大数据课程论文综述

|  |  |
| --- | --- |
| **论文题目** | 短文本分析综述 |
| **学 院** | 计算机学院 |
| **组 员** | 杨小雷 赵鹏飞 刘江涛 |

**短文本分析综述**

**摘要** 所谓的短文本，指的是长度比较短，字符数通常在百个字符以内的文本。处于信息大爆炸时代的现代信息社会，大量的短文本正以迅猛的势头闯进了人们的生活当中。短文本的形式多种多样，电子邮件、微博、手机短信、新闻标题摘要、书评影评等都属于短文本，可以说短文本涵盖了人们生活的方方面面，蕴含着无穷的价值等待人们挖掘。短文本分析与长文本相比，短文本具有一些独有的特征，比如长度短、包含的信息单一、文本特征向量稀疏等，使得传统的文本处理方法难以满足对短文本处理的要求，短文本对文本处理提出了更高的要求。

本文第一、二章详细介绍了短文本的概念、特点、研究意义及相关知识，阐述了国内外研究现状，由杨晓雷负责编写；第三章详细介绍了短文本分析的五种方法，并阐述了每种方法的优点及不足之处，该部分由刘江涛负责编写；第四章通过情感分析的实验验证了短文本分析的有效性，并研究了特征维度、权重计算方法等对结果的影响，由赵鹏飞负责编写。

**Summary of the short text analysis**

**Abstract** A short text is a text of short length, usually less than 100 characters. In the modern information society in the era of information explosion, a large number of short texts are breaking into people's lives with a rapid momentum. There are various forms of short text books, such as E-mail, microblog, mobile phone short message, news headline summary, book review and film review, etc., all belong to short text books. It can be said that short text books cover all aspects of people's life, and contain infinite value waiting for people to explore. Compared with long text, short text has some unique features, such as short length, single information contained, sparse text feature vectors, etc., which makes it difficult for traditional text processing methods to meet the requirements of short text processing, and short text processing puts forward higher requirements.

The first and second chapters of this paper introduced the concept, characteristics, research significance and relevant knowledge of the short text in detail, and elaborated the research status at home and abroad which was compiled by Yang Xiaolei. Chapter three introduces five methods of short text analysis in detail, and expounds the advantages and disadvantages of each method. which was compiled by Liu Jiangtao.

The fourth chapter verified the validity of the short text analysis through the experiment of emotion analysis, and studied the influence of feature dimension, weight calculation method and other methods on the results, which was compiled by Zhao Pengfei.

目录

[第1章 绪论 6](#_Toc533631946)

[1.1 研究背景 6](#_Toc533631947)

[1.2 国内外研究现状 7](#_Toc533631948)

[1.3 主要研究内容 8](#_Toc533631949)

[1.4 本文结构 9](#_Toc533631950)

[第2章 相关知识介绍 10](#_Toc533631951)

[2.1 隐性语义模型 10](#_Toc533631952)

[2.1.1 隐性语义分析（LSA）模型 10](#_Toc533631953)

[2.1.2 超空间模拟语言（HAL）模型 10](#_Toc533631954)

[2.1.3 神经语言模型（NLM） 11](#_Toc533631955)

[2.1.4 段向量（PV） 12](#_Toc533631956)

[2.2 半显性语义模型 12](#_Toc533631957)

[2.3 显性语义模型 13](#_Toc533631958)

[2.3.1 显性语义分析（ESA）模型 13](#_Toc533631959)

[2.3.2 概念化（Conceptualization） 14](#_Toc533631960)

[第三章 分析方法 15](#_Toc533631961)

[3.1 特征表示 15](#_Toc533631962)

[3.1.1 向量空间表示模型（VSM） 15](#_Toc533631963)

[3.1.2 主题模型表示方法 16](#_Toc533631964)

[3.1.3 分布式表示方法 17](#_Toc533631965)

[3.2.1 利用外部知识库进行扩展 17](#_Toc533631966)

[3.2.2 利用元数据进行特征扩展 18](#_Toc533631967)

[3.3 特征选择 18](#_Toc533631968)

[3.3.1 文档频率和逆文档频率 18](#_Toc533631969)

[3.3.2 互信息 19](#_Toc533631970)

[3.3.3 信息增益 19](#_Toc533631971)

[3.3.4 卡方统计量 19](#_Toc533631972)

[3.4 传统学习算法 20](#_Toc533631973)

[3.4.1 SVM 20](#_Toc533631974)

[3.4.2 朴素贝叶斯 21](#_Toc533631975)

[3.5 LSTM进行短文本分析 22](#_Toc533631976)

[3.5.1 LSTM 22](#_Toc533631977)

[3.5.2 GRU 24](#_Toc533631978)

[3.5.3 BiGRU-Attention 24](#_Toc533631979)

[第四章 方法阐述与实验部分 27](#_Toc533631980)

[4.1 方法阐述 27](#_Toc533631981)

[4.2 实验设计 28](#_Toc533631982)

[4.2.1 实验数据 28](#_Toc533631983)

[4.2.2 评估方法 28](#_Toc533631984)

[4.2.3 参数规范化 29](#_Toc533631985)

[4.3 实验结果 30](#_Toc533631986)

[4.4 结果分析 33](#_Toc533631987)

[第五章 总结 35](#_Toc533631988)

[第6章 致谢 36](#_Toc533631989)

[参考文献 37](#_Toc533631990)

# 第1章 绪论

## 1.1 研究背景

随着电子商务和在线交流方式的兴起，互联网改变了我们传统的生活方式，成为了一个人们交流情感、传达思想的重要平台。每天从社交媒体上交互的数据已经远远地查处了人们的想象，越来越多的用户可以随时随地的发表自己的评价，分享自己的经历，表达自己的观点，Facebook、Teitter、新浪微博以及腾讯微信等社交媒体已然成为了一个海量信息的熔炉，成为了一个公共事件、舆情研究，甚至个人情绪观点的有价值的数据源泉。互联网普通用户每天通过社交媒体产生TB级的海量短文本数据，其中的所包含的信息量是传统媒体所无法比拟的。

通常，互联网上的短文本数据包括：网页搜索片段、即时通讯、电子布告栏系统标题、新闻标题、论文摘要、网络聊天记录、博客与新闻的评论，电商网站的用户评论、Twitter信息以及微博信息等等。随着短文本数量的剧增，如何利用机器学习，自然语言处理，大数据技术等方法挖掘这些短文本数据的主题信息对于建立用户画像，舆情分析和信息决策等具有非常重要的意义。例如：挖掘Twitter的信息的主题能够掌握公众的观点和兴趣；挖掘微博信息的主题能够帮助用户发现和跟踪热门话题，进而增加自己对事件的认知和理解；挖掘电商网站评论的主题信息对于了解用户对商品的态度，进行用户的情感分析，从而能够帮助商家更好地改善产品质量等具有重要意义。挖掘海量短文本的主题对于改善用户体验等具有巨大的商业价值和社会价值。

传统的文本分析方法在分析长文本主题方面已经达到令人满意的效果，原因是长文本字数较多，词语信息丰富。然而，运用相同的方法分析短文本的主题信息是效果很差的，原因是短文本长度非常短小，一般不会超过140个词。例如：我们日常使用的短信不会超过70字符；电子布告栏系统和新闻的标题则不会超过30字符；Twitter网站限制推文不能超过140个字符。在人们实际的生活和交流中，字数远远少于各媒介所规定的的最大字数。除了内容少之外，短文本数据还具有信息量大，噪声大、书写不规范、特征过度稀疏等特点。一定程度上，依据词频和足够的单词量进行主题挖掘的文本分析方法，在处理短文本时会因为数据问题而无法保证主题挖掘的准确性。

综上所诉，短文本的主题挖掘具有广阔的发展前景和实际价值，提高短文本的主题挖掘效果，是文本领域重要的课题，也是我们本文的主要研究内容。

## 1.2 国内外研究现状

近年来，短文本分析的研究已经成为文本挖掘和自然语言处理领域的重要研究内容，国内外学者所提的模型、方法层出不穷。

为了降低短文本的特征稀疏性，许多学者通过引入WordNet、维基百科、主题相关的辅助文本等方法对短文本特征信息进行扩充或转换，从而减低词语的稀疏性和模糊性，提高文本分析效果。Frey B J 等人提出利用维基百科，对短文中的特征词进行检索，然后基于检索结果对之前的短文本进行融合，形成新的文本向量的方法。Hong等人提出一种将包含同样词语的推特消息进行聚合，然后在传统的主题模型上进行训练的方法。Weng等人将每个用户的推特信息聚合为一篇长文档，然后使用传统的主题模型LDA进行主题挖掘。Gabrilovich等人通过使用Wikipedia灯幅短文本的表达来增加特征语义，改善文本的相似性。Hu等人通过将原始文本的内部语义和来自外Wikipedia和WordNet的外部概念相结合，改善短文本的挖掘效果。Sahami等使用搜索引擎增加额外的信息，并引入基于网页的核函数的方法进行短文本之间的相似性计算。范云杰等人引入特征拓展的思想，基于维基百科提出了一种中文短文本分类模型，该模型依照维基百科特征扩展词表对向量空间的概念进行了扩充，并运用支持向量算法构建分类器，通过与传统方法的对比实验证实了该方法性能的可靠性。赵辉等人针对短文本特征概念模糊、特征词汇稀少的问题，计算概念相关度，扩展特征信息，结合上下文解决了短文本中词语产生歧义的问题，并通过实验对算法进行验证。郭永辉针对有歧义信息的扩展问题，提出一种基于图的特征扩展方法，并以新浪微博为实验数据，设计了中文短文本分类系统。Jin等提出了二元潜在狄利克雷分配（dual latent dirichlet allocation，DLDA）主题模型，该模型通过将长文本知识引入到短文本中，然后将长文本与短文本联合起来学习长文本辅助问题和短文本目标主题。

除了引入相关信息提高主题挖掘效果外，许多学者提出了能够直接应有于短文本主题挖掘的方法。Yan等提出了Biterm Topic Mode (BTM)，该模型直接对语料库中的共现词建模，从而发现短文本的主题。邱云飞等人针对短文本聚类面对的主要挑战，提出了一种k-means算法的语义结合改进算法，缓解了短文本聚类的特征稀疏问题和聚簇可理解性的问题，并通过实验证实了算法的可靠性。Seifzaden等人利用统计学和语义分析法进行短文本主题挖掘。

短文本主题挖掘存在的另一个重要问题是估计短文本的主题数目。短文本主题挖掘实际上是短文本主题聚类过程，在传统的文本聚类应用过程中，许多的聚类方法都需要一个较强的假设，即类别的数目是由用户提前给定的，为一个固定的常数值，一般情况类别的数目不会随样本的变化进行自适应的变化。在现实生活中，确定数据集的类别个数是一个棘手的问题，当用户获得大规模的文本数据集之后，首先必须对文本进行快速的浏览阅读，了解文本的大致内容，其次不得不通过相关的数据分析方法和工具分析和推断类别的个数，在这一过程中，会消耗大量的时间，金钱，人力等资源。除此之外，类别数目的准确度也非常重要，错误的类别估计会大大影响文本挖掘的过程，降低挖掘结果的可信度。显然这种通过用户估计类别数目是不切合实际的。所以拒绝人为的提前设定，利用机器学习的合理模型对文本类别数目进行估计也是文本研究领域一个重要的问题。在利用模型进行主题数据自动学习的初期研究中，Smyth通过将不同的主题数目取值作为模型的输入，然后计算不同取值下似然函数的值，最后以似然函数最大的主题值作为该数据集的主题数目，这一一尝试的遍历方法，增加了时间开销，简单暴力但不讨巧。此外，Huang等人在非参模型之狄利克雷过程模型（Dirichlet Process Model，DPM）研究的基础上提出的两个具体特征选择的狄利克雷过程回合模型DPMFS（Dirichlet Process Mixture model with Feature Partition）和DPMFP（Dirichlet Process Mixture model with Feature Selection），这两个模型解决了类别的数目人为提前设定的问题，实现了类数目的自动学习和推断。Yin和Wang将整个短文本语料库视为一个主体分布，提出了利用吉布斯采样的狄利克雷多项混合（Gibbs Sampling for Dirichlet Multinomial Mixture,GSDMM）模型。Zheng等人针对短文本提出了一种自动聚类并且提取热点主题的方法，该方法能够生成样本的特征向量，建立空间模型，应用余弦相似度和K均值方法进行聚类，并提取出每个聚类的热点主题。该模型能够自动推断文本集的主题数目并能够跟踪主题的动态变化过程，极大地提高了短文本的聚类效果。

## 1.3 主要研究内容

主要研究内容如下：

（1）短文本的概念、其研究背景以及意义，介绍短文本分析的研究现状，研究了短文本分析的相关技术分类

（2）针对短文本半结构化，长度短，噪声大的特点，从特征表示、特征扩展、特征选择、聚类分类四个方面介绍短文本分析的方法

（3）从最前沿的研究知识介绍短文本最新的研究实验以及实验分析

（4）从对前人的总结的基础上，对短文本的前景进行展望。

## 1.4 本文结构

本文的组织结构包括五个章节，各个章节的主要内容如下：

第1章：绪论。阐述了论文的研究背景以及研究意义；总结了国内外对此课题的研究进展；简要说明本章的论文的结构安排。

第2章：相关知识介绍。阐述了短文本分析的模型类别，从隐性、半显性、显性三种分析模型进行了详细的说明介绍。

第3章：短文本分析方法。从特征表示、特征扩展、特征选择、LSTM进行短文本分析、传统学习算法五个方面对短文本分析技术进行了介绍。

第4章：方法阐述与实验部分。从实验的角度，详细的介绍了基于机器学习的细粒度的短文本情感分析方法。

第5章：总结，对上述四个章节进行总结，并对短文本的前景进行了展望。

# 第2章 相关知识介绍

本章根据短文本分析所需要知识源的属性，将短文本分析模型分为三类：隐性（implicit）语义模型、半显性（semi-explicit）语义模型和显性（explicit）语义模型。其中，隐性和半显性模型试图从大量文本数据中挖掘出词与词之间的联系，从而应用于短文本分析。而显性模型使用人工构建的大规模知识库和词典辅助短文本分析。

## 2.1 隐性语义模型

隐性语义模型产生的短文本表示通常为映射在一个语义空间上的隐性向量。这个向量的每个维度所代表的含义人们无法解释，只能用于机器计算。下面将介绍几种代表性的隐性语义模型。

### 2.1.1 隐性语义分析（LSA）模型

隐性语义分析（Ldatent Semantic Analysis, LSA）模型：最早的基于隐性语义的文本分析框架为LSA模型，也被称为Latent Semantic Indexing(LSI)模型。LSA模型旨在用统计方法分析大量文本从而推出词与文本的含义表示，其思想核心是在相同语境下出现的词具有较高的语义相关性。具体而言，LSA模型构建一个庞大的词与文本的共现矩阵。对于每个词向量，它的每个维度都代表一个文本；对于每个文本向量，其每个维度都代表一个词。通常，矩阵每项的输入是经过平滑或转换的共现次数。常用的转换方法为TF-IDF。最终，LSA模型通过奇异值分解（SVD）的方法将原始矩阵降维。在短文本的情境下，LSA模型有两种使用方式。首先，在语料足够多的离线任务上，LSA模型可以直接构建一个词与短文本的共现矩阵，从而推出每个短文本的表示。其次，在训练数据量较小的情境下，或者针对线上任务（针对测试数据），可以事先通过标准的LSA模型方法得到每个词向量，然后使用额外的语义合成方法获取短文本向量

### 2.1.2 超空间模拟语言（HAL）模型

超空间模拟语言（Hyperspace Analogue to Language,HAL）模型：一个与LSA模型类似的模型是HAL模型。HAL模型与LSA模型的主要区别在于前者是更加纯粹的词模型。HAL模型旨在构建一个词与词的共现矩阵。对于每个词向量，它的每个维度代表一个语境词。该模型统计目标词汇与语境词汇的共现次数，并经过相应的平滑或转换（如TF-IDF、Pointwise Mutual Information(PMI)等）得到矩阵中每个输入的值。。通常，语境词的选取有较大的灵活性。例如，语境词可被选为整个词汇，或者除停止词外的高频词。类比LSA模型，在HAL模型中可以根据原始向量的维度和任务要求选择是否对原始向量进行降维。由于HAL模型的产出仅仅为词向量，在短文本分析这一任务中需采用额外的合成方法（例如向量相加）来推出短文本向量。

### 2.1.3 神经语言模型（NLM）

神经语言模型（Neural Language Model,NLM）：近年来，随着神经网络和特征学习的发展，传统的HAL模型逐渐被NLM取代。与HAL模型通过明确共现统计构建词向量的思想不同，NLM旨在将词向量当成待学习的模型参数，并通过神经网络在大规模非结构化文本的训练来更新这些参数以得到最优的词语义编码。

最早的概率性NLM由Bengio等提出。其模型使用前向神经网络（Feedforward Neural Network）根据语境预测下一个词出现的概率。通过对训练文本中每个词的极大似然估计，模型参数（包括词向量和神经网络参数）可使用误差反向传播（BP）算法进行更新。此模型的缺点在于仅仅使用了有限的语境。后来，Mikolov等提出使用递归神经网络（Recurrent Neural Network）来代替前向神经网络，从而模拟较长的语境。此外，原始NLM的计算复杂度很高，这主要是由于网络中大量参数和非线性转换所致。针对这一问题，Mikolov等提出两种简化（去掉神经网络权重和非线性转换）的NLM，即Continuous Bag of Words（CBOW）和Skip-gram。前者通过窗口语境预测目标词出现的概率，而后者使用目标词预测窗口中的每个语境词出现的概率。

另一类非概率性的神经网络以Collobert和Weston的工作为代表。其模型Senna考虑文本中的n元组。对每个n元组中某个位置的词（如中间词），模型选取随机词来代替该词，从而产生若干新的n元组作为负样本。在训练中，一个简单的神经网络为n元组打分，训练目标为正样本得分与负样本得分间的最大间隔排序损失（max-margin ranking loss）：

总而言之，NLM同HAL模型相似，所得到的词向量并不能直接用于短文本分析，需要额外的合成模型依据词向量得到短文本向量。

### 2.1.4 段向量（PV）

段向量（Paragraph Vector，PV）：PV是另一种基于神经网络的隐性短文本分析模型。PV可被视作CBOW和Skip-gram的延伸，可直接应用于短文本向量的学习。PV的核心思想是将短文本向量当做“语境”，用于辅助推理（例如，根据当前词预测语境词）。在极大似然的估计过程中，文本向量亦被作为模型参数更新。PV的产出是词向量和文本向量，对于（线上任务中的）测试短文本，PV需要使用额外的推理获取其向量。

## 2.2 半显性语义模型

半显性语义模型产生的短文本表示方法也是一种映射在语义空间里的向量。与隐性语义模型不同的是，半显性语义模型的向量的每一个维度是一个“主题”（topic），这个主题通常是一组词的聚类。人们可以通过这个主题猜测这个维度所代表的含义。但是这个维度的语义仍然不是明确、可解释的。半显性语义模型的代表工作是主题模型（topic model）。

主题模型：最早的主题模型为LSA模型的延伸。LSA模型尝试通过线性代数（奇异值分解）的处理方式发现文本中的隐藏语义结构，从而得到词和文本的特征表示；而主题模型则尝试从概率生成模型（Generative Model）的角度分析文本语义结构，模拟主题这一隐藏参数，从而解释词与文本的共现关系。

最早的主题模型Probabilistic LSA（PLSA）模型由Hofmann等提出。PLSA模型假设文本具有主题分布，而文本中的词从主题对应的词分布中抽取。以d表示文本，w表示词，z表示主题（隐藏参数），Z表示主题结合，则文本和词的联系概率p(d,w)的生成过程可被表示如下：

虽然PLSA模型可以模拟每个文本的主题分布，然而其没有假设主题的先验分布（每个训练文本的主题分布相对独立），它的参数随训练的个数呈线性增长，且无法应用于测试文本。

一个更加完善的主题模型为LDA模型（Latent Dirichlet Allocatiton Model）。LDA模型从贝叶斯的角度为两个多项式分布添加了狄利克雷先验分布，从而解决了PLSA模型中存在的问题。在LDA模型中，每个文本的主题分布为多项式分布Mult()，其中从狄利克雷先验Dir()中抽取。同理，对于主题的词分布Mult()，其参数从狄利克雷先验Dir()获取。

总之，通过采用主题模型对短文本进行训练，最终可以获得每个短文本的主体分布，以作为其表示方式。这种方法将短文本转为了机器可计算的向量。

### 2.3 显性语义模型

近年来，随着大规模知识库系统的出现（如Wikipedia、Freebase、Probase等），越来越多的研究关注于如何将短文本转化为人和机器都可以理解的表示方法。这类模型成为显性语义模型。与前两类模型相比，显性语义模型最大的特点就是它所产生的短文本向量表示不仅是机器可以计算的，也是人类可以理解的，每个维度都有明确的含义，通常是一个明确的“概念”。这意味着机器将短文本转为显性语义向量后，人们很容易就可以判断这个向量的质量，并发现其中的问题，从而方便进一步的模型调整与优化。

### 2.3.1 显性语义分析（ESA）模型

显性语义分析（Explicit Semantic Analysis，ESA）模型：在基于隐性语义的模型中，向量的每个维度都没有明确的含义标注。与之相对的是显性语义模型，向量空间的构建由知识库辅助完成。最具代表性的显性语义模型为ESA模型。ESA模型同LSA模型的构建思路一致，旨在构建一个庞大的词与文本的共现矩阵。在这个矩阵中，每个输入为词与文本的TF-IDF。然后，在ESA模型中词向量的每个维度代表一个明确的知识库文本，如标题。此外，原始的ESA模型没有对共现矩阵进行降维处理，因而产生的词向量具有较高维度。在短文本分析这一任务中，需要使用额外的语义合成方法推导出短文本向量。

### 2.3.2 概念化（Conceptualization）

概念化（Conceptualization）：另一类基于显性语义的短文本理解方法为概念化。概念化旨在借助知识库推出短文本中每个词的概念分布，即将词按语境映射到一个以概念为维度的向量上。这一任务中，每个词的候选概念可从知识库中明确获取。

最早的概念化方法由Song等提出。其模型使用知识库Probase，获取短文本中每个词与概念间的条件概率p(concept|word)和p(word|concept)，从而通过朴素贝叶斯方法推出每个短文本的概念分布。这一单纯基于概率的模型无法处理由语义相关但概念不同的词组成的短文本。为了解决无法识别语境的问题，Kim等对Song的模型做出了改进。新的模型使用LDA主题模型，分析整条短文本的主题分布，进而计算p(concept|word,topic)。

另一个基于Probase的短文本分析框架为Hua等提出的Lexical Semantic Analysis (LexSA)。LexSA将短文本分析系统划分为分词、词性标注和概念识别三个步骤，并在每个步骤使用新的模型消除歧义。在分词和词性标注环节，作者分别使用图模型推出短文本的最优分词方式和词的词性；在概念识别环节，每个词表示成以概念为维度的向量。为了进一步强联LexSA中各环节的相互作用关系，Wang等提出为短文本构建统一的候选词关系图，并使用随机游走的方法推出最优的分词、词性和词的概念。

# 第三章 分析方法

短文本产生与网络环境中，在移动互联网不断发展的现在，它逐渐具有了很多区别与普通文本的特点，同时也具有很高的研究价值。例如微博数据，电商平台的评论信息，电影评论等。这些环境产生的短文本往往具有长度端，信息含量少，数量达，具有动态性实时性，同时也具有文本不规范，文本呈现板结构化的特点，内容上具有很多个人的感情色彩。针对这些特点，进行分析和处理时需要进行有别与传统文本分析的工作，下面就针对这些特点详细分析处理的方法。

## 3.1 特征表示

特征表示的过程就是将自然语言处理成计算机可以进行处理的结构化的形式。目前主要的方法有向量空间表示模型（VSM）以及主题表示模型。

### 3.1.1 向量空间表示模型（VSM）

VSM（Vector Space Model）将文本映射成为一个特征向量，ti表示一组互不相同的特征词，w表示该文本中这个特征词的权值。整个向量的长度等于特征词的数量。

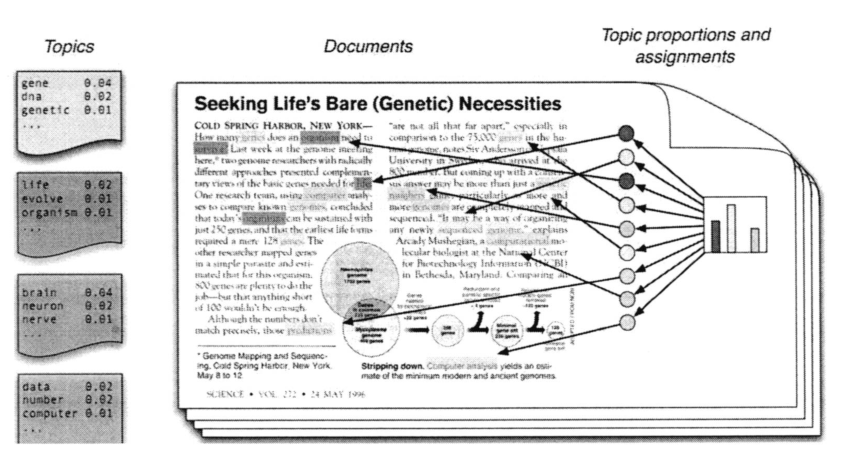
每个权值的计算常用的有TF-IDF算法，𝑡𝑓(𝑡,𝑑) 表示单词t在文档d中出现的次数，i𝑑𝑓(𝑡,𝑑) 为逆文档频率。在计算出权重之后还需要对其进行标准化，以避免文本长度对权值的影响。公式如下。

从IDF-IDF公式公式我们可以发现，数据集中包含某一特征词的文档越多，就可以推测它对于区分某类文档的能力越弱，其计算得到的权值也月小。而当某一文档中某一个特征词出现的频率越高，就可以推测它对与这类文档具有更高的区分能力，权值也就越大。在得到了文档的向量空间表示之后可以通过计算向量之间的夹角来衡量两个文档之间的相似度，计算时通常利用向量的余弦值来表示。

这种文本表示方法的特点是方法简单，易于处理。但由于短文本的特点，即具有零散性，高噪声，碎片化的特点，所以它容易造成语义信息的丢失。同时向量长度与特征词的数量直接相关，容易造成维度灾难，以及特征向量的稀疏，所以最终分类或者聚类的效果并不好。

### 3.1.2 主题模型表示方法

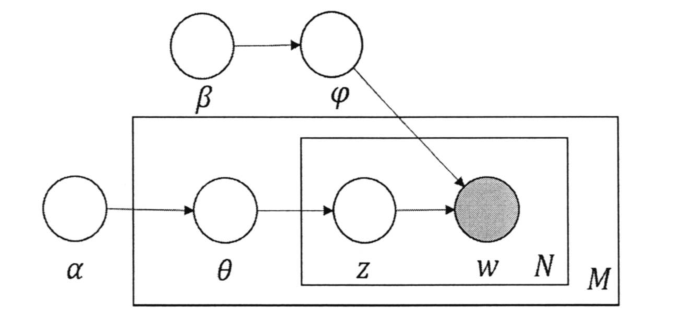
LDA(Latent Dirichlet Allocation)，即隐含狄利克雷分布主题模型，由普林斯顿大学的David.Blei在2003年提出，它是一个三层的贝叶斯模型。它对文档-主题，主题-词语两个层次进行建模。他提出在一篇文章形成的过程时，首先从服从Dirichelet分布多个隐含主题中产生一个主题。然后在从这个主题中产生词，从而产生整篇文章。其模型示意图如下。



LDA模型是生成模型，它通过设定隐含参数来随机产生可观测的数据，进而估计实际数据的联合概率分布。对于一篇文章来所，其生成的过程如下：

1. 根据文档的主题分布，选择一个主题
2. 根据这个主题在词表中的策瑜的分布，产生一个单词
3. 重复第二步知道产生文档中的所有单词

这个过程中国主题数量需要事先确定，每篇文档的主题分布是多项式分布。每个主题中，单词的分布也是多项式分布。整个过程的概率图模型表示如下



### 3.1.3 分布式表示方法

词语的分布式表示（Distributed Representation）也称为词向量表示方法，不同于前面所说的VSM，它是将词汇映射成为一个有固定长度的向量。而这个向量的长度远远小于数据集中字典的大小，这些向量通常只有几十到几百维，这些向量构成了向量空间，而每一个词就是这个空间中的一个点。这不仅降低了数据的维度，同时也给系统带来了词汇的语义信息。通过计算这些词向量的距离就可以判断这些词在语义上的相似度。

词向量并不是人为设定的，它是利用机器学习算法学出来的。常用的方法有LSA(Latent Semantic Analysis)以及LDA(Latent Dirichlet Allocation)。这些模型可以发现文本中的主题分布，同时可以将文档中的词汇利用主题向量来表示，向量中每个维度表示该词汇对于这个主题的贡献程度。它将词汇映射到了主题构成的空间中，给词汇引入了语义。

词向量也可以作为语言模型的副产品出现，在利用神经网络训练语言模型时，同时获得了词向量。2000年Xu首次提出利用单层神经网络训练语言模型的方法。Bengio在2006年构造了三层神经网络训练语言模型。而Mikolov在2010年提出了利用循环神经网络训练语言模型的方法，其更加充分的利用了文本中上下文信息来训练语言模型。

而伴随这这些语言模型，也产生了相应的词向量。这些词向量中最为经典的就是有Google团队实现的Word2vec。这个算法能够快速的从语料库中学习得到词向量。通过这种办法得到的词向量，能够方便的计算词之间的相似度。

3.2 特征扩展

由于短文本的长度较短，所以在进行特征表示之后特征会很短，同时对于一义多词显的无能为力。所以需要进行特征扩展，增加特征空间的数据量，提升短文本分析的效果。

### 3.2.1 利用外部知识库进行扩展

利用现有的语义网络或者分类知识库可以有效增加文本的信息量，这些外部资源按照一定的格式保存定义了词语之间的语义关系。这些知识库包括WordNet,HowNet等。其中WordNet是由Princeton 大学的心理学家，语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典。它不是光把单词以字母顺序排列，而且按照单词的意义组成一个“单词的网络”。 它是一个覆盖范围宽广的英语词汇语义网。名词，动词，形容词和副词各自被组织成一个同义词的网络，每个同义词集合都代 表一个基本的语义概念，并且这些集合之间也由各种关系连接。在中文语言处理领域也有相似的知网（HowNet），可以提供相似的功能。

Sheng利用知网获取文本中的上下为关系以及关系强度，扩展文本的特征。Sun等人利用知网来计算词语语义上的相似度，进而计算短文本之间的相似度，实现分类。Wang.P等人通过维基百科中的词汇概念以及链接中的信息计算词语的相似度，并利用消歧义页以及上下文信息解决了一词多义的问题。Li提出了“Bag of Concepts”的概念进行文本表示，相比传统的“bag of words“,它能够更好的处理一词多义问题。

### 3.2.2 利用元数据进行特征扩展

短文本中往往具有区别普通文本的特征，加入这些元数据信息可以扩展短文本特征。B.Sriram等人提出了一种利用微博元数据进行特征扩展的方法，他们共提取了微博数据中的8中特征，包括作者信息，发布时间，标志符号等。

## 3.3 特征选择

在进行特征表示之后，特征维度往往会很大，特征还具有非常稀疏的特点。同时这些特征词会包涵很多噪音，造成系统鲁棒性不足，所以特征选择在短文本分析中也十分重要。

### 3.3.1 文档频率和逆文档频率

词频（TF）表示给定的词语在文档中出现的频率，语料库中包含某个词语x的文档的数量称为该词的文档频率。如果仅适用这两个值过滤词语，在一定程度上可以过滤无效词语，但没办法表示某个词对某个类别的重要性，所以用到了IDF逆文档频率。IDF逆文档频率公式如下。

𝑁表示预料库中文本的总数，𝑁(𝑥)表示语料库中包含词x的文本总数。将TF与IDF结合起来就可以衡量词语对某个文档的重要性，但没办法表示某个词对某个类别的重要性。同时还可以比较词语在某个类别文档中的频率，如果低于一定阈值就可以过滤掉。

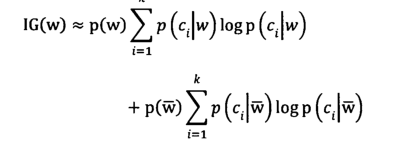
### 3.3.2 互信息

互信息（Mutual Information）是信息论中信息量的一种度量，它可以度量两个任意集合之间的相关性。它一般用于衡量文本向量中特征词和文本类别，特征词与特征词，文本与文本之间的关联性。如果一个特征词在某个类别中出现的频率很高但在其他类别的中出现频率很低则这个特征词与该类别之间的互信息就很大。所以在降维时就可以保留互信息大的特征。

文本空间有n个维度，有n个特征词（t1,t2…tn），有k个类别（c1,c2…cn）,那么特征词语类别之间的互信息公式为。

### 3.3.3 信息增益

信息增益（Information Gain）用来分析词语给系统带来的信息量。单词为w类别为ci，p(w)是在语料库中w出现的概率。P(ci|w)表示包含单词w的文档中数据ci的概率，p(w ̅)表示文档不包含w的概率。可以表示特征词对于分类系统整体的贡献，无法区分该该特征词对于不同类别的权重。



### 3.3.4 卡方统计量

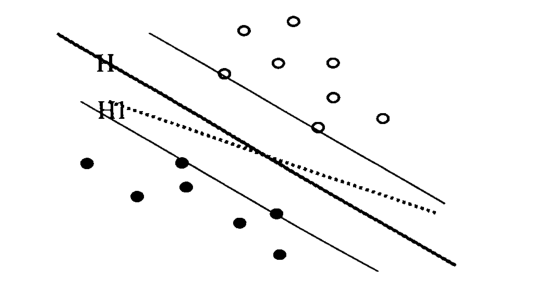
是统计学中用于度量事件独立性的一种方法，该方法基于统计学中的“假设检验”思想。假设词w和类别c是独立的，然后用卡方分布计算检验值，如果检验值偏离阈值越大，就越有信心否定原假设, 词语w和类别c的CHI值计算公式如下。

N表示语料库中文档总数量，A是包含词语w且属于类别c的文档数量，B是包含词语w但不属于类别c的文档数量，C是不包含w但属于类别c的文档数量，D是既不包含词语w也不属于类别c的文档数量。

## 3.4 传统学习算法

### 3.4.1 SVM

支持向量机（Support vector machines）是一种二分类模型，它的目的是发现一个超平面对数据样本进行分割，所得到的超平面与两类样本的距离间隔最大。SVM求解超平面的过程可以转化为一个凸二次规划问题进行纠结。超平面示意图如下所示。



在上图中超平面H对样本的分割能力就要优于H1,两类样本中距离超平面的最小值更大。支持向量机不仅能够处理线性可分的数据，它还引入了核函数的概念，将在原始空间中线性不可分的数据扩展到高维的线性可分的空间中，然后求解超平面实现分类。超平面公式如下。

其中x表示样本特征，w和b就是需要求解的超平面参数。求解超平面的搓成就是确定w与b使得样本点距离超平面的距离最大。这样就越有自信将样本进行分类。这里使用来表示样本点与超平面的距离。而样本类别y为1或者-1。当超平面函数的符号与y一致时表示分类正确。所以可以统一两个判断函数为如下所示。

观察公式发现对w和b按比例进行缩放时超平面并没有改变，所以这里引入几何距离的概念，给定数据x，与其类别y和超平面，则可以计算这个点到超平面的几何距离为。

定义为训练数据中所有样本点的几何距离的最小值。

则SVM的优化问题就转换为了最大化的过程，如下所示。

这个问题还可以转化成一个更易于求解的对偶问题，如下所示。

### 3.4.2 朴素贝叶斯

根据贝叶斯定理，设实验样本空间为S，A为E的事件，,…,为S的一个划分，且P(A) > 0,P(B) > 0(I = 1, 2, …, n)则

贝叶斯分类的思想是设x=[]为一个数据项，其中为x的一个特征属性，类别集合为C = {}，则x所属的类别c可以通过下面的公式计算得出。

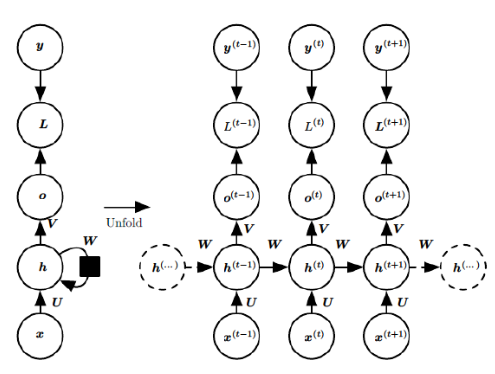
根据贝叶斯定义上面的公式可以转化成下面的公式。

同时朴素贝叶斯假设各个特征属性都是条件独立的，则可以进一步将条件概率化成特征项乘积的形式，如下所示。

在实际使用的过程中由于P(x)的值都是相同的，所以在判断时只需要最大化分子即可。

## 3.5 LSTM进行短文本分析

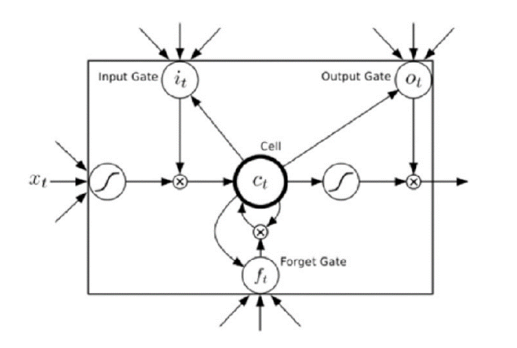
循环神经网络（Recurrent Neural Network, RNN）是为了处理序列数据而设计的神经网络，它不仅将数据传输给网络的下一层，同时还会输出一个隐藏状态，它的结构是一个链状的结构，让它可以处理很长的序列数据。序列中每个节点的状态不仅与当前节点的输入有关，还与之前节点的状态有关，这就让RNN具备了记忆的特性。其结构如下所示。



但它在处理较长序列时容易发生梯度弥散，所以提出了LSTM(Long Short Term) 网络，它是RNN的一种特殊类型，可以学习长期信息。同样LSTM也有很多不同的变体，下面下面主要介绍几种使用广泛的方法。

### 3.5.1 LSTM

在自然语言处理方面，循环神经网络在捕捉任意长度序列的信息有类似的记忆功能。传统的循环神经网络在输入数据过长时，因为非线性变换，在尾部的梯度进行反向传播时不能将梯度传给句子的起始位置，经过训练会出现一个不可忽视的问题：梯度弥散，梯度弥散是一个需要解决的问题。针对这个问题人们提出了一个机遇循环神经网络的新的网络模型——LSTM，LSTM可以理解为升级版的循环神经网络。长短时记忆网络解决了循环神经网络中存在的梯度弥散的问题，LSTM提出了一个叫做记忆单元的新结构。LSTM由四个部分组成：输入门、输出门、遗忘门和记忆单元。LSTM的网络及具体结构如下图。



下面就几个公式介绍一下LSTM具体的工作流程：

第一步：新输入前状态决定C哪些信息可以舍弃与运算，对部分信息进行去除：

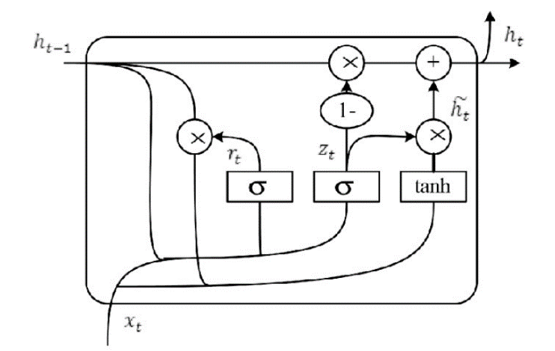
第二步：新输入前状态决定C哪些信息想要保存，新信息添加时的系数（对比）单独新数据形式的控制参数，用于对进行更新。

第三步：根据就的控制参数，新生成的更新控制参数，组合生成该时刻最终控制参数。

第四步：根据控制参数产生此刻的新的LSTM输出。

### 3.5.2 GRU

GRU是长短时记忆网络的一个重要变种，它从两个方面做出了改变，首先是遗忘门和输入门的结合（不是独立，不是互补），将其合成了一个单一的更新门。其次，控制参数与输出结合，直接产生带有长短记忆能力的输出。而GRU同样还混合了细胞状态和隐藏状态，和其他一些改动。GRU与LSTM效果类似，GRU比标准LSTM简单，GRU所使用的参数要少于LSTM中使用的个数。具体结构如下图所示。



下面就几个公式介绍一下GRU具体的工作流程：

第一步：新输入前状态和权重对进行更新。

第二步：新输入前状态和权重计算，从新得到的内容中抽取信息：

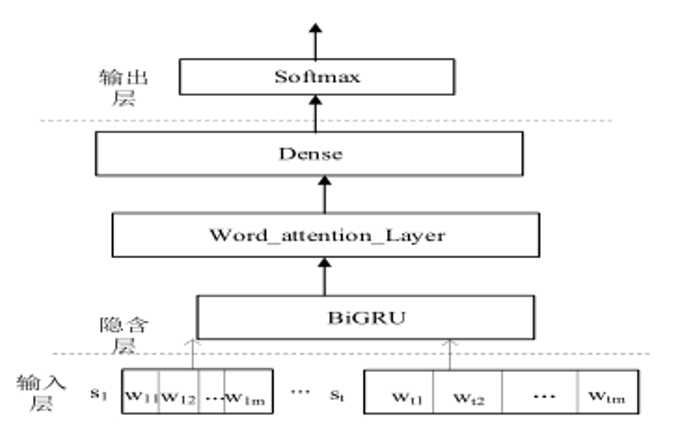
第三步：新输入前状态和权重W计算，为0~1之间的值，是对的一个约束，对的幅度的调整

第四步：根据更新的前状态和计算得到的产生此刻的新的GRU输出。

GRU计算的过程中不去要C，GRU将C合并到了中，在GRU中做不断地更新，更新的方式是遗忘多少就添加多少。

### 3.5.3 BiGRU-Attention

BiGRU-Attention模型共分为三部分：文本向量输入层、隐含层和输出层。其中隐含层由BiGRU层、attention层和Dense层三层构成。BiGRU-Attention模型结构如下图所示。



因为输入层即对数据进行数据预处理，所以这里重点介绍隐含层和输出层

隐含层的计算主要分为两个步骤完成：

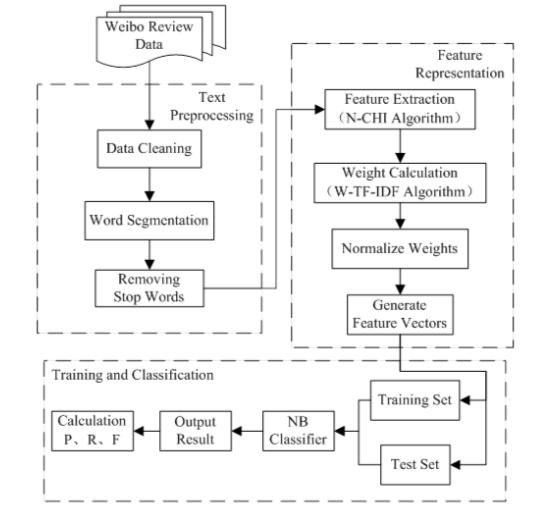
(1)计算BiGRU层输出的词向量。文本词向量为BiGRU层的输入向量。BiGRU层的目的主要是对输入的文本向量进行文本层次特征的提取。根据BiGRU神经网络模型图，可以把BiGRU模型看做由向前GRU和反向GRU两部分组成，在这里简化为下式。在第i时刻输入的第j个句子的第t个单词的词向量为c\_ijt，通过BiGRU层特征提取后，可以更加充分地学习上下文之间的关系，进行语义编码，具体计算攻击如下所示。

计算每个词向量应分配的概率权重。这个步骤主要是为不同的词向量分配相应的概率权重，(2) 进一步提取文本特征，突出文本的关键信息。在文本中，不同的词对文本你情感分类起着不同的作用。地点状语、时间状语对文本情感分类来说，重要程度极小；而具有情感色彩的形容词对文本情感分类却至关重要。为了突出不同词对整个文本情感分类的重要度，BiGRU-Attention模型中引入了attention机制层。Attention机制层的输入为上一次层中经过BiGRU神经网络层激活处理的输出向量，attention机制层的权重系数具体通过以下几个公式进行计算。

其中：为上一层BiGRU神经网络层的输出向量，表示权重系统，表示偏置系数，表示随机初始化的注意力矩阵。Attention机制矩阵由attention机制飞陪的不同概率权重与各个隐层状态的乘积的累加和，使用softmax函数做归一化操作得到。

# 第四章 方法阐述与实验部分

## 4.1 方法阐述

实验所用基于机器学习的细粒度的短文本情感分析对原有情感分析做出了改进，处理流程如下图所示：

首先进行的是文本预处理。为了提取特征并进行分析，首先要对文本进行预处理，去除多余的干扰信息，保留有用的信息。文本预处理是情感分析的基础和重要环节，预处理的质量对于情感分类和后期信息处理非常重要。文本预处理的内容包括数据清理、分词和删除停顿字等。数据清理是清理掉与主体相似的干扰项，分词和删除停顿字目标是尽可能地完全删除数据集中的冗余信息，并保留最有价值的内容。

接下来是特征提取、权重计算、规范化、生成特征向量。如果所有词都直接作为情感分析的特征出现，将导致数据量庞大，效率低，因此，需要选取情感倾向较强的词作为特征，并计算合适的特征维度，这样可以有效避免因特征维度过大而导致的过度拟合，提高分类的效率。常用的权重计算方法有频率法、TF-IDF法等，情感分析一般TF-IDF及其改进算法。选择的特征和计算的权值并没有直接输入到分类器中，而是储存在特征向量中，文本d的n维特征向量表示为：d = [(t1,w1,(t2,w2,(t3,w3,...,(tn,wn))。ti表示文本的第n个特征，wi表示第n个特征的权值。

最后是构造、训练分类器，分析结果。有很多方法来构造分类器,如基于统计方法,基于机器学习的方法,基于深度学习的方法。结合文本分类的特点，采用NB构造分类器进行分类，与其他较为成熟的算法相比，取得了较好的分类结果，公式如下:

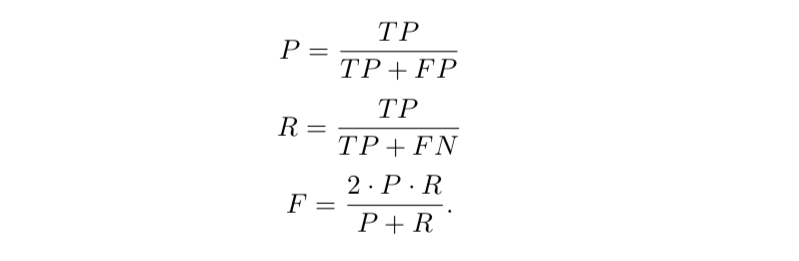


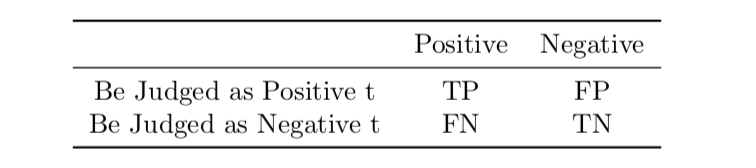
## 4.2 实验设计

### 4.2.1 实验数据

实验使用COAE 2014（汉语倾向性分析与评价的第六个语料库）task3和task4语料库作为原始数据，数据集包含了被标注为正、负两种情绪的微博评论，分别是1138篇正评论和1193篇负评论。经过处理后，有1048篇正面评论，993篇负面评论，包括6222字。其中90%的数据为训练集，剩余10%为验证集。通过对知网的情感词典进行增删修改，生成了本实验使用的情感词典，包括7164哥积极词汇和12060个消极词汇。

### 4.2.2 评估方法

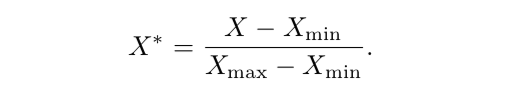
在实验中，我们使用精度（P）、召回率（R）和F测度（F）作为评价标准。计算方法如公式所示：

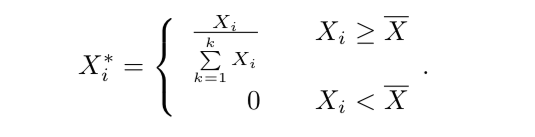
含义如下表所示：

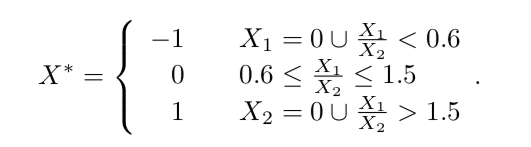
### 4.2.3 参数规范化

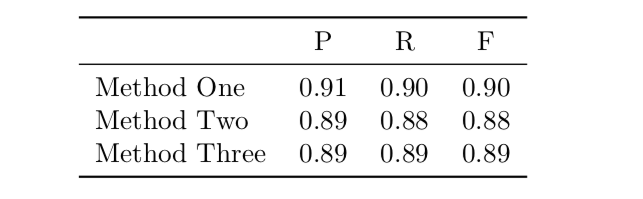
参数规范化的目的是将权值更好地应用于分类器，规范化的方法有很多，方法的选择对数据处理的效率和准确性有很大影响。故对参数规范化方法进行实验比较；

方法一：线性规范法通过线性变换将原始数据缩放到范围[0,1]。设X\*是规范化后的数据，X是原始数据，Xmax和Xmin分别是原始数据集的最大值和最小值，规范化公式如下所示：



方法二：将一类特征的权重除以每一类特征的权重之和，将特征权重规范化到[0，1]，同时考虑低词频时的影响。当特征权重大于特征项，对所有类别的平均值进行规范化，否则为零。规范化公式如下所示：

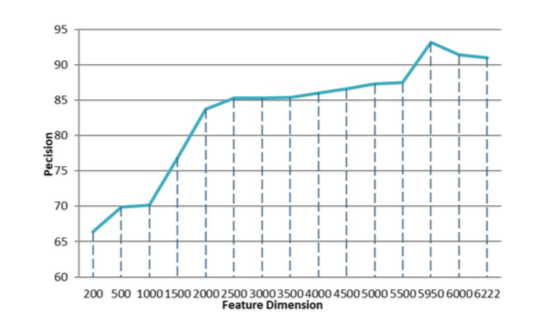
方法三：因为最终的目的是判断文本倾向，权重范围缩小至[-1,0,1]。X1为特征t在正类中的权重，X2为特征t在负类中的权重。规范化公示如下所示：

使用上述三种规范化方法分别对特征权重进行处理，的到的结果如下表所示：

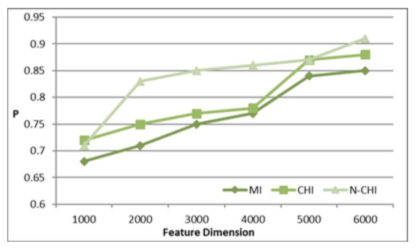
可以看出，第一种方法具有较高的P、R、f值且该方法在权重上保留了情词的优势，而第二种方法忽略了情词的功能。当处理低频词,范围太大,从而导致更多的重量是0,虽然第三种方法更直观的让体重接近的分类,但该方法忽略了功能词的重量大小,和范围选择有一些局限性,可伸缩性不强。因此本文选择方法一作为参数规范化化的方法。

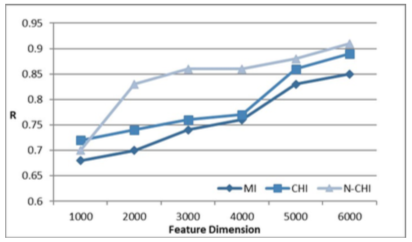
## 4.3 实验结果

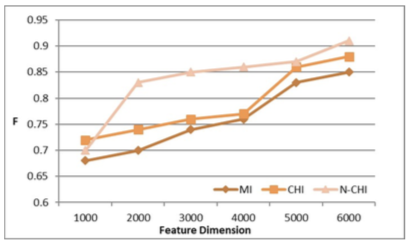
实验一：特征维度的选择。在分类训练过程中，特征维的选择一直是非常重要的一部分。当特征词的数量较少时，这些特征不能充分表达文本中所表达的情感，导致正确率较低。如果特征太多，在训练过程中容易导致过度拟合。通过大量实验发现，当特征维数在5800-5900之间时，分类精度最高，分类效果最好。因此，我们选择5950作为特征的维数，实验结果如下图所示。



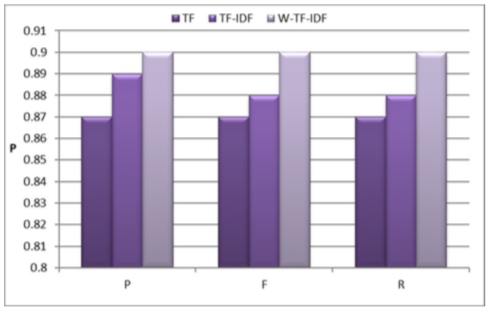
实验二：对N-CHI算法改进。本实验进行了三次对比试验设计，三组实验均采用TF-IDF作为权重计算:第一组采用互信息作为特征选择方法，命名为MI;第二组使用CHI作为特征选择方法，称为CHI;第三组采用改进的N-CHI作为特征选择方法，称为N-CHI。随着尺寸的增大三种方法的，P、R、F测度实验结果分别如下图所示。





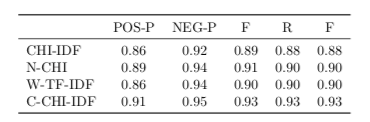


实验二：对TF-IDF算法改进。CHI作为特征选择的方法:第一组以词频作为权重计算方法，称为TF;第二组采用TF-IDF作为权重计算方法，称为TF-IDF;第三组采用改进的W-TF-IDF作为权重计算方法，称为W-TF-IDF。实验结果下所示。



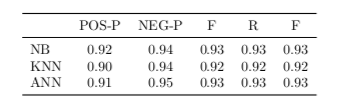
从图中可以看出，对权重计算方法的改进虽然促进效果不明显，但P、R、F测度均有所提高。

实验四：改进验证算法。本文对特征选择和权重计算方法进行了改进。为了验证其改善效果，设计了四组实验:第一组没有任何改善，称为CHI-IDF;第二组是只改进特征提取的方法;第三类是只改进权重计算的方法;第四组对文本进行了两次改进，命名为C-CHI-IDF，实验结果如下表所示。



从上表可以看出，无论是正精度、负精度还是整体精度，在改进特征提取和权重计算方面都取得了较好的效果。

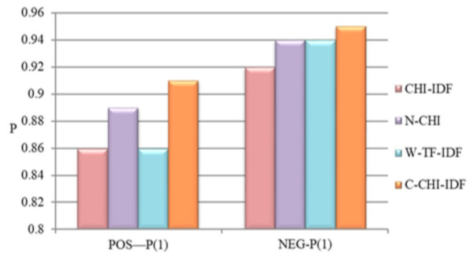
实验五：选用不同的分类算法构造分类器，会得到不同的分类结果。根据文本情感分析的分类任务,用改进的C-CHI-IDF方法选择朴素贝叶斯(NB),哪个更适合文本分类和资讯和人工神经网络(ANN)来构建分类器,并比较分类结果,分类结果如下所示。



可以看到，在样本量较小的情况下，NB与KNN的分类阳离子效应没有显著差异。但是，数据量继续增加。NB的分类效果明显高于KNN，且NB阳性的分类效果更好。神经网络的分类准确率与其他两种分类准确率没有显著差异，但对于样本不平衡的神经网络具有更好的性能。

## 4.4 结果分析

通过以上实验结果和简要分析，可以看出基于机器学习的细粒度文本情绪分类方法是有效的。为了更直观地反映实验结果，我们将一些数据转换成图形的形式。如下图所示：



从图中我们还发现，虽然否定类的句子数量远低于主动类，但是否定类方法的正确率明显高于主动类方法的正确率。这说明消极句中的词更容易被选为特征，而在汉语中，人们在表达消极情绪时更容易携带消极词汇。

# 第五章 总结

本论文综述首先分析了短文本的特点，并针对这些特点展开讨论。

第一章介绍了短文本的研究背景，以及研究现状。随着电子商务，以及互联网的迅速发展，短文本分析显得越来越重要。人们每天在社交媒体中发布的信息具有很高的价值，这些信息成为公共事件、舆情研究，甚至个人情绪观点的有价值的数据源泉。

最近几年在相关领域的研究也不断推陈出新，出现很多新的模型和方法。针对短文本的特点，学者提出了WordNet、维基百科、主题相关的辅助文本等方法对短文本特征信息进行扩充或转换。同时也有学者提出了能够直接应有于短文本主题挖掘的方法。

第二章介绍了与短文本分析相关的一些知识。分析了隐性语义分析（LSA）模型的，神经语言模型（NLM），段向量（PV），半显性语义模型，显性语义模型等模型。这些模型将文本信息映射到了不同的向量空间中，利于分析。

第三章介绍了短文本分析的主要方法。根据短文本分析的步骤可以主要分成特征表示，特征扩展，特征选择过程，以及最终的机器学习方法。分别介绍了SVM以及朴素贝叶斯两种传统的机器学习方法，并着重分析了基于LSTM及其变种的深度学习模型，发现深度学习模型在短文本分析领域具有巨大的潜力。

第四章针对最新的研究成果进行详细分析和实验，阐述了文章提出的方法，并进行了不同的实验，根据实验结果评估了不同方法的实际表现。实验结果显示了基于机器学习的细粒度文本情绪分类方法的有效性。

# 第6章 致谢

首先感谢张老师一学期以来的教导，在课堂中学习了知识，扩展了视野。同时也感谢组员们的努力，在解决问题的过程中相互支持与鼓励。

# 参考文献

[1] Frey B J , Dueck D . Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315(5814):972-976.

[2] Hong L , Davison B D . Empirical study of topic modeling in Twitter[C]// Sigkdd Workshop on Social Media Analytics. 2010.

[3] Weng J, Lim E P, Jiang J, et al. TwitterRank:finding topic-sensitive influential twitterers[J]. 2010.

[4] Gabrilovich E. Feature generation for textual information retrieval using world knowledge[J]. Acm Sigir Forum, 2007, 41(2):123-123.

[5] Hu X, Sun N, Zhang C, et al. Exploiting internal and external semantics for the clustering of short texts using world knowledge[J]. 2009.

[6] Sahami M, Heilman T D. A web-based kernel function for measuring the similarity of short text snippets[C]// Proc of World Wide Web Conference. 2006.

[7] 范云杰, 刘怀亮. 基于维基百科的中文短文本分类研究[J]. 数据分析与知识发现, 2012, 28(3):47-52.

[8] 一种基于维基百科的中文短文本分类算法[J]. 图书情报工作, 2013, 57(11):120-124.

[9] 面向短文本分类的特征扩展方法[D]. 哈尔滨工业大学, 2013.

[10] Ou J, Liu N N, Kai Z, et al. Transferring topical knowledge from auxiliary long texts for short text clustering[C]// Acm International Conference on Information & Knowledge Management. 2011.

[11] Cheng X, Yan X, Lan Y, et al. BTM: Topic Modeling over Short Texts[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(12):2928-2941.

[12] 邱云飞, 赵彬, 林明明,等. 结合语义改进的K-means短文本聚类算法[J]. 计算机工程与应用, 2016, 52(19):78-83.

[13] Seifzadeh S , Farahat A K , Kamel M S , et al. Short-Text Clustering using Statistical Semantics.[J]. 2015.

[14] Elkan C. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution[C]// International Conference on Machine Learning. 2006.

[15] Mccallum A K, Mccallum A K, Thrun S, et al. Text Classification from Labeled and Unlabeled Documents using EM[J]. Machine Learning, 2000, 39(2-3):103-134.

[16] Smyth P. Model selection for probabilistic clustering using cross-validated likelihood[J]. Statistics & Computing, 2000, 10(1):63-72.

[17] Antoniak C E. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems[J]. Annals of Statistics, 1974, 2(6):1152-1174.

[18] Ishwaran H, James L F. Gibbs Sampling Methods for Stick-Breaking Priors[J]. Journal of the American Statistical Association, 2001, 96(453):161-173.

[19] Neal R M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models[J]. Journal of Computational & Graphical Statistics, 2000, 9(2):249-265.

[20] Guan Y, Huang R, Wang Z. Document clustering via dirichlet process mixture model with feature selection[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2010.

[21] Huang R, Guan Y, Wang Z J. Dirichlet Process Mixture Model for Document Clustering with Feature Partition[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(8):1748-1759.

[22] Yin J, Wang J. A dirichlet multinomial mixture model-based approach for short text clustering[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2014.

[23] DEERWESTER, DUMAIS, S. T, et al. Indexing by latent semantic analysis[J]. Journal of the Association for Information Science & Technology, 2010, 41(6):391-407.

[24] Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence[J]. Behavior Research Methods Instruments & Computers, 1996, 28(2):203-208.

[25] Turney P D, Pantel P. From Frequency to Meaning: Vector Space Models of Semantics[J]. Journal of Artificial Intelligence Research, 2010, 37(1):141-188.

[26] Bollacker K, Evans C, Paritosh P, et al. Freebase:a collaboratively created graph database for structuring human knowledge[C]// Sigmod Conference. 2008.

[27] Collobert R, Weston J. A unified architecture for natural language processing:deep neural networks with multitask learning[C]// International Conference on Machine Learning. 2008.

[28] Li S, Wu C, Hai L, et al. FPGA Acceleration of Recurrent Neural Network Based Language Model[C]// IEEE International Symposium on Field-programmable Custom Computing Machines. 2015.

[29] Le Q, Mikolov T. Distributed representations of sentences and documents[C]// International Conference on International Conference on Machine Learning. 2014.

[30] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2012, 3:993-1022.

[31] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis[C]// Proc International Joint Conference on Artificial Intelligence. 2016.

[32] 翟延冬, 王康平, 张东娜, et al. 一种基于WordNet的短文本语义相似性算法[J]. 电子学报, 2012, 40(3):617-620.

[33] 王连喜. 微博短文本预处理及学习研究综述[J]. 图书情报工作, 2013, 57(11):125-131.

[34] Raffel C , Ellis D P W . Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems[J]. 2015.

[35] 韩忠明, 张玉沙, 张慧, et al. 有效的中文微博短文本倾向性分类算法[J]. 计算机应用与软件, 2012(10):89-93.

[36] 郭永辉. 面向短文本分类的特征扩展方法[D]. 2013.

[37] 王盛, 樊兴华, 陈现麟. 利用上下位关系的中文短文本分类[J]. 计算机应用, 2010, 30(3):603-606.

[38] 王月. 基于互信息的短文本分类技术[D]. 东北大学, 2012.

[39] 江大鹏. 基于词向量的短文本分类方法研究[D]. 2015.

[40] Zhao H , Du L , Buntine W , et al. MetaLDA: a Topic Model that Efficiently Incorporates Meta information[J]. 2017.

[41] 彭泽映, 俞晓明, 许洪波, et al. 大规模短文本的不完全聚类[J]. 中文信息学报, 2011, 25(1):54-60.