

# 硕士学位论文

## 基于机器学习的微博评论信息倾向性分析 的研究

### RESEARCH ON THE ANALYSIS OF WEIBO COMMENTS TENDENCY BASED ON MACHINE LEARNING METHODS

汪淳

哈尔滨工业大学  
2016 年 6 月

国内图书分类号: TP393  
国际图书分类号: 621.3

学校代码: 10213  
密级: 公开

## 工学硕士学位论文

# 基于机器学习的微博评论信息倾向性分析 的研究

硕 士 研 究 生: 汪淳

导 师: 李东教授

申 请 学 位: 工学硕士

学 科: 计算机科学与技术

所 在 单 位: 计算机科学与技术学院

答 辩 日 期: 2016 年 6 月

授予学位单位: 哈尔滨工业大学

Classified Index: TP393  
U.D.C: 621.3

Dissertation for the Master Degree in Engineering

**RESEARCH ON THE ANALYSIS OF WEIBO  
COMMENTS TENDENCY BASED ON MACHINE  
LEARNING METHODS**

<b>Candidate:</b>	Wang Chun
<b>Supervisor:</b>	Prof. Li Dong
<b>Academic Degree Applied for:</b>	Master of Engineering
<b>Speciality:</b>	Computer Science and Technology
<b>Affiliation:</b>	School of Computer Science and Technology
<b>Date of Defence:</b>	June, 2016
<b>Degree-Confering-Institution:</b>	Harbin Institute of Technology

## 摘 要

本文的重点研究对象是微博评论信息的倾向性分析，主要的研究内容是以新浪微博中的评论为研究对象并且对评论中的情感倾向进行研究。本研究是将几种特征进行提炼、融合通过改进的机器学习方法来增强分类效果。

情感分析在舆论监控、商品检验有着广泛的应用。基于此，本文提出一个设想，把评论划分成三种类型：垃圾评论、主观评论、客观评论。针对不同类型的评论选取相应的方式进行分析，主观评论褒贬倾向是此研究分析的重点。

本文首先对评论数据进行清理，剥离垃圾评论以及客观性评论。其中利用几种特征的有效融合并结合朴素贝叶斯、阈值划分等技术方法判断垃圾评论、客观评论，大大降低了文本的噪声。

其次，针对文本褒贬倾向性分析，通过比较几种特征提取方法，并在其基础上改进情感词的选取方式和权值计算方式，构成新的文本向量空间。通过集成学习方法以及投票方式将传统的机器学习算法进行融合，达到更好的分析效果。本文实现了针对情感词的特征提取和权值计算的性能提升，使用 AdaBoost、Random Subspace、融合分类器组合方式提升传统机器学习方法，提高评论分析的准确率。

最后，通过性能评估方法说明本文的方法对评论分析具有很好地效果，同时本文针对评论的情感转移以及情感载体进行分析，判断褒贬数据集的情感载体异同，起到舆情分析预警作用。

**关键词：**微博评论；机器学习；情感分析；特征融合；集成学习

## Abstract

This paper focuses on research object of orientation analysis of Weibo comment. The main research content is Xinlang microblogging reviews for the study and the emotional tendency. The research refines several features of the study and through improved integration of machine learning methods to enhance the accuracy of classification.

Sentiment analysis has been widely used in monitoring public opinion and commodity inspection. Based on this, the paper proposes an idea that divides the comment into three types, which contains spam comments, subjective or objective comments, and appraise or critical comments. To select the appropriate mode for analysing different types of comments different types of comments.

Firstly, the paper clean up the data of comment that peels off the spam comments and objective comments. The paper uses effective integration of several features, Naive Bayes and the method of dividing threshold to determine spam and objective comments. This method greatly reduces the noise of the comment.

Secondly, contrary to the comments of appraise or critical, by comparing several feature extraction methods and improve emotional words based on its selection method and calculated weights method to constitute a new comment vector space. By combing machine-learning methods and voting method to merge traditional machine learning methods. This method can achieve a better analysis effect. This paper implements the emotion words based on its selection method and calculated weights method, and use AdaBoost, Random Subspace and fusion classifion to enhance traditional machine learning methods, which can improve the accuracy of the analysis of comments.

Finally, through the performance evaluation methods describe the method for the analysis of public opinion has a good effect. The paper expands emotional shift and emotional carrier for the comments, juding similarities and differences of emotional carriers of appraise and critical dataset, which can play a role in early warning public opinion.

**Keywords:** Weibo comment, machine learning, sentiment classification, feature fusion, ensemble learning

# 目 录

摘 要.....	I
Abstract .....	II
第 1 章 绪 论 .....	1
1.1 课题背景及意义 .....	1
1.2 国内外研究现状 .....	2
1.3 本文的主要研究内容及论文结构 .....	4
第 2 章 微博评论分析的相关理论与技术 .....	6
2.1 社交媒体平台概况 .....	6
2.2 微博评论相关技术和资源 .....	6
2.2.1 预处理技术 .....	7
2.2.2 文本的向量表示 .....	7
2.2.3 多类词典的构建 .....	8
2.2.4 情感分类的评价标准 .....	9
2.3 本章小结 .....	10
第 3 章 微博非情感评论信息的识别 .....	11
3.1 微博垃圾评论信息识别 .....	11
3.1.1 微博垃圾评论信息介绍 .....	11
3.1.2 基于垃圾线索的微博垃圾评论识别 .....	12
3.1.3 实验结果与分析 .....	16
3.2 微博主客观评论信息识别 .....	18
3.2.1 微博主客观评论信息介绍 .....	18
3.2.2 基于机器学习方法的主客观评论识别 .....	19
3.2.3 实验结果分析 .....	23
3.3 本章小结 .....	24
第 4 章 微博评论信息倾向性分析 .....	25
4.1 微博评论情感识别介绍 .....	25
4.2 特征选择 .....	25
4.3 特征提取及权重的计算 .....	27
4.3.1 传统的特征提取方法 .....	27

4.3.2 特征提取的改进 .....	28
4.3.3 权重计算 .....	30
4.4 基于投票与集成学习的融合分类器 .....	31
4.4.1 传统机器学习方法 .....	32
4.4.2 集成学习方法 .....	34
4.4.3 基于投票与集成学习的融合分类器 .....	37
4.5 实验结果与分析 .....	42
4.5.1 实验数据 .....	42
4.5.2 特征提取与权重计算改进实验结果 .....	42
4.5.3 分类器改进的实验结果 .....	43
4.5.4 实验结果分析 .....	46
4.6 本章小结 .....	49
<b>第 5 章 微博评论倾向性载体分析 .....</b>	<b>51</b>
5.1 微博评论话题转移 .....	51
5.2 文本表示与 LDA 主题模型 .....	51
5.2.1 文本表示 .....	51
5.2.2 LDA 主题模型 .....	52
5.3 实验结果与分析 .....	53
5.3.1 实验数据 .....	53
5.3.2 实验结果与分析 .....	54
5.4 本章小结 .....	56
<b>结 论 .....</b>	<b>57</b>
<b>参考文献 .....</b>	<b>59</b>
<b>攻读学位期间发表的学术论文 .....</b>	<b>63</b>
<b>哈尔滨工业大学学位论文原创性声明和使用权限 .....</b>	<b>64</b>
<b>致 谢 .....</b>	<b>65</b>

# 第1章 绪 论

## 1.1 课题背景及意义

目前社会上的交流、沟通手段都在不断地发展以及扩充。互联网用户的数值以每年翻倍式增长,使用人数占全世界比例从1995年只有0.6%左右增长到2014年占有比例为39%。世界民众的使用人口分布逐年变化,目前中国使用互联网人数已约占世界总使用人数的三分之一,其比例还在不断上升。而微博的迅速发展和崛起,使得微博成为了邮箱和QQ等交流工具后一个新的互联网上的重要的应用。这种民众间的信息交互已经成为人们在生活上不可或缺的交流方式,其影响力也在不断的扩大,不仅停留在民众等普通用户的层面上,更加不断深入到了企业、政府等社会组织上的方方面面,由于网络监管力度不够成熟和完善,网络上的信息真实性与可信度不高,存在着很多安全隐患如:谣言、假消息肆意散布。网络环境需要有效的清理与改善,并且政府等部门也正在密切关注大众对热点事件的反响。网上的报告显示<sup>[1]</sup>,到2015年6月,中国已经有了2.04亿人在使用微博,占总使用人数约有30%。而大约1.62亿通过移动端微博来交流沟通,比例达到27.3%,和2014年相比上升了约有10个百分点。从图1-1和图1-2可以看出,互联网对人们生活越来越重要,微博对人们的日常生活的重要性不断加重。因此如何判断微博社交平台上群众观点的倾向性,已经成为舆情分析急需解决的问题之一。

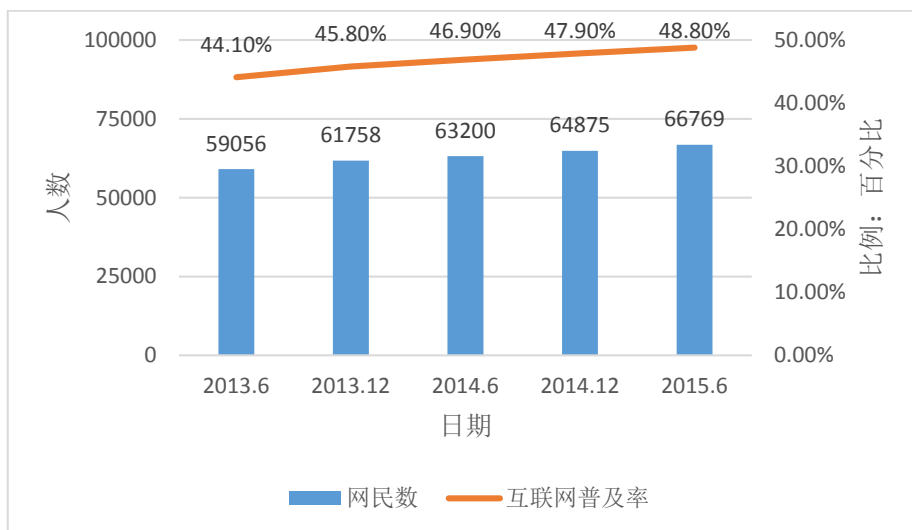


图 1-1 互联网网民规模及普及率



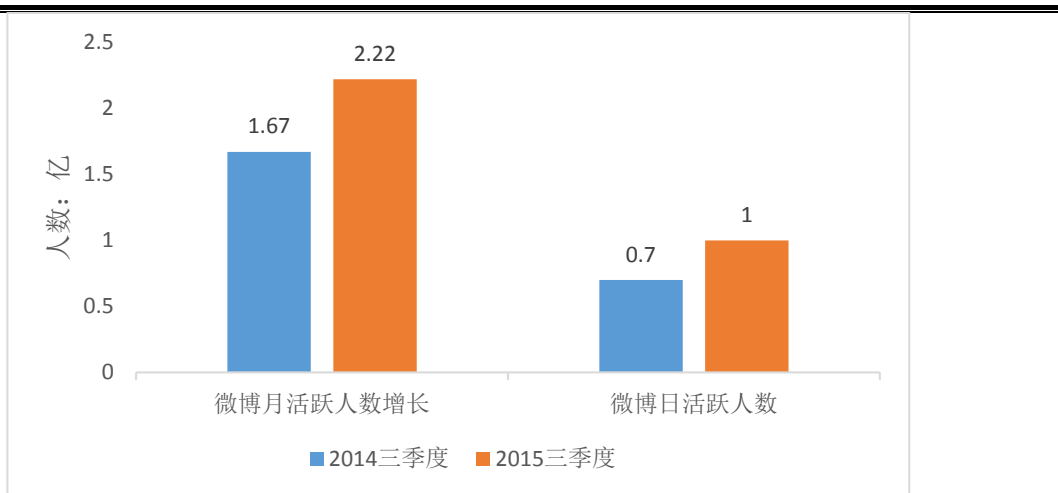


图 1-2 微博人数增长趋势

由于微博用户数量的庞大增长,微博在各个领域迅速的成为舆论的发起点,并且很明显的影响大众对于突发事件的看法,例如:汶川地震事件,不仅展现出了微博平台的影响程度,同时对此事件起到了监督以及增强救援作用。微博平台的快速、简单的信息传播方式,使得大家能够快速的实现即时信息的交流。然而评论信息却并没有什么特定的规律,其中包含了大量的没有章法的信息,所以微博评论信息研究价值非常高。而在不同的话题上,人们对话题的反应都有所区别,每个人都描述自己对同一事件的观点,进而发表自己的评论。通过评论内容汇总形成文本流<sup>[2]</sup>,研究文本流中的情感倾向,若这些文本流所表现出的情感有所偏差,那么说明事件有很大的影响力或者事件有需要关注的问题点,需要把握好当前时下人们对于热门事件具体看法。政府等部门可以根据情感的流向来进行舆情监控,提供一定的决策方式,对人群进行一定的心理疏导,防止有过激的行为和预防暴力事件出现。

因此,无论是对于舆情的监督还是商品评论的统计等,都需要针对评论获得情感倾向性,而仅仅通过人工进行统计,获取结果的效率过低,所以自动分析情感倾向的方式是十分必要的。

## 1.2 国内外研究现状

目前情感分析主要是分为两种方式,分别为无监督和有监督方式。国外的研究成果,大部分针对 Twitter 上的每个用户情感倾向性做研究,文献<sup>[3-5]</sup>通过匹配一定的规则来辨别情感倾向性,其主要是抽取评论信息特点,将评论信息按不同情况划分,记录评论的特征,将数据集按照规则匹配来划分类别,没有用到机器学习等方法。Liu B.<sup>[6]</sup>通过相关领域的知识来构建词典库,想通过特定方式来实现不同领域之间的评论分析,在跨领域研究上进一步发展,得到了

较好的分类效果。Turney<sup>[7]</sup>通过词典中的贡献值来判断文本的情感倾向，但是效果以及规则之间很难相兼容，所以过于依赖情感词典是不可取的。至于情感词典的无监督学习，国外的早期运用有 Neviarouskaya 等人<sup>[8]</sup>设计的模型，其针对 Twitter 上的简单句式评论取得较好效果，但是对于复杂句式，由于人工选取构建的情感词典人力有限，效果很不理想，可见只是单独运用情感词典过于片面。

Prasad 等人<sup>[9]</sup>运用贝叶斯的机器学习方法将博客分类，也取得了一定的成果，这是较早的机器学习运用到情感分类。Pang 等人<sup>[10]</sup>对英文的有关电影评论情感值进行判断，其效果在早期较为显著。Davidov 等人<sup>[11]</sup>对于 Twitter 中的有关于文本的内容进行人工标注，提取出了其中的主题标签以及表情符号，通过 K 近邻算法分类器，来实现情感分类。Go 等人<sup>[12]</sup>实现使用距离的学习方法来形成训练数据，通过主题和关键词分析并结合相应学习方法得到了较好的结果。在学习方法方面不断成熟条件下，Li S.<sup>[13]</sup>将句子分为主观、客观两类，通过有监督的学习进行测试，同时采用了半监督的学习方法和其进行比对，获得了对句子主客观分类的较好效果。Jiang L 等人<sup>[14]</sup>将 Tweets 文本分类的问题看做一个普遍的分类问题进行处理，运用机器学习的处理方式分类主客观内容，通过提取一些相应的特征，来实现其分类目标。Barbosa 等<sup>[15]</sup>将 Tweets 的多种特征进行比较并且按照权重抽取特征，得到特征权值大小的排序，对特征选取做出了很大贡献。Agarwal 等人<sup>[16]</sup>选取了针对 Twitter 自身特点，由于口语以及符号占有一定比例，较为有针对性抽取了自身特征。Wilson T.<sup>[17]</sup>对于短语形式的情感分析提出了自己方法，通过自己设定的分类器进行情感倾向性的判断。在针对 Twitter 的特殊性上，Mohammad 等人<sup>[18]</sup>进行一系列有关于特征的工程性建设，有较强的针对性并且综合了多方面的优势，对 Twitter 上评论分析有很好的工程性判断。

目前中文微博的研究也正在不断发展，虽然还处于一定的起步阶段，但相关的研究内容正在不断丰富。国内的研究方式则大致上有两种。

基于情感词典以及规则的无监督学习方法上，徐琳宏<sup>[19]</sup>通过词汇和对于句子的结构分析，构建了一个最初步的情感词典，通过情感词典的匹配来分析句子情感极性，是研究的初步阶段。李钝、曹付元等<sup>[20]</sup>对词典中词在短语中的语义进行分析，获得语义倾向度，通过短语权值计算句子的情感倾向。柳位平等<sup>[21]</sup>通过词语之间的相似度计算，设计每个词的情感值，使用 Hownet 情感词语，构建情感词典，计算中文文本情感值，从而判断情感倾向。张成功等人<sup>[22]</sup>提出了基础词典、领域词典、网络词典、修饰词词典等内容并且将词的极性排序，通过规则分析评论，有一定的效果，但是针对性过强，不适合推广。党蕾<sup>[23]</sup>

将否定词、副词以及句法关系运用到情感分析之中，通过句法树和主题抽取等方式计算句子的情感极性值，此方法对词典的依赖性过强。赵妍研等<sup>[24]</sup>使用句法的自动识别来计算情感单元值，利用了距离和位置的分析和路径匹配算法。这种做法可以更好的消除因句法分析误差而产生的错误的识别问题，对于英文句法分析有一定的帮助，但是对中文还没有进一步证实。

机器学习等方法是目前国内比较主流的研究方向，同时对于特征的选取相关技术也在不断进步。刘志明等<sup>[25]</sup>使用了三种机器学习方式，例如 **SVM** 等。并且对不同的特征提取方式进行了研究，从而选取了信息增益的方式，验证了微博评论的适用性非常依赖于评论的风格和主题内容。而王志涛<sup>[26]</sup>在新浪微博分析上使用支持向量机的机器学习方式，得出和主题相关内容和主题无关内容有一定差别，文本特征的选取较为重要特征，针对性强的文本特征效果可以有更好效果。闻彬，何婷婷等<sup>[27]</sup>对情感词赋予语义上的相似度，将语义方面的表示含义投射到情感倾向之中，达到分类效果。谢丽星<sup>[28]</sup>提出了根据层次结构设计框架来分析微博情感，其中包括了提出部分微博文本的属性作为特征，例如：情感词典、表情符号等，通过层次结构可以进一步融合不同特征，可以使微博文本的分类更准确。傅向华等<sup>[29]</sup>针对博客上的文档进行分类，首先运用 **LDA** 进行话题抽取与剥离，其次使用 **Hownet** 词典对文本段落进行情感打分，通过段落的情感值来合成最后博客的文本倾向度，有 90% 左右的准确度，但是博客的篇幅比较大并且用词更规范也更容易进行分析，并不一定适用于评论的分析。周胜臣等<sup>[30]</sup>提出了微博相较于传统文本的区别点和共同点，例如主客观分类、主题的抽取等，无论是对目前方法的应用，还是对后期的研究都有很大帮助以及指导作用。李泽魁、赵妍研等<sup>[31]</sup>将目前情感分析的研究做出了一个全面的综述，包括对词特征、词组特征、数值特征等特征的组合进行测试，得出了一些比较有利的特征组合，能够为后续研究提供较大帮助。

通过对比来看，中文和英文无论是在语法、词性还有用法习惯上都有一定的差距，中文还涉及到英文无需的分词技术等，所以中文分析要更加复杂。另外，中文评论的表情符号等，在英文评论中也没有太具体的体现，可见中文文本的研究还需要找出更好的解决方案。

### 1.3 本文的主要研究内容及论文结构

本课题首先将评论信息进行分类，主要分为垃圾评论、主客观评论以及具有情感倾向评论，针对不同类别采取各自的研究方式，从而对评论进行系统分析。除了不同类型的评论划分，同时对**SVM**等分类器进行深入研究、找出其中的不足，通过集成和融合的方式提高分类器性能，最终研究出评论的自动分析

方法，高效的实现社会类微博的评论分析。

本文共有五章，每个章节内容如下所示：

第一章，主要是介绍本研究课题的课题来源和背景意义，介绍了国内外的研究现状和情感分析方面已有的研究成果。

第二章，通过微博特点和形式的介绍，结合文本处理的相关知识，阐述情感分析的定义及用到的技术等。

第三章，第一部分介绍垃圾评论的特点和垃圾评论的分析方法，提出基于垃圾线索的微博垃圾评论识别方法。第二部分介绍了主客观评论的定义以及主客观评论的识别流程，使用通过机器学习和特征组合方法实现主客观评论识别，并通过实验数据验证效果。

第四章，对于微博主观性评论，通过分析其特征，提出改进选取特征的方法。利用集成和投票方式来提升分类器的性能，从而能够更好的服务于情感倾向性的分类。最后对算法之间的比较和实验结果的分析，证明本文研究内容的可行性。

第五章，针对微博评论倾向性的载体进行分析，使用LDA模型抽取评论的主题关键词，实现了评论载体转移检测，从而起到发出舆情预警等作用。

第六章，总结本文工作、对下一步工作进行规划。

## 第2章 微博评论分析的相关理论与技术

### 2.1 社交媒体平台概况

目前社交媒体有微信、QQ、微博等形式的沟通工具，工具的使用促成了人们对社交媒体相关研究，通过分析媒体上的评论，找出其中的商业价值以及对国家有用的信息。微博已经成为了中国互联网的最重要的应用平台之一，全世界有很多微博形式的社交媒体，例如 Twitter、Youtube、新浪微博、腾讯微博等。由于新浪微博是目前发展最为迅速并且使用数量最大，所以本文的研究主要是针对新浪微博下的评论。微博目前有以下几个特点：

#### （1）评论短小、简洁

新浪微博的内容发布以及评论限制于 140 字以内，用户的用词以及知识领域的深度很随意。随着用户的不断增加，微博已经成为一种很普及的聊天与交流方式，通过不同移动设备如：手机等，能够快速、有效的了解实时信息。

#### （2）更新迅速、有很强的开放性及时效性

热点新闻会非常迅速的通过平台展现出来，实时的更新信息能让所有用户都参与进来，覆盖范围很广并且交流更加容易，跟传统的报纸等信息交流方式上，有很大的提高，逐渐替代了一些传统的交流平台。

#### （3）信息传播快、受众广

人们对于新鲜的事件，总是想快速的告诉他人，促成了人们更加追求快速传播信息的方式。通过用户之间的关注，能够更新动态信息并且进行评论，表达自己的意愿和观点，加强了人们之间的沟通和交流，使得用户量不断增加。

因为微博的用户量非常的庞大，其中的商机<sup>[32]</sup>和对于政府的重要性也在不断地上升，一个微博信息包括了大量评论，从评论中可以很好地看出目前民众对于此事件的看法和热度。评论信息中既包含了垃圾评论，其中有广告、谣言、辱骂等信息，也包含了对于发生事件客观性观点。从这些评论中提取出主观评论才能更好的分析人们针对突发性事件的看法，政府和相关部门才能更好的做出对应举措和对相应的事件进行有效处理。

### 2.2 微博评论相关技术和资源

评论信息的特点主要是短小、符号较多、表达不规范并且其中冗余信息较多，对情感分析十分有影响，需要预处理技术去除冗余信息，再通过文本向量



表示等方式来提炼评论文本信息。

### 2.2.1 预处理技术

#### (1) 文本的编码规范

评论数据十分复杂，其中的编码格式多种多样，这样会使得在文本分析上造成不必要的缺失，所以对要处理的文本进行统一编码十分有用，实验编码格式统一为“utf-8”编码，为后续研究做好铺垫。

#### (2) 文本降噪

评论信息降噪必不可少，由于评论之中常常会出现一些无关紧要的词，例如：“的、和”等，同时可能出现一些不正常的特殊符号如：“&”等符号，去除噪声可以明显增加分词的质量，去噪最主要的方法就是构造停用词词典。

停用词词典的构建。通过构建停用词典来剔除其中的特殊符号、特殊字符、无用标点符号等。停用词的获取上，对研究无用的符号进行统计，并且利用网上的停用词数据，形成最终版本的停用词词表，为“stopwords.txt”。

#### (3) 文本切分和词性标注

相较于英文文本，中文文本是一句或者几句话方式呈现的，每个词与词之间是没有空格的，所以需要运用分词工具，这其中选用的分词工具是 python 的 jieba 分词包，其中包括了分词、词性标注、提取部分关键词等功能。jieba 包的分词效果如图 2-1 所示。

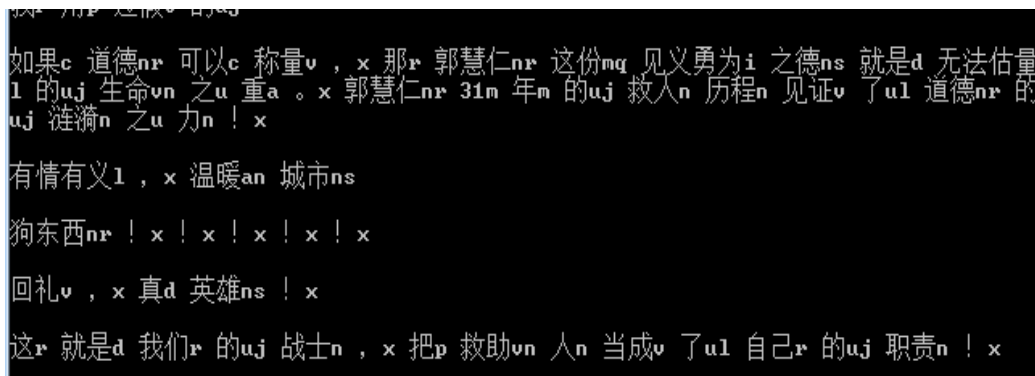


图 2-1 jieba 分词工具分词效果

### 2.2.2 文本的向量表示

文本内容只是一串字符串，并不能作为机器学习方法的输入数据，所以要将数据转化成为向量表示。

文本的向量表示<sup>[33]</sup>，经过分词后将文本串划分为词的组合表示，形式如 $W_n = \{word_1, word_2, word_3, \dots, word_n\}$ ，所有的评论数据，可以表示成为  $P =$

$\{W_1, W_2, W_3, \dots, W_n\}$ , 而这其中每一个评论的标签, 分为四类 0, d, 1, -1。进一步表示成为向量  $M = [m_1, m_2, \dots, m_n]$ , 将评论分为正样本  $m_i = 1$ , 负样本  $m_i = -1$ , 客观样本  $m_i = 0$ , 垃圾样本  $m_i = d$ 。

向量空间模型 VSM(Vector Space Model), 是将文本字符串表示成空间中的多维向量, 这其中用词或者词频等内容作为特征项, 同时每个特征项都对应特征值  $sc$ , 特征向量表示为  $F = [\langle fe_1 \rangle : \langle sc_1 \rangle, \langle fe_2 \rangle : \langle sc_2 \rangle, \langle fe_3 \rangle : \langle sc_3 \rangle, \dots, \langle fe_n \rangle : \langle sc_n \rangle]$ , 其代表特征及相对应的特征值的组合, 构造向量空间是十分重要的一环。

### 2.2.3 多类词典的构建

本文所使用的词典包括褒义词典、贬义词典、否定词词典、关联词词典、表情符号词典、程度词词典。各类词典是通过开源的词典以及自己总结的有用词项合并而成。

HowNet 开源词典对中文总结的程度级别词语、负面评价词语、负面情感词语、正面评价词语、正面情感词语、主张词语。将词语库分别形成自己的词典, 将评价性词语和情感词语合并, 形成了 HW\_POS、HW\_NEG、HW\_DE、HW\_OP 词典库。

同时参照网络公开的《大连理工情感词典词库》<sup>[34]</sup>, 其中包括了词语的种类、词语的情感类别, 将其中符合本文所需要使用的词语挑选出来, 形成 DL\_NEG, DL\_POS 词典库。

台湾大学的 NTUSD 情感词典, 也包括了正、负情感倾向的词库, 包括了 8276 个负向情感词和 2812 正向情感词, 形成 NTUSD\_POS、NTUSD\_NEG 词典库。

清华大学李军<sup>[35]</sup>构建的有关中文正负倾向情感词典, 对每个情感词都详细描述了词性等信息, 为后续的情感词典构建提供了很有意义的帮助, 通过总结和归纳形成 QHCD\_NEG、QHCD\_POS 词典库。

最后通过抓取的几万条评论通过 jieba 分词包进行分词, 进行人工标注将每条评论中有很明显倾向的词进行统计, 观察每个词出现的频率, 将出现次数较多并且能明显表示正负倾向含义的词挑选出来, 将两部分合并, 构建出情感词库, 记为 MSD\_NEG, MSD\_POS。

否定词词典, 否定性评论中情感词的情感极性会反转或者削弱, 为了找出情感反转的评论, 建立否定词典。其中包括了例如: “不是”等 63 个词, 构建否定词库记为 NEGTION 词库。关联词词典, 为了更细致的分析句子中的关联关系和情感之间的权值, 设置了关联词词库, 例如: “虽然, 但是”等关联词,

构建关联词词库，记为 TURN 词库。各类词典的统计如表 2-1 所示。

表 2-1 各类词典统计

词典名称	词语数目	词语举例
HW_POS	4566	成器、聪明
HW_NEG	4370	垃圾、仇视
NTUSD_POS	2810	上好、钦佩
NTUSD_NEG	8276	损坏、极恶
DL_POS	13052	好人、厉害
DL_NEG	14322	蠢、废物
QHCD_POS	5567	自豪、抵抗
QHCD_NEG	4468	诬陷、渎职
MSD_POS	1065	好人、光荣
MSD_NEG	2125	垃圾、滚
NEGTION	63	不、并无
TURN	15	尽管..但是、虽然..但是
HW_OP	103	非常、过于

将各个词典进行统计，尤其是将正负情感倾向词典进行有选择性的合并，最后形成上述的 6 个词典。

## 2.2.4 情感分类的评价标准

本文用查全率(Precision)、查准率(Recall)、F 值(F-Measure)三个指标来评估分类方法的效果，公式 (2-1)、(2-2)、(2-3) 分别为查全率的计算公式、计算查准率公式以及计算分类性能的综合指标 F 值。

$$P = \frac{A}{A+B} \times 100\% \quad (2-1)$$

$$R = \frac{A}{A+C} \times 100\% \quad (2-2)$$

$$F = \frac{2 \times P \times R}{P+R} \quad (2-3)$$

其中，A 表示分类结果为正向并且实际标注为正向的评论个数，B 表示分类结果为正向但实际标注上不是正向的评论个数，C 表示分类结果不为正向但实际标注为正向的评论个数，D 表示分类结果不为正向并且实际标注也不为正向的评论个数。其中的具体表示如表 2-2 所示。



表 2-2 文本之间关系

	实际为正的数量	实际不为正的数量
分类后为正的数量	A	B
分类后不为正的数量	C	D

通过上述几个公式可以检验设计的分类方法的效果，并进一步改进。

## 2.3 本章小结

本章主要是微博系统以及微博相关技术和资源的介绍，首先是微博评论的预处理，其中包括编码格式、去除噪声、分词和词性标注等内容。其次介绍了向量空间的相关概念。再次，介绍了不同来源的情感词典，并对每个情感词典进行了统计，从而合并形成本文研究所需的各类词典。最后是分类效果的评价标准描述。

## 第3章 微博非情感评论信息的识别

虽然微博评论的研究方法趋于成熟,但是针对评论中的垃圾评论以及客观性评论,这种非感情评论的提取等研究并不充分。这些垃圾评论和客观性评论,直接干扰了评论情感的分析效果。评论的情感倾向分析首先应该排除其中的垃圾评论和客观评论,从而提取出有用的主观性文本。本章主要是正负情感分析之前的垃圾评论识别、主客观评论分类。通过特征提取以及机器学习等方法的运用,将垃圾评论和主客观评论剥离和抽取。

### 3.1 微博垃圾评论信息识别

#### 3.1.1 微博垃圾评论信息介绍

##### (1) 微博垃圾评论的特点介绍

目前垃圾评论识别研究已经比较成熟,但是垃圾评论的形式在不断变化,每个时间段呈现的主要形式并不相同。这类评论内容没有明显规律,而且评论中主题的相关信息也比较少,特征出现较为稀疏。微博的垃圾信息大致可以分为两类,第一类属于和主题不相关,对他人进行语言攻击、辱骂并且对于不相关的话题进行评论,这其中可能包含暴力、色情等信息内容,会使用户体验极差,对浏览评论的用户起到不良的影响。第二类主要是通过图片、链接等方式推销产品、发布不健康网页链接。

针对垃圾评论分析,其特点如下所示:评论短小、主观意向较重、口语化用词较多;受当时的热点话题影响;含有超链接、图片评论、特殊字体。

##### (2) 微博垃圾评论的分类介绍

微博的垃圾评论形式多样,每种特征对于评论的影响程度也不尽相同,所以不同特征对评论影响程度也不一样。垃圾评论示例如表3-1所示。前两个为垃圾评论、后两个为正常评论:

垃圾评论分类识别需要匹配不同特征词典所包含内容,例如情感词典:“赞扬”、“好样”、“畜生”、“土匪”、“报应”等。表情符号词典:“[鼓掌]”、“[good]”、“[给力]”等。广告词和明显垃圾词汇词典:“转发”、“京东”、“淘宝”、“全网热卖”等。

根据某一个特征在评论中出现的量级,分析该特征在评论分类中重要程度,进一步计算每种特征的贡献值,通过贡献度的综合打分来计算评论的总分,利

用人为选定的阈值，将测试文本划分类别。

表 3-1 微博垃圾评论示例

微博示例	评论类别
转发微博@中国文明网 @文明贵阳 @修文文明	垃圾
在淘宝好好卖，赶紧来买啊 <a href="http://www.taobao.com">http://www.taobao.com</a>	垃圾
有一种领导叫顺丰家的领导，手动点赞，棒棒哒	正常
张瑞和他的义工伙伴们用自己的方式承担起了爱心奉献社会的责任，他们温暖着一座城，传递着更多的爱	正常

### 3.1.2 基于垃圾线索的微博垃圾评论识别

本文在垃圾评论的识别主要分为 4 个步骤：首先，对垃圾评论进行收集，进行预处理。第二步，是针对目前收集的垃圾评论，提取相应的特征，针对不同的垃圾特征计算其相应的贡献度。第三步，通过计算权值，设定阈值。第四步，通过测试数据集验证本文方法的效果，其流程图如图 3-1 所示。

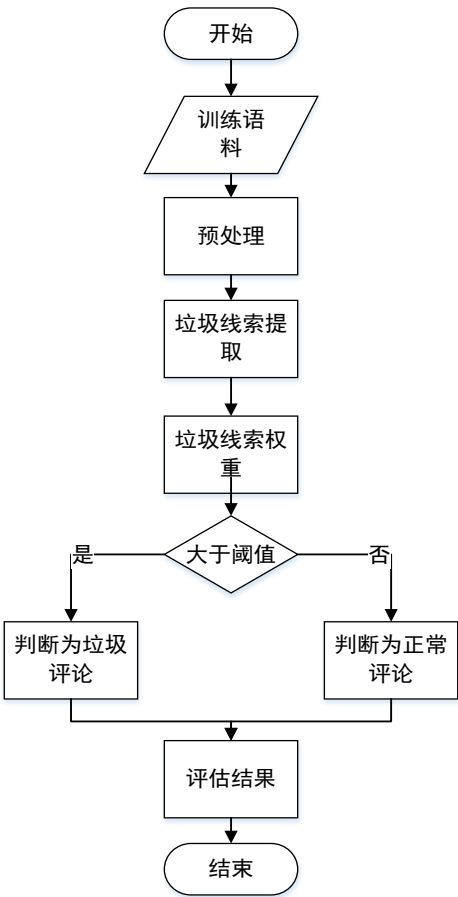


图 3-1 垃圾评论流程图

垃圾评论具有与微博内容相似度较小并且其形式具有多样化、多变化的特点，例如之前的广告链接等明显形式演变为图片等具有一定伪装、不易发觉的非链接形式。目前的垃圾评论中，一部分伪装成图片，还有部分垃圾评论的信息伪装到常规性评论之中，针对这些缺陷和特点，本文提出了一些特征，通过特征组合对评论进行打分，同时划定阈值，将评分高过阈值的评论设定为垃圾评论，通过实验分析效果。

### 3.1.2.1 特征项的选取

以垃圾线索为根据的垃圾评论识别中，特征项的选取尤为重要，分类的效果主要是取决于特征项的选取是否有效，选择的特征对于评论的覆盖范围应该能够达到 90% 以上，才有一个非常好的效果。本文从文本相似度、情感的表达等几个方面挖掘特征项。选择的特征如下：

#### (1) 文本相似度特征

文本的相似度<sup>[36, 37]</sup>是表明是否属于一个主题非常重要的特征，相似度是很多研究的重中之重，垃圾评论与微博主题的相似度会非常的小。本文选取了余弦相似度、编辑距离两种方法，对评论文本和对应微博的相似程度进行度量。

余弦相似度：首先，通过 VSM(Vector Space Model)将训练文本投射到多维空间，将自然语言形式文本转化为机器可以处理的方式表示。其次，余弦相似度是多维空间的相互计算来得到相似度值，相似度值表示实际语义的相似程度。最后，余弦相似度计算的结果越大那么表明评论与微博主题越相似。其计算公式 (3-1) 如下所示：

$$Sim(M, C) = \frac{\sum_{i=1}^n A_{im} \times B_{ic}}{\sqrt{\sum_{i=1}^n A_{im}^2} \times \sqrt{\sum_{i=1}^n B_{ic}^2}} \quad (3-1)$$

公式中的  $A_{im}$  表示属于  $m$  的词在文本数据集出现的概率， $B_{ic}$  表示属于  $c$  的词在文本数据集中出现的概率， $n$  为微博和其评论中出现的词的总数。但是由于评论比较简短，所以只用余弦相似度进行计算并不准确，随后又采用了编辑距离的动态匹配相似度。

编辑距离动态相似度匹配，通过两个不同字符串的计算次数找出其间的差度  $n$ ，那么相似度就可以说是  $\frac{1}{n+1}$ ，动态的编辑距离相似程度的匹配公式 (3-2) 如下所示：

$$L_{a,b}(m, n) = \begin{cases} \max(m, n) & m(m, n) = 0 \\ \min \begin{cases} L_{a,b}(m-1, n) + 1 \\ L_{a,b}(m, n-1) + 1 \\ L_{a,b}(m-1, n-1) + 1 \end{cases} & a_m \neq b_n \end{cases} \quad (3-2)$$

通过公式得到不同字符串的编辑距离，计算相似度值如公式 (3-3) 所示。

其中  $L(a,b)$  是两个字符串的最小编辑距离。

$$Sim(a,b) = 1 - \frac{L(a,b)}{\max(|a|,|b|)} \quad (3-3)$$

通过上述所说的两种计算相似度的方法，分别得到两个得分值，取其中分值最大的作为最终相似度得分。

#### (2) 情感词缺失特征

对于发表的微博，人们都喜欢表达自己观点，社会类微博中人们的情感倾向较高，情感表达突出。垃圾评论的特征较为明显，例如“转发微博”，这种评论没有表达任何意思，所以应该判定为垃圾评论。上述第二章已经介绍过几个词典，包括了情感词典、能够表达倾向的网络常用词以及微博表情词典，统计这些词的总和作为情感特征词典。词典举例，例如：“好人”、“赞成”、“给力”、“[蜡烛]”、“[爱心]”等。通过匹配方式，观察某一评论是否符合此特征。

#### (3) 评论中的超链接

在垃圾评论中，会有一定量的评论中含有超级链接，但是这类评论已经在减少，因为各个网站已经开始屏蔽这类评论。含有超级链接的评论中广告居多，在本文数据中，所有包含超链接的数据 80% 以上都是垃圾数据，例如“还在等什么，欢迎前来购买夏装短袖 <http://detail.tmall.com/>”，还有一种形式是“亲们本旺铺现在推出所有童装商品以最优惠的价格一件代发批发诚招各地代理哦抢购吧 O 网页链接”。绝大数正常评论都不会有超级链接，观察评论中是否有出现超链接。

#### (4) 名词的个数

正常评论之中，对社会事件的描述很少，大多民众都是发表观点，所以不会很正式的采用许多名词，但是垃圾评论中名词很多，例如“我是塔莉妈妈 H2O 水喷雾，专注肌肤代谢，为你的脸排毒减压，爱美的你不妨试试我哦”，使用的名词个数显然过多，采用名词个数不小于 4 作为定界。

#### (5) 句子的长度

非垃圾评论之中，民众大多不会通过很长的评论来描述事件，而广告或者商品都需要细致描述，所以句子过长有很大的几率是垃圾评论。例如：“你还在像以前一样花几倍的价钱去发廊染发并且有时候还染不出自己想要的颜色吗、那你就亏大发了、现在都流行自己 DIY 头发颜色了、在网上看到喜欢什么颜色的头发在我这都可以染出来哦、有兴趣来看一看哈”，通过上述的分词软件，去除停用词，统计词的个数，本文采用不小于 18 或者小于 2 个词作为阈值。

#### (6) 广告词和明显的垃圾词汇

此特征主要是针对打广告的垃圾评论，但是由于数据的限制，广告词收集的程度可能还不够多，评论中的广告词越多，说明为垃圾评论的可能性越大。

本文搜集了一些常用的广告词，例如“淘宝”、“京东”、“图片评论”、“扫码”等词构成广告词词典。看评论是否会出现此类特征。

#### (7) 针对别人评论的回复

在评论中，尤其是针对争议或者是热门事件，评论一般的情感极值都比较尖锐，但是人与人之间互动所产生的评论，对事件本身的相关性较小，而且很容易造成人身辱骂、互相抨击等情况，有几率成为垃圾评论。例如：“@、回复@、//@”等表示内容的标签，有可能为垃圾评论，但是要区别回复对象，若回复对象是相应微博作者，那么就为正常评论，否则记为垃圾评论。同时统计评论被别人回复时的数量，如果评论的被评论数较多，基本为非垃圾评论。

#### 3.1.2.2 文档频率-词频差 diff

本文选取文档频率的改进词频差作为特征权重的表示，它能够有效的表示出不同特征对于类别划分的重要程度。

文档频率（DF，document frequent）是一种常用的特征选择方法。文档频率是通过 DF 得分值来显示某一特征在数据集中出现的概率。通过得分值来表示此特征的贡献程度，其值越高，则说明此特征的有效性越好。其计算公式(3-4)如下所示。

$$d_i = \frac{N_0(i)}{N_{all}} \quad (3-4)$$

但是区分某一特征在两种类别中贡献度的大小，使用文档频率的词频差能更好表示。其公式如下（3-5）所示。

$$d_i = \frac{N_0(i)}{N_{0-all}} - \frac{N_1(i)}{N_{1-all}} \quad (3-5)$$

其中， $N_0(i)$ 表示特征  $i$  在垃圾评论中的个数， $N_{0-all}$ 表示垃圾评论的总个数。 $N_1(i)$ 表示特征  $i$  在正常评论中的个数， $N_{1-all}$ 表示正常评论的总个数。

#### 3.1.2.3 垃圾评论识别计算步骤

利用人为设定的不同特征，构造一个特征匹配库，每种特征都有其各自的权重。通过权重向量计算相似度，根据相似度来描述评论的无用程度。使用预先设定好的经验阈值划分评论，如果大于阈值则为垃圾评论，否则为正常评论。具体步骤如下所述。

---

#### 算法 3-1：垃圾评论划分步骤

---

**输入：**评论数据集 S

**输出：**评论分类：垃圾评论 R，正常评论 N。

**Step1:** 通过人工设定的几种不同特征，形成特征匹配库。将初始权值向量预先设定为空向量，记为  $\text{Score}(S)$ ，通过词频差的方式引入每种特征的权值，

---

将权值归一化后，记为 $W_i$ 。

Step2: 输入评论集  $S$ ，将评论匹配所有的特征库，如果命中了特征模式库中的特征，将该特征的权重 $W_i$ 按顺序加入到  $\text{Score}(S)$ 向量中。

Step3: 评论表示为  $\text{Score}(S)$ 向量后，与特征组合的总特征向量计算相似度，相似度值  $SC$  作为阈值设定的根据，使用确定的阈值  $\text{Th}(S)$ 划分类别。

Step4: 若  $SC \geq \text{Th}(S)$ 则为垃圾评论  $R$ ，否则划分为为正常评论  $N$ 。

### 3.1.3 实验结果与分析

#### (1) 数据来源

社会类的数据源中，垃圾评论与正常评论的评论个数并不相同，所以数据源主要来自于几个主题的合并，主题包括了“老母亲感动中国”、“黑工厂加工”等。选取了垃圾评论 2346 条，正常评论 2300 条。测试数据，垃圾评论为 923 条、正常评论 1000 条。具体条目如表 3-2 所示。

表 3-2 微博垃圾评论示例

数据集	垃圾评论	正常评论	合计
训练数据	2346	2300	4646
测试数据	923	1000	1923

#### (2) 实验结果与分析

通过训练集的匹配，统计每项特征在数据集中出现的频数，随后通过文档词频差的计算方法来计算特征的权值。其结果如下表 3-3 所示。

表 3-3 特征在垃圾评论中统计示例

垃圾线索	垃圾评论	正常评论	权值
文本相似度	1678	938	0.3074
情感词典	739	231	0.2164
超链接	931	57	0.3721
名词的个数	1076	449	0.2632
句子的长度	1476	1124	0.1405
广告词明显垃圾词汇	1211	76	0.4832
回复性评论	1324	158	0.4954

从上表看出垃圾评论的回复性评论、广告词和明显的垃圾词汇例如：“图片评论”等占绝大多数，情感词方面有部分的正常评论中包含客观性的评论，而

垃圾评论中出现较多情感词来混淆视听，情感词的表征能力变得较弱，同时句子的长度有很多正常评论也很长，区分能力最弱。本文选用上述的几种特征作为划分阈值的根据。下表 3-4 为归一化后约取小数点后两位的各项特征权值。

表 3-4 特征在垃圾评论中统计示例

垃圾线索	权值
文本相似度	0.15
情感词典	0.10
超链接	0.16
名词的个数	0.12
句子的长度	0.06
广告词明显垃圾词汇	0.21
回复性评论	0.22

测试集通过阈值划分后结果如表 3-5 所示。P、R、F 为第二章所述的性能评价标准。

表 3-5 分类结果统计

阈值	P	R	F
0.1	0.6894	0.9653	0.8043
0.2	0.7987	0.9588	0.8714
0.3	0.8516	0.9447	0.8916
0.4	0.9024	0.9219	0.9108
0.5	0.9228	0.8429	0.8810

从上表可以看出，分类准确率随着阈值的改变而变化，查准值呈现上升的趋势，但是增加趋势不断减缓。而查全值不断降低，但是降低速度随着阈值的增加不断加快。也就是说当阈值过低时，很多正常评论都被误认为垃圾评论，而当阈值过高时，由于门槛太高，会导致很多的垃圾评论没有进入到垃圾类别之中。通过 F 值的计算结果可以看出其在一个峰值到到最高，然后开始下降，



那么平均分类性能也就是在此阈值时效果最好，从而说明方法总体的性能是先升后降的，分类准确率在达到一个阈值时效果最好，本文采用此阈值为最终经验阈值。

不同阈值所得垃圾评论分类结果如图 3-2 所示。

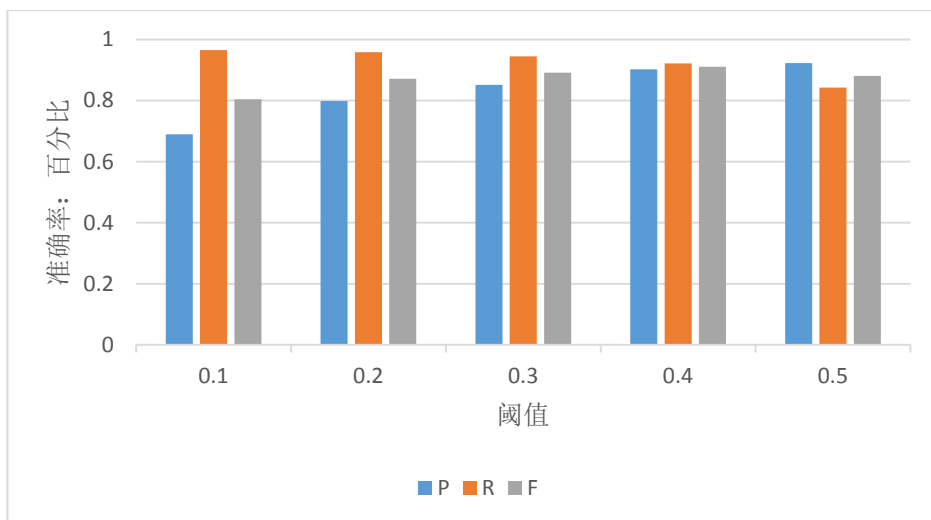


图 3-2 测试数据结果统计

通过上述实验可以看出当阈值为 0.4 时，F 值指标达到最高，所以本文将阈值设定为 0.4 时，该方法对垃圾评论的识别能够达到最好。

## 3.2 微博主客观评论信息识别

在去除垃圾评论后，情感分析的二值分类之前，应该先判断评论中的主客观评论，这样可以使正负情感分析达到一个更好的效果。通过主观性评论的特征组合，结合机器学习的方法，将主客观评论分离。为后续的主观评论分析起到很好的铺垫作用。

### 3.2.1 微博主客观评论信息介绍

#### (1) 主客观评论的特点介绍

**主观性评论的定义。**主观性质的评论说明评论者不是对微博进行客观性质的描述，是对发生的事件表达自己的观点，能够体现自身主观意识，主要包含两类性质评论，分别是评价性质、推测性质。评价言论主要有观点、意见、个人主观上的判断。推测言论主要是依据发生事件的后续影响表达自己对未发生事件看法。主观性评论例如：“普通人也有大能量！希望我们的生活当中能看到更多这样的好人。”

**客观性文本的定义，**主要为评论者是对此事件客观描述，但是却并不表达

自己对事件的看法与意见，不具有任何的喜好。例如：“山西晋城大学毕业生张瑞在父母的帮助下，发起成立晋城义工联合会。”

### （2）主客观评论的分类介绍

上文较为详细的描述了主客观评论的特点，主客观评论的分类主要是将评论分为客观性评论、主观性评论。主客观评论如表3-6所示。

表 3-6 微博垃圾评论示例

微博示例	主客观分类
蒙牛冰淇淋月饼包装很漂亮，味道也超赞	主观
[赞]正是有了千千万万这样平凡而又伟大的人，我们的社会才会更加和谐，我们的国家才会更加强大	主观
山西晋城大学毕业生张瑞在父母的帮助下，发起成立晋城义工联合会。	客观
19日凌晨5点，三江镇官路村内一家四口被困。海南省军区战士张鹏带领9名官兵赶赴现场。	客观

主客观分类首先构造训练数据，然后根据所提取的特征，构造几种词典。例如情感词典：“好人”、“英雄”等。程度词典：“非常”、“绝对”等。指示性动词词典：“感觉”、“认为”等。

## 3.2.2 基于机器学习方法的主客观评论识别

主客观性评论分类，在选取特征方面，不仅要考虑到主观评论特点，还要将其与客观评论相区分。针对主观评论特点，本文提出了一些特征，通过这些特征的组合，将文本表示成向量。用机器学习的方法对主客观评论进行识别，通过实验验证效果。

### 3.2.2.1 主客观评论的识别流程

主客观评论的识别上，本文分为四个基本的步骤：第一步，将全部数据集预处理，其方法第二章描述过。第二步，是针对目前收集的主客观评论，提取相应的特征，表示成向量空间。第三步，通过朴素贝叶斯算法构建训练模型。第四步，用测试数据验证结果。其具体流程如图 3-3 所示。

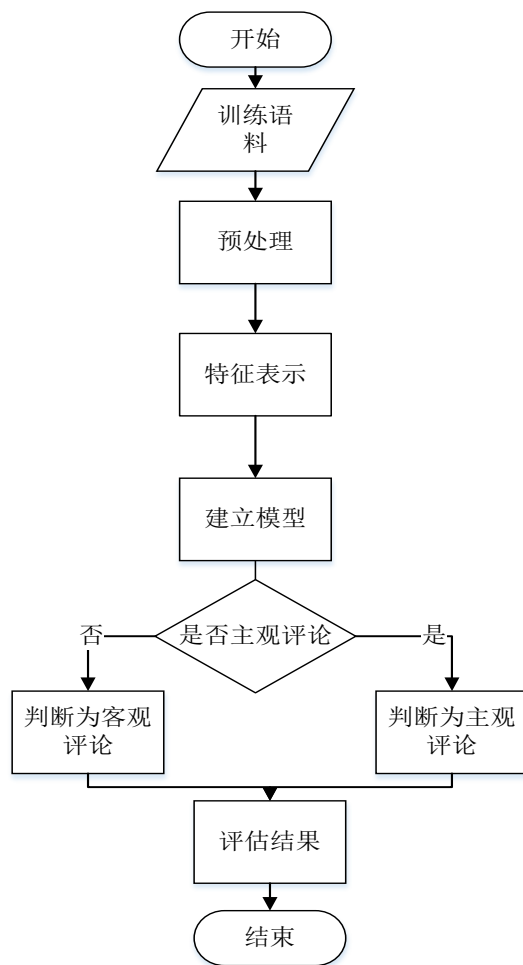


图 3-3 主客观评论流程图

### 3.2.2.2 主客观评论的特征选取

主客观评论的特征较为突出，本文主要针对主观评论分析，所有特征基本覆盖主观性评论。以下为几种特征描述。

#### (1) 情感词特征

类似于垃圾评论的分类，主客观评论的分类中情感词典也占有很大影响因素。人民对于社会发生的事件情感倾向较高并且较为明显，在客观评论之中情感词出现的次数非常少甚至没有，例如“山西晋城大学毕业生张瑞在父母的帮助下，发起成立晋城义工联合会。”这种评论虽然和主题有关，但是没有表达任何意思，所以应该判定为客观评论。第二章已经介绍过几个词典，包括情感词典。例如：“好人”、“美丽”等。

#### (2) 程度词特征

社会事件评论中不乏情绪比较激动并且想通过语言文字表达的人，大多都会选取程度性的副词作为修饰，在修饰的过程中可能会使用例如：“非常”、“过

分”、“很”这种表示程度的副词。例如：“蒙牛冰淇淋月饼包装很漂亮，味道也超赞。”可以看出评论者用程度副词加强情感。程度副词词典的构建，在第二章进行了详细的描述。

### （3）主张词特征

民众关心的事件中可能会使用一些表明主张性的词语。例如“感到”、“觉得”等词。主观性质的评论会可能会有明显的主张性词语。例如：“我觉得他是一个好干部。”然而客观性评论不会有这类词语。第一、二人称等可能会与表达主观感受的词语相连接使用。此特征的选择在一定程度上可以区分两类评论。

### （4）感叹词特征

客观性事件的描述，一般都不会包含感叹词，因为感叹词基本不会对事实性的描述起到任何的帮助。所以在主观性的评论之中，感叹词使用次数比较多。例如：“这就是无名英雄啊”。感叹词词典的构建，例如：“啊”、“唉”、“呸”等。

### （5）带有感情的特殊符号特征

主观性评论之中，用词较为随意，并且人们对于事件的发表观点，可能会通过一些标点符号来代替本身词语要表达的含义，然而客观性的评论之中一般不会含有这类符号，因为其并不能帮助描述客观事实。例如：“[赞][赞][赞]我们的楷模！！[给力][给力][给力]仁者无敌！！！”这其中“！！！”对于语言的表达非常明显，强烈的表达了对于好人的尊敬之情。评论的标点符号既包括了规范的标点也有表示更加强烈语气的形式，例如：“！”、“？？？”。

### （6）关联词特征

微博的评论内容中有很多复杂句式，这其中会有关联词将不同的简单句式合并，形成复杂句式。然而这些复杂句式一般都是表达主观倾向，其中有明显表示转折关系、假设关系等关联词，对主观性评论的识别有一定作用。例如：“虽然不能一棒子打死，但是国家养的这些人基本就是混吃等死，祸害人民。”很明显表达出了评论者的情感倾向。

### （7）微博表情符号特征

微博的评论和正常的文本并不一样，其中的评论也并不规则，很多评论者采用了表情符号来代替文本性质的描述，所以微博表情符号是很重要的影响因素。将微博抓取于下来的表情进行总结，形成微博表情词典，例如：“[给力]”、“[蜡烛]”、“[good]”。

## 3.2.2.3 提取特征算法和朴素贝叶斯分类器

### （1）提取特征算法

---

**算法 3-2:** 提取特征算法

---

**输入:** 输入文本 M

---

**输出：**表示特征的集合  $Q$

Step1: 将待分类文本分词等操作，保留其词性的识别，将整体文本划分成  $n$  个特征词的组合。

$$M = \{m_1/C_1, m_2/C_2, m_3/C_3, \dots, m_n/C_n\}$$

Step2: 通过每种不同特征的词典，建立特征词表  $L$ ， $L$  可以表示成。

$$L = \{L_1, L_2, L_3, \dots, L_n\}$$

Step3: 将特征词典表  $L$  应用到整个数据集  $M$  上，通过匹配方式获取到特征的集合  $Q$ 。

$$Q = \{p_1, p_2, p_3, \dots, p_n \mid p_i \in M \cap p_1 \in L\}$$

## (2) 朴素贝叶斯分类器

朴素贝叶斯(NB, Naive Bayes)的基本原理，在贝叶斯上改进和发展延续后得到朴素贝叶斯算法，而贝叶斯如公式 (3-6) 所示。

$$P(C_i|X_i) = \frac{P(c_i)P(x_i|c_i)}{P(x_i)} \quad (3-6)$$

这里  $C_i$  表示类别， $X_i$  表示文档内容，而贝叶斯定理中的所有属性的联合概率不能适用于有限制的训练集进行建模。假设每个特征  $d_i$  对于分类结果都是独立的产生影响，不受其他因素的影响，具有独立性。经过重写后得到公式(3-7)。

$$P(C_i|X_i) = \frac{P(c_i)P(x_i|c_i)}{P(x_i)} = \frac{P(c_i)}{P(x_i)} \prod_{i=1}^n P(x(d_i)|c_i) \quad (3-7)$$

其中  $P(c_i)$  代表  $C_i$  在总评论上的出现概率， $P(x_i)$  代表第  $i$  个评论用恒定不变数值进行计算，通过对上边分子的计算，找到  $P(c_i)$  的最大值。

为了每个特征之间互相不干扰，所以应该在估计时采用“拉普拉斯修正”进而平滑，上式中  $P(x_i|c_i)$  改变成公式 (3-8)。

$$P(x_i|c_i) = \frac{1+(x(d_i)|c_i|)}{2+|D_{c_i}|} \quad (3-8)$$

其中  $|(x_i|c_i)|$  和  $|D_{c_i}|$  分别表示特征在类别中的个数以及类别的个数。

只要能求出计算后的最大值，就可以对主客观评论进行分类。对应的公式 (3-9) 可以写为。

$$h_{nb}(x_i) = \underset{c_i \in C}{argmax} \{P(c_i) \prod_{i=1}^n P(x(d_i)|c_i)\} \quad (3-9)$$

此为最后的朴素贝叶斯的表达式，可以看出因为条件概率，训练不仅需要规模比较庞大，而且对于训练集的要求也较为严格，虽然没有考虑到每个特征之间的影响，但是效果在分类器中还是不错的。

### 3.2.3 实验结果分析

#### (1) 实验设置

实验数据的来源：目前还没有公开标注好的数据源，所以采用了自行抓取的评论数据。

训练数据，从新浪微博爬取了 18 个有关社会舆论的微博评论，在筛选直观垃圾评论之后，从 2 万余条剩余数据中，挑选了 4200 条作为主观数据，4172 条作为客观数据，共 8372 条作为训练集。

测试集：从中选取了三个主题，删除垃圾评论后共计 2147 条主观评论、447 条客观评论，但是由于客观评论和主观评论过于不平衡，在同一话题下查找了 1012 条表示客观性的微博，作为最后的测试数据。具体统计如表 3-7 所示。

表 3-7 微博评论示例

数据集	主观句	客观句	合计
训练数据	4200	4172	8372
测试数据	2147	1459	3606

#### (2) 实验结果与分析

利用特征组合将文本表示为向量，放入朴素贝叶斯分类器，形成最终模型。通过测试数据测试分类模型后，得到的结果如下表 3-8 所示。

表 3-8 测试得到的主客观评论示例

	实际为主观评论	实际为客观评论
分类为主观评论	1738	523
分类为客观评论	409	936

实验的检验标准是第二章所说的三项检测标准值。得到的结果如下表 3-9 所示。

表 3-9 主观评论各项评价价值

实验数据	查准率	查全率	F 值
主观评论	80.95%	76.86%	0.7885

通过观察结果来看，其中有一部分的客观评论不容易分辨，可能有很多客观性的语句也会使用情感词，例如：“2015 年 5 月的新安江正值汛期，水急浪高，一名儿童在河边玩乐时不幸落入水中，路过此处的王吉权不顾个人安危，迅速跳入水中救人。”中“不幸”、“玩乐”等是表达情感的词项，所以很容易误判。同时，程度副词虽然在主观特征用的较多，但是例如“由于水流太急，他屏住呼吸，潜入深水区艰难搜索，仅用不到 60 秒便将落水儿童救出水面，随后悄然离开”中“太”、“仅”等可能在某些评论中起到了负向作用。由于是针对

主观评论进行的抽取，所以对于客观评论的判断上会很大程度上降低准确率。本文所用的方法可能不是最好的主客观分类方法，但考虑社会类微博数据比较繁杂，规律并不明显，F 值为 0.7885，可以说比较理想，对后续的褒贬情感分析起到了很好的筛选作用。

### 3.3 本章小结

微博评论的类型有多种多样，这其中包含了垃圾评论、客观评论、正负情感倾向评论。本章主要是对垃圾评论以及主客观评论分析进行研究，提出各自类别的特征组合。垃圾评论是以垃圾线索和文档频率差等方法来识别和分类，而主客观评论主要是特征组合和机器学习相结合的方法进行分析，通过测试验证本文方法有效性。

## 第4章 微博评论信息倾向性分析

### 4.1 微博评论情感识别介绍

随着目前的微博影响力不断加大，人们认识到微博的作用越来越广泛，随之产生的商业作用以及对正负舆论的导向都值得不断加深研究，挖掘其中的潜在价值。

目前情感分析主要是两种方法，其一是一次性分为三种倾向，其二是通过两次每次分为两种类别。本文评论分类框图如图4-1所示。采用了两步两分类的情感分析方法。首先使用了第3章提出的主客观分类技术筛选出主观评论，接着对这些主观评论进行情感分类。

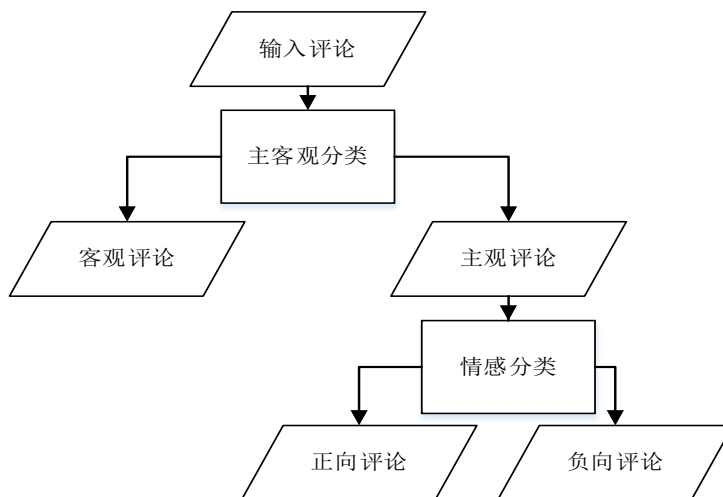


图 4-1 评论分类框图

情感倾向分析的学习方法，目前大致有两种方法，第一种是基于词典规则匹配的方法，通过各个词典的运用以及制定相应规则来判别情感倾向。第二种是基于机器学习的有监督、半监督学习等方法，通过人工标注的训练集来学习未标注的测试集，使分类器达到分类效果。

### 4.2 特征选择

评论特征选取，有很多特征的组合，通过了解和翻阅前人对特征的使用，例如：单词特征、词性特征、词频特征等。本文选取了其中最有效的特征组合



方式，包括单词特征、词频特征等。

### （1）Unigram词特征

Unigram 词特征主要是一元词特征。一元词特征是特征选取中最为重要的特征表示，许多研究是在其基础上的改进和变形。4.3 节将详细描述本研究针对一元特征提取方法上的改进。

### （2）二元词特征

二元特征主要是词与词之间的结合，例如：“托举生命传递正气”，可以分成：“托举生命 生命传递 传递正气”本文主要是选取了前 800 个左右准确率最高并且出现次数最多的二元词组合作为特征组合的一部分。

### （3）数值特征

数值特征包含了情感词的词频差、否定词数目、动词数目、程度词数目、感叹词数目、疑问词数目等，本文选取了情感词词频差、程度词词频、疑问词词频、感叹词词频分别作为一维特征向量。其中情感词数目，例如：“中国好医生”，其中词频数为2，其余特征以此类推，将每维个数归一化后，填入到向量空间之中。

### （4）微博表情符号特征

针对目前微博评论特点，表情符号的使用已经很频繁，也代表了情感倾向性，例如：“伤心”、“开怀大笑”、“棒”等。所以可以将表情符号加入到特征向量之中。但是由于微博表情占整体评论集的比例较小，所以其只能作为辅助特征进一步构成向量空间。

本文共找到 192 个表情符号，其中能够表达情感倾向的表情进行了提取，抽出了其中的 91 个带有情感倾向的表情符号。其表示形式为“[蜡烛]”、“[太开心]”、“[good]”等。本文将微博独有的表情分为两种，其中表正向情感，例如：“[good]”等，表负向情感表情：“[鄙视]”等，分别为 38、53 个。表示成为向量空间中的两维向量，其值表示为每种类别情感符号个数归一化后的值。

### （5）句式特征

简单型评论类别。该类评论是比较短并且倾向性比较明显的评论，所以对这种类型的评论不用做过多的处理，提取特征并且运用机器学习模型进行学习分类，直接判断此类的情感值构建一维向量。

复杂型评论类别。该类评论句子种类繁多，并且有很多干扰句子，尤其是对于社会类事件情感的分析其中句式和词频都比较复杂。对于否定性质的和带有总结、转折关系的句子往往是评论中的主要情感倾向的代表，否则就为各个句子情感倾向总和，而评论情感值加和的计算采用情感词典判断方法，通过否定词或者程度副词辅助，获得情感值形成一维向量。

### 4.3 特征提取及权重的计算

情感分类中，特征选择的重要性不言而喻，向量空间的构成是由多个特征之间拼接构成，评论的复杂性决定评论会有大量的不同特征，而这些特征之间会相互影响，所以特征的提取就尤为重要，而小比例的特征对分类结果有很明显的影响，大多数的特征只是无用特征，甚至如果向量空间过大，会导致分类结果更差。

#### 4.3.1 传统的特征提取方法

特征的选择是首要任务，在能够最大限度的保留有用信息的前提下，尽可能排除干扰特征，将向量空间的大小保持在一个合理的范围，并且每一维度上都是最有用的信息，能够提高分类准确率和运行效率。特征提取的方法主要是通过公式计算每个特征对于样本分类的贡献程度，最后将选取的特征拼接形成向量空间。下面是几种常用的特征选择算法，例如信息增益(IG)、互信息(MI)、卡方统计(CHI)特征提取方法。

##### (1) 信息增益

信息增益 (IG, Information Gain) <sup>[38]</sup> 是描述每个特征所包含的在数据集上的信息内容也就是信息熵(Entropy)，它表示了一个特征对于文本集合的影响大小，通过信息量的增加或者减少，计算数据集的信息内容的变化，如果数据集的信息内容有所变化，那么变化的信息就是特征所代表信息量，反之，则说明特征对分类没有效果，属于无用特征。通过计算的信息熵值来判断某个特征对数据集划分的作用程度，信息熵值越高说明分类作用越大，反之越小。信息增益公式如下 (4-1) 所示。

$$IG(t) = -\sum_{i=1}^n P(c_i) \log P(c_i) + P(t) \sum_{i=1}^n P(c_i|t) \log P(c_i|t) + P(\bar{t}) \sum_{i=1}^n P(c_i|\bar{t}) \log P(c_i|\bar{t}) \quad (4-1)$$

其中n代表所有的类别相加。 $P(t)$ 代表特征t在数据集中出现概率， $P(c_i)$ 代表数据集中的 $c_i$ 类出现的概率。 $P(c_i|t)$ 代表数据集包含特征t同时是 $c_i$ 类的概率。 $P(c_i|\bar{t})$ 代表数据集中不包含特征t但为 $c_i$ 类的概率。 $P(\bar{t})$ 代表特征t不在数据集中的概率。

信息增益的方法，有其自身的缺陷，因为其只计算了某一个特征在整个文本集中的贡献度，没有计算不同类别之间的贡献度。

##### (2) 互信息

互信息 (MI, Mutual Information) <sup>[39]</sup> 通过计算每个特征与每个类别之间的关联性，来表示特征与类别之间的影响程度。如果计算得到互信息值越大，表

示相应的特征在不同类别上的划分起到越大作用，否则作用较小。互信息值计算如公式（4-2）所示。

$$MI(t, c) = \log \frac{P(t|c)}{P(t)} \quad (4-2)$$

具体的估值计算如公式（4-3）所示。

$$MI(t, c) = \log \frac{A \times N}{(A+C) \times (A+B)} \quad (4-3)$$

这其中，A代表包含特征t同时是c类评论数量，B代表包含特征t但不是c类评论数量，C代表不涉及特征t但是c类评论数量，D代表不包含特征t同时不是c类评论数量。特征t按公式（4-4）所示计算互信息值。

$$MI(t) = \sum_{i=1}^n \log \frac{P(t|c)}{P(t)} \quad (4-4)$$

$P(t)$ 代表特征t在文本集中出现的概率， $P(t|c)$ 代表包含t同时是c类的可能性大小。

互信息的缺陷在于其没有将特征出现的可能性纳入考虑范围之中，所以其很容易造成不选择频率较高并且有很好区分能力的特征项，而选择出现次数相对少的干扰特征。

### （3）卡方统计

卡方统计（CHI, Chisquare Statistics）<sup>[40]</sup>是一种主要的选取特性的算法，也是一种关于特性和分类的互相扰动方法。该方法预先为某一特征和类别之间满足一阶自由度的 $\chi^2$ 分布， $\chi^2$ 值的大小意味着某一特征对类别划分起作用的大小， $\chi^2$ 值越大代表着该特征相关性越强，反之越小。而如果卡方值为0，则证明两者之间没有关联性。卡方统计如公式（4-5）所示。

$$\chi^2(t, c) = \frac{N \times (AD - BC)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (4-5)$$

卡方统计的公式可以化简为公式（4-6）所示。

$$\chi^2(t, c) = \frac{(AD - BC)^2}{(A+B) \times (C+D)} \quad (4-6)$$

卡方统计的劣势为卡方检验没有针对性的选择词频高的特征，所以会有出现次数低的特征影响程度反而更大的可能。

## 4.3.2 特征提取的改进

本节主要针对Unigram词特征提取进行改进。特征提取主要是通过人为选取阈值或者人工选取的前n个特征权重较大特征，形成一个完备的特征集。本文先对上述的三个特征提取方法进行了比较，实验选取了同样的训练集与测试集，选取相同的维度以及同一种分类器。其结果如表4-1所示。

表 4-1 特征提取方式实验结果

类型	选用模型	ACC 准确率
IG	SVM	75.89%
MI	SVM	69.91%
CHI	SVM	73.46%

从上表的结果可以看出，信息增益的特征提取方法对本文所使用数据效果最为出色，所以本文采用了信息增益作为基础特征提取方法。

但是由于目前评论的复杂性，其中会出现大量的无效词语，选出识别效果最好的词语才能进一步提高分类的准确性。对于训练集来说，冗余和无用的特征词占了一定比例，那么能够表达情感的特征词语就会被削弱。由于信息增益没有针对情感词上的辨别，所以虽然挑出了一部分特征，但是效果并没有达到最好。在情感分析上，情感词对评论极性的判断是最有效的，所以特征词提取中应该增加情感词的挑选比例。

本文在信息增益的挑选方式上进行了一定的改动，使其可能更加容易的挑选出情感词，本文将其命名为IG-M（Information Gain-Motion）方法，其作用为将情感词挑选作为一定程度上的优先选择。如公式（4-7）所示。

$$IG-M(t) = \begin{cases} IG(t) + \alpha \times \overline{IG(\bar{q})} & \text{if } t \in q \\ IG(t) & \text{if } t \notin q \end{cases} \quad (4-7)$$

其中 $q$ 情感词集合，用匹配方式判定其是否为情感词语。 $IG(t)$ 代表 $t$ 的IG值， $\overline{IG(\bar{q})}$ 代表没有表达情感的特征IG值的平均值，然后通过 $\alpha$ 来调控挑选的情感词，将其选定为0.2。

（4-7）公式表示了非情感词不增加其权重，而增强情感词的贡献度，这样对出现次数稍多有一定情感倾向的词语能更好挑选出来。通过IG-M(t)方法计算每个特征词的信息值后，将其按信息值从大到小排序后，选取分数最高的N个词作为最后Unigram的特征词集。

N大小的确定，对实验至关重要，词项太少会导致分类精度明显降低，而太多又会导致过拟合。实验中，观察在不同N的情况下实验结果的准确率，可以看出高频词在1800词左右准确率提升到一个最高点，然后基本维持稳定，可以确定在训练集不变的情况下，1800词左右能使分类器达到最佳效果。

示例如图4-2所示。

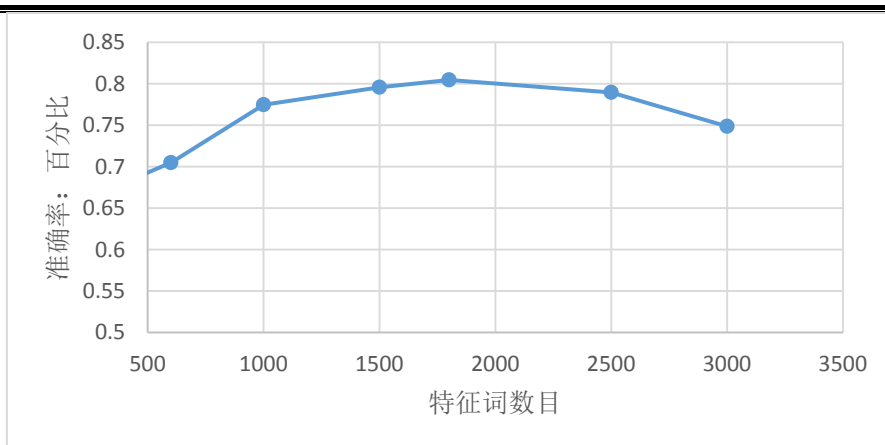


图 4-2 特征数目下的准确度

### 4.3.3 权重计算

在提取的一元特征表示的向量空间中，每个特征对于文本集分类的影响程度并不相同，所以采用权重计算的计算方法来衡量特征的影响因子，权重计算有布尔表示等方法。本文使用TF-IDF计算方法，并且针对其进行了改进。

#### (1) TF-IDF

目前TF-IDF(Term Frequency-Inverse Document Frequency)的技术方法是权重计算中最广为使用的，其具体公式(4-8)如下所示。

$$q_i = TF_i \times IDF \quad (4-8)$$

其中 $TF_i$ 代表特征 $i$ 在数据集中出现的可能性比例。 $IDF$ 为逆向文档频率，其说明了在文本集中包括了特征 $i$ 的文本内容的比例。此方法将文档频率和逆文档频率相结合。其中 $TF$ 、 $IDF$ 如公式(4-9)、(4-10)所示。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (4-9)$$

$$idf_i = \frac{|D|}{|j:t_i \in d_j|} \quad (4-10)$$

$|D|$ 代表所有数据的总数， $n_{i,j}$ 表示在数据集中 $i$ 总共出现次数， $\sum_k n_{k,j}$ 代表数据集中的不同特征相加。

#### (2) TF-IDF改进

本文采用了TF-IDF的方法作为特征权值的计算方式，此方法可以将每个类别中能够明显区别于其他类别的特征词提取出来，但是传统的TF-IDF方法有其自身的缺陷。首先，它在是否为情感词上没有区分能力，但是情感词和普通词汇明显有区分类别的能力差异<sup>[41, 42]</sup>，并且位置因素没有考虑在内，比如在程度副词后加入情感词，那么相应的情感词得分应该有所提高。其次，TF-IDF方法可能错误的认为一些稀疏并且无用词语的贡献度过高，不同类别之间的特征

表示能力也会因此减弱。

针对这些不足，同时查阅文献，通过提高情感词及其组合权重，同时改进特征权重的计算方式，提出了TF-IDF-M方法，其中包括公式（4-11）、（4-12）、（4-13）。

$$gn_{i,j} = \begin{cases} n_{i,j} \times 1.1, & i \in Q \cap S_{i-1} \in C \\ n_{i,j}, & \text{其他} \end{cases} \quad (4-11)$$

$$tf_{i,j} = P(n_k|c_i) = \frac{gn_{i,j}}{\sum_k gn_{k,j}} \quad (4-12)$$

$$IDF = \log(1 + \frac{P(n_k)}{P(n_k)^{\sim}}) \quad (4-13)$$

其中Q代表情感词词典、C代表程度副词词典， $S_{i-1}$ 表示特征i前一个词的内容， $gn_{i,j}$ 函数表示若一个句子中程度词在情感词之前，加大权值1.1倍，否则维持原先权值， $P(n_k|c_i)$ 在 $c_i$ 类别中含有特征i的概率， $n_{i,j}$ 表示在 $c_i$ 类别中包含特征i的总数。 $P(n_k)^{\sim}$ 表示不在 $c_i$ 类中特征i出现的概率。这样既可以增强情感词与副词之间搭配权重，还能减弱稀疏特征的作用。选用此权值计算方式，构成向量空间。

本文提出的TF-IDF-M主要的改进点是将出现次数可能相同，但是其分类效果完全不同的特征区分开来，并且对情感词和修饰词的修饰项组合，用以上方法加重情感词的表示权重。

#### 4.4 基于投票与集成学习的融合分类器

目前，在文本分类的方法当中，情感倾向性分析主要是将评论进行多分类，目前大多数是将情感的倾向分为中性评论、正向评论、负向评论。第一步，将数据集用预处理技术处理，并且形成向量空间。第二步，形成训练模型。第三步，将测试数据放入模型，预测分类。本文采用了SVM、KNN、朴素贝叶斯机器学习方法，朴素贝叶斯已经在上一章详细介绍。本文采用的是基于投票与集成学习方法的融合分类器，通过每种不同分类器优点相叠加，形成一个更有效的分类模型。

融合分类器原理图如图4-3所示。

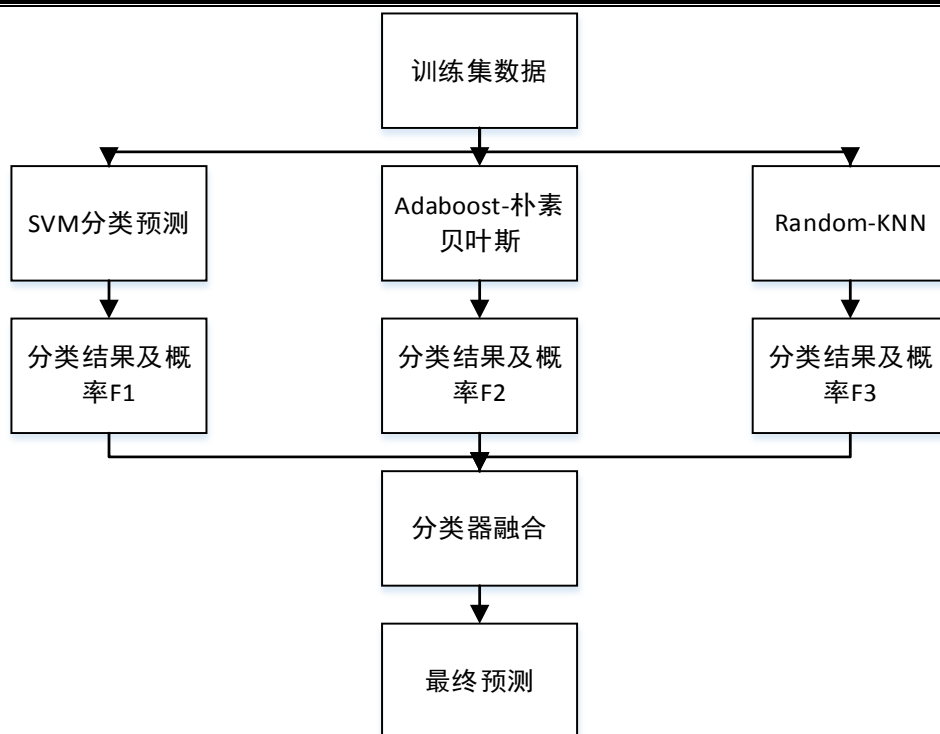


图 4-3 融合分类器原理图

#### 4.4.1 传统机器学习方法

##### (1) SVM 支持向量机

SVM(Support Vector Machine)支持向量机大部分是将评论二分类的模型，基于多维空间中通过线性分类器切割平面进行分类，并且通过核函数的不断改变，使其成为非线性分类器，有很好的分类效果。其基本思想是：将相应向量投射到多维空间之中，由迭代和检验找到一个最优的可以将空间切割的平面，记为超平面，此平面将向量空间划分成为两个部分，进而找出一个超平面能最大限度的使两种类别分离开来，并且对训练集局部扰动的容忍性最好。意味着此超平面能够达到一个最好的划分效果。

如图 4-4 所示，此为通过超平面进行划分的支持向量机示意图，其中有很多平面可以划分空间，其超平面的为公式 (4-14) 所示。

$$\mathbf{w}^T \cdot \mathbf{x} + b = 0 \quad (4-14)$$

其中  $\mathbf{w}=(w_1, w_2, w_3, \dots, w_d)$  是代表切割空间平面的法向量， $b$  是所有超平面距离原点的偏移量。整个超平面是由法向量和位移距离合一表示。向量空间每个点到超平面的位移量都可以表示成公式 (4-15)。

$$\gamma = \frac{|\mathbf{w}^T \cdot \mathbf{x} + b|}{\|\mathbf{w}\|} \quad (4-15)$$

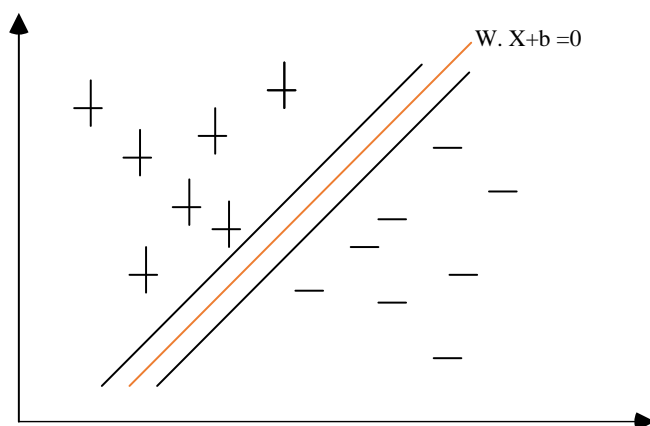


图 4-4 多个超平面划分支持向量机

如果超平面可以将训练样本分类正确,那么就有 $y_i = +1$ ,有 $\mathbf{w}^T \cdot \mathbf{x}_i + b > 0$ ;如果 $y_i = -1$ ,有 $\mathbf{w}^T \cdot \mathbf{x}_i + b < 0$ ,那么就有公式(4-16)所示。

$$\begin{cases} \mathbf{w}^T \cdot \mathbf{x}_i + b \geq +1, & y_i = +1 \\ \mathbf{w}^T \cdot \mathbf{x}_i + b \leq -1, & y_i = -1 \end{cases} \quad (4-16)$$

本文采用libsvm<sup>[43]</sup>的软件包,文本分类主要是采用其中的线性核函数,将SVM进行参数调优,形成最终分类器。

## (2) KNN情感分类

K近邻(KNN, K-NearestNeighbor)<sup>[44]</sup>是一种常用的监督学习方法,对给定的测试样本通过某种距离度量找出训练集中距离相对最小的k个训练数据,通过形成的k个簇对待测文本进行预测分类,在分类任务下,一般都选取“投票法”,意思是选择k个样本中划分最多的类别作为最终分类预测结果。

而对于k的选择也至关重要,当k取不同值的时候,会有明显的结果区别,并且采取不同的距离计算公式,分类效果也有一定差异。这些方面都会导致分类效果有一定区别,示例如图4-5所示。

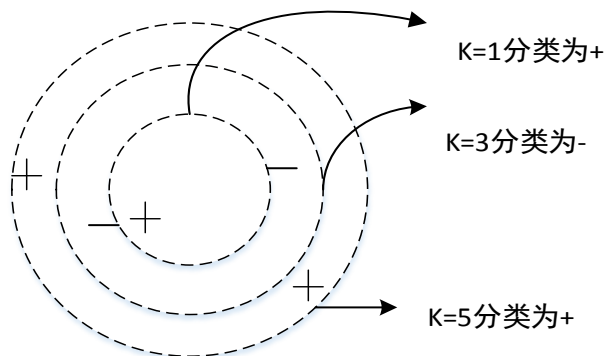


图 4-5 k值不同下的分类示意图

本文使用的是余弦相似度来计算相似度值,其公式为(4-17)所示:



$$\cos(q_1, q_2) = \sum_{i=1}^n \frac{w_{1i}}{\sqrt{\sum_{k=1}^n w_{1k}^2}} \times \frac{w_{2i}}{\sqrt{\sum_{k=1}^n w_{2k}^2}} \quad (4-17)$$

通过公式计算以后,将不同类别训练数据和测试数据之间公式所得值加和,可以得到所有类别的得分值,得分大的类别即为最终类别。

#### 4.4.2 集成学习方法

集成学习方法也是将文本表示成向量的学习方式,通过每个基分类器得出的结果,用特定方式将n个基分类器得到的分类结果集成起来,从而求出最后的分类结果。集成学习主要分为两大类别,有单一的基础分类算法自身提升,也包含多类的基础分类算法的融合。

本文的多分类器集成如图4-6所示。

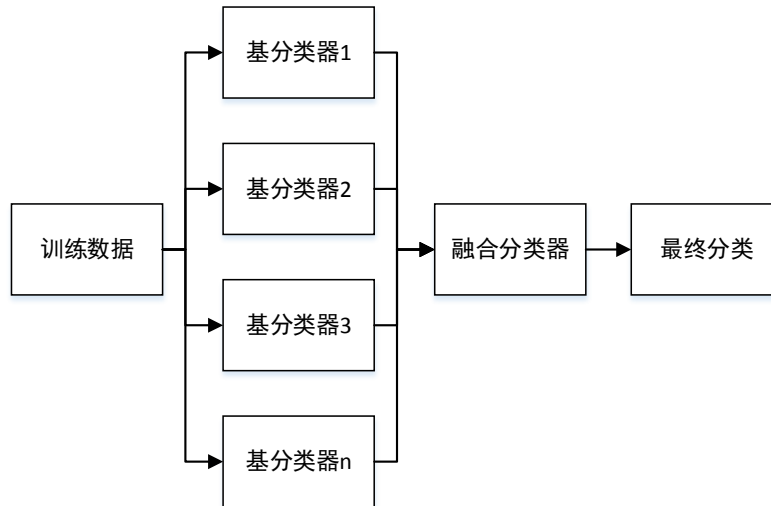


图 4-6 多分类器集成学习示意图

集成学习有几个前提条件。第一点,不同分类器对于同样数据集分类有所区别,否则多分类器的集成就毫无意义。第二点,基分类器的分辨准确率应该大于等于 50%,已经在数学推导上得到了证明,在这里就不详细描述<sup>[45]</sup>。集成学习在统计学的角度、计算方式的角度、描述问题的角度上都对单一分类器有很好的提升。

##### 4.4.2.1 集成分类器结构

目前的集成分类器的结构,主要是分类器之间的串联、分类器之间并行、混合式结构<sup>[46-48]</sup>。

###### (1) 分类器间的串联集成

串联结构优点是,后续分类器能够很好的借鉴之前分类器,对前面分类效果不好的样本更有针对性的增强学习。缺点是,训练时间加长,若出现一点差

错，就会影响后续模型运转。其中Boosting算法是代表，而这其中AdaBoost算法是最典型的代表。

#### (2) 分类器并行集成

优点主要是能够很有效的提高学习效率，实现分类器之间互补。其缺点是对融合器设计过于复杂，Bagging方法是其代表算法。

#### (3) 混合式集成

混合式集成方法在很多领域也有很好的表现，本文主要是借助于这种混合式集成方法思想，从而提高整体分类器的准确率。示意图如图 4-7 所示。

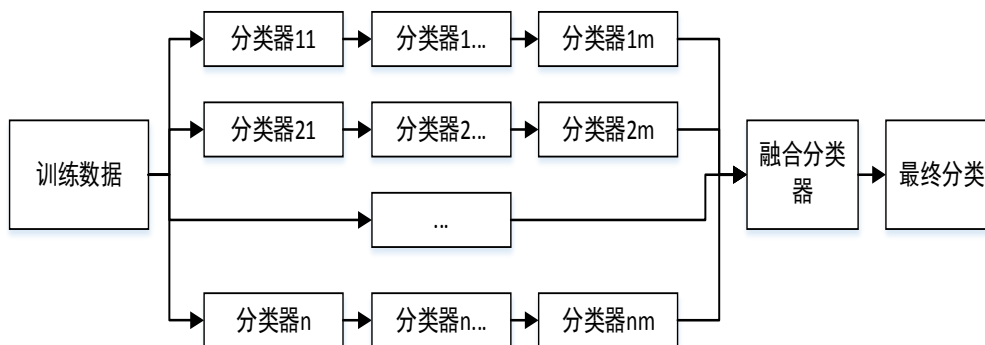


图 4-7 多分类器混合式集成示意图

#### 4.4.2.2 构造基分类器

基本的分类器构造，主要是以下的几种方式进行综合集成，每种算法之间应该有一定的区分性。本文将三种方法全部采用，从而找出一种较好的提升精度的方式。

##### (1) 不同分类器组合

虽然训练集相同，但是对向量空间处理方式，由于每种分类器的原理不相同，所以必然会出现分类效果相异的情况。分类器之间的集合方法有很多，例如：投票法、乘积法、最大值法、最小值法、模糊乘积法等，而本文采用的是投票法，通过不同分类器的得到的标签和分类器权值，采用权值 $\beta_i$ 相加，取M个类别中值大的一类， $la_{i,j}$ 为情感标签。可以表示成公式（4-18）。

$$Q(x) = \max_j \sum_{i=1}^M (\beta_i la_{i,j}) \quad (4-18)$$

##### (2) 训练子集划分-AdaBoost算法

训练集的划分是随机抽取一部分作为子集，不同训练子集虽然使用同样的分类算法但是得到的分类器效果显然不同。因此每个分类器之间也会有一定的差异，AdaBoost算法是典型的代表。

AdaBoost 算法是 Boosting 算法的一种变形<sup>[49, 50]</sup>，而 Boosting 算法是几种增强学习的方法之一。AdaBoost 算法在这之上进行了一定的调整，开始时，对于训练集中每个样本都设定一个权值，在下次迭代的过程中，通过迭代后更新

样本权值，将训练集中每个训练样本重新设定其相应的贡献得分，错误的样本增加其贡献得分，对分类正确的数据降低贡献得分，那么错误的样本就有很大的几率被抽出当做训练样本，下一个分类器就可以将错误样本作为训练集，以此方式生成多个分类器，最后的融合分类器是由此前的  $n$  个分类器组合而成。AdaBoost 算法的优点很明确，比较简单，不用考虑过拟合的问题。其示例如图 4-8 所示。

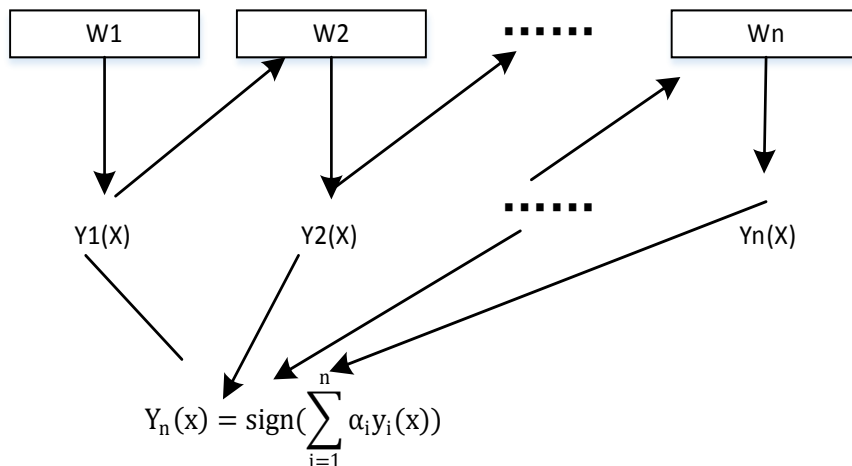


图 4-8 AdaBoost原理

### (3) 特征子集划分-Random Subspace算法

至于特征项划分，首先建立最基础的向量空间，从中再选定部分的特征来重构训练数据集向量，通过此方式构建不同训练空间，形成有差异性的分类模型。此方法可以一定程度上缩小向量空间，但是此方法大多适用于向量空间较大的情况，Random Subspace是其中代表。

随机子空间（Random Subspace）算法与Boosting的分类原理不同<sup>[51, 52]</sup>，上一种方法是对于训练样本上的数据进行划分，而随机子空间主要是通过随机选择不同的特征子集。其具体的做法是针对特征划分，形成不同的子集，对每个特征子集构建自己的向量空间，根据生成的 $n$ 个子特征空间，通过单一分类模型的组合，形成最后的综合分类器。

其原理图4-9如下所示。

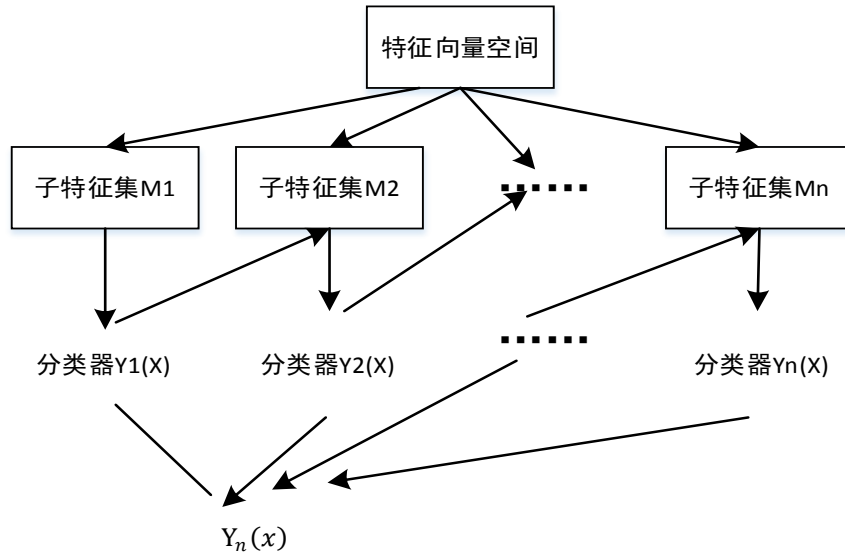


图 4-9 随机子空间原理图

#### 4.4.3 基于投票与集成学习的融合分类器

##### (1) AdaBoost-朴素贝叶斯算法

经过SVM、NB、KNN等算法的测试，观察出SVM的表现效果最好，在SVM基础之上进行改进已经比较困难，所以本文没有针对SVM进行提升学习，而是通过调节部分参数来使其达到一个最好的效果。

而朴素贝叶斯分类效果并没有SVM效果好，朴素贝叶斯分类器的理论是基于所有特征之间没有关联性，但这是不切实际的，而且分类决策也有一定的误差。既然朴素贝叶斯分类器没能够达到SVM的效果，那么就可以想办法在其算法基础上进行改进，使其能够提升一定分类精度。AdaBoost可以一定程度上弥补误差率的计算，但是这种做法的提升会因不同的基分类器得到不同的效果。

AdaBoost-朴素贝叶斯算法步骤如下所示。

---

##### 算法 4-1: AdaBoost-朴素贝叶斯算法

---

**输入:** 输入文本  $Y=(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i)$ , 基分类器: 朴素贝叶斯

**输出:**  $Y_n(x) = \text{sign}(\sum_{i=1}^M \alpha_i y_i(x))$

Step1: 将每个样本的权值设定  $w_i = \frac{1}{N}$ ,  $i=1,2,3,\dots,N$ , 其中为训练集样本数总和。

Step2: for  $t=1,\dots,M$ :

对于朴素贝叶斯分类器  $y_t$ , 计算其误差,

---

$$\varepsilon_t = \sum_{i=1}^N w_i^{(t)} I(h_t(x_i) \neq y_i)$$

通过误差值得出每个朴素贝叶斯的得分值

$$\alpha_t = \ln \left\{ \frac{1 - \varepsilon_t}{\varepsilon_t} \right\}$$

对每个样本权值更新

$$w_{t+1,i} = \frac{w_{t,i}}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} = \frac{w_{t,i} \exp\{-\alpha_t y_i h_t(x_i)\}}{Z_t}$$

将系数归一化得到  $Z_t$ ,  $\sum_{i=1}^N w_i = 1$

Step3: 最后得到  $M$  个朴素贝叶斯分类器:

$$Y_n(x) = \text{sign} \left( \sum_{i=1}^M \alpha_i y_i(x) \right)$$

AdaBoost-朴素贝叶斯情感分类框架如图4-10所示。

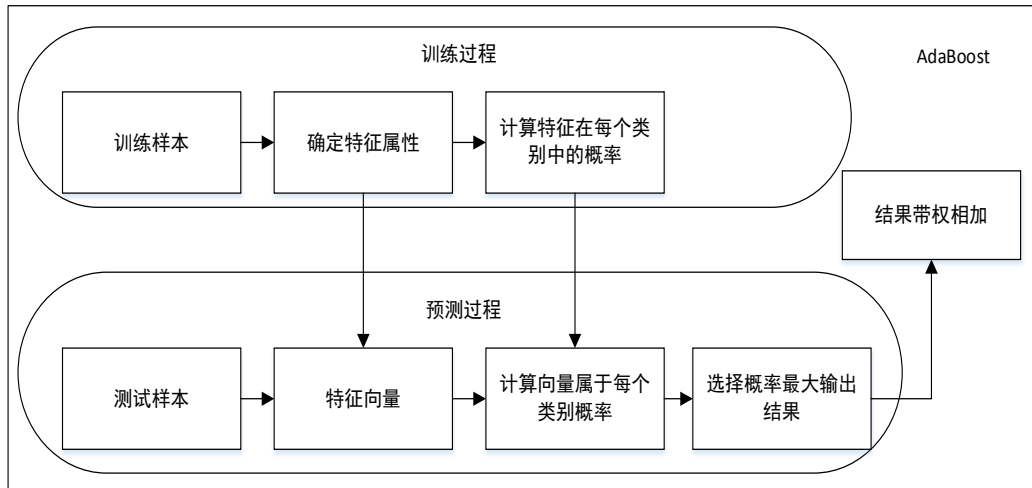


图 4-10 AdaBoost-朴素贝叶斯框架图

通过增强学习的方法希望尽可能的接近 SVM 的分类准确率，使得最后的三种分类器的融合可以在效果差不多情况下再次提高。通过和下文所述的 Random Subspace 提升学习方法对比，选定更好的一组方法作为选定的集成学习方法。

## (2) Random Subspace-KNN算法

上述说过，经过支持向量机、朴素贝叶斯、KNN等算法的测试，可以得出 KNN 的分类效果是最差的，但是如果能通过提升方法进行一定的增强，就会有更好的效果。和朴素贝叶斯和 SVM 算法的缺点不同，KNN 算法是一种惰性学习，没有主动去学习训练集，所以 KNN 算法在大数据的情况下效果较慢，但是为了保持准确度，这种方法可以接受的。其缺点是，在样本平衡性不好情况下，一

种类别样本过多，可能会导致分类效果大大下降，所以本文的训练集进行了人为的平衡，尽量的使两类的样本数量相同。

使用AdaBoost、Random Subspace方法分别对KNN方法进行增强测试，期望能够尽可能的接近SVM的分类效果，通过提升学习方法间的对比，选定更好的集成方法作为提升。Random Subspace-KNN情感分类框架如图4-11所示。

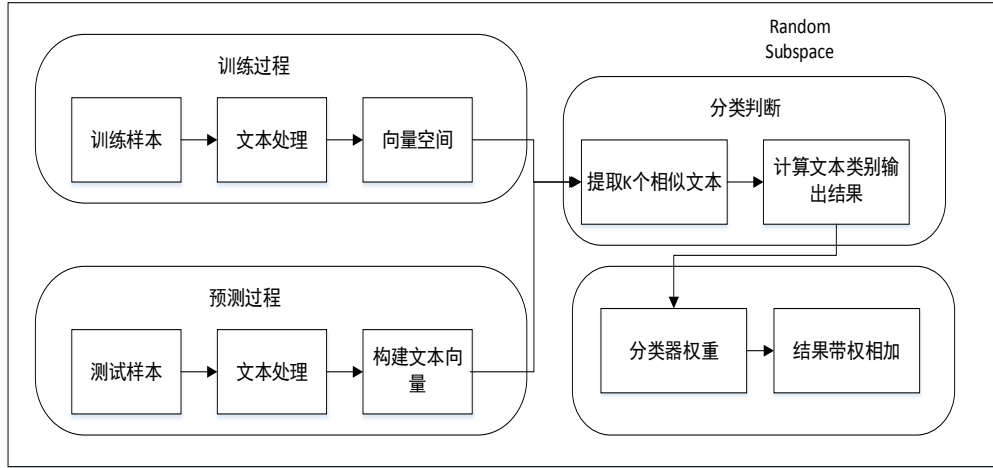


图 4-11 Random Subspace-KNN框架图

Random Subspace-KNN算法步骤如下所示。

#### 算法 4-2: Random Subspace-KNN 算法

**输入:** 输入文本  $Y=(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i)$ ，基分类器: KNN 分类器，随自己空间比率  $k$ 。

**输出:**  $Y_n(x) = \arg \max_{i \in C} (\sum_{i=1}^M \alpha_i)$

Step1: for  $t=1, \dots, M$ :

根据比例，随机生成一些子向量空间  $V_i$ ，每个都是  $K$  维，同时将其将  $V_i$  在原本向量空间标记位置信息

$$V_t = RS(V, k)$$

Step2: 根据所有得到的特征子空间  $V_1, V_2, \dots, V_m$ ，数据集  $Y$  通过不同的特征子空间得到不同子数据集  $Y_1, Y_2, \dots, Y_m$

Step3: 将 KNN 分类器和数据集  $Y_1, Y_2, \dots, Y_m$ ，形成  $m$  个 KNN 分类器，

$$h_i = KNN(Y_i)$$

Step4: 对于测试样本  $X$  根据特征子空间  $V_1, V_2, \dots, V_m$ ，同样是划分为不同子数据集  $X, X_2, \dots, X_m$

Step5: 最后得到  $M$  个 KNN 分类器，每种类别分别相加得到类别得分:

$$Y_n(x) = \arg \max_{i \in C} (\sum_{i=1}^M \alpha_i)$$

### (3) SVM 情感分类步骤

上述已经介绍过 SVM 算法的原理，通过对训练数据的学习构造分类器，从而预测测试样本分类。SVM 情感分类框架图如 4-12 所示。

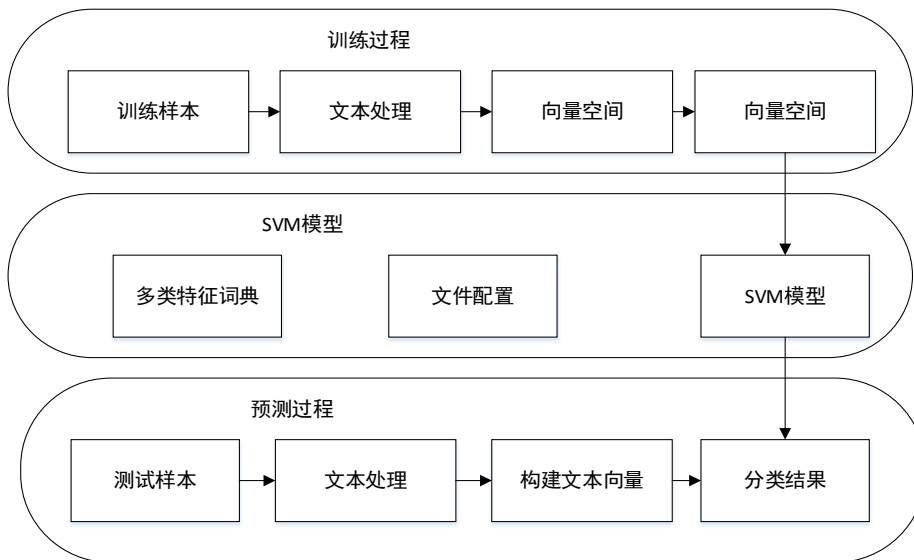


图 4-12 SVM 框架图

SVM 分类器步骤如下所示。

#### 算法 4-3: SVM 分类器步骤

**输入：**输入文本训练集  $Y=(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i)$

**输出：**类别标号以及每个类别贡献值

**Step1:** 将训练数据集通过预处理来处理

**Step2:** 通过上述构建向量空间的方法将文本表示成向量空间，但是要满足 LibSVM 的输入格式。

**Step3:** 通过调节 SVM 分类器参数  $c$ 、 $g$  来使得分类器的效果最好，调出最优的分类器。

**Step4:** 通过  $c$ 、 $g$  生成的分类器训练训练样本，生成分类器模型

**Step5:** 测试样本以同样的输入格式形成向量空间。

**Step6:** 将测试样本分类后，输出其类别标号以及每个类别贡献值。

### (4) 分类器融合

分类器融合的思想上边已经逐一介绍过，本文采用的是基于带权重的投票方式。每种算法都有本身固有的缺陷，对于朴素贝叶斯来说就算使用增强学习，让其逼近 SVM 分类的准确度，但是其固有属性之间的相互独立，还是会影响其分类结果。对于 SVM 来说，由于数据缺失造成的问题是无法克服的，尤其评论



信息较短并且覆盖面较小，导致分类精度有所偏差。KNN算法会因为类别或者数据集的不均匀导致结果不理想。

通过集成学习，能够增强较弱的分类器的效果，尽可能的拉近三个分类器之间的准确率，使得在二次融合之前单个分类器的分类效果尽可能好。为了尽量减小每种分类器固有的缺陷所带来的影响，将多个分类器融合，使其覆盖范围更广，形成基于投票和集成学习方法的融合分类器。本文想通过这种二次提升的办法来增加机器学习算法的分类精度。融合分类器的框架图如图4-13所示。

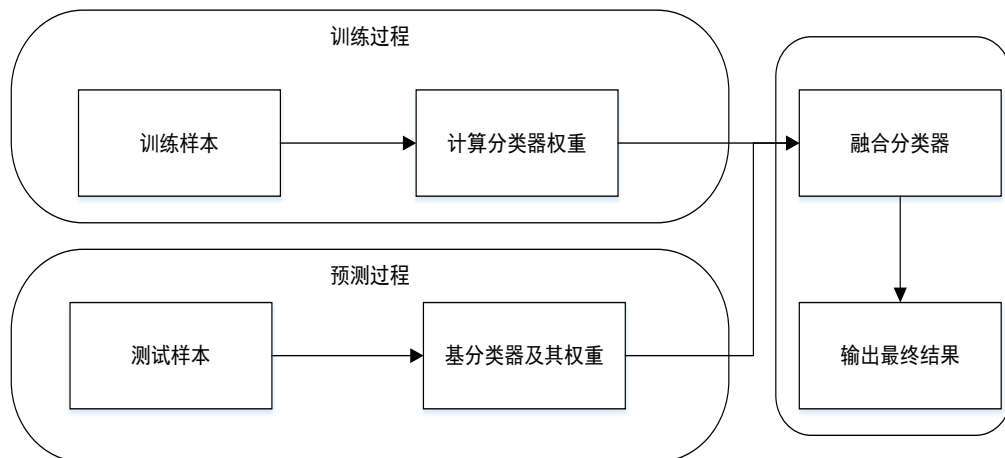


图 4-13 融合分类器框架图

分类器权重根据三种分类器输出的类别及其对应的概率计算，针对每个类别概率相加，相加后权值最大类别为最终类别。

融合分类器算法步骤如下所示。

#### 算法 4-4：融合分类器算法

输入：输入文本  $Y=(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i)$

输出： $Q(x) = \max_j \sum_{i=1}^M (sc_i la_{i,j})$

Step1: 将每个分类器的权值设定  $w_i = \frac{1}{M}$ ,  $i=1,2,\dots,M$ ,  $M$  为分类器个数。

Step2: 通过分类器 1 预测结果,  $F_1 = \langle la_1, sc_1 \rangle$ ,

分类器 2 预测结果,  $F_2 = \langle la_2, sc_2 \rangle$

分类器  $M$  预测结果,  $F_m = \langle la_m, sc_m \rangle$

根据每个分类器的得分更新分类器的权值

Step3: 最后得到:

$$Q(x) = \max_j \sum_{i=1}^M (sc_i la_{i,j})$$



## 4.5 实验结果与分析

本章提出了特征挑选和计算相应权值上的改进以及基于投票和集成学习的融合分类器来进一步提高判断情感倾向性的准确率。通过阅读文献和参考资料选取了其中比较有效的特征进行组合，使用单一的分类算法作为本文的Baseline，实验表明本文提出的方法有一定分类效果上的提高。

### 4.5.1 实验数据

本文的实验数据是通过爬虫在网上抓取的评论，其中包含了多个社会舆论方面的微博及其评论，因为每个话题下微博评论数量不一致，所以本文尽可能抽取同样的数据。通过垃圾评论的处理后，共有2万条左右评论数据，经过人工标注倾向性，剔除其中的干扰数据。经过主客观评论分离后，统计了15844条的训练数据，这其中正负倾向评论大体各占一半。测试数据共有三个数据集，测试集A是经过合并后的数据集，这部分测试集和训练集比较相似，共有2000多条。测试集B和训练集相似度比较小，也为2000条左右。单一话题数据集包括了2000条的训练集和400条测试集。其具体内容如表4-2所示。

表 4-2 数据集示例

数据集	训练集条数	测试集条数	话题个数
1	15844	A2147	多个
2	15844	B1657	多个
3	2000	C400	单个

### 4.5.2 特征提取与权重计算改进实验结果

本文对特征提取方法的改进，可以让信息增益方法，更加注重对于情感词的挑选，将训练语料中没有情感的字比重降低，本文将其命名为IG-M。计算特征词权重时，通过TF-IDF的改进，使用权重的表示区分出现次数相同，但是作用明显不同的特征，同时在程度词与情感字的组合出现时，加大情感字的贡献度，让其有更好的表示分类能力，将其命名为TF-IDF-M。本文的对比试验主要是针对第一个数据集，Baseline为传统的信息增益和TF-IDF并且使用SVM分类器实验结果。第二组是使用改进的TF-IDF-M的结果。第三组是使用IG-M、SVM的分类结果。第四组本文方法为IG-M、TF-IDF-M以及SVM分类器。几组实验都使用相同的特征组合。

四种方法的实验效果如表4-3所示。

表 4-3 不同方法比较结果

实验方法	POS_P	POS_R	POS_F	NEG_P	NEG_R	NEG_F
Baseline	0.8495	0.7884	0.8178	0.7734	0.8380	0.8044
TF-IDF-M	0.8648	0.8153	0.8393	0.7991	0.8521	0.8247
IG-M	0.8625	0.8213	0.8414	0.8036	0.8481	0.8253
本文方法	0.8756	0.8361	0.8554	0.8193	0.8622	0.8402

从表4-3可以看出，两种改进方法对分类效果都有一定的提高，但是通过F值比较得出两种改进方法的结合比TF-IDF-M和IG-M各自单独使用效果更好，说明本文方法确实有效。

对于不同数据集，使用本文的方法进行测试，包括数据集1、2、3。实验结果如表4-4所示。

表 4-4 不同测试集的比较结果

数据	POS_P	POS_R	POS_F	NEG_P	NEG_R	NEG_F
测试集 1	0.8756	0.8361	0.8554	0.8193	0.8622	0.8402
测试集 2	0.7028	0.6760	0.6891	0.7167	0.7414	0.7288
数据集 3	0.9046	0.8452	0.8739	0.8551	0.9150	0.8840

从上表可以看出在同一个训练集下，测试集 1 和测试集 2 的各项指标有比较明显的差距，测试集 1 是和训练集话题及评论风格非常相近的数据，所以其结果更好。测试集 2 是和训练集较为不相像的评论，主要为商品评论，而商品评论和本文所用的社会舆情类性质的评论集有所差异。数据集 3 的训练集和测试集是同源数据，其结果较好于其余两个测试集，可见测试集与训练集出于同源可以有更好的分类效果。

#### 4.5.3 分类器改进的实验结果

训练集和测试集在本节开始已经描述过，测试主要是针对数据集1，并且本文的实验都是基于多特征组合方式构成向量空间，特征选择和权重计算都使用本文提出的改进方法。通过对三种不同分类器的实验结果观察，可以得出SVM的测试结果要稍好于NB、KNN分类器，如表4-5所示。

表 4-5 实验结果示例

分类器类型	ACC 准确率
SVM	84.82%
朴素贝叶斯	80.58%
KNN	77.22%

### (1) SVM分类器

SVM分类器的计算，本文采用的是libsvm的SVM工具包，SVM工具包在训练模型之前要进行参数调优，不断测试c和g的参数，通过寻求最优的参数调节最大化的提升分类器性能。

其中c和g的参数的示例如图4-14所示。

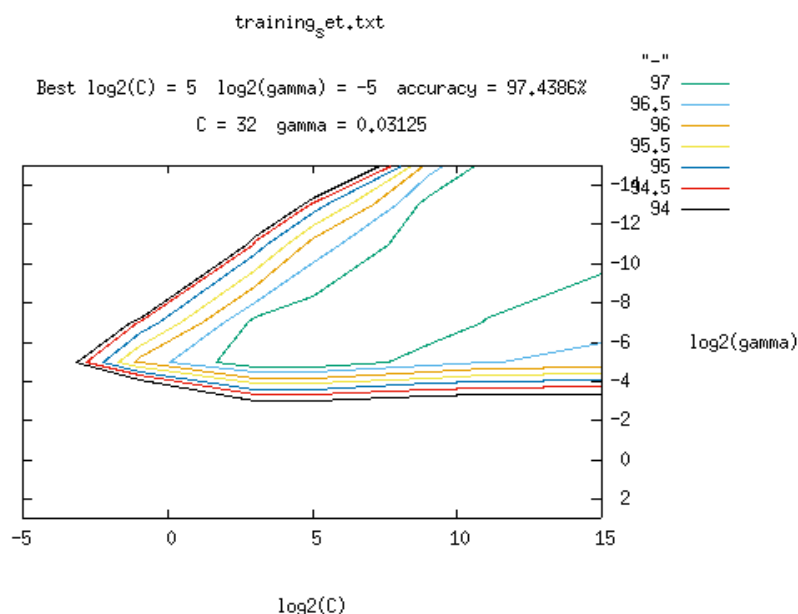


图 4-14 SVM参数调优示例图

### (2) 朴素贝叶斯与AdaBoost、随机子空间结合

本文希望能通过增强学习的方法将朴素贝叶斯分类器和SVM的分类效果拉近，可以减少因一种分类器效果远好于其他分类器，造成的多数投票反而会降低分类效果的情况。将朴素贝叶斯通过随机子空间和AdaBoost分别进行提升，其实验结果如表4-6所示。

从下表可以看出，在分别对朴素贝叶斯使用AdaBoost、Random Subspace方法上得到的结果中，AdaBoost的提升方法更有优势，准确率上的提升约有1%-2%，可以一定程度上拉近和SVM分类器的效果，所以选用了AdaBoost-朴素贝叶斯作为投票集成的基分类器之一。

表 4-6 朴素贝叶斯提升实验结果

分类器	POS_P	POS_R	POS_F	NEG_P	NEG_R	NEG_F
Bayes	0.8458	0.7805	0.8118	0.7664	0.8350	0.7992
AdaBoost-Bayes	0.8588	0.8023	0.8296	0.7869	0.8471	0.8159
Random-Bayes	0.8480	0.7840	0.8147	0.7697	0.8370	0.8019

### (3) KNN与AdaBoost、随机子空间结合

本文希望能通过增强学习的方法将KNN分类器和SVM的分类效果拉近,使用两种增强学习方法进行集成学习,其中KNN的k值选取,采用10次10折交叉验证,得到k值为30,其实验结果如表4-7所示。

表 4-7 KNN 提升实验结果

分类器	POS_P	POS_R	POS_F	NEG_P	NEG_R	NEG_F
KNN	0.7954	0.7701	0.7825	0.7410	0.7686	0.7545
AdaBoost-KNN	0.7906	0.7639	0.7770	0.7348	0.7636	0.7489
Random-KNN	0.8341	0.7898	0.8101	0.7681	0.8129	0.7899

从上表可以看出,对KNN算法使用AdaBoost的增强学习是失败的,其中的正向以及负向的F值都有所降低,说明并不适合KNN算法,但是Random的效果更好,准确率提升了约有3%左右,尤其是在负向评论上提升效果更好,F值都有明显的提高,说明Random Subspace适合KNN算法的提升,所以选用Random Subspace-KNN作为基分类器。

### (4) 基于投票与集成学习的融合分类器

通过实验来看,采用基于投票的融合分类方法来实现分类器的增强是有效的。将三种没有做集成学习的传统分类器通过投票融合,得到实验结果 T-F1。实验对比结果如下表 4-8 所示。

表 4-8 实验结果示例

分类器	POS_P	POS_R	POS_F	NEG_P	NEG_R	NEG_F
SVM	0.8756	0.8361	0.8554	0.8193	0.8622	0.8402
AdaBoost-Bayes	0.8588	0.8023	0.8296	0.7869	0.8471	0.8159
Random-KNN	0.8341	0.7898	0.8101	0.7681	0.8129	0.7899
T-F1	0.8846	0.8378	0.8605	0.8227	0.8732	0.8472
融合分类器	0.8881	0.8465	0.8668	0.8311	0.8763	0.8531

根据上表的结果来看，本文的基于投票和集成方法在分类效果上有一定的提升，其中正负两类倾向的F值都有所增加，其效果好于单独分类器的使用。并且相较于没有提升的传统分类器的简单组合在各项评价标准上有更好表现，准确率总共提升了1.2%左右。可见本文方法对于情感倾向性分析起到一定增强作用。

#### 4.5.4 实验结果分析

本文对分类的提升主要是分为两部分，其一是信息增益和TF-IDF的改进，其二是针对分类器的改进。

第一部分是针对信息增益和TF-IDF的改进，根据数据集1的测试情况来看，可以增加大约4个百分点的准确率。通过图4-15可以明显的看出，改进以后的方法对于初始的Baseline有一定的提高。

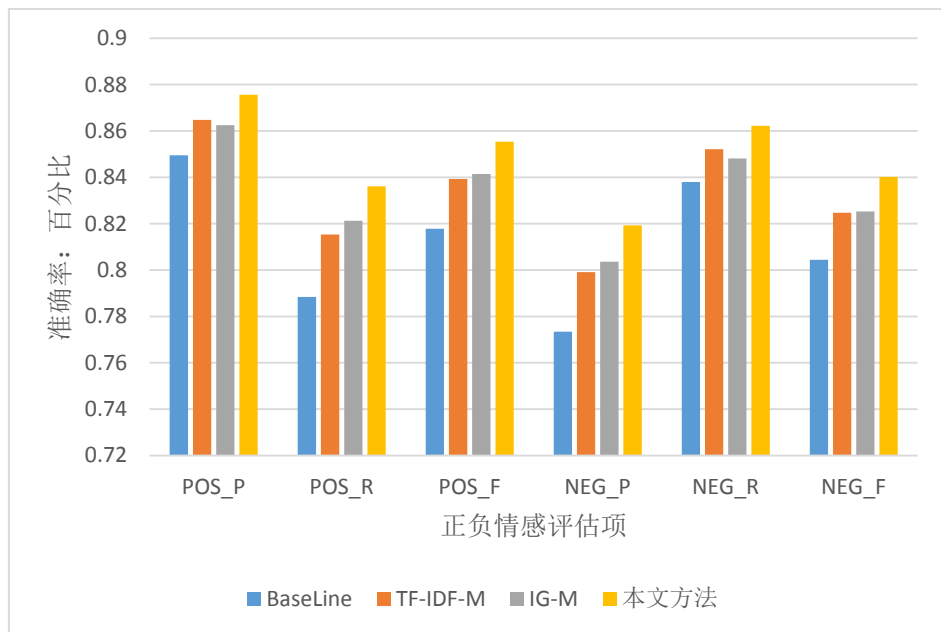


图 4-15 在数据集1上的方法比较

从上图可以看出，无论是在召回率、准确率、F值上都有一些明显的提高，虽然提高的效果没有飞跃式的增长，但是可以说明本文的方法是较为有效的，在一定的向量空间中可以包括更多的情感词信息，在增强准确率的基础之上，还能有效的控制计算效率。

第二部分是分类器的融合，希望能够通过提出的方法使得不同分类器之间的缺陷在一定程度上得到克服，其在准确率上大约提升了百分之1.4，通过集成学习的分类器比单个分类器效果更好一些。

##### (1) 基于投票与集成的融合分类器与SVM比较

从图 4-16 可以看出 SVM 与本文提出的融合分类器的效果不同。

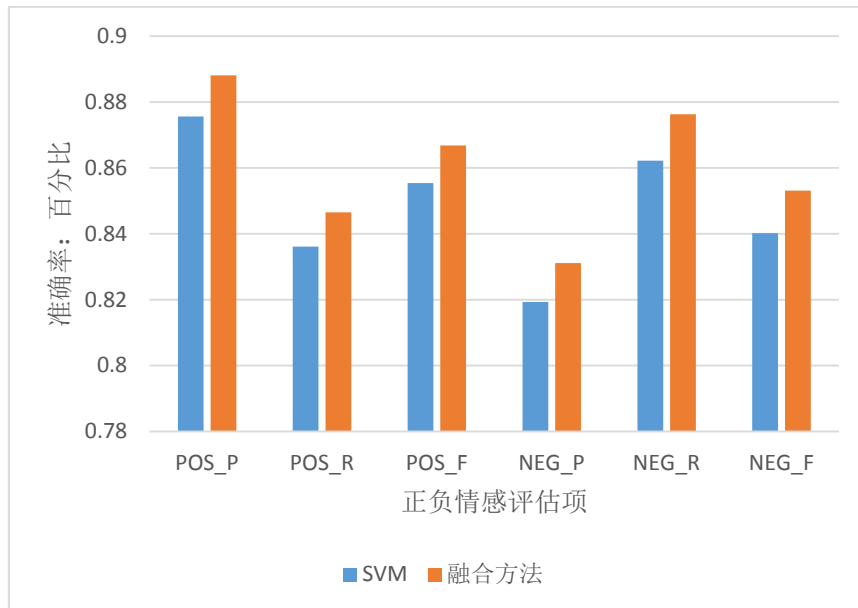


图 4-16 融合方法和SVM比较

融合后的分类器的预测效果更好，并且正样本和负样本的评价标准中 F 值都要略高一些，证明本文提出方法效果更好。

(2) 基于投票与集成的融合分类器与朴素贝叶斯比较，如图 4-17 所示。

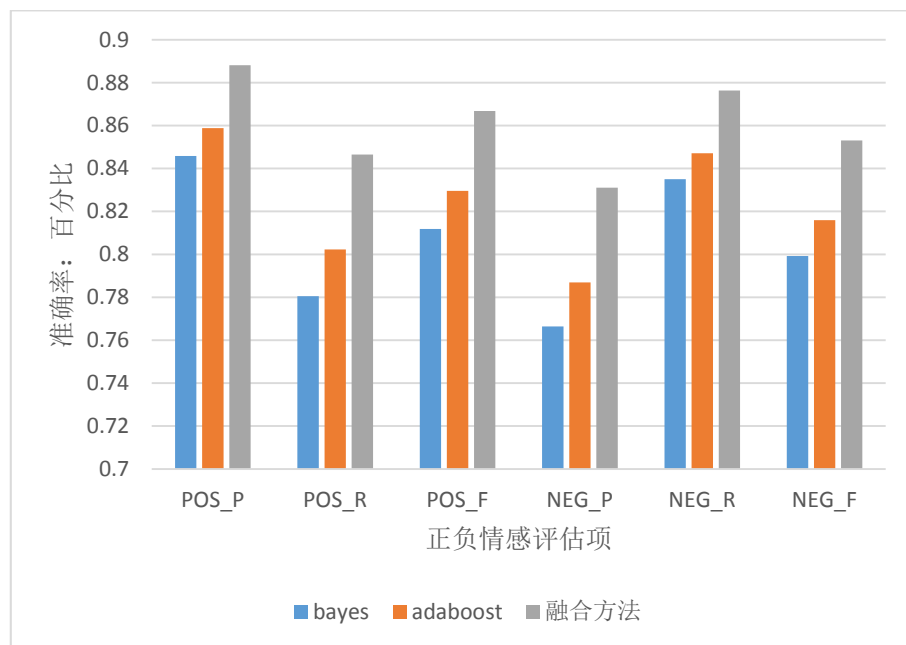


图 4-17 融合方法、朴素贝叶斯、朴素贝叶斯增强比较

通过观察 F 值，得出 AdaBoost-朴素贝叶斯相比传统朴素贝叶斯在分类效果上有了一定提高，而融合分类器的提升效果更明显，说明本文方法确实有效。

(3) 基于投票与集成的融合分类器与 KNN 方法比较，如图 4-18 所示。

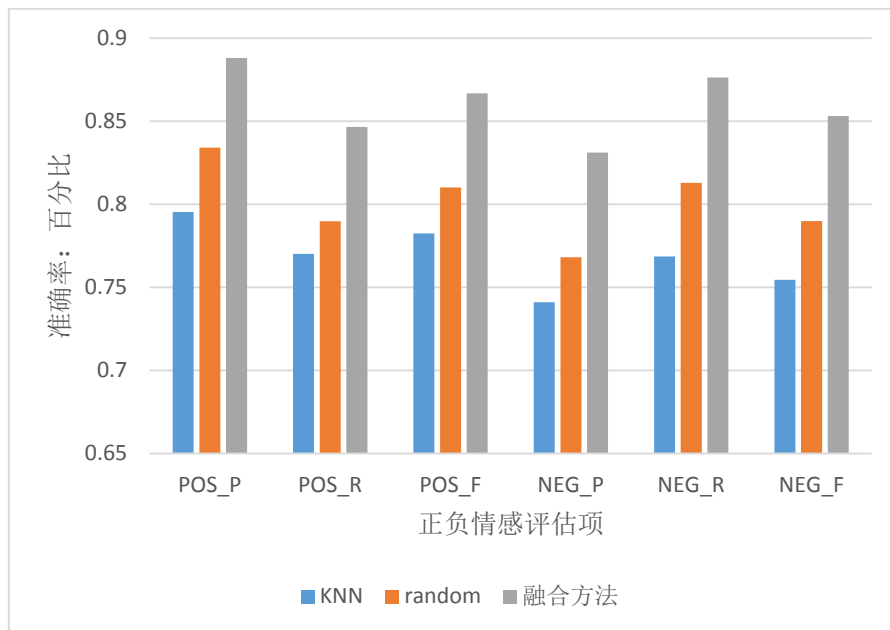


图 4-18 本文方法、KNN、KNN增强比较

KNN 分类器对负向评论的分类效果稍显弱一些，通过增强学习后，明显提高了一些分类精度，并且融合分类器预测的效果有显著的提升。

通过第一部分和第二部分的结合，分类结果有所提高。通过 F 值可以观察到，针对特征提取及其权值计算的改进并且结合本文提出的融合分类器，较 Baseline 有了一定的提高，在准确率上约有 5% 的提升，验证了本文方法有很好的效果。本文方法 1 为特征改进方法和 SVM 相结合，本文方法 2 为改进方法和融合分类器相结合。其相应的实验数据统计如下图 4-19 和表 4-9 所示。

表 4-9 实验结果示例

实验方法	POS_P	POS_R	POS_F	NEG_P	NEG_R	NEG_F
Baseline	0.8495	0.7884	0.8178	0.7734	0.8380	0.8044
本文方法 1	0.8756	0.8361	0.8554	0.8193	0.8622	0.8402
本文方法 2	0.8881	0.8465	0.8668	0.8311	0.8763	0.8531

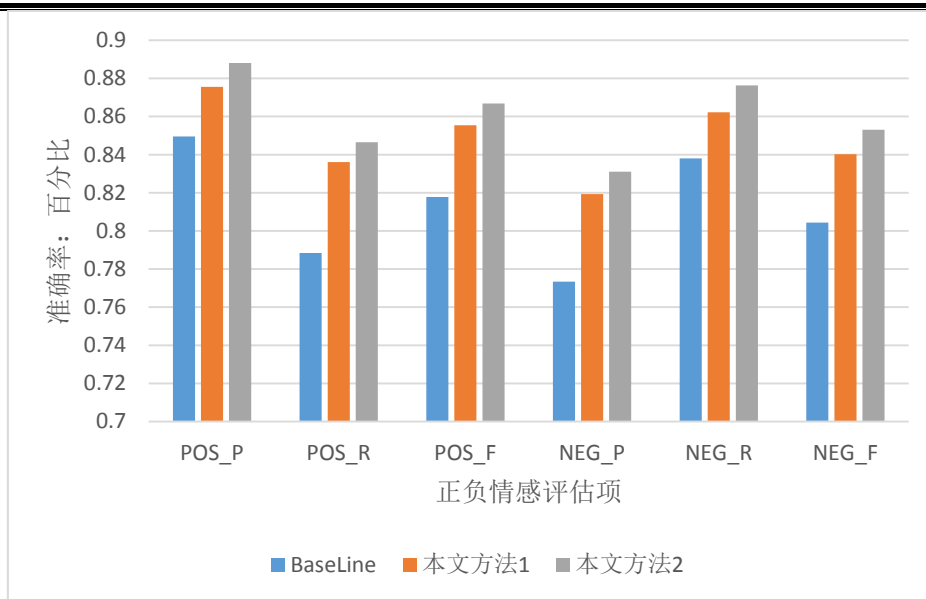


图 4-19 本文方法 1、本文方法 2、Baseline 之间比较

通过数据分析评论情感倾向性，可以看出主要是因为以下几个原因会明显的影响分类的精度。

#### (1) 情感词及其权重

情感词及其权重的运用十分重要，非情感词可能会对于分类没有任何帮助甚至可能来干扰分类的效率，例如：在同一分类之中出现“社会”、“政府”之类词项，在训练集中因为出现在正向较多，但是测试集中贬义评论出现更多，很容易造成分类的不精确。

#### (2) 语料库

语料库对于实验效果也有很大的影响，对于不同的训练集和测试集在相同的方法下达到而效果可能会不一致，尤其是在测试集和训练集不相近的情况下，那么领域词就会导致训练集对于测试集失效，所以只有在相同领域下才能有更好的分类结果

#### (3) 分类器性能

分类结果也会在一定程度上受分类器的性能的影响。不同分类器所运用的机器学习算法也不相同，每种算法都有各自的优缺点，也就是说同一测试集使用不同分类器得到的结果自然不同，为了达到更好的效果，本文使用了集成组合的方式运用分类算法来达到理想的效果。

## 4.6 本章小结

本章主要研究评论的情感倾向性判断的方法，第一部分提出了 IG-M 特征提取方法和 TF-IDF-M 权重计算方法。其中 IG-M 特征对于传统的基于信息增



益的特征提取方法进行改进，能够挑选出更多有效情感词。TF-IDF-M 则对原有的 TF-IDF 方法进行了改进，在情感词和修饰词同时出现的时候，改进的方法可以有效的增加情感词的权重，从而进一步提高微博评论的情感分类效果。第二部分针对不同分类算法的缺陷，提出了基于投票与集成学习的融合方法来使得分类器能够进一步提高分类准确率。通过实验证明，本文方法取得了一定的效果。

## 第5章 微博评论倾向性载体分析

上述的情感倾向性分析，主要是针对单一评论的情感倾向性的研究，但是评论的载体不尽相同，尤其是热点事件的褒贬倾向评论可能所针对的对象并不一样，所以只是针对每个评论的倾向性判断往往不够，可以从整体方面来研究褒贬倾向评论，本文采用LDA主题模型，将其应用到评论的情感载体的抽取上，观察爆发意见的评论集的具体载体。

### 5.1 微博评论话题转移

微博评论话题转移主要是针对民众共同探讨的话题，热点话题大部分评论围绕一定的主题内容，但是具体的评论对象却是有所差别。评论载体的研究是一种细粒度的倾向性分析，评论载体也可以看做倾向性分析的一部分。本文对经过判断情感倾向后的数据集上，分析褒贬评论对应的情感载体，看其中部分评论是否发生了情感对象及话题上的转移。

### 5.2 文本表示与 LDA 主题模型

本文采用上一章描述的情感倾向性分析方法，将评论数据集分类成两部分，一部分为褒义性评论，一部分是贬义性评论。这两个文本集的倾向就比较明显，而其中的情感词已经清晰。本文载体分析主要是分为两部分，第一部分是针对性的表示文本，第二部分是LDA主题模型抽取评论载体。

#### 5.2.1 文本表示

从褒贬数据集中分别抽取其中的情感词，情感词典在第二章已经有所描述，这里用到了正负情感词典、程度词词典、微博表情符号词典，将其合并形成情感词典D。通过从文本中挑出特定词性的特征词来表示评论。

文本表示方式如下所述。

---

**算法 5-1：文本表示方式**

---

**输入：**评论集合 X，评论集合 Y，情感词典 D，程度词 C

**输出：**文本表示集合  $Q_x$ ， $Q_y$

**Step1:** 将文本中的每个评论去除停用词，删除不必要的干扰词。

---

Step2: 遍历评论集合  $X$ ，将评论集合中每条评论和情感词典匹配，将情感词抽出，放入情感词集合  $X_q$ 。并且按顺序抽取情感词、名词、程度词，表示成为文本集合  $Q_x$

Step3: 遍历评论集合  $Y$ ，将评论集合每条评论和情感词典匹配，将情感词抽出，放入情感词集合  $Y_q$ 。并且按顺序抽取情感词、名词、程度词，表示成为文本集合  $Q_y$ 。得到  $Q_x$ 、 $Q_y$  两个文本集。

通过上述方法，将每条评论都以情感词、程度词、名词、地名等表示出来，作为LDA主题模型的输入。

### 5.2.2 LDA 主题模型

LDA主题模型是一种无监督的机器学习方法，目的是抽取文档的主题内容，通过词的层次结构来表示文档集合的主题。具体的概率表示模式如图5-1所示。

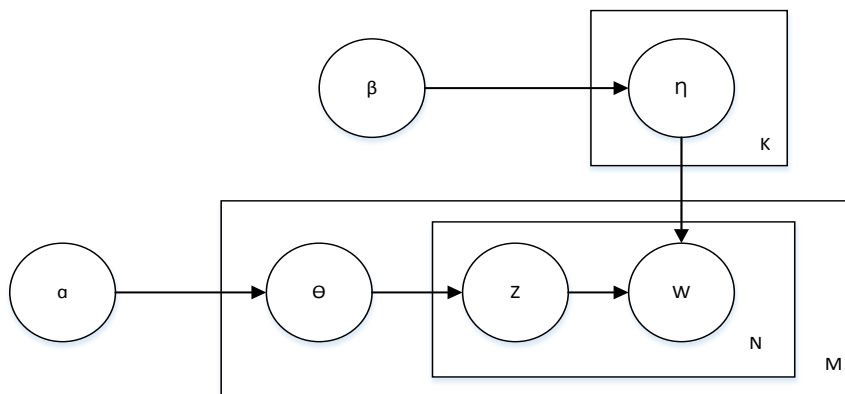


图 5-1 LDA模型

表5-1所示为每个参数所代表的含义：

表 5-1 参数含义

参数	说明	参数	说明
$\alpha$	主题在文档上的比值	$W$	文本集的词
$\beta$	主题和单词之间的参数	$N$	单词总数
$\theta$	使用矩阵表示每个主题概率	$M$	文本集总数
$\eta$	单词在主题上的概率	$K$	主题个数
$z$	文本集单词主题		

LDA的计算过程如下所示：

通过  $\theta$  和  $\eta$ ，计算  $w$  的概率如公式（5-1）所示：

$$p(w|\theta, \eta) = \sum_z p(z|\theta) p(w|z, \eta) \quad (5-1)$$

通过  $\alpha$  和  $\eta$ ，得到  $\theta$ 、 $z$ 、 $w$  的概率如公式（5-2）所示：

$$p(\theta, z, w|\alpha, \eta) = p(\theta|\alpha) \prod_{n=1}^N p(Z_n|\theta) p(w_n|z_n, \eta) \quad (5-2)$$

通过 $\alpha$ 和 $\eta$ ，得到 $w$ 的概率如公式（5-3）所示：

$$p(w|\alpha, \eta) = \int p(\theta|\alpha) \prod_{n=1}^N p(w_n|\theta, \eta) d\theta \quad (5-3)$$

将前三个公式融合，得到一个文本集概率如公式（5-4）所示：

$$p(w|\alpha, \eta) = \int p(\theta|\alpha) \prod_{n=1}^N \sum_{Z_n} p(Z_n|\theta) p(w_n|z_n, \eta) d\theta \quad (5-4)$$

最后得到每个文本集的概率矩阵制，如公式（5-5）所示：

$$p(D|\alpha, \eta) = \prod_{d=1}^D \int p(\theta_d|\alpha) \prod_{n=1}^N \sum_{Z_{dn}} p(Z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \eta) d\theta_d \quad (5-5)$$

由 LDA 主题模型生成的文件中包括了：文档及相应主题关系文档，主题与单词关系文档、文档、主题及单词关系文档等。本文主要是运用其中的主题与单词的概率关系矩阵。

LDA 主题模型主要是应用在文本话题中关键词的抽取，而其优势是在好的文本环境下可以很有效的抽出话题的关键词，但是其劣势也很明显，如果文本集过于繁杂而且主题不明确，那么模型会失效。而本文之所以采用 LDA 进行主题抽取，前提就是在已有的情感倾向性分析基础上，因为同一话题和类别下大部分带有倾向性的评论，其所针对的情感载体都有相似之处。可以通过 LDA 主题模型进行主题的抽取，从而看出群众的评论的情感对象和是否发生话题的转移。

### 5.3 实验结果与分析

LDA 模型主要分两种方法用参数估计，分别是 EM 和 Gibbs，而本文使用的是吉普斯采样的方式，尤其是目前研究计算量不大，比较适合吉普斯采样。因为评论的个数并不多，所以本文选取的主题数 K 主要是 2 到 10，在这个区间取每个整数值进行实验。

#### 5.3.1 实验数据

首先评论集合分为两类数据，其中一组是包含了 2000 条左右评论，其中正负评论分别大概为 1000 左右。如表 5-2 所示。

表 5-2 数据集

数据集	个数
数据集 P	1001
数据集 N	954

### 5.3.2 实验结果与分析

通过上述实验的进行实验观察其评价对象是否发生变化，其中数据集的微博内容简化为“老母亲养残疾儿子 20 年，不花政府一分钱。”经过上一章的分类方式分成数据集 P、数据集 N。通过情感词提取和 LDA 模型的主题抽取得到结果如表 5-3、5-4 所示。示例为抽取 5 个主题，每个主题的词数为 20。

表 5-3 数据集 P 主题抽取

T1	T2	T3	T4	T5
母亲	伟大	爱	感动	母爱
好	母爱	泪	中国	心
政府	母亲	妈妈	赞	[good]
感动	说	好	伟大	不言弃
儿子	哭	永远	话题	伟大
能	世界	无私	网页	给力
部队	回复	不弃	链接	罗长姐
负担	想	无言	央视	回复
没有	无私	只有	新闻	鲜花
愿	责任	不离	致敬	温暖
应该	真	伟大	平安	中
母亲	加油	母亲	好人	太阳
真的	真的	父母	爱心	人性
当兵	坚持	这位	可怜	支持
老人	奉献	看过	老奶奶	鼓掌
人民	善良	希望	心酸	无疆
但	天下	感人	传递	老母
因为	健康	故事	一生	无边
孩子	赞一个	孩子	除了	太
太	这位	世上	微	悲伤

LDA 的主题抽取为隐式抽取，只能通过自己归纳和总结出主题内容，如果只对评论进行主题抽取，可以看出两类评论中大多数都围绕一个主题发表评论。

通过上一章的情感倾向性分类方法，在已知将评论分成两类基础上，对微博评论情感倾向进行统计，得出民众更倾向于何种情感。但是从上两张表可以看出，褒义和贬义数据集评论对象并不相同，从表 5-3 看出同一个话题和评论类别下，每个主题的分度度很小，其主要围绕“母爱、母亲”话题，评论的载体和微博本身并没有很大的差异度。

然而从表 5-4 所示负向评论主要是围绕“政府、部队、国家”的话题，但是微博本身并没有出现这几种评论载体，可见负向话题的载体并不是和微博内容一样，产生了大部分评论的载体转移。

表 5-4 数据集 N 主题抽取

T1	T2	T3	T4	T5
国家	政府	部队	政府	国家
不	却	不	部队	没有
为什么	黑	负担	怒	怒骂
责任	想	不管	不管	无耻
应该	相关	说	黑	泪
负责	搞	受伤	作为	黑
宣传	老人	增加	做	老人
承担	忽略	真	母亲	新闻
母亲	泪	儿子	无能	报道
当兵	承受	愿	事	伤残
社会	百姓	意外	好	事情
却	ZF	国家	回复	怒
干嘛	妈妈	真的	年	高级
钱	鄙视	当兵	背后	伤心
部门	希望	添麻烦	军队	脸
不该	感动	老人	却	军人
相关	添麻烦	母亲	脸	好意思
傻逼	负责	精神病	泪	脸
无奈	中国	好意思	责任	添麻烦
治疗	精神病	添	中国政府	不是

本文的测试数据的主要内容为正向主题，但是评论中出现了大量集中的对政府等部门的责备。从情感的变化来看负向情绪有增加的趋势，对于这种潜在的舆情爆发点，可以进行监测，一旦超过临界点，就可以发出舆情预警，例如：部队或者政府应该加强退伍军人的保障，做好善后处理等。

舆情言论的分析可以对事件可能引起的爆发点起到一定监督作用，政府等部门也可以快速的做出反应和应对措施。这种抽取的方法可以引申出几个关注的问题：

- (1) 评论集体的情感转移、舆论爆发点及对象挖掘
- (2) 政府对爆发性事件进行及时的舆论监控
- (3) 为检测到的恶性事件提供有效的建议和合理的应对措施

通过情感倾向性的分析和 LDA 模型的抽取主题，能够在舆情分析上有更

好的帮助以及指导作用。

## 5.4 本章小结

本章主要是描述了情感倾向分析的载体转移，通过特定方式的文本表示和 LDA 主题模型抽出评论集合的主题来分析评论集的评价对象，观察看其是否有评价载体的转移。通过分析本文所举示例，能得出某类争议性事件的评论集合很可能发生载体转移，从而起到舆论预警、提供合理建议及措施等作用。

## 结 论

目前对于网络环境的改变,很多信息交流平台已经替代了传统的交流方式,其中以新浪微博等平台作为代表,民众在平台基础上交流沟通。本文的研究对象主要是从新浪微博上抓取的有关于社会现象的微博评论。

本文主要研究目的是对不同类型的评论进行分类,其中以主观评论的褒贬倾向作为重点研究内容。通过各自的特征组合来实现垃圾、主客观评论分类。对倾向性分析,采用了特征组合和IG-M以及TF-IDF-M方法来构建向量空间,结合SVM、AdaBoost-NB、Random Subspace-KNN融合的集成分类器来提高分类精度。

首先本文采用了两步两分类的研究方式,根据已有研究,此方法能够更好地将三分类问题通过分步解决,分类精度能够更好。垃圾评论的分类主要是采用了基于统计的无监督分类,通过对垃圾线索的统计,根据每种特征的权重来界定阈值,实现对测试样本分类。主客观评论采用了特征组合方式,选取了几种比较明显的特征,用朴素贝叶斯分类器分类,测试得到准确率在80%左右。

其次,评论褒贬倾向分析中,确定了构造向量空间的特征,作为特征组合。同时本文提出了针对情感词的信息增益和TF-IDF上的改进,使得带有情感倾向的词起到更大作用,作为更明显的特征项加入到向量空间之中。随后,通过AdaBoost和Random Subspace来提升朴素贝叶斯和KNN分类器的分类效果,采用投票方式融合不同的基分类器从而实现二次提升。

最后,本文针对褒贬评价对象进行研究,利用LDA模型抽取主题关键词,通过人工概括方法研究评论集载体,对舆情分析有一定指向性作用。

本文研究虽得到了一定效果,但是还需要进一步改进。

(1) 本文的语料库进行了平衡调整,但是有很多情况是不平衡语料库的导致的误差,所以应该寻求一些方法来解决不平衡语料上的分类,这也是后续的研究方向。

(2) 目前的研究方法主要是有监督的研究方式,但是其主要运用的领域词太过关键,如果不是同一领域或者相似度太低,那么模型就会失效。本文使用了word2vec词向量方法来构造向量空间,虽然得到了一定效果,但是结果并不是很理想。相同的特征组合针对不同训练集可能效果并不一样,然而深度学习方法可以不用人为提取特征,能够对任何训练数据统一进行学习。目前深度学



习的方法发展的还不够成熟，在使用同样数据的情况下，本文使用了theano深度学习框架进行实验，但是效果并没有传统机器学习方法好，然而这种统一化构造方式是十分值得深入研究。

## 参考文献

- [1] 樊博.2015微博用户发展报告[R]. 北京: 新浪微博数据中心, 2015.
- [2] 王琳, 冯时, 徐伟丽, 等. 一种面向微博客文本流的噪音判别与内容相似性双重检测的过滤方法[J]. 计算机应用与软件, 2012(08):25-29.
- [3] Joshi A, Balamurali A R, Bhattacharyya P, et al. C-Feel-It: a sentiment analyzer for micro-blogs, 2011[C]. Association for Computational Linguistics, 2011.
- [4] Das A, Bandyopadhyay S. Dr Sentiment knows everything!, 2011[C]. Association for Computational Linguistics, 2011.
- [5] Chesley P, Vincent B, Xu L, et al. Using verbs and adjectives to automatically classify blog sentiment[J]. Training, 2006,580(263):233.
- [6] Liu B. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012,5(1):1-167.
- [7] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, 2002[C]. Association for Computational Linguistics, 2002.
- [8] Neviarouskaya A, Prendinger H, Ishizuka M. SentiFul: A lexicon for sentiment analysis[J]. Affective Computing, IEEE Transactions on, 2011,2(1):22-36.
- [9] Prasad S. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods[Z]. Technical Report, 2010.
- [10] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008,2(1-2):1-135.
- [11] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using twitter hashtags and smileys, 2010[C]. Association for Computational Linguistics, 2010.
- [12] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[J]. CS224N Project Report, Stanford, 2009,1:12.
- [13] Li S, Huang C, Zhou G, et al. Employing personal/impersonal views in supervised and semi-supervised sentiment classification, 2010[C]. Association for Computational Linguistics, 2010.
- [14] Jiang L, Yu M, Zhou M, et al. Target-dependent twitter sentiment classification, 2011[C]. Association for Computational Linguistics, 2011.

- [15] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data, 2010[C]. Association for Computational Linguistics, 2010.
- [16] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data, 2011[C]. Association for Computational Linguistics, 2011.
- [17] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis[J]. Computational linguistics, 2009,35(3):399-433.
- [18] Mohammad S M, Kiritchenko S, Zhu X. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets[J]. arXiv preprint arXiv:1308.6242, 2013.
- [19] 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007(01):96-100.
- [20] 李钝, 曹付元, 曹元大, 等. 基于短语模式的文本情感分类研究[J]. 计算机科学, 2008(04):132-134.
- [21] 柳位平, 朱艳辉, 栗春亮, 等. 中文基础情感词词典构建方法研究[J]. 计算机应用, 2009(10):2875-2877.
- [22] 张成功, 刘培玉, 朱振方, 等. 一种基于极性词典的情感分析方法[J]. 山东大学学报(理学版), 2012(03):47-50.
- [23] 党蕾, 张蕾. 一种基于知网的中文句子情感倾向判别方法[J]. 计算机应用研究, 2010(04):1370-1372.
- [24] 赵妍妍, 秦兵, 车万翔, 等. 基于句法路径的情感评价单元识别[J]. 软件学报, 2011(05):887-898.
- [25] 刘志明, 刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012(01):1-4.
- [26] 王志涛, 於志文等. 基于词典和规则集的中文微博情感分析[J]. 计算机工程与应用, 2015(08):218-225.
- [27] 闻彬, 何婷婷, 罗乐, 等. 基于语义理解的文本情感分类方法研究[J]. 计算机科学, 2010(06):261-264.
- [28] 谢丽星, 周明, 孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报, 2012(01):73-83.
- [29] 傅向华, 刘国, 郭岩岩, 等. 中文博客多方面话题情感分析研究[J]. 中文信息学报, 2013(01):47-55.
- [30] 周胜臣, 瞿文婷, 石英子, 等. 中文微博情感分析研究综述[J]. 计算机应

- 用与软件, 2013(03):161-164.
- [31] 李泽魁, 赵妍妍, 秦兵, 等. 中文微博情感倾向性分析特征工程[J]. 山西大学学报(自然科学版), 2014(04):570-578.
- [32] 金永生, 王睿, 陈祥兵. 企业微博营销效果和粉丝数量的短期互动模型[J]. 管理科学, 2011(04):71-83.
- [33] Salton G, Wong A, Yang C. A vector space model for automatic indexing[J]. Communications of the ACM, 1975,18(11):613-620.
- [34] 徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008(01):116-122.
- [35] Li J, Sun M. Experimental study on sentiment classification of Chinese review using machine learning techniques, 2007[C]. IEEE, 2007.
- [36] 郭庆琳, 李艳梅, 唐琦. 基于VSM的文本相似度计算的研究[J]. 计算机应用研究, 2008(11):3256-3258.
- [37] 刁兴春, 谭明超, 曹建军. 一种融合多种编辑距离的字符串相似度计算方法[J]. 计算机应用研究, 2010(12):4523-4525.
- [38] Lang K. Newsweeder: Learning to filter netnews, 1995[C]. 1995.
- [39] 成卫青, 唐旋. 一种基于改进互信息和信息熵的文本特征选择方法[J]. 南京邮电大学学报(自然科学版), 2013(05):63-68.
- [40] 徐明, 高翔, 许志刚, 等. 基于改进卡方统计的微博特征提取方法[J]. 计算机工程与应用, 2014(19):113-117.
- [41] 覃世安, 李法运. 文本分类中TF-IDF方法的改进研究[J]. 现代图书情报技术, 2013(10):27-30.
- [42] 马雯雯, 邓一贵. 新的短文本特征权重计算方法[J]. 计算机应用, 2013(08):2280-2282.
- [43] Chang C, Lin C. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011,2(3):27.
- [44] Joachims T. Text categorization with support vector machines: Learning with many relevant features[M]. Springer, 1998.
- [45] Dietterich T G. Ensemble learning[J]. The handbook of brain theory and neural networks, 2002,2:110-125.
- [46] 蒋艳凰, 杨学军. 多层组合分类器研究[J]. 计算机工程与科学, 2004(06):67-69.
- [47] 杨利英, 覃征, 王向华. 多分类器融合实现机型识别[J]. 计算机工程与应

- 用, 2004(15):10-12.
- [48] 林煜明, 朱涛, 王晓玲, 等. 面向用户观点分析的多分类器集成和优化技术[J]. 计算机学报, 2013(08):1650-1658.
- [49] Freund Y. Boosting a weak learning algorithm by majority[J]. Information and computation, 1995,121(2):256-285.
- [50] Liu M. Fingerprint classification based on Adaboost learning from singularity features[J]. Pattern Recognition, 2010,43(3):1062-1070.
- [51] Xia J, Dalla Mura M, Chanussot J, et al. Random subspace ensembles for hyperspectral image classification with extended morphological attribute profiles[J]. Geoscience and Remote Sensing, IEEE Transactions on, 2015,53(9):4768-4786.
- [52] Ho T K. The random subspace method for constructing decision forests[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1998,20(8):832-844.

## 攻读学位期间发表的学术论文

汪淳等. 基于网络舆情倾向性分析的机器学习方法研究 [J]. 智能计算机与应用,2016 年（已收录，待发表）

## 哈尔滨工业大学学位论文原创性声明和使用权限

### 学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于机器学习的微博评论信息倾向性分析的研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：汪淳

日期：2016 年 6 月 30 日

### 学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

(1) 学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；(2) 学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；(3) 研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：汪淳

日期：2016 年 6 月 30 日

导师签名：李东

日期：2016 年 6 月 30 日

## 致 谢

毕业在即，研究生生涯即将告一段落，两年的学习过程不仅增强了自身的知识储备，同时也锻炼了自身的学习和动手能力。通过半年的时间，完成了自己的毕业设计，通过做毕设的过程也对自己的研究方向有了更深刻的知识理解。通过不断地调研、修改、理解、调整能够完成最后的论文撰写。在这期间，很多人帮助了我，我要对他们表达万分的感谢。

首先，要感谢导师李东老师。老师平时为人谦逊、对我们所遇到的问题都会细心帮助，无论是在学术上还是在平时的学习生活之中都给予了我很大的帮助。在我有疑问的时候，会非常认真并且耐心的回答我的问题，同时在论文的书写上也给予了我很大的帮助。他非常认真地学习和做研究的态度以及其对于学生的孜孜不倦的教诲都是将会影响我以后的人生，是我做事的标杆，在此，向其表达深深的感谢之情。

其次，要感谢实验室中张羽、张宏莉等老师，再有问题和困难的时候会无私的帮助。通过认真地做事风格和对于知识上的仔细教导，才能有我现在所学的知识，通过这两年的关心和传授知识以及指导方向，使我能够在和谐上进的氛围之中学习知识、提升自我。

再次，我要感谢同寝室和帮助过我的同学，他们在我有不懂的知识或者有困难之时，伸出援助之手，不求回报、互相互助。不仅让我认识了很多新的朋友，同时也在我写论文期间尽其所能帮助我、对不理解的知识对我详细说明，对他们表达感激之情。

最后，在此要感谢我的家人，是他们对于我给予了最大的支持，是他们对我的激励与支持让我有克服困难的动力。学习生涯已经结束了，对于马上要步入社会，我会秉承着老师们对我的教导和做人的道理为社会尽自己的一份力。祝所有帮助过我的人幸福快乐。