见附件。

实验使用的数据集分为三类，分别为book、dvd和music三类评论文本数据。三类数据集的情况如表2所示：

表2 实验语料数据情况说明表

## 4.2 特征数量的大小对情感分类影响实验

实验对book类别训练集进行测试，分别使用预处理处理后得到的特征集合的10%，15%与20%作为选取的特征。文本的特征权值计算使用TFIDF方法，使用SVM分类器进行训练集的分类训练。在不同特征数量的情况下，训练集的最优分类率如表3所示：

表3不同特征数量最优分类率表

## 4.3 不同权值计算方法对情感分类结果影响实验

实验对book类别训练集进行测试，分别使用预处理处理后得到的特征集合的10%作为选取的特征，文本的特征权值计算分别使用布尔方法、TF方法和TFIDF方法，使用SVM分类器进行训练集的分类训练。在不同权值计算方法的情况下，训练集的最优分类率如表4所示：

表4不同权值计算方法下最优分类率表

## 4.4 SVM分类器对于不同类型文本数据分类效果实验

实验分别对book类别、dvd类别和music类别训练集进行测试，分别使用预处理处理后得到的特征集合的10%作为选取的特征，文本的特征权值计算使用TFIDF方法，使用SVM分类器进行训练集的分类训练。在不同类型数据情况下，训练集的最优分类率如表5所示：

表5 不同文本数据使用SVM分类最优分类率表

## 4.5测试文本数据分类实验

结合上述三个实验结果，分别对book类别、dvd类别和music类别测试集进行分类实验。实验过程中使用预处理处理后得到的特征集合的10%作为选取的特征，文本的特征权值计算使用TFIDF方法，使用SVM分类器进行训练集的分类训练。三种不同文本测试数据的分类效果如表6所示：

表6测试文本数据分类结果表

## 5总结

本文主要是对基于机器学习方法的情感文本分类做了初步的研究，

取得了一定的研究成果，但得到的结果还有很大的提升空间，就本文而言，在许多方面还需要进一步的研究和改进。

情感分类是一个具有挑战性的研究课题，其问题更加复杂，同一方法应用于不同语料所产生的结果存在较大差异，本文中只在3个类别的数据集上进行了实验，应将该方法应用于更多的数据集进行对比分析研究。此外，分类器的融合方法还有多种，可以进一步实验其他融合策略对于情感分类性能的影响。

从整个实验结果来看，TFIDF权值计算相较于其他两种更有利于文本的情感分类。SVM分类器作为文本情感分类器对于不同类型的文本数据，其分类效果不同，但总体上取得了较好的效果。

## 6参考文献

[1]赵妍妍，秦兵，刘挺．文本情感分析[J]．软件学报，2010：21(8)：1834～1848．

[2]Huang X J, Zhao J. Sentiment analysis for Chinese text[C]．Proceedings of theCommunications of CCF,2008：4(2)．

[3]Picard R W．Affective Computing[M]．Cambridge：MIT Press，1997．

[4]Hatzivassiloglou V，Mckeown K R．Predicting the semantic orientation of adjectives[C]．Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics．Association for Computational Linguistics，1997:174-181．

[5]Tumey 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]Dave K，Lawrence S，Pennock D M．Mining the peanut gallery：Opinion extraction and semantic classification of product reviews[C]．Proceedings of the 12th internationalconference on World Wide Web．ACM，2003：519-528．

[7]Sista S．P，Srinivasan S．H．Polarized lexicon for review classification[C．Proceedings of the International Conference on Artificial Intelligence．2004：867-872．

[8]Faye B，Graeme H．Collocations as cues to semantic orientation[C]．Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text，2004．

[9]王根，赵军．中文褒贬义词语倾向性的分析[C]．第三届学生计算语言学研讨会论文集．2006：81～85．

[10]李钝．基于中文文本的网络内容分层监测研究[D]．北京：北京理工大学，2007.

[11]Salton G，Wong A，Yang C S．A vector space model for automatic indexing[J]．Communications ofthe ACM，1975，18(1 1)：613-620

[12]Zheng Z，Srihari S N．Text Categorization Using Modified-CHI Feature Selection and Document／Term Frequencies[C].Proceedings of the2002 International Conference On MachineLearning And Applications，2002：143-146．

[13]Joachims T．Text categorizationwith support vector machines：Learning with many relevant features[C]．Proceedings of the lOth European Conference on Machine Learning．Chemnitz，Germany，1998：137-142．

## 附件 人员分工

表7 人员分工情况表

|  |  |
| --- | --- |
| 姓名 | 主要工作 |
| 蒋浩浩 | 实验程序编写，总结修改实验报告 |
| 卢丽华 | 查阅相关资料与论文，分析并撰写实验报告等 |
| 聂卉 | 实验程序编写，并测试等 |
| 薛景 | 实验测试，并统计数据、作图以及完成PPT等 |