分类号_____ 学号 M200972519

学校代码 10487 密级_____

华中科技大学

硕士学位论文

基于情感词典的中文微博情感倾向分析研究

学位申请人：陈 晓 东

学 科 专 业：计算机应用技术

指 导 教 师：李玉华 副教授

答 辩 日 期：2012 年 1 月 12 日

A Thesis Submitted in Full Fulfillment of the Requirements
for the Degree of the Master of Engineering

Research on Sentiment Dictionary based Emotional Tendency Analysis of Chinese MicroBlog

Candidate : Chen Xiaodong

Major : Computer Application Technology

Supervisor : Prof. Li Yuhua

Huazhong University of Science & Technology

Wuhan 430074, P.R.China

January, 2012

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到，本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　月　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于　保密□ ，在_____年解密后适用本授权书。

不保密□。

（请在以上方框内打“ √ ”）

学位论文作者签名：　　指导教师签名：

日期：　年　月　日　　日期：　年　月　日

摘 要

近年来微博的出现，极大丰富了人们的生活。其简短写作，便捷发布，实时交互的特点深受大众欢迎。越来越多的用户乐于在微博平台上分享信息，交流观点和情感。通过对这些信息展开情感分析，可以实现微博营销、品牌宣传、客户关系管理、舆情监控等。当前微博情感分析研究大多是针对于英文微博的，而中文微博的情感分析研究还处于起步阶段。

情感分析主要是判别微博文本的情感倾向性，即属于正面、负面、中性。根据中文微博的自身特点，在传统文本情感分析的已有基础上，展开对微博的情感倾向分析。首先，对当前已有情感词汇资源加以总结和整理，并运用了扩展的情感倾向点互信息算法(Semantic Orientation Pointwise Mutual Information, SO-PMI)对新浪微博语料进行实验，自动获得领域情感词，构建了一个面向中文微博的情感词典。其次，基于中文微博表达多元化的特点，对微博文本进行了相应预处理，并采用微博消息文本中的情感词作为特征选择方法，对微博消息文本中存在的否定词、程度副词、感叹句、反问句、以及微博表情符号等进行相应分析处理。最后对整条微博消息作加权计算获得其情感倾向性，实现了一个面向中文微博的情感倾向分类系统。

实验数据选用数据堂的新浪微博语料，对来自科技、体育、娱乐三个领域的微博消息进行人工标注后，实验验证了该方法的可行性。实验结果显示：该方法获得的最高准确率为 74.2%，平均准确率为 70.5%，取得了较好的效果，对中文微博的情感倾向分析进行了初步探索。

关键词：微博，情感词典，情感倾向，权值计算，自然语言理解

Abstract

In recent years, microblog has greatly enriched people's life. Due to its brief writing, convenient publishing and real-time interacting, microblog becomes very popular. More and more people are actively sharing information with others and expressing their opinions and feelings on microblog. Analyzing emotion hidden in these information can benefit microblog marketing, branding, customer relationship management and monitoring public opinions. Currently, most of the emotional analysis is on English microblog, while Chinese microblog emotional analysis is still at the initial stage.

Emotional analysis is to identify the emotional tendencies of the microblog messages, that is to classify users' emotions into positive, negative and neutral. By learning from the traditional text emotional analysis, we analyze the emotional tendencies of microblog based on the characteristics of Chinese microblog. Firstly, summarize and organize the existing resources. Then, use the extended Semantic Orientation Pointwise Mutual Information (SO-PMI) to perform experiments on the Sina microblog and build an emotional dictionary for Chinese microblog. Based on the diversity of expression forms of Chinese microblog, we conduct some preprocessing on the microblog text. We use the emotional words in the microblog text as feature selection method, and process the negative words, adverbs of degree, exclamatory sentence, rhetorical question, and emotional signs in the microblog accordingly. And finally obtain the emotional tendencies by computing the weighted sum of various aspects of microblog messages. A system of emotional tendencies analysis for Chinese microblog is implemented in this paper.

The experimental data is selected from Sina microblog corpus in datatang. We manually annotate the microblog messages in domains of science and technology, sports, entertainment. Experimental results show that the method can achieve the accuracy up to 74.2%, and the average accuracy is 70.5%. The experiment validates the effectiveness of our method, by which we have performed a preliminary exploration of the emotional tendencies analysis of Chinese microblog in this paper.

Key words: MicroBlog, Sentiment dictionary, Emotional tendency, Weight calculating Natural language understanding

目 录

摘　要..... I

Abstract..... II

1 绪论

1.1 研究背景.....(1)

1.2 研究的目的与意义.....(2)

1.3 国内外研究现状.....(3)

1.4 论文的主要研究内容.....(7)

1.5 论文的组织结构.....(7)

2 相关介绍与理论概述

2.1 微博相关概述.....(9)

2.2 文本预处理技术.....(12)

2.3 特征选择.....(15)

2.4 本章小结.....(18)

3 微博情感词典的构建

3.1 情感词典相关介绍.....(19)

3.2 基础情感词典.....(21)

3.3 网络情感词典.....(22)

3.4 微博领域情感词典.....(23)

3.5 本章总结.....(28)

4 微博情感倾向分析

4.1 情感倾向的含义.....(29)

4.2 有情感词的微博情感分析.....(30)

4.3 无情感词的微博情感分析.....(33)

4.4 情感倾向加权计算.....(35)

4.5 本章小结.....(36)

5 实验结果与相关分析

5.1 实验数据介绍.....(38)

5.2 实验性能评估指标.....(39)

5.3 实验设计与结果分析.....(40)

5.4 本章小结.....(43)

6 总结与展望

6.1 全文总结.....(44)

6.2 进一步的研究方向.....(45)

致 谢.....(46)

参考文献.....(47)

1 绪论

1.1 研究背景

21 世纪是一个充满竞争的信息时代，信息将成为这个时代最为宝贵的社会资源。近年来，伴随着 Web2.0 与 3G 技术的高速发展，互联网以其丰富的内容，快捷的方式，呈现给我们一个前所未有的崭新世界。在过去，互联网更多的是基于 Web1.0 的应用，这种以数据为核心的网络，一般通过静态网页的方式呈现出来，多为网站工作人员的贡献，有限的交互活动极大限制了用户加入。而 Web2.0 是以用户为出发点的，人人都可以是贡献者，用户可以随时随地通过网络发布信息，传达自己的声音，表达自己的观点，享受便捷的互联网服务。由于这些信息是用户自己所提供，更加符合用户感受，故能充分调动他们的参与积极性。快速的信息交流将彻底改变互联网形式，使用户具有更多平等地获得信息的机会，而不再受到现有资源积累的限制，这给工作和生活带来了极大改变。

互联网技术的迅猛发展，使过去 Web1.0 时代的静态网页信息传达者变成了如今无数个活生生的用户，一大批基于互联网的社交网络平台步入了高速发展期。与此同时，论坛、博客、微博等得到了更为广阔的发展空间，这些都给人们的生活和社会运行方式带来了深深的影响。但在 Web2.0 时代，最具有影响力的产品无疑是微博，它实现了把信息发布与社会网络紧密结合在一起。自 2006 年问世至今，微博作为一个新兴的科技信息产物，目前在全球已成为一个能高度互动的信息转播平台。从国外的 Twitter、Plurk，到中国的饭否、新浪微博等，短时间里以惊人的速度发展并拥有了大量用户。据来自近日召开的第十一届中国网络媒体论坛公布的最新数据，我国微博用户数已突破三亿，这将是一个非常庞大的用户群。而在过去的 2010 年底，中国微博用户数仅仅是 6311 万。在国内，短短两年时间里，微博从互联网的新秀跃升为互联网的基础应用之一，以微变革的力量，打开了一个大时代之门。

1.2 研究的目的与意义

微博虽微，但传播信息的功能，却不容小视。尤其是手机发微博，每天 24 小时都可能有信息传播。据来自新浪微博官方数据显示，截至 2011 年 9 月底，新浪微博用户平均每天发布微博数达 8600 万条。更值得关注的是，微博正在成为继手机、邮箱、QQ 或 MSN 之后的又一个重要联系工具，“加微博”成为用户沟通新方式，甚至微博已出现在名片上、成为个人身份的重要组成部分。微博表现出强大的影响力，正吸引着越来越多的用户加入，其覆盖人群得到不断泛化。微博用户从最初的以 IT、媒体、城市白领、年轻人等为主向社会大众蔓延，如今还被各类机构如企业、政府、社会组织等所接受。无论是微博消息发布数目，还是使用频率，抑或应用创新，都呈现一条明显的上扬线，开始向大众化和主流化迈进。

微博用户数快速增长的同时，微博应用领域也在不断拓展，从最初的圈子应用发展为如今互联网的主流应用，信息的发布与获取、社交拓展、社会事务讨论、商业营销、社会公益、微博问政……微博开始全面渗透社会各领域。微博使用的低门槛、低成本，即时、简洁的信息发布方式，使越来越多的用户更热衷于在微博上发表自己的意见，表达自己的态度、观点及情感。比如对政治人物、娱乐人物、热点人物的个人喜好；对各类机构最新消息发布后的个人见解；对某个企业品牌的热衷或者厌恶；对某类产品的用后评价与建议；对社会突发事件的看法，自身立场；甚至更有用户通过微博直接表达自己的心情，做网络情感宣泄等等。

如此庞大的微博信息流揽括了众多话题，也许这些信息看似琐碎，而且非常不规则，可事实上蕴藏着巨大的潜在价值。微博平台上的各种互动，往往与用户的心理有关，用户一旦在微博中发言，便有了立场和倾向，这就可以对其做情感分析。通过情感分析的结果：名人可做自身形象维护；企业可做微博营销、客户关系管理以及品牌宣传；商家可通过用户产品体验后的评价做产品改进，从而提高市场占有率；政府机构可掌握突发事件后的社会群体心理，进行舆情监控；除此之外，还可对特定的高压力人群做情感分析，从而给他们提供有针对性的心理疏导等等。

因此，如何从微博消息中获取用户的情感倾向，并服务于生活是一个很有价值的工作。面对海量的微博数据，仅仅依靠人工浏览来获取用户情感将是一件十分繁琐与困难的事情。遗憾的是，目前极少有针对中文微博做情感倾向方面的相关研究。基于此，本文将对如何获取微博消息中的情感倾向进行分析，选取当前发展最为迅速的新浪微博，实现一个面向中文微博的情感倾向分类系统。

1.3 国内外研究现状

文本情感分析，又称为倾向性分析、意见挖掘、情感分类等。简言之，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。随着互联网的普及应用，特别是社交网络的迅猛发展，产生了大量由用户参与的、例如对人物、事件、产品等有价值的评论信息。这些评论信息表达了用户的各种感情色彩与情感倾向性，譬如喜、怒、哀、乐、批评、赞扬、支持、反对等。目前，文本情感分析已成了自然语言处理研究领域的热点之一。由于微博消息的主体内容仍然是文本，故针对微博做情感倾向分析的基础应来自于文本情感分析。下文首先介绍国内外在文本情感分析中所应用的主要方法，后介绍国内外学者在微博上的情感分析研究情况。

1.3.1 文本情感分析研究现状

通过目前收集到的国内外刊物及会议论文来看，关于文本情感分析方法的文献大致分为两类：

（1）使用情感词典及与其关联信息分析文本情感

国外的文本情感分析研究开始于 20 世纪 90 年代末，早期的 Riloff 和 Shepherd[1]对基于语料数据做了构建语义词典的相关研究。Hatzivassiloglou 和 McKeown[2]在大规模语料数据集上考虑了形容词的语义情感倾向性受连词的制约，尝试对英文的词语做情感倾向性判断。后来，越来越多的研究思路考虑了情感词或词组与特征词的依靠。Turney 等[3]使用点互信息(Pointwise Mutual Information ,PMI)方法扩展基准的褒贬词汇，然后把极性语义(ISA)算法用于分析文本的情感，在处理通用语料数据时达

到了 74%。Tsou 等[4]对词语的语义倾向性做计算, 同时将极性元素分布、密度与语义强度对新闻语义倾向进行统计, 以此衡量出大众对政治人物的评价。

近些年的研究中, Miao 等[5]在之前研究基础上为了能更好的解决特征的倾向分析，提出了一个基于产品评论的四元组抽取概念，从而实现了特征级别的分析。由于类型不同的句子表达情感的方式存在差异，Narayanan 等[6]对条件语句进行分析, 基于时态信息对相关句子进行类别标注, 结合各种特征表示信息, 提出基于分句、结果句和整个句子的分类方式, 收到了很好的效果。

中文方面，国内的徐琳宏、林鸿飞[7]从句子的词汇和结构作考虑，提取影响语句情感的 9 个语义特征，采用手工与自动获取相结合的方法，构建情感词汇本体库，对情感分析研究做了初步的尝试。李钝、曹付元等[8]从语言学的角度出发，采纳“情感倾向定义”权重优先的计算方式得到短语中词语语义倾向度，并分析词语的组合方式特点，提出中心词概念对词语的倾向性做计算，从而识别出短语的倾向性及其强度。该方法为更大粒度的文本情感分析打下了基础，具有一定的价值。

近年来，闻彬，何婷婷等[9]提出一种基于语义理解的文本情感分类方法，通过在情感词识别中引入情感义原，赋予概念情感语义，对概念的情感相似度重新定义，得到词语情感语义值。分析文本情感倾向性是否受语义层副词的出现规律的影响。该方法对有效地判定文本情感倾向性得到了一定的提高。赵妍妍等[15]提出了一种基于句法路径的情感评价单元自动识别方法, 该方法自动获取句法路径来描述评价对象及其评价词语之间的修饰关系，并通过计算句法路径编辑距离来改进情感评价单元抽取的系统性能，在对电子产品领域的应用，取得了较好的实验效果。王素格等[16]提出了基于赋权粗糙隶属度的文本情感分类方法，该方法利用特征倾向强度，定义赋权粗糙隶属度，应用在真实汽车评论语料上，取得了不错的分类性能。

总体来看，使用情感词典及与其关联信息来分析文本情感，其优点是应用在词语特征级，句子级，粒度细，分析精准。但受到自然语言处理技术及相关抽取技术的限制，该方法容易丢失数据集中隐藏着的重要模式，使得未来研究工作中还有很大的提高空间。

（2）使用机器学习方法分析文本情感

这类方法常用的机器学习模型有：中心向量分类法，朴素贝叶斯(Naive Bayes)，最大熵(Maximum Entropy)，K 最近邻分类和支持向量机(SVM)。在国外，Pang 等[10,11]用机器学习的方法对电影评论进行情感极性分类，分为正向情感和负向情感，他分别采用了朴素贝叶斯、最大熵、 支持向量机三种分类方法做实验，并将与他们手工分类结果做比较，发现支持向量机方法在这种机器学习方法中效果最好，分类精确度达到 80%。Whitelaw 等[12]通过提取电影评价信息中带形容词的词组，结合标准词袋特征表示，使用向量空间模型来表示文本，用支持向量机作分类，区分出带正面与负面评论信息，将准确率提高到了 90.2%。Moens[14]用机器学习方法分别对法语、荷兰语、英语做情感分析实验， 结果显示三种语言的情感分类准确率分别达到 68％、70％、83％。由此，可见机器学习方法在外文情感分析中展示出了一定的优势。

中文方面，文献[17,18,19]都是基于机器学习的方法分析文本情感，唐慧丰等[17]通过用名词、副词、形容词、动词做不同的文本表示特征，以信息增益、文档频率、CHI 统计量和互信息做不同的特征选择方法，分别以中心向量法、贝叶斯分类、K 最近邻和支持向量机做不同的文本分类方法做对比实验，其结果显示：在足够大的训练集与选择合适特征的情况下，采用 n-Gram 特征表示、信息增益特征选择和支持向量机分类方法，能取得较好的情感分类效果。夏火松等人[18] 通过 TF-IDF 权重计算方法，使用基于 RBF 核函数的支持向量机方法的分类器，对携程网客户评论做分析，研究了停用词表在情感分类中的影响。乔向杰等[20]基于模糊推理方法，得到学生对学习事件的期望度推理，并运用贝叶斯网络对其构建的模型做实验，验证其模型的合理性。

纵观近期的研究发展，基于机器学习方法的情感分析关键在于特征信息的有效提取。优点是知识获取客观，准确性较高，缺点是对训练语料依赖性比较高，训练周期相对较长。总体来看，使用机器学习方法并不比运用情感词典及与其关联信息方法具有明显的优势。另外，针对中英文上的研究方法还有很多不同点，随着语义信息的加入以及训练语料集的发展，机器学习方法应该有较好的发展空间。

1.3.2 微博情感分析研究现状

对于微博的情感分析，在英文方面主要是针对 Twitter 上用户发布信息作为语料展开分析，文献[21-25]都是选用 Twitter 信息做情感语料展开研究工作。Jiang 等[25]运用五折交叉验证的方法对 1939 条 Twitter 文本信息做训练和测试，从实验结果分析得知，扩充情感词典特征与主题相关特征，其分类结果有很大的提升。Dmitry Davidiv 等[24]把 Twitter 文本信息中的相关标签和表情符号作为标注，设计一个有监督的类似 K 近邻的分类器，实现了微博情感分类。

Barbosa 和 Feng[23]利用三个针对 Tweets 消息进行情感分析的网站（即 Twendz、Twitter Sentiment、TweetFeel）所提供的情感分析工具对 Tweets 进行情感分析得到初步情感标签，并制定一些规则来对 Tweets 进行预处理，去除不一致情感的 Tweets，对于每个用户仅保留一条 Tweets 消息及去掉包含排名靠前的情感词的 Tweet。经过预处理后带有情感标签的 Tweets 被用作训练数据。针对 Tweets 的情感分类问题，他们采用二部法：即采用抽象特征训练分类器进行主客观性分类和采用相同特征但修改词的情感极性的权重来进行情感极性分类两种方法。实验结果表明：在上述两大类特征中，按照作用大小排序依次为：负面情感极性最大，其次是正面情感极性、动词、表示正面情感的表情符号，最后为大写字母开头的词的个数。

Go 和 Bhayani[22]提出了一种距离监督学习方法对 Twitter 中的消息进行情感分类。给定一个检索词，消息自动被分为正面或负面情感。文中抽取 Twitter 中含有表情图标的消息作为训练集，利用朴素贝叶斯，最大熵以及支持向量机等分类算法进行了实验，达到 80%以上的精度。

中文微博在最近两年呈现出高速发展趋势，针对中文微博的情感研究目前还处于起步阶段。从查阅的文献来看，极少有这方面的文章。谢丽星[26]选择了 4 种特征共用，采用支持向量机的方法对新浪微博数据展开情感分析研究。实验结果显示，使用主题无关与主题相关的特征时所获得的最高准确率分别为 66.467%与 67.283%。

总体来看，由于中英文微博本身存在着较大差别，中文文本比英文文本要复杂得多，对中文文本的处理还涉及到分词技术，分词质量的高低往往很大程度上会影响分类效果。另外，中文微博消息所包含的文本内容是丰富多彩的，有着新的多特

征符号加入，这跟传统中文文本也存在着较大差异，这些都是研究工作的挑战。如何将这些年在文本情感分析领域的研究积累，运用到中文微博上，探索针对中文微博的情感倾向分析是很有价值的研究工作。

1.4 论文的主要研究内容

微博情感分析关键是如何判别微博消息的情感倾向性，首要条件是构建一个合适的情感词典，依靠情感词语、微博表情符号及语气句子等作为特征提取方法，对不同情况下微博消息做相应处理，最后进行加权计算，由最终的权重结果判别出微博消息的情感极性。主要研究内容有以下几点：

1. 微博情感词典的构建

研究情感词获取方法，尽可能构建一个足够大、覆盖面广的情感词典应用于微博消息文本的特征提取中。一方面对当前已有情感词汇资源进行总结和整理，另一方面采用扩展的情感倾向点互信息算法(Semantic Orientation Pointwise Mutual Information, SO-PMI)，从微博语料集中自动获取领域情感词，构建了微博情感词典。

2. 微博情感倾向的判别

基于中文微博表达多元化的特点，先对微博消息文本进行了相应的预处理，并采用微博消息文本中的情感词作为特征选择方法，分别从微博消息文本中包含情感词和不包含情感词两个方面展开分析，实现了一个面向中文微博的情感倾向分类系统。

1.5 论文的组织结构

第一章 绪论

本部分给出了研究背景以及目的意义，并分析了文本情感分析的国内外发展现状以及国内外学者在中英文微博情感分析领域的工作，提出了研究内容和方向，最后介绍了本文的章节结构。

第二章 相关介绍与理论概述

本部分首先对微博的定义与发展、相关符号特征、当前研究中存在的难点作了分析，其次对文本预处理技术做了介绍，最后详细讲述了后续章节会应用到的特征选择方法。

第三章 微博情感词典的构建

本部分先对情感相关术语做了介绍，其次对基础情感词典，网络情感词典的来源做了描述，最后介绍了如何通过扩展的 SO-PMI 算法构建领域情感词典。

第四章 微博情感倾向分析

本部分从两个方面对微博消息文本展开分析，在微博消息文本含有情感词的情况下，对否定词、程度副词、感叹句的处理分别做了阐述；在微博消息文本不包含情感词时，从微博表情符号、反问句方面做了相应处理，最后给出了情感倾向加权计算方法。

第五章 实验与分析

本部分首先介绍实验数据的来源、语料预处理及评估方法，然后展示实验结果并进行相关分析。

第六章 总结与展望

本部分主要对本文工作进行总结，并阐明了下一步工作的方向。

2 相关介绍与理论概述

2.1 微博相关概述

2.1.1 微博定义及发展

微博是微型博客的简称，英文名称为 MicroBlog。它是一个基于用户关系的信息传播、分享以及获取的平台，用户可以通过多种渠道（如 WEB，WAP 以及各种客户端组件，即时通讯等）即时更新信息，每次更新内容将限制在一定数目内（中文微博通常为 140 字左右），它具有便捷性、原创性、互动性、传播速度快及内容碎片化等特点。

相关资料显示，最早的也是最具影响力的微博是 Twitter。2006 年，由比兹·斯通(Biz Stone)、埃文·威廉姆斯 (Evan Williams)等人在美国旧金山成立。用户可以在 Twitter 上直接发布不超过 140 字的短文本信息，类似于 Facebook 的状态更新服务。由于它对所有人免费开放，许多网民认为 Twitter 上的信息交流更为便捷、自由与即时，这是其它社交网站无法比拟的，越来越多的人注册为 Twitter 用户。特别是名人的加入助推了 Twitter 的发展，例如在美国总统选举中，奥巴马和约翰·麦凯恩都通过 Twitter 来获得更多的支持者，从而扩大自己的影响力。另外，在一些重大事件上，Twitter 成为最快的新闻发布渠道，它能比 Google 等搜索引擎提供更快的结果。

在国内，Twitter 的成功吸引了很多人的目光。2007 年 5 月，王兴创办饭否网，标志着微博进入中国，随后各种类“Twitter”网站相继推出，如叽歪网、腾讯滔滔、嘀咕网、做啥网等。微博作为从国外引入的“舶来品”，在发展初期，由于其产品功能和服务不够成熟，关注度比较低，用户偏少，一些网站艰难维持，甚者停止运营。但是伴随互联网的高速发展，特别是移动互联网的兴起，微博获得了前所未有的发展空间。2009 年 8 月，新浪率先推出了“新浪微博”内测版，随后国内几大综合门户网站网易、搜狐、腾讯等相继推出。一时间微博呈现出井喷式发展，中国也真正进入了微博时代。

2.1.2 微博文本中的符号

微博是名副其实的短文本写作，通常一条微博消息的文本内容不超过 140 字。微博与博客在写作方式上存在差异，博客的写作是一个深思熟虑的长时间过程，更新比较慢，博主写作时比较注重文章的理论深度，遣词造句以及文章的完整性。而微博写作非常自由，表现更加随和，多为随时随地记录发生在自己身边的事件，或天气情况、或表达自己情绪及观点等。另外，微博消息跟传统文本也有着较大差异，为了最大化的表达信息，不仅用户使用观点鲜明的语言来表达，微博平台还提供了多元化的表达形式，以此丰富微博的文本内容。总体来看，微博文本中包含了大量信息，在这些多元化的表达之中比较明显的特征有：

1．网页链接：由于微博提供了分享视频、网页、图片等功能，通常在用户分享后的文本末尾会跟随出现一个以“http”开头的地址，例如：http://t.cn/SJq4yH，这类文本符号在本文的情感倾向分析中是没有用处的，应该在文本预处理阶段过滤掉。

2．标签符号：通常微博应用最广泛的标签符号有四类。下面将作分别介绍：

@：代表 at，意思是“对某人说”或者“需要引起某人的注意”。相当于超链接，点击时可跳转到被@的某人的微博。转发某条微博时，系统会自动在转发内容前加上“@微博用户昵称”。

#：两个#框起来的文字，可以理解为“话题”，也可以理解是为某条微博贴的一个标签，方便它与其他提到该关键字的内容相互关联起来，点击后将跳转到包含该关键字的微博搜索结果页面。因此，可以利用该功能抽取包含特定话题的所有微博消息。

//：一般是由微博系统自动添加的，出现在再一次转发已转发并带有评论的微博时，主要起分隔针对同一微博的多人多次评论的作用。某种程度上，通过观察转发次数的多少，可以反映出该条微博的讨论热烈程度。

V: 代表该用户是通过微博官方认证的，是特殊身份的象征，通常为政界、商界、娱乐界等名人，使用橘红色字体标注在微博用户昵称的右侧。

3．表情符号：微博消息中存在着大量表情符号，这些表情符号由微博平台所提供，用户可以自主选择。通过观察大量的微博数据表明，很多微博用户喜欢使用表情符号来表达自己当时的心情。因此，表情符号在本文情感倾向分析中起着重要作用。微博消息中的表情符号经抓取后转变为括号包含的文本内容。例如：表情符号  经抓取后转变为了[哈哈]。新浪微博平台提供的表情符号如图 2.1 所示：

图 2.1 新浪微博表情示例图

2.1.3 中文微博研究中的困难

1．中文微博文本的自身特点

通常中英文微博的文本内容都限制为 140 个字符，在英文微博中大约等于 25 个英语单词，而 140 个汉语字符大约等于 85 个英语单词，所以中文微博的信息量是英文微博的 3-4 倍。在微博消息中，英文微博用户更倾向发布简单信息和状态，而中文

微博中，用户可以发布更有深度的内容（评论、新闻、分析等），这导致中文微博包含的情感信息比英文微博更多更复杂。另外，中文微博文本与传统中文文本相比，微博用语多为非书面语言，口语化严重，大多不规范、语句结构杂乱，这在自然语言理解上给情感分析带来难度。

2．情感词典的构建

由于汉语表达比较灵活，同样的词语，短语存在多义性，甚至同一个词语既有褒义又有贬义，根据所处的语境不同所表达的感情倾向往往不相同，给感情色彩的判别带来了偏差。微博中大量网络用语的出现表现尤为明显，这对判断情感倾向同样造成了困难，构建一个适用于微博的情感词典是一个难点。

3．中文微博的数据获取

目前，还没有一个公共地、统一地、可供用作测试的微博语料集。研究人员还得依赖于微博平台官方提供的 API 接口获取数据，而当前大多数微博都只开放部分 API 接口，并对用户的访问权限进行了一定的限制。如某个时间段内，对用户应用请求次数进行了约束，所以难以获取比较完整的数据。而有些数据在对微博的情感倾向分析中是很有帮助的，这给深入研究带来了一定难度。

2.2 文本预处理技术

2.2.1 中文分词

在英文微博的情感分析研究中，其微博消息多为英文文本。英文是以词为单位，英文词语之间依靠空格隔开，单独的词可以独立表达一个意思。而中文是以字为基本书写单位，单个字往往不足以表达一个意思，通常认为词是表达语义的最小元素。在汉语中，一句话的意思通过一段连续的字符串来表达，字符串之间并没有明显的标志将其分开，计算机如何正确识别词语是非常重要的步骤。例如：一条英文微博消息 “I love this movie.”,其汉语意思为：“我喜欢这部电影。”计算机处理过程中，可以依靠空格识别出“movie”是一个词，但不能识别的“电”和“影”是一个词，只有将“电影”切分在一起才能表达正确意思。因此，这就须对中文字符串进行合理的切分，可

认为是中文分词。下面将对分词技术特点与分词系统作分别介绍：

1. 中文微博的情感倾向分析首要解决的就是对文本内容进行分词。如何实现准确、快速的分词处理是自然语言处理领域研究中的一个难点。当前主要的分词处理方法分为：基于字符串匹配的分词方法、基于统计的分词方法和基于理解的分词方法。这三类分词技术代表了当前的发展方向，有着各自的优缺点，总体来看：

基于字符串匹配的分词方法优点是：分词过程是跟词典作比较，不需要大量的语料库、规则库，其算法简单、复杂性小、对算法作一定的预处理后分词速度较快。缺点是：不能消除歧义、识别未登录词，对词典的依赖性比较大，若词典足够大，其效果会更加明显。

基于统计的分词方法优点是：由于是基于统计规律的，对未登录词的识别表现出了一定的优越性，不需要预设词典。缺点是：需要一个足够大的语料库来统计训练，其正确性很大程度上依赖训练语料库的质量好坏，算法较为复杂，计算量大，周期长，但是都较为常见，处理速度一般。

基于理解的分词方法优点是：由于能理解字符串含义，对未登录词具有很强的识别能力，能很好的解决歧义问题，不需要词典及大量语料库训练。缺点是：需要一个准确、完备的规则库，依赖性较强，效果好坏往往取决于规则库的完整性。算法比较复杂、实现技术难度较大，处理速度比较慢。

2．常用的中文分词系统

中文分词技术是对汉语文本进行处理的基础要求，长期是自然语言处理领域的研究热点，目前已取得了很多成果，出现了一大批实用的、可靠的中文分词系统。其代表有：基于 lucene 为应用主体开发的 IKAnalyzer 中文分词系统、庖丁中文分词系统， 纯 C 语言开发的简易中文分词系统 SCWS，中国科学院计算技术研究所推出的汉语词法分析系统 ICTCLAS，哈尔滨工业大学信息检索研究室研制的 IRLAS，另外国内的北大语言研究所、清华大学、北京师范大学等机构也推出了相应的分词系统。

林林总总的分词系统各有其特点，比如 IKAnalyzer 实现了以词典分词为基础的正反向全切分算法，更多的用于互联网的搜索和企业知识库检索领域；庖丁中文分

词系统致力于成为互联网首选的中文分词开源组件，它追求分词的高效率和用户良好体验;而简易中文分词系统 SCWS 目前仅用于 UNIX 族的操作系统;哈工大 IRLAS 主要采用 Bigram 语言模型，大大提高了对未登录词识别的性能。目前来看，表现最为抢眼的无疑是中国科学院研制 ICTCLAS，该分词系统综合性能十分突出，在国内外权威机构组织的多次公开评测中都取得优异成绩，已得到国内外大多数中文信息处理用户的支持。

本文研究选用中国科学院研制的分词系统 ICTCLAS 2011 JAVA 版，该分词系统支持多种环境下的应用开发。具备统一的语言计算理论框架，采用了层叠隐马尔可夫模型，将汉语词法分析的所有环节都放到了一个完整的理论框架中，不仅分词速度快，还准确率极高。最新版本的 ICTCLAS 2011 新增功能有：支持多线程调用、UTF-8、对繁体中文的识别处理，支持 Big5 编码、支持 Windows7 及大用户词典等，其性能也更加稳定。由于微博消息的表达多元化，其文本内容可能包含有中文繁体、简体、甚至有英文，各种微博符号，也及来自不同微博平台数据的不同编码方式等。另外，ICTCLAS 2011 还支持用户自定义词典，则可以将自定义的情感词典加入分词词典中，并设置相应的优先级，从而提高情感词的分词正确率。

2.2.2 去除停用词

停用词也被称为功能词，与其它词相比通常是没有实际含义的。由于本文是针对微博的情感倾向分析研究，故停用词一般是指在文本内容中出现频率极高或者极低的介词、代词、虚词、以及一些与情感无关的字符。由于微博表达的多元化，大多数的微博平台都支持图片、文本、视频、表情等。因而微博文本中不仅包含针对传统文本信息的停用词处理，还包含其它一些对情感无关符号处理。例如微博消息中常见的“@、V、#、http://”等。这些字符在微博文本中起辅助作用，但在情感分析研究中没有实际意义。若计算机对其处理不但是没有价值的工作，还会增加运算复杂度，通常文本的停用词处理中可采用基于词频的方法将其除去。文献[27]中王素格与魏英杰构造五种不同的停用词词表作为候选特征依据，对汽车语料作情感分类研究，考察对最终分类结果的影响，其结果表明无停用词表即全部作为候选特征与选

用除了动词、副词、形容词的停用词表对情感分类的结果比较好。但是微博消息是一个表达多元化的信息文本，跟汽车语料数据存在着一定的差异。另外，部分标点符号在情感分析中有着重要作用，故为了避免处理上的不当造成情感特征丢失，本文专门针对微博构建了一个停用词表。

2.3 特征选择

2.3.1 常用的特征选择算法

文本处理中，特征选择是一个非常重要的步骤，其目的在于从原始特征信息中，挑选出最具有代表性的、分类性能优异的特征进行分类，故合适的特征选择方法将很大程度上决定最终分类效果的好坏。通常情况下，文本分类领域都是选取词语作为特征，通过计算词语与类别之间的关系，度量各个词语对类别的贡献度大小，从而归属于某个类别。由于在文本数据处理中，往往词语的数量非常多，若把所有的词语都选为特征项，则特征空间的维度过大，不仅会增加运算量，还会影响分类的精度。因此须对文本内容作降维处理，合适的特征选择方法就尤为重要。

情感分类中的特征选择，一方面需要去除与情感无关、类别关联度较小的特征，排除不必要干扰。另一方面，特征选择方法要能获取与情感分类有关联的特征信息，才能提高情感倾向性判别的准确性。因此，必须针对微博消息选择合适的特征抽取方法，才能提高情感识别的分类效果。目前，常用的特征选择方法有：词频法、文档频次法、信息增益法、互信息法等。下面将依次介绍：

1．词频法

词频法（Word Frequency, WF）：词频是指一个词语在文本中出现的次数，一般由统计获得，通常特征选择的时候可将词频低于某个阀值的词语删除，从而减小特征空间的维数。

2．文档频次法

文档频次法(Document Frequency, DF)是指整个数据集中，有多少个文档包含了某个特征项，占数据集中总文档数目的比值，其计算公式如式(2-1)所示：

$$DF(t_i)=\frac{N_{t_i}}{N_{all}} \tag{2-1}$$

公式中，N_{t_i}为出现特征项t_i的文档数，N_{all}为整个数据集中的总文档数。该方法通过对每个特征项在数据集出现的频率进行统计，然后根据预先给定的特征向量维数或者设定的阀值，去除掉那些 DF 值小于某个阀值或大于某个阀值的特征项。其思想在于这两种状态代表两种极端情况，若 DF 值过小，表明包含某特征的文档数目过少，该特征项没有代表性。反过来，若 DF 值过大，这表明包含某特征项的文档数目过多，该特征项没有区分度。

3．信息增益

信息增益（Information Gain, IG）是指某个特征在文档中出现或不出现对判断文本隶属类别所能提供的信息量大小。信息增益借助了信息论中熵的概念，定义为信息熵的有效减少量，即不考虑任何特征时与考虑该特征时两文档的熵值之差。其计算公式如式(2-2)所示：

$$\begin{aligned} Gain(t_i) &= Entropy(S)-Entropy(S_i) \\ &=\left\{-\sum_{j=1}^{|c|}P(c_j)\times\log P(c_j)\right\}-\left\{P(t_i)\times\left[-\sum_{j=1}^{|c|}P(c_j|t_i)\times\log P(c_j|t_i)\right]\right. \\ &\left.+P(\bar{t}_i)\times\left[-\sum_{j=1}^{|c|}P(c_j|\bar{t}_i)\times\log P(c_j|\bar{t}_i)\right]\right\} \end{aligned} \tag{2-2}$$

公式中，$P(c_j)$表示c_j类文档在训练文档集中出现的概率，$P(t_i)$表示训练文档集中包含特征项t_i的文档频率，$P(c_j|t_i)$表示文档包含特征项t_i时属于c_j类的条件概率，$P(\bar{t}_i)$表示训练文档集中不包含特征项t_i的文档频率，$P(c_j|\bar{t}_i)$表示文档不包含特征项时t_i属于c_j类的条件概率。信息增益是一个统计量，用于度量特征对分类贡献的大小，其值越大，该特征就越重要，越有助于分类，故应选择信息增益值较大的候选特征。

4．互信息法

互信息（Mutual Information, MI）在统计语言模型中被广泛运用。它是用来度量两个随机变量之间的关联性。在分类系统中体现的是特征项与类别之间的依赖程度。若相互之间依赖程度越大，其特征项就越重要。

特征 t_i 与类别 c_j 之间的互信息公式如式(2-3)所示：

$$MI(t_i, c_j) = \log \frac{P(t_i, c_j)}{P(t_i)P(c_j)} = \log \frac{P(t_i|c_j)}{p(t_i)} \tag{2-3}$$

公式中，$P(t_i|c_j)$ 为特征 t_i 在类别 c_j 中出现的概率，$P(t_i)$ 为特征 t_i 出现的概率。当 $MI(t_i|c_j) = 0$ 时，表明特征 t_i 与类别 c_j 不相关，两者之间是相互独立的。如果（WF）的特征值越高，其两者时间的关联性越大。

在一个包含 m 类别的文档集中，特征项 t_i 对整个文档集的互信息值计算公式如式(2-4)所示：

$$MI(t_i, c_j) = \sum_{j=1}^{m} P(c_j) \log \frac{P(t|c_j)}{P(t)} \tag{2-4}$$

2.3.2 特征选择方法优缺点比较

词频法（WF）实现简单，应用中常把低频词语删除掉，以达到降维的目的。但存在的问题是有些低频词对分类也会有很重要作用，故不宜大面积的删除。

文档频次法（DF）最大优势是计算量小，速度快、在实际运用中有着很好的效果。它的时间复杂度和文本数量成线性关系，能够适用于任何语料，非常适合超大规模数据集的分类处理，因此是特征降维的常用方法。缺点是没有考虑特征项与类别之间的关联性，对于低频词的删除，虽然降低了特征空间的维数以及干扰噪声，可是有的低频词也会带有很大的信息量，特别是在情感分类中，往往有着非常重要作用。

信息增益法（IG）是一种基于熵的评估方法，度量标准就是某个特征项能为类别预测提供的信息量多少，信息量越多，其贡献越大,对整个分类系统就越重要。该方法不足之处在于，当类分布和特征值分布高度不平衡的情况下，会出现特征稀疏的

问题，此时信息增益值由不出现的特征决定，使信息增益的效果降低，表现为分类性能下降。

互信息法（MI）体现了特征项与类别的相关程度，特征项与类别的互信息越大，表示其贡献就越大。该方法只考虑了特征t_1在某一个类别c_1中出现的情况，而没有考虑在另一类别c_2中没有出现的情况。不足之处还在于其值受特征项边缘概率的影响。

2.3.3 微博的特征选择方法

传统的文本分类大多是把测试数据集中的文档归入预先设定好的文档类别中去，比如：“体育、艺术、军事、经济、政治、文学等”，这可通过文本的主题、属性及内容来划分。文本的情感分类则是特殊的文本分类，需要从语义级别上做考虑，根据文本内容所能体现出的观点、态度、立场等相关情感信息做倾向性分类。

微博消息的文本内容虽然限制为 140 个字符，但是包含的信息却是丰富多彩的，有文字、链接、表情、标签符号等，如何从短文本信息中获取情感信息是非常关键的。比如：从文本内容获取具有情感倾向的词语与短语、或从自然语言处理领域做基于语义的文本理解、抑或通过微博文本中的表情符号获取情感倾向性等。

2.4 本章小结

本章首先介绍微博的相关定义与发展、相关符号特征，并分析了当前微博情感研究中存在的困难。其次，对文本预处理中的分词技术进行比较，最后对后续章节中涉及到的特征选择方法做了详细介绍。

3 微博情感词典的构建

3.1 情感词典相关介绍

3.1.1 情感相关术语说明

情感的英文名称是 emotion，《心理学大辞典》中的定义是：“情感是人对客观事物是否满足自己的需要而产生的态度体验”。另外，普通心理学课程中定义：“情绪和情感都是人对客观事物所持的态度体验，只是情绪更倾向于个体基本需求欲望上的态度体验，而情感则更倾向于社会需求欲望上的态度体验”。在微博平台上，用户参与某个话题的讨论或就某个事件发表自己的观点、立场、见解等，这些信息大多具有明显的情感特征的，包含个人价值感与社会道德感。

情感词是指在文本中具有情感倾向性的词语，它可以是名词、动词、形容词、副词以及一些习惯性用语或短语等。一般情况下，文本内容表达的情感倾向主要通过情感词来体现，故它也是情感倾向性判断的重要依据之一。情感词通常具有明显的感情色彩，比如表达心情的愉快、高兴、压抑、沉闷，或是表达思想的积极、正直、堕落、腐化等，通常可将情感词分为正面情感词与负面情感词，也叫褒义词与贬义词。

情感倾向是指某个词语或短语与其所要表达内容主题意思的偏离度，它是对表达自身观点、态度、立场等语言的一种量度。它通常由两个标准来衡量：一个是指偏离的方向，即该情感词所要表达意思是正面的，还是负面的。另一个指偏离的程度它是指情感词在表达正面或负面状态下的强弱程度，不同的情感词其表达的情感强弱程度通常是不一样的。

情感词典是情感词的集合。从狭义上讲，指包含有感情色彩的词语集合，从广义上看可以是包含有感情色彩的短语或者句子。情感词典包含两个部分，一个褒义词典和一个贬义词典。

3.1.2 情感词典的重要性

文本的情感倾向大多通过情感词语来体现，情感词典能否覆盖全面在一定程度上影响着情感分类效果，故情感词典的构建是情感分类研究的基础。一条微博消息是 140 字内的短文本，如何识别其情感倾向，情感词典就显得尤为重要。

使用情感词典作为特征选择实质就是把微博文本中的情感词语作为特征项进行提取。该实现方法简单易行，把情感词典直接导入分词系统自定义词典，并设置相应分词词典的优先选择顺序，这样得到的分词结果更为准确。另外，针对微博的特征选择过程可以和分词过程一起进行。这样做的优点有：实现简单，复杂度低，处理速度快，可以大大减少针对文本的降维运算，同时直接提取情感词，能大大提高情感倾向判别的准确度。

3.1.3 微博情感词典的组成

目前，文本情感分析研究领域还没有一部完整且通用的情感词典。若构建一个面向中文微博的情感词典，一方面须对当前的已有相关资源进行总结与整理，另一方面需要构建一个基于微博的领域情感词典。在国外，哈佛大学于 1966 年整理开发 General Inquirer（GI）词典，该词典不但将每个词的义项列出，还对情感属性进行了标注。在英文情感研究中，它是很多学者常常选用的资源之一。而国内的情感分析研究起步不久，目前能应用的资源有董振东先生开发的知网[28](HowNet)；张伟、刘缙等人[29]编著的《学生褒贬义词典》；史继林、朱英贵[30]编著《褒义词词典》；杨玲、朱英贵[31]编著的《贬义词词典》；哈尔滨工业大学信息检索实验室整理的《同义词词林扩展版》[32]以及台湾大学整理的中文情感词典（NTUSD）等。

情感词的收集是一个不断积累的过程，采用手工标注需要阅读大量的文本，非常费时费力。目前，通用方法都是对大规模语料集进行统计分析，预先对有代表性的词语采用人工标注方法选为基准词，然后对候选词作语义相似度计算来获取新情感词，从而扩展情感词典的覆盖面。在本文中，将基于上述已有资源的前提下，还将构建一个面向中文微博的情感词典，主要组成结构图见图 3.1 所示：

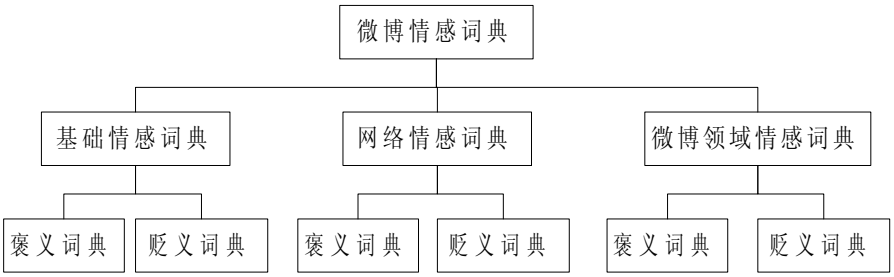

图 3.1 微博情感词典组成图

3.2 基础情感词典

3.2.1 知网

知网（HowNet）是中国科学院的董振东教授花了逾十年时间创建的一个知识系统。其官方定义为：是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

2007 年 10 月 22 日，知网在其官网发布了“情感分析用词语集（beta 版）”，共有 12 个文件。其中“中文情感分析用词语集”与“英文情感分析用词语集”各为 6 个，一共包含词语 17887 个。该词语集最大的特点在于作者已经根据词语情感倾向将其分为了 6 类，分别为“正面评价”词语、“负面评价”词语、“正面情感”词语、“负面情感”词语、“主张词语”以及“程度级别”词语。

中文微博作为互联网产品，其文本信息表达多元化，在微博消息中常常有消息发布、商品评价、话题讨论、也有情感宣泄等。故本文选用知网“情感分析用词语集”中两个词语集，即正负面情感词语与正负面评价词语。将其中一些不常见情感词语去掉后，具体数目见表 3.1 所示：

表 3.1 知网“情感分析用词语集”情感词数目表

词语集名称	词语（个数）
“正面情感”词语	755
“负面情感”词语	1218
“正面评价”词语	3360
“负面评价”词语	3028

3.2.2 台湾大学 NTUSD

NTUSD 的英文全称是 National Taiwan University Sentiment Dictionary，它是由台湾大学整理并发布的情感词典，分为繁体中文和简体中文两个版本。两个版本都包括有 2810 个正面情感词语和 8276 负面情感词语。本文选用中文简体版 NTUSD 作为基础情感词典的扩充，并将正面情感词语加入褒义词典，负面情感词语加入贬义词典。

3.2.3 其它情感词典

其它情感词典包括张伟、刘缙等[28]编著的《学生褒贬义词典》、史继林、朱英贵编著[29]《褒义词词典》以及杨玲、朱英贵编著的《贬义词词典》[30]。它们包括的情感词语数目见表 3.2 所示：

表 3.2 其它情感词汇数目表

词典名称	褒义词（个数）	贬义词（个数）
学生褒贬义词典	730	939
褒义词词典	5067	
贬义词词典		3495

3.3 网络情感词典

互联网的高速发展，带来了许多网络用语。这些网络上的非正式语言跟传统词语有着很大区别，它们往往具有强烈的感情色彩。有的是过去已经存在的词语，因为某个事件或某些热门话题而演变成了带有感情色彩的词语。比如：“宝马女、凤姐……”。有的则是过去不存在，新出现的网络新词，大多为谐音、错别字改成、字母缩写、也有象形字词。比如：“稀饭（喜欢）、JJWW(唧唧歪歪)、SP(支持，support)、弓虽……”。这些词语是已有情感词典中不存在的，但在情感倾向判别过程中有着重要作用。

微博作为一个新兴产物，跟深思熟虑后的博客写作不一样。微博更多的是口语化的表达，大量的网络用语充斥其间。这些网络用语大部分都是带有感情色彩的，因此针对微博构建网络情感词典的是非常重要的。网络情感词的收集是一个漫长的

过程，目前还没有现成的情感词典可用，因此通过社交网络、BBS、博客、评论、微博、收集并标注具有感情色彩的词语加入微博情感词典之中是必要补充。

3.4 微博领域情感词典

3.4.1 领域情感词典构建重要性

由于基础情感词典囊括的情感词是有限的，而中文的表达是变化万千的。特别是在微博中，口语化的表达常常带来很多新的情感词汇，通过已有的情感词典是无法辨别的，但是这些词汇在分析情感倾向时非常重要。比如：微博中经常出现一个名词“临时工”，这个词本身是不具有感情色彩的，可是发现用户在微博中所用到这个词时，往往是表达负面情感。这种例子还有很多，为了能识别这些新情感词，提高情感分类的准确性，可以构建一个面向微博的领域情感词典。

微博领域情感词典的构建首先是情感词的获取，如何判断一个词语是褒义词或贬义词，目前的研究有两种思路：一种是基于语义计算，例如：朱嫣岚等人[33]采用HowNet 语义相似度的方法，计算目标词语跟基准词之间的紧密程度，得以判定情感极性。而路斌等人[34]则通过对《同义词语林》的计算来判断词语的正负情感倾向。另外一种方法是基于统计分析，例如：Turney 等人[3]采用统计的方法，根据计算目标词与基准词之间的点互信息值，确定两个词之间的紧密程度，从而获取目标词的情感倾向。国内的王素格等人[35]通过词频概率估计找出区别类别能力强的词语，跟构建的情感词表相结合，提出了基于同义词的情感倾向判别方法，获取情感词。

情感分析研究中一个很重要的工作就是通过分析词语相似度来确定词语的情感倾向。目前，情感词极性判断的两种方法关键之处都在于对基准词的选取，基础词是否是基于领域的、感情色彩是否明显等因素将很大程度上影响情感倾向判断的效果。为了避免当前研究的不足，本文将采用微博语料集做情感基准词的选取，通过基于扩展的点间互信息（PMI）的方法计算候选词与基准词的相似度，从而判断候选词的情感倾向，构建一个微博领域情感词典作为本文情感词典的必要扩充，最大程度上提高微博情感倾向分类的正确性，并通过对比实验来做验证。

3.4.2 SO-PMI 算法

互信息是非常重要的信息度量，其相关理论已在 2.3 节作了介绍。在实际情况中，应用最为广泛的通常是点间互信息（PMI），主要用于计算词语间的语义相似度，基本思想是统计两个词语在文本中同时出现的概率，如果概率越大，其相关性就越紧密，关联度越高。

两个词语 *word*1 与 *word*2 的 PMI 值计算公式如式(3-1)所示为：

$$PMI(word1, word2) = \log_2\left(\frac{P(word1 \& word2)}{P(word1)P(word2)}\right) \tag{3-1}$$

公式 3-1 中 $P(word1 \& word2)$ 表示两个词语 *word*1 与 *word*2 共同出现的概率，$P(word1)$ 与 $P(word2)$ 分别表示两个词语单独出现的概率。若两个词语在数据集的某个小范围内共现概率越大，表明其关联度越大；反之，越小。 $P(word1 \& word2)$ 与 $P(word1)P(word2)$ 的比值是 *word*1 与 *word*2 两个词语的统计独立性度量。通过对其作 $\log_2$ 的计算，其值将转化为 3 种状态：

$$PMI(word1, word2)\begin{cases} >0; & \text{两个词语是相关的；值越大，相关性越强} \\ =0; & \text{两个词语是统计独立的，不相关也不互斥} \\ <0; & \text{两个词语是不相关的，互斥的} \end{cases}$$

在本文中，两个词语的共现概率以及单个词语出现的概率都可以通过对语料集的统计得到。其方法采用 2.3 节中介绍的文档频次法。

定义 N 表示语料集中的文档总数， $df(word1)$ 是语料集中词语 *word*1 的文档频次，即语料集中包含有词语 *word*1 的文档数。则可以得到词语 *word*1 在语料集中的概率计算公式如式(3-2)所示：

$$P(word1) = \frac{df(word1)}{N} \tag{3-2}$$

同理词语 *word*2 在语料集中的概率计算公式如式(3-3)所示：

$$P(word2) = \frac{df(word2)}{N} \tag{3-3}$$

同样定义 $df(word1 \& word2)$ 是两个词语 $word1$ 与 $word2$ 在语料集中的同时出现的文档频次，则其两词语共现的概率计算公式如式(3-4)所示：

$$P(word1 \& word2) = \frac{df(word1 \& word2)}{N} \tag{3-4}$$

此时，公式的 PMI 计算公式将转化为(3-5)式所示：

$$PMI(word1, word2) = \log_2 \frac{N \times df(word1 \& word2)}{df(word1) \times df(word2)} \tag{3-5}$$

公式 3-5 给出了两个词语的相关性计算，可以得到两个词语的相似度度量，因此可以将 PMI 方法引入计算词语的情感倾向（Semantic Orientation，简称 SO）中，从而达到捕获情感词的目地。

基于点间互信息 SO-PMI 算法的基本思想是：首先分别选用一组褒义词跟一组贬义词作为基准词，假设分别用 $Pwords$ 与 $Nwords$ 来表示这两组词语。这些情感词必须是倾向性非常明显，而且极具领域代表性的词语。若把一个词语 $word1$ 跟 $Pwords$ 的点间互信息减去 $word1$ 跟 $Nwords$ 的点间互信息会得到一个差值，就可以根据该差值判断词语 $word1$ 的情感倾向。其计算公式如式(3-6)所示：

$$\begin{aligned} SO-PMI(word1) = & \sum_{Pword \in Pworss} PMI(word1, Pword) \\ & - \sum_{Nword \in Nwords} PMI(word1, Nword) \end{aligned} \tag{3-6}$$

通常情况下，将 0 作为 SO-PMI 算法的阀值，由此，可以将得到三种状态：

$$SO-PMI(word1) \begin{cases} >0; & \text{为正面倾向，即褒义词} \\ =0; & \text{为中性倾向，即中性词} \\ <0; & \text{为负面倾向，即贬义词} \end{cases}$$

通过 SO-PMI 算法，可以获得候选词的情感倾向，进而得到具有感情色彩的词语，根据其点互信息值分别归入情感词典中的褒义词词典或贬义词典。

3.4.3 SO-PMI 算法扩展

目前，SO-PMI 算法在英文情感倾向分析中得到大量应用，但是在汉语中直接应用会带来一些缺陷。这是由于中英文的文本内容存在着较大差异。汉语的表达非常

灵活，丰富多彩。不同的人表达相同意思，其遣词造句的差异性是很大的。如果严格按照 PMI 算法统计，会带来数据稀疏问题。在极端情况下，由于一些词语出现的次数较少，为造成有的情感词的 SO-PMI 值为 0 的情况，而被过滤掉。因此本文将引入同义词对 SO-PMI 算法做扩展。

通过把基准词的同义词与候选特征词共现的次数，加入 SO-PMI 计算，这样基准词 *word* 在所有语料数据集中的概率 $P(word)$ 就扩展为基础词 *word* 及其同义词的在语料数据集中的概率 $P(words)$， $P(word1 \& words)$ 也因此成为 *word*1 与 *word* 及其同义词共现的概率。本文选用的扩展同义词来自哈工大整理的《同义词词林扩展版》，通过扩展后的 SO-PMI 算法能够有效的避免数据稀疏问题。

3.4.4 领域情感词典的生成

微博领域情感词典生成主要分为两个部分，第一个是基准词的选取，第二个是对候选词的情感倾向识别过程。下面将作具体介绍：

1．基准词语生成过程

当前研究中，有很多学者将词语放在搜索引擎中去检索(如 Google)，通过返回相关页面的 hits 数进行排序，选取 hits 数最高的词语作为 SO-PMI 算法中的基准词。由于微博文本跟传统文本存在一定差异，故本文直接选用微博语料数据，采用人工方法挑选基准词。

首先选择 50 万条来自不同领域的新浪微博语料，包括科技、体育、军事、环境、娱乐等多领域相关话题。首先将这些数据进行文本预处理，即分词、去噪、停用词去除，最后通过词频统计并排序。文献[36]论证了采用 40 对情感词的情况下，正确率能提高到 81.37%，本文将词频统计后的词语排序，手工挑选褒义词和贬义词各 20 个作为基准词，其余 20 对通过同义词扩展所得。基准词选取满足条件是：首先根据词频统计排序结果由高到低选取，另外这些词语必须具有明显感情倾向。为了尽量避免个人思维差异，这部分工作由 5 名志愿者参与完成，对标注存在有争议的词语采用投票制，以票数最多的为准。

2．情感词的自动识别与领域情感词典的生成

在微博领域情感词典构建中，SO-PMI 算法对微博语料中的领域情感词识别步骤如下：

输入：微博语料

输出：情感词

第一步：运用中文分词工具对微博语料进行预处理。

第二步：提取微博语料中的候选词，将得到的候选词与基础情感词典逐个扫描，判断是否为已有情感词，如果属于已有情感词则结束，否则进步下一步。

第三步：候选词分别与预先选定的基准词中的褒义词组和贬义词组做 SO-PMI 计算，获取情感倾向点互信息值，根据 3.4.2 节的给出的公式(3-6)判断其情感极性。

第四步：根据第三步所判断的情感极性把候选词归入相应的领域情感词典。否则，视为中性词，应舍弃。

第五步：结束。

微博领域情感词的获取流程图如图 3.2 所示：

根据本节介绍的 SO-PMI 算法，对 50 万条微博消息展开实验。把实验结果中一些情感极性不明显的词语以及与基础情感词典中重复的情感词去除后，最终得到正面情感词语 1482 个，负面情感词语 1650 个，组成领域情感词典。

总体来看，对于微博消息预处理后的候选词，可能是基础情感词典中已有情感词，也可能是不存在的情感词，通过 SO-PMI 算法获取其感情倾向，一方面是对基础情感词典的必要补充。另一方面，由于 SO-PMI 方法对低频词判别还存在一定的误差，这又可以通过基础情感词典与网络情感词典来弥补。

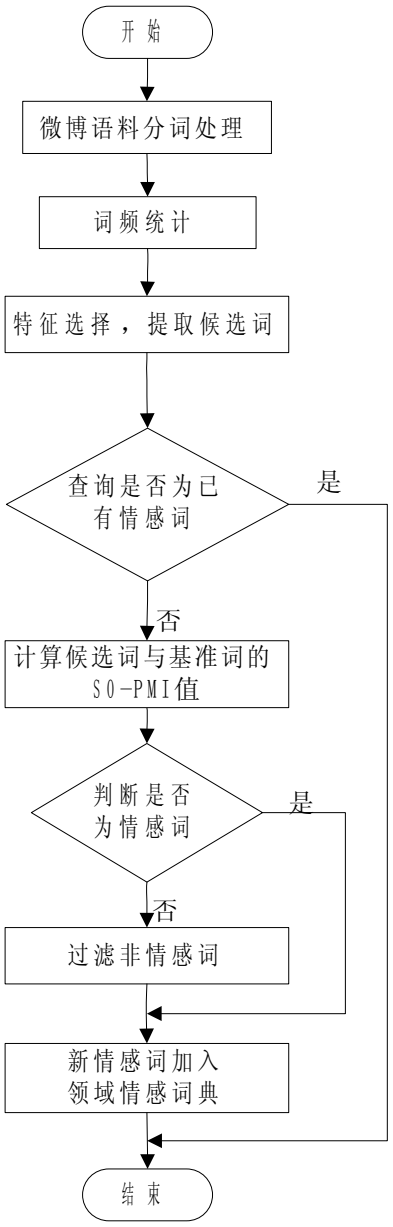

图 3.2 微博领域情感词获取流程图

3.5 本章总结

本章介绍了情感相关的术语概念，情感词典在情感倾向分析中的重要性，还对微博情感词典的组成进行说明。详细描述了怎么构建微博领域情感词典，先对 SO-PMI 算法做了详细介绍，在对 SO-PMI 算法做了相应扩展，并应用于微博语料集中得到了一个关于微博的领域情感词典，作为微博情感词典的重要组成部分。

4 微博情感倾向分析

4.1 情感倾向的含义

根据本文 3.3.1 节对情感倾向的定义，情感倾向可认为是主体对某一客体主观存在的内心喜恶，内在评价的一种倾向。它由两个方面来衡量：一个情感倾向方向，一个是情感倾向度。

情感倾向方向也称为情感极性。在微博中，可以理解为用户对某客体表达自身观点所持的态度是支持、反对、中立，即通常所指的正面情感、负面情感、中性情感。例如“赞美”与“表扬”同为褒义词，表达正面情感，而“龌龊”与“丑陋”就是贬义词，表达负面情感。微博消息的情感极性示例如表 4.1 所示：

表 4.1 不同情感极性的微博消息示例

名称	微博消息示例
正面情感	1、九寨沟山美水美人更美。 2、九寨沟的海别有景致，五彩斑斓，流连忘返。只有亲临其境，才能感受这种美妙！
负面情感	1、看天气预报，九寨沟那边好冷哦！这可就麻烦了，要带什么衣服呢？ 2、二三说九寨沟不适合我……是因为我看起来不能够修身养性么……
中性情感	1、今天要去九寨沟。 2、云南、鼓浪屿、西藏、海南、黄山、九寨沟。选一地儿十一走起来

情感倾向度是指主体对客体表达正面情感或负面情感时的强弱程度，不同的情感程度往往是通过不同的情感词或情感语气等来体现。例如：“敬爱”与“亲爱”都是表达正面情感，同为褒义词。但是“敬爱”远比“亲爱”在表达情感程度上要强烈。通常在情感倾向分析研究中，为了区分两者的程度差别，采取给每个情感词赋予不同的权值来体现。本文为了便于处理，将微博情感词典中的所有褒义词权值设置为 1，而贬

义词设置为-1。

由 2.1 节关于微博介绍所知，每条微博消息都是 140 字以内的短文本信息，多元化的表达不但包含文本，还有各种特征符号存在。针对微博消息做情感分析时，情感词是判断情感倾向的基础，通常先是获取微博消息中所包含的情感词，然后对这些情感词做权值计算，可得到该条微博消息的情感倾向。实际应用中，一条微博消息存在两种情况，即可能包含有情感词，也可能没有包含情感词。因此，本文将从两个方面展开分析。

4.2 有情感词的微博情感分析

通过分析微博消息，发现否定词的修饰会使情感词语的情感极性发生改变。比如："我今天身体不舒服……唉……"，该微博消息中"舒服"是褒义词，由于否定词"不"的修饰，使其情感极性发生了改变，转变成了负面情感。

另外，当程度副词修饰情感词、以及情感词所在位置是感叹句时，该情感词的情感倾向程度发生了变化。比如：一条针对"九寨沟"话题的微博消息："今天要坐 12 个小时的车到九寨沟，痛苦……极度疲惫中……"，"疲惫"是一个贬义词，前面一个程度副词"极度"的修饰使得"疲惫"原来的情感倾向程度发生了变化，这比没有修饰之前更加强烈。同样感叹句也有类似的情况，因此，为了准确表达微博用户的情感倾向，需做相应的权值调整。

4.2.1 情感词的获取

情感词的获取方法有很多种，本文选用基于情感词典的方法。通常的处理方法是先把情感词典中所有的情感词构建为一张情感词表，然后对微博消息进行中文分词处理，将处理后得到的候选特征词依次与预先构建好的情感词表逐个查找，若能找到，则是情感词，并读取情感极性及相应权值，否则，不是情感词，则进入下一个候选特征词，直至整条微博判断结束。实际应用中，情感词典中的情感词数量往往比较多，若每条微博都查询处理，所带来的时间开销会很大。为了避免这种情况，文献[37]中采用了索引技术的方法来解决，本文将作同样处理。

首先将情感词典中的所有情感词按照其字长排序，然后将排序后的情感词表按情感词的字长构建一张索引表。这样，在对微博消息做情感词查找时，先查询索引表，在查询情感词表，从而达到缩短时间开销的目的。

4.2.2 否定词的分析

否定词是副词的一种，它是表示否定意义的词语，在文本中具有独特的语法意义和影响。分析表明，被否定词修饰的情感词往往会改变情感极性。当一个否定词修饰一个正面情感词，则原本表达的正面情感就会转变为负面情感，反之则反。例如：“我今天是怎么了，感觉有点不开心！”，“开心”原本是正面情感词，否定词 “不”的修饰使该微博消息转变成了负面倾向的情感，针对这种情况应作相应处理。

由于汉语中存在多重否定现象，即当否定词出现奇数次时，表示否定意思；当否定词出现偶数次时，表示肯定意思。故在这里，引入一个“滑动窗口”的概念，窗口的大小以一个句子为标准，可根据文本中的标点符号作为判断依据，假设情感词所在位置的前面窗口范围内出现奇数个否定词时，须将原情感词的情感极性取反。实际算法处理中，可将原情感词的权值乘以-1 即可。本文单独构建了一个否定词典，并设置其权值为-1，常见的否定词如表 4.2 所示：

4.2　否定词表

个数	权值	否定词
19	-1	不、没、无、非、莫、弗、毋、勿、未、否、别、無、休 不曾、未必、没有、不要、难以、未曾

4.2.3 程度副词的分析

程度副词也是副词的一种，副词一般用于修饰或限制动词与形容词，表示范围、程度等。“程度”是指某个量处于相应层次序列中的某个层级上，是量的层级表现。由于微博消息多为即时性发布，具有文本内容少，信息含量广的特点。非书面化形式的写作，带来了大量的程度副词去限制或者修饰用户在观点、立场、态度等方面的表达。例如：“九寨天堂酒店，果然比凯宾斯基好很多。酒店里有小森林，环境超级好。号称“森林里的酒店，酒店里有森林”。睡了，晚安，明天一早去九寨沟。明

天回来上图……”。该条微博消息中使用两个程度副词“很”与“超级”来修饰褒义词“好”，表达用户对“九寨天堂酒店”的强烈好感，相比运用陈述语句来表达用户好感有着很大区别。

由此可知，程度副词的加入使用户在的情感倾向强弱程度上发生了变化，仍需做相应处理。同否定词的处理一样，同样引入一个“滑动窗口”概念，构建一个程度副词表，并根据程度副词的情感倾向强弱程度设置相应的权值。

本文中的程度副词来源于 3.3.1 节介绍的知网，选用“情感分析用词语集（beta 版）”中的“中文程度级别词语”共 219 个，蔺璜等人提出了把程度副词划分四个等级即极量、高量、中量和低量，并为每个程度副词定义了一个系数量，被程度副词修饰后的情感词其权值应做相应调整。本文将把“中文程度级别词语”已有的情感强弱标注改造为四个等级，同样赋予相应的权值。程度副词示例表如表 4.3 所示：

表 4.3 程度副词示例表

量级	权重	程度副词示例	个数
极量	2	极其、最、非常、极度、绝对、无可估量、至极无以伦比、卓绝、过度……	99
高量	1.75	不过、不少、出奇、大为、分外、格外、何等颇为、太、特别、着实……	42
中量	1.5	大不了、更、更加、还、还要、较、较为、愈加进一步、足足、越发……	37
低量	0.5	半点、不怎么、轻度、弱、丝毫、稍微、略略不丁点儿、略微……	41

4.2.4 感叹句的分析

感叹句是以抒发感情为主的句子，它所抒发的感情有赞美、愉悦、愤慨、叹息、惊讶、哀伤等，句末通常都用感叹号来标识。微博消息中的感叹句多为用户所表达情感的增强，其情感倾向程度发生了变化。通常感叹句是依附于它所在情感句的情感极性，可以是对正面情感或者负面情感的程度加深。本文将包含有情感词的感叹句权重设置为增强 2 倍关系。

感叹号“！”又叫感情号，主要用在感叹句的句末，表示强烈的感情。某种程度

上说，它是感叹句存在的标志。可以做简单处理，将“！”的权值设为 2，在对感叹句具体处理时，首先读入文本预处理后字符串 S_1 中的特征词 w，判断 w 是否为“！”，若不为“！”，则读取特征词 w 的下一个特征词 w_1；若为“！”，则向前查找最近的情感词，若情感词存在，将其权值乘以“！”权值。若没有情感词，则舍弃，继续后续处理。

4.3 无情感词的微博情感分析

由于中文微博表达的多元化，有的文本内容中不包含情感词，同样可能存在着情感倾向。大量微博语料显示，大部分用户都非常喜欢使用微博官方提供的表情符号来直观的表达情绪。例如：“这是一个小孩的画画，  ”。该微博消息中，没有一个情感词，用户通过一个表情符号“  ”来表达自己持赞美的态度，该条微博实际存在着情感倾向。另外，类似感叹句一样，微博中的疑问句往往也带有一定感情色彩，因此对微博中疑问句进行分析也能获取情感倾向。本文将分两个方面处理：微博表情符号处理和疑问句的处理。

4.3.1 表情符号的分析

在中文微博中，大多数微博平台都提供了丰富多彩的表情符号，供用户选择。由于表情符号暗含了感情色彩，一些用户常常使用合适的表情符号来直接表达心情。在微博消息中，表情符号的加入不但使文本信息充满了个性化色彩，而且还为分析用户情感倾向带来了帮助。基于此，可以构建一个基于微博表情符号的情感倾向判别表，同时标注情感极性与权值大小，以此为依据判别微博消息的情感倾向。

本文以新浪微博为例，该平台提供了大量的表情符号，人工选取微博消息中应用最为广泛的表情符号，它们主要来自于“默认”、“心情”两个类别，其中表示正面情感的 37 个，负面情感的 49 个。依据情感程度把正面和负面各分为两个等级。具体情感表情符号如表 4.4 所示：

表 4.4 新浪微博表情符号情感表

名称	个数	权值	内容
正面情感	12	2	[好得意],[哈哈],[太开心],[鼓掌],[ok],[good][耶],[赞],[给力],[威武],[爱你],[haha]
	25	1	[bobo 抛媚眼],[红包],[呵呵], [嘻嘻], [可爱],[亲亲],[抱抱],[钱],[酷],[心],[蜡烛],[蛋糕],[话筒],[礼物],[熊猫],[兔子],[奥特曼],[互粉],[手套],[吃饭],[思考],[顶],[握手],[右抱抱],[左抱抱]
负面情感	19	-2	[怒火], [闭嘴], [鄙视], [泪], [生病], [吐], [怒], [悲伤],[抓狂],[阴险],[怒骂], [伤心], [失望], [挖鼻屎], [愤怒], [最差]
	30	-1	[可怜],[吃惊], [害羞],[偷笑],[懒得理你],[右哼哼],[左哼哼],[嘘],[衰],[委屈],[打哈气],[疑问],[馋嘴],[汗],[困],[花心],[哼],[晕],[猪头],[不要],[弱], [挤眼],[睡觉],[书呆子],[黑线] [拜拜],[感冒],[拳头],[围观], [囧],[神马], [浮云]

4.3.2 疑问句的分析

疑问句通常分为两种，一种是有疑而问，也叫询问句，另外一种是无疑而问，通常叫反问句。第一种疑问句跟本文情感分析没有多大关系，主要考虑第二种。反问句的目的往往是加强语气，把原本的思想表达更加强烈、鲜明。它通常比陈述句表达更为有力，感情色彩也更加明显。

微博消息中的反问句大多带有强烈的感情色彩，而且通常以表达负面情感的居多，多为对某个事件、产品、某个人物、某个机构等等的质疑，其语气程度比较强烈。故在微博消息不包含情感词存在的情况下，可以通过判断疑问句是否为反问句而得到其情感倾向。反问句的存在可以通过反问标志来判断，本文从微博语料中挑

选出大量反问句并对反问标记词进行抽取，获得部分反问句标记词。类似感叹号，疑问句的出现有个明显的标志，就是在句尾有疑问号“?”的出现，这给反问句的判断提供了帮助。

对微博消息中反问句处理，跟感叹句的处理一样，反问句的存在与否以“？”作为标志。首先找到“?”，继而根据“?”向前查找是否存在反问标记词。如果找到，则证明是反问句，读取“？”权值。否则，舍弃，继续后续处理。本文中将“?”的权值设置为-2，即当存在反问句时，其权值直接为-2。反问句标记词示例表如表 4.5 所示：

表 4.5 反问句标记词示例表

权值	反问标记词
-2	为什么、凭什么、难道、何必、怎能、怎么能、怎么会 怎会、哪能、能不、能没、不都、不也、不就、谁叫 谁让、就算、这算、还算、就不、还不、莫非……

4.4 情感倾向加权计算

本章对微博消息的情感倾向性从两个部分展开分析，多个方面处理。但在实际情况中，有的微博消息既有情感词也有表情符号，也可能包含各种修饰词，甚至语气句子等多种情况出现。因此，为了简化运算，缩短时间开销，可把表情符号按其情感极性并入微博情感词典。经过文本预处理后的微博文本，首先识别不同极性类别的特征项，通过构建好的微博情感词表、否定词词表、程度副词词表以及反问句标记词表做相应处理，获取该条微博中每个特征项的权值，最后作求和运算，获得整条微博消息的情感倾向值，进而判别出情感倾向性。

由于每条微博消息都是不超过 140 字的短文本信息，所以以句子为单位，以标点符号为分割标志，将每条微博消息文本分割为 n 个句子 S_1、S_2、$S_3 \ldots S_n$,提取每个句子中的情感词 w_i，如果出现程度副词 w_a 修饰情感词 w_i 或者该句子是包含情感词的感叹句时，该情感词的情感倾向权重计算公式如式(4-1)所示：

$$O_{w_i} = M_{w_a} \times S_{w_i} \tag{4-1}$$

公式(4-1)中，M_{w_a}表示程度副词，或者感叹号“!”的权值，S_{w_i}是句子中情感词w_i的权值。

当出现否定词w_b修饰情感词w_i时，为了实现其情感极性取反，则情感词的情感倾向权值公式如式(4-2)所示：

$$O_{w_i} = M_{w_b} \times S_{w_i} \tag{4-2}$$

公式(4-2)中，M_{w_b}表示否定词的权值，S_{w_i}是句子中情感词w_i的权值。

句子S_i中可能包含k个情感词，即为W_1、W_2…、W_k，故该条句子的情感倾向度计算公式如式(4-3)所示：

$$O_{S_i} = \sum_{i=1}^{k} O_{w_i} \tag{4-3}$$

故含有n条句子的微博消息d_i最终情感倾向计算公式如式(4-4)所示：

$$O_{d_i} = \sum_{i=1}^{n} O_{S_i} \tag{4-4}$$

根据公式(4-4)得到的最终情感倾向值O_{d_i}，将会出现下列三种情况：

$$O_{d_i}\begin{cases} > 0; \text{ 为正面情感} \\ = 0; \text{ 为中性情感} \\ < 0; \text{ 为负面情感} \end{cases}$$

故根据最终的情感倾向值O_{d_i}所处的不同情况，可以识别出该条微博消息中的文本内容所体现出的情感倾向是属于正面、负面、或者中性的。

4.5 本章小结

本章先是对情感倾向的含义进行了说明，然后分两个方面对微博情消息展开分析：一方面是微博消息文本含有情感词的情况下，怎么提取情感词，以及对程度副

词、否定词、感叹句的具体处理方法；另一方面是微博消息文本中无情感词的情况下，从微博文本自身特征着手处理，对微博表情符号以及反问句方面做出相应判断。最后对微博情感加权计算进行了说明，实现了微博消息的情感倾向分类。

5 实验结果与相关分析

5.1 实验数据介绍

在国内，关于中文微博的情感分析研究刚刚起步，目前还没有现成的标准数据集可直接运用。本文实验数据选用于数据堂提供的新浪微博真实数据。数据堂[38]是国内专业的科研数据共享服务平台，致力于为国内外高等院校、科研机构、 研发企业及相关科研人员提供科研数据支持。目前，在国家科技部的大力支持下，数据堂对各个领域的数据进行收集、加工、整理，并通过统一的平台提供丰富的数据资源服务，为科研工作带来了极大的便利，其影响力越来越大。

本章为了避开实验结果过分依赖于特定领域，所选用测试数据分别来自于科技、体育、娱乐三个类别的多个话题微博语料，每个类别 1000 条，共计 3000 条。首先对这些数据进行人工标注。人工标注方法是：选用 5 名志愿者分别对 3000 条微博消息进行主观情感倾向判断，并作正面、负面、中性的标注。最后，由一名志愿者对前 5 个志愿者的标注结果进行统计，以每条微博消息的情感倾向支持票数最多的为准，这样可以尽量避免个人思维的差异。人工标注后的统计情况如表 5.1 所示：

表 5.1 微博消息数据表

类别	正面情感数目	负面情感数目	中性情感数目	总数目
科技	253	325	422	1000
体育	335	501	164	1000
娱乐	415	458	127	1000
总计	1003	1284	713	3000

表中的统计数据表明，各类别的样本数据集在情感倾向性上分布并不均衡，在科技领域话题类，近半数用户都处于中立立场，客观陈述发言比较多。体育和娱乐两个类别中，保持中立立场的人相对较少，绝大数用户都表达了明显情感倾向，而且以负面情感居多，体育类尤为明显，娱乐类较为平衡。一定程度上，对于不同领域的话题，用户往往表现出更为明显的情感倾向性。

5.2 实验性能评估指标

实验结果评估是一个非常重要的部分，其评价指标应考虑多方面因素的均衡性，满足客观、公正的评价。但是在实际中往往无法做到定量衡量。由于微博情感分析属于自然语言处理研究范畴，可以借用信息检索领域广泛的评估指标对本实验结果进行测评分析。其参照物是人工标注的结果（其条件是假设人工标注的结果都是正确的并排除在认知方面的个人差异），如果实验结果越接近人工标注的样本则证明越准确。文献[39~41]给出了两种主要的评估指标准确率与召回率，下面将对两种评估方法做介绍，并给出计算公式。

准确率 (Precision)：记为 p，它考察的是情感分类模型的准确性，所反应的是通过分类实验后，判断为该类的正确数目占判断属于该类别数目的比值，其数学公式如式(5-1)所示：

$$\text{准确率}=\frac{\text{判断正确的类别数目}}{\text{判断为该类别的数目}} \tag{5-1}$$

召回率(Recall)：记为 r，它考察的是情感分类模型的完备性，所反应的是通过分类实验后，判断为该类的正确数目占本应判断为该类别数目的比值，其数学公式如式(5-2)所示：

$$\text{召回率}=\frac{\text{判断正确的类别数目}}{\text{应判断为该类别的数目}} \tag{5-2}$$

设分类后判断为正面情感且正确的数目是 a_1,判断为负面情感且正确的数目是 a_2，判断为中性情感且正确的数目是 a_3；分类后判断为正面情感数目是 b_1，分类后判断为负面情感数目是 b_2，分类后判断为中性情感数目是 b_3；实验中的测试数据正面情感数目为 c_1，实验中的测试数据负面情感数目为 c_2，实验中的测试数据中性情感数目为 c_3。则所有测评指标的计算公式如表 5.2 所示：

表 5.2 测评指标公式表

名称	正面情感数据	负面情感数据	中性情感数据
准确率	$p_1=\frac{a_1}{b_1}\times100\%$	$p_2=\frac{a_2}{b_2}\times100\%$	$p_3=\frac{a_3}{b_3}\times100\%$
召回率	$r_1=\frac{a_1}{c_1}\times100\%$	$r_2=\frac{a_2}{c_2}\times100\%$	$r_3=\frac{a_3}{c_3}\times100\%$

相关文献表明，准确率与召回率看似没有必然的关系，但在实际中往往不能两全其美，当准确率高时，召回率低，当准确率低时，召回率高。这两个指标在一定程度上是相互制约的关系，故需要引入一个合适的度来衡量，寻求两者之间的一个平衡点。跟信息检索领域中说选用的评估指标一样，选用一个综合度量指标 $F-Measure$ 作为两者的调和平均数来衡量[42]。其公式如式(5-3)所示：

$$F-Measure=\frac{(\beta^2+1)\ \mathrm{p*r}}{\beta^2*p+r} \tag{5-3}$$

公式(5-3)中，$F-Measure$ 是一个综合度量指标，P 是准确率，r 是召回率，其中 β 是一个可调整系数。通过不同的取值去评估两者之间的关系，若 β 大于 1 时，p 对 $F-Measure$ 的影响比较大；若 β 小于 1，r 对 $F-Measure$ 的影响较大。由于 p 与 r 同样重要，故均衡两者的影响程度，选用 β 为 1，则 $F-Measure$ 的计算公式如式(5-4)所示：

$$F-Measure=\frac{2pr}{(p+r)} \tag{5-4}$$

此时的 $F-Measure$ 也称为 F1 值，表明准确率与召回率得到同等的看待。

5.3 实验设计与结果分析

本文的主要研究内容是构建了微博情感词典，并选用该词典的情感词作为特征选择，对情感词的相关修饰词、以及语气句子展开分析，引入了微博消息中的表情

符号作为补充判断条件。为了验证这些处理方法在微博情感倾向判别上的分类效果，分别设计了三个实验来验证。

1. 微博情感词典由三个部分组成，基础情感词典和网络情感词典是对已有情感词汇资源的总结与整理，而领域情感词典是通过扩展的 SO-PMI 算法获取情感词后所构建的。首先验证微博领域情感词典在情感倾向判别中的重要性，可分为：

实验一. 情感词典包括基础情感词典和网络情感词典，采用 4.2 节介绍的有情感词的处理方法情况下进行实验，其实验结果如表 5.3 所示：

表 5.3　实验一的实验结果

类别	评估指标	正向情感	负向情感	中性情感
科技	准确率 p	0.665	0.683	0.603
	召回率 r	0.644	0.591	0.678
	F 值	0.654	0.634	0.638
体育	准确率 p	0.682	0.722	0.485
	召回率 r	0.686	0.704	0.692
	F 值	0.684	0.712	0.570
娱乐	准确率 p	0.641	0.671	0.622
	召回率 r	0.646	0.600	0.843
	F 值	0.643	0.634	0.716

实验二. 情感词典包括基础情感词典、网络情感词典以及领域情感词典，采用 4.2 节介绍的有情感词的处理方法情况下进行实验，其实验结果如表 5.4 所示：

表 5.4　实验二的实验结果

类别	评估指标	正向情感	负向情感	中性情感
科技	准确率 p	0.701	0.721	0.659
	召回率 r	0.695	0.692	0.682
	F 值	0.698	0.706	0.670
体育	准确率 p	0.702	0.736	0.536
	召回率 r	0.720	0.728	0.685
	F 值	0.711	0.732	0.601
娱乐	准确率 p	0.690	0.723	0.631
	召回率 r	0.682	0.666	0.835
	F 值	0.686	0.693	0.719

2. 导入微博情感词典，即以词典中所有的情感词作为特征选择，并运用第 4 章介绍的全部处理方法做验证实验。

实验三. 在实验二的基础上，引入对微博表情符号与对反问句的判别方法，即增加了无情感词情况下的处理方法以检测对分类结果所带来的变化，其试验结果如表 5.5 所示：

表 5.5 实验三的实验结果

类别	评估指标	正向情感	负向情感	中性情感
科技	准确率 p	0.716	0.726	0.680
	召回率 r	0.727	0.732	0.668
	F 值	0.721	0.729	0.674
体育	准确率 p	0.710	0.742	0.636
	召回率 r	0.736	0.750	0.689
	F 值	0.723	0.746	0.661
娱乐	准确率 p	0.723	0.736	0.675
	召回率 r	0.718	0.688	0.850
	F 值	0.720	0.711	0.752

通过三个相关实验，由试验结果可得到如下结论：

1．通过领域情感词典加入，微博消息的情感倾向判别准确率得到了一定程度的提高。这说明判别效果的好坏与情感词典中所包含的相关领域情感词有较大关系，领域情感词覆盖面越广，越能从文本中提取到更多的情感特征，使更加充分的把握文本内容整体情感倾向性，从而获得准确的分类效果。另外，通过对比结果发现，在对负面情感的判断上，收到了相对较好的效果。这是由于在情感词典中，含有负面情感倾向的词语相比较多，这就更能提取较多的情感特征，故情感词的覆盖面是一个非常重要的因素。

2．在微博消息没有包含情感词的情况下，通过引入表情符号，以及对反问句子展开判断，其准确性获得了一定程度的提高，表明这两者在情感倾向分类中都有着重要的补充判别作用。表情符号表达情感较为直接，但是往往与讨论的话题有关，通常在一些娱乐，体育等热点话题中比较多，而反问句主要体现在一些比较有争议性的话题中居多。

3. 通过分类结果中判别为正面情感倾向的微博消息分析。部分判断错误的微博条目中，有较多的以反讽语气、暗喻、借代等表达形式出现，其文本内容本身为正面情感，但所表达的意思往往是负面情感。这要求在进一步的研究中应更多考虑对特殊句式的识别。

4.在中性情感的判别上准确度不够高。由于汉语表达灵活，部分情感词在不同的语境中所表达的情感倾向不一样，本文的微博情感倾向判断主要是基于情感词的提取，这就造成有的情感词只是客观陈述表达的情况下，也会被判别为存在情感倾向而导致分类失败。

5. 有的微博消息中，用户常常讲述两个不同的话题，表达两种不同的情感倾向，但这就造成其前后两部分存在着相反的情感极性。这是由于不同用户所表达的语言习惯不一样，有的喜欢在开始就表明态度，其情感重心在前，有的用户喜欢在后面作总结，其情感重心自然在后。这给计算机判别带来了困难，在处理中往往因为前后极性相反被判别为中性情感。为了避免这种情况发生，可以通过统计寻找规律，尽量考虑大多数用户的语言表达习惯，也提高分类的准确性。

6. 部分微博消息的文本内容由多个句子组成，且句式结构较为复杂，当所包含的情感词较多情况下，对其作情感倾向加权计算。倘若情感极性相反的词语相当，则容易出现正负情感词权重相抵消的情况，这就要求为不同情感词设置不同的权值来加以区别，从而获得更高的准确率。

5.4 本章小结

本章选用数据堂提供的新浪微博语料，运用本文设计的情感倾向判别方法对人工标注后的微博消息验证实验。实验结果显示：微博领域情感词典的加入，提高了微博消息的准确率。另外，增加对微博表情符号及反问句子的判别，一定程度上提升了情感倾向判断的召回率，收到了初步的效果。

6 总结与展望

6.1 全文总结

本文针对中文微博做情感倾向分析研究，先是概述了当前国内外学者在文本情感分析领域的研究现状以及中英文微博情感分析领域所取得的最新成果。然后对研究对象中文微博做了简要介绍，对微博研究的难点进行了分析说明，并对情感研究领域的基本理论及相关技术情况进行了阐述。情感分析是微博研究中的一个重要课题，本文将传统文本情感分析领域的研究成果进行分析总结，并从微博文本内容自身特点出发，借鉴已有成果，结合多个方面的技术进行扩展，力求找到一个好方法对微博消息作情感倾向判别分析。

主要研究工作可概括如下：

（1）构建了微博情感词典。一方面，对现有情感词汇资源进行了总结与整理，如：知网情感分析用词语集、NTUSD、学生褒贬义词典、褒义词词典、贬义词词典、网络情感词语等，另一方面采用扩展 SO-PMI 算法应用于新浪微博语料，自动获取情感词组成领域情感词典，扩大了情感词典的覆盖面，构建了一个面向微博的情感词典。

（2）从两个方面针对微博消息展开情感分析。一方面在传统文本情感分析方法的基础上，考虑了微博消息文本中包含情感词的情况下，周围修饰词所带来的影响，对文本中存在的否定词、程度副词、感叹句等做了相应的分析处理。另一方面，在微博消息文本不包含情感词的情况下，引入对微博表情符号及反问句的分析处理。

（3）实现了一个中文微博的情感倾向分类原型系统，并设计了相关验证实验以及对实验结果展开了分析。根据中文微博表达的多元化特点，先对微博文本进行了相应预处理，采用微博消息文本中的情感词作为特征选择，运用本文从两个方面提出的处理方法，对整条微博消息作加权计算获取情感极性，实现了一个面向中文微博的情感倾向分类系统。实验数据选自数据堂的新浪微博语料，通过对人工标注后

的微博消息进行验证实验。实验结果显示：该方法获得的最高准确率为 74.2%，其平均准确率为 70.5%，取得了一定的效果，对中文微博的情感倾向分析做了初步探索。

6.2 进一步的研究方向

伴随微博的快速发展，越来越多的学者投入微博研究中，而微博情感分析是一个基础性工作，也是一个非常有价值的工作。本文对微博消息的情感倾向判别方法进行初步探讨，取得了一定的效果，但是还远远不够。进一步的研究须改进的地方有很多，具体来说，有以下几个方向：

（1）构建覆盖面更广的情感词典

微博是一个开放、自由的信息分享平台，覆盖的信息包罗万象，讨论的话题涉及到各行各业，故情感词同样应涉及到不同知识领域；另外，互联网中网络新词的不断涌现，其中大部分词语往往都带有感情色彩，而这些词语多为已有情感词典未包含的，但在情感倾向判别中却有着重要地位。由此可知，在把情感词作为特征选择方法的情况下，只有当情感词典中所包含的情感词越多，覆盖面越广，才能更有助于情感倾向的判断。

（2）特殊句型的正确识别

汉语的表达方式灵活多变。某些时候，微博文本内容中无明显的情感特征标志，当作者运用一些特殊的修辞手法，比如比喻、借代、夸张、反复、引用等，来表达自身情感时，计算机往往不容易正确识别。这涉及到自然语言的语义理解，故可对大量微博语料统计分析，找出更多特殊句型存在的标志，有利于分析上下文的情感极性，做相应权值调整，以获得更高的准确率。

（3）引入微博中的更多特征

由于本文受微博语料所限，没有获得微博用户相互之间的各种关系数据，未来的研究中可以更多的考虑微博中关注、转发、广播、粉丝圈等社交网络关系，找出更多有利于情感判断的特征信息，进而提升实验效果。

由于时间仓促和笔者水平有限，论文中存在错误与纰漏之处在所难免，敬请各位老师和同学批评指正。

致　　谢

本论文得以顺利完成，首先要感谢我的导师李玉华副教授的悉心指导。她在研究方向，课题选择上为我指明方向，常常给予我许多积极地鼓励。李老师在学术上有着严谨的科学态度，对专业知识孜孜不倦追求，非常严格地要求自己。生活上，她平易近人、和蔼可亲，更是一位乐于助人的知心人。作为学生的良师益友，李老师有太多值得我学习的地方，再次衷心感谢李老师在我研究生学习期间的关怀与帮助。

其次，衷心感谢实验室的卢正鼎老师、李瑞轩老师，辜希武老师、文坤梅老师、李开老师两年多来在专业学习和日常工作中对我的关怀和帮助。感谢答辩秘书辜希武老师在整个答辩过程给予我的指导，使得我能够顺利的完成答辩。

同时，我还要感谢室友陈锐、包敏，王坚，师兄段东圣、同学温爱民等人在论文写作期间，给我提供的支持与帮助。我们之间的多次讨论，给我论文写作提供了思路源泉，让我得以顺利完成。另外，要谢谢计算机学院硕士 0908 班的所有同学以及实验室的各位朋友们，谢谢你们陪我走过了研究生阶段的点点滴滴。正因为有你们的相伴，才使我的学习生活增添了许多色彩，在这两年半的时间里不仅仅收获了知识，还收获了许许多多珍贵的友谊。

特别感谢我的家人，在漫长的求学生涯中，您们一直关心着我，鼓励着我，教育开导我，为我提供了最好的成长环境，给予了我最大的关爱和支持。衷心感谢我的父母，您们教育我如何做人做事的道理让我终身受益，这么多年的辛劳和付出，儿子永远铭记于心。

最后，感谢论文评审委员会的各位老师在百忙之中对我论文的悉心指正。

参考文献

[1] RiloffE, Shepherd J. A corpus-based approach for building semantic lexicons. in: Proceedings of the secong conference one mpirical methods innatural language processing. 1997. 117~124

[2] Hatzivassiglouv, McKeownK R. Predicting the semantic orientation of adjectives. in: Proceedings of the 35th annual meeting of the European Chapter of the ACL. Morristown, NJ,USA: ACL, 1997. 174~181

[3] Turney PD,Littman M l. Measuring Praise and Critism Inference of Semantic Orientation from As sociation. ACM Trans on Information Systems,2003, 21(4): 315~346

[4] BKY Tsou, RWM Yuen, OY Kwong, et al. Polarity classification of celebrity coverage in the Chinese press . in: Proceeding of the 2005 International Conference on Intelligence Analysis. Virginia, USA,2005.

[5] Qingliang Miao ,Qiudan Li, Ruwei Dai. AMAZING:A sentiment mining and retrieval system. Expert Systems with Applications: An International Journal, 2009,36(3): 7192~7198

[6] Ramanathan Narayanan, Bing Liu, Alok Choudhary. Sentiment Analysis of Conditional Sentences. in: Proceedings of the 2009 Conference on EMNLP. Morristown, USA: ACL, 2009. 180~189

[7] 徐琳宏, 林鸿飞. 基于语义特征和本体的语篇情感计算. 计算机研究与发展，2007,44(3):356~360

[8] 李钝, 曹付元, 曹元大等. 基于短语模式的文本情感分类研究. 计算机科学，2008,35(4):132~134

[9] 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. 计算机科学, 2010,37(6):261~264

[10] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. in: Proceedings of the Conference

on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia, USA: 2002. 79~86

[11] Bo Pang, Lillian Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. in: ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Morristown,USA: 2005.115~124

[12] Casey Whitelaw, Navendu Garg, Shlomo Argamon. Using appraisal groups fo rsentiment analysis. in: Proceedings of the 14th ACM international conference on Information and knowledge management. New York,USA: 2005. 625~631

[13] CHENG X. Automatic topic term detect ion and sentiment classificatio n for opinion mining:[Master Thesis]. Saarbrucken, Germany: The University of Saarland , 2007.

[14] Moens E. A machine learning approach to sentiment analysis in muhilingual Web texts. Inf Retrieval, 2009,12(8): 526~558

[15] 赵妍妍, 秦兵, 车万翔等. 基于句法路径的情感评价单元识别. 计算机研究与发展, 2011,22(5):887~898

[16] 王素格, 李德玉, 魏英杰. 基于赋权粗糙隶属度的文本情感分类方法. 计算机研究与发展, 2011,48(5):855~861

[17] 唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究. 中文信息学报, 2007,21(6):88~94

[18] 夏火松, 陶敏, 王一等. 停用词表对基于 SVM 的中文文本情感分类的影响. 情报学报, 2011,30(4):347~352

[19] 陈俊杰, 张大炜, 李海芳. 融入模糊理论的 SVM 在图像情感识别中的应用研究. 计算机科学, 2009, 36(8):288~290

[20] 乔向杰, 王志良, 王万森. 基于 OCC 模型的 E-learning 系统情感建模. 计算机科学, 2010,37(5):214~218

[21] Ravi Parikh, Matin Movassate. Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques. CS224N Final Report, 2009.1~18.

[22] Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision. Technical report. Stanford Digital Library Technologies Project, 2009

[23] Luciano Barbosa , Junlan Feng. Robust Sentiment Detection on Twitter from Biased and Noisy Data.In Coling 2010 (poster paper), 36~44

[24] Dmitry Davidiv, Oren Tsur, Ari Rappoport. Enhanced Sentiment Learning Using Twitter Hash-tags and Smileys. In Coling 2010 (poster paper), 241~249

[25] Long Jiang, MoYu, Ming Zhou, et al. Target-dependent Twitter Sentiment Classification. in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Oregon,USA: 2011. 151~160

[26] 谢丽星. 基于 SVM 的中文微博情感分析的研究:[硕士学位论文]。北京:清华大学, 2011.

[27] 王素格, 魏英杰. 停用词表对中文文本情感分类的影响. 情报学报, 2008,27(2): 175~179

[28] HowNet [R/OL]. HowNet's Home Page. http //www.keenage.com. 2011,12,10

[29] 张伟, 刘缙, 郭先珍. 学生褒贬义词典. 北京:中国大百科全书出版社,2004.

[30] 史继林, 朱英贵. 褒义词词典. 成都:四川辞书出版社,2005.

[31] 杨玲, 朱英贵. 贬义词词典. 成都:四川辞书出版社,2005.

[32] 同义词词林扩展版. http://www.ir-lab.org .2011,12,15

[33] 朱嫣岚, 闵锦, 周雅倩等. 基于 HowNet 的词汇语义倾向计算. 中文信息学报, 2006,20(1):14~20

[34] 路斌, 万小军, 杨建武. 基于同义词词林的词汇褒贬计算. 见:中国计算技术与语言问题研究—第七届中文信息处理国际会议论文集. 武汉:中国中文信息学会, 2007.

[35] 王素格, 李德玉, 魏英杰等. 基于同义词的词汇情感倾向判别方法. 中文信息学报, 2009,23(5):68~74

[36] 郭叶. 中文句子情感倾向分析:[硕士学位论文]。北京:北京邮电大学, 2010.

[37] 庞俊. 基于确定话题和情感极性的博客文本聚类研究:[硕士学位论文]。武汉:武汉理工大学, 2010.

[38] 数据堂. http://www.datatang.com/.2012,1,2

[39] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 2002, 34(1): 11~12. 32~33

[40] Mobasher B, Dai H, Luo T, et al. Discovery of aggregate usage profiles for web personalization. in: Proceedings of the WebKDD 2000 workshop at the ACM SIGKDD 2000. Boston,USA: 2000.142~151

[41] Megerian S, Koushanfar F, Qu G, et al. Exposure in wireless sensor networks: theory and practical solutions. Wireless Networks.,2002,12(5)443~454

[42] Guangxia Li, Steven C.H.Hoi, Kuiyu Chang, et al. Micro-blogging Sentiment Detection by Collaborative Online Learning. in: Proceedings of the 2010 IEEE International Conference on Data Mining. Sydney, Australia: 2010. 893~898

[43] Yang Shen, Shuchen Li, Ling Zheng. Emotion Mining Research on Micro-blog. in: Web Society (SWS) 2009. Lanzhou, China: 2009. 71~75

[44] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, et al. Sentiment analysis of Twitter data. Proceedings of the Workshop on Languages in Social Media, 2011, 30 ~38

[45] Mike Thelwall, Kevan Buckley, Georgios Paltoglou. Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 2011,62 (2): 406~418

[46] Dipankar Das, Sivaji Bandyopadhyay. Extracting emotion topics from blog sentences: use of voting from multi-engine supervised classifiers. in: Proceedings of the 2nd international workshop on Search and mining user-generated contents. New York,USA: 2010. 119~126

基于情感词典的中文微博情感倾向分析研究

作者：陈晓东
学位授予单位：华中科技大学