基于 SVM 的中文微博情感分析的研究

Sentiment Analysis of Chinese Micro Blog using SVM

（申请清华大学工学硕士学位论文）

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计算机科学与技术

研 究 生 ： 谢 丽 星

指 导 教 师 ： 孙 茂 松 教 授

二〇一一年四月

基于SVM的中文微博情感分析的研究

谢丽星

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，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；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘 要

微博自诞生以来，其应用价值迅速获得认可，并被用户所广泛接受。越来越多的用户注册了微博账户，通过微博来分享消息，表达观点和情感。微博影响的大幅增长，吸引了一大批学者对微博进行各种研究，而情感分析就是其中较为重要的课题。情感分析主要是进行情感极性的判定，即判断一条微博消息表达情感的正、负、中性。到目前为止这些研究主要是针对英文微博的，针对中文微博的研究工作尚处于起步阶段。

中文微博的用户不仅数量多，而且增长速度快，中文微博消息每天更是在大量更新，因此针对中文微博的情感分析变得尤为迫切和重要。本文通过从新浪提供的 API 抓取数据，对微博的链接、表情、情感词及上下文等主题无关的特征的有效性及多种分类方法进行了研究，最终选定 4 种特征共用及基于 SVM 的方法对微博消息进行了情感分类。实验结果表明，该方法使用主题无关特征时获得的最高准确率为 66.467%。此外，本文还就主题相关的特征对情感分类进行了初步尝试，获得的最高准确率为 67.283%。

关键词：新浪微博 情感分析 SVM

ABSTRACT

Since its birth, Micro blog's application value has quickly gained recognition and been widely accepted. More and more people register to micro blog services and share their opinions and emotions through them. As a result of the rapidly increasing number of micro blog updates, researches on micro blog have attracted more and more attention. Sentiment analysis, which is one of the most important research topics, aims at mining the polarity of micro blog updates, namely classify the emotion expressed by the updates into positive, negative or neutral. However, all of these studies focus only on English micro blog, and so far research work on Chinese micro blog is still at the initial stage.

Chinese micro blog users not only have large quantities, but also grow with a fast pace. Chinese micro blog updates updated a lot every day. Thus, sentiment analysis of Chinese micro blog seems particularly urgent and important. In this paper, we get the raw data through Sina's API and study the effectiveness of the target-independent features, including links, emoticons, sentiment words and context. In the mean while, we compare the performance of various classification methods. Finally we find out combining the 4 features and SVM based method to classify the micro blog updates gains best performance with accuracy rate of 66.467%. In addition, this paper makes a preliminary attempt to take target-dependent features into consideration when doing sentiment classification and the best accuracy is 67.283%.

Keywords：Sina Micro Blog, Sentiment Analysis, SVM

目 录

第1章 引言

1.1 研究背景

近年来，随着互联网的日益普及和互联网技术的蓬勃发展，互联网已经发生了翻天覆地的变化。十年前，那个近乎静态的互联网上的主角是网页和信息，而今天互联网上的主角却是一个个活生生的用户。在网络中生活、社交、展示自己、发出自己的声音，成为互联网用户日益增长的需求，并成为新一代网民对网络生存方式的共识。基于此，从天涯论坛到猫扑，从校内网到开心网，从饭否网到新浪微博，一个个信息发布和社交网络平台接踵而至，先后登场。但真正把信息发布与社交网络紧密结合发挥到极致的非微博莫属。

微博(Micro Blog)，顾名思义，是微型博客的简称。早在 2000 年，杰克•多尔西就有了一个关于实时发布信息、快速写作并与朋友互动的想法，即将短信与博客相结合，这便是微博最初的雏形。微博，是一个基于用户关系的信息共享、传播及获取平台，用户可以通过电脑网页、手机客户端等方式登录微博服务，发表 140 字以内的文字更新信息。同时，用户还可以在微博上关注好友、名人等动态，了解时事，回复、转发、评论他人的消息，拓展自己的社交圈等。微博目前可分为两大市场，一类是个人用户微博，另一类是企业客户的微博。

总的来说，微博具有即时通信和社会化媒体两大基本特点。

在即时通信方面，微博与传统博客相比，具有“短、灵、快”的特点。对于传统的博客书写，用户需要考虑标题、文章组织、语言修辞等内容，往往需要酝酿很长时间才能写出一篇完整的博客。而且博客反映出来的生活、人物性格的真实性也会由于这种酝酿失真。而在微博时代，用户只需三言两语就可以记录下生活的点点滴滴，包括自己在做什么、在想什么、对事物的看法与感悟等。显然，微博这样简单便捷的表述方式更能展现真实自我，而这种即时表述也更加迎合我们快节奏的生活，这也是微博迅速崛起风靡全球的原因之一。

在社会化媒体方面，因为微博的方便易用，每个用户都可以成为信息发布者，成为新媒体，经营自己的品牌，发出自己的声音。而因为微博上人与人之间的“关注”关系，微博上的信息传播更快更广，呈现“病毒式”传播的特点。微博满足了每个人展示自己、网络社交的基本需求，使媒体平民化、大众化，降低了内容门槛，而也因为人的参与，微博成为了一个最具个性化的媒体平台。

由于微博日益流行，越来越多的微博服务商向互联网用户提供微博服务，越来越多的互联网用户注册微博，通过微博发出自己的声音。微博正在以其独特的魅力，以不可想象的速度影响着人们的生活。以新浪微博[1]为例，它是中国最具影响力的微博，处于国内领先地位，它主要从以下三方面影响人们的生活：

（1）信息的大量传播：截止到 2010 年年底，已经有超过 60,000,000 用户注册了新浪微博，每天用户通过新浪微博发布的消息超过 25,000,000 条。

（2）更快的信息发现及传播：很多热门事件经常是第一时间从微博爆料出来并得到了广泛关注，例如前不久的“3Q 大战”（360 与腾讯 QQ）、“大小恋”（大 S 与汪小菲的恋情）等。

（3）与世界的紧密连接：在微博上用户可以关注名人、名企，了解名人的生活、想法，关注企业的动态；除此之外，用户还可以拓展自己的社交圈，结识更多的朋友和社团。

由于微博的巨大影响力，吸引了越来越多的用户，他们在微博上大量自由地发表自己的观点及情感，比如对某些名人的喜欢或憎恶、对某些电影的评论、对某些品牌的评价及建议、对某些时事的看法等。这些信息看似琐碎，其实具有潜在的商业价值，如帮助我们预测电影票房、改进影片及产品、了解用户体验等。除此以外，情感分析的技术还有助于文本摘要、问答系统等研究工作。

遗憾的是，目前还没有针对中文微博的情感分析方面的研究工作，而现在市面上的搜索引擎，包括新浪公司自己提供的搜索引擎都是基于关键词的，没有考虑任何观点及情感分析方面的因素。而依据上文所说，了解用户的情感及观点意义重大。因此，针对中文微博的情感分析成为较为迫切的需求，这能有助于我们更好的了解用户的情感及观点，从中发掘商业价值，增强用户体验。

基于此，我们提出本课题，主题相关的新浪微博的情感分析，准备以课题背景为契机，分析中文微博的情感，以实现一个具有实用价值的情感分析系统。

[1] Available at http://t.sina.com.cn

1.2 发展现状

谈及微博的发展，不得不提及 Twitter[2]。微博的核心概念最早由 Twitter 发明。2006 年 3 月，博客技术先驱 blogger.com 的创始人埃文•威廉姆斯 (Evan Williams) 创建了新兴公司 Obvious (不久改名为 Twitter) ，把人们的眼光引入了微博世界。Twitter 是一个社交网络及英文微博服务，用户可以使用即时通信、电邮、Twitter 网站或者客户端输入最多 140 个英文字符的更新。2007 年初，Twitter 凭借其方便快捷、创新的交互方式及社会化的巨大魅力，迅速进入人们的视野，并成为当年用户数增长最迅猛的社交网络之一。同年，它在德克萨斯州奥斯汀举办的南非西南会议赢得了部落格类的网站奖。此后，从 2008 年到 2009 年，两年时间中，Twitter 经历了爆炸性的增长，成为美国、欧洲乃至全球的家喻户晓的媒体明星。截止到 2010 年年底，Twitter 的用户总数接近 2 亿，市场估值约 37 亿，Twitter 开创了微博时代的先河，成为微博史上不折不扣的传奇。

Twitter 的巨大成功引起了中国国内投资者及创业者的高度关注。不久，一批创业团队开始打造中国人自己的微博网站了。

首当其冲的是著名的“饭否网”。“饭否”一词源自著名词人辛弃疾的词章：“廉颇老矣，尚能饭否？”。这两个字在中国人看来具有浓重的中国文化色彩，就像人们碰面时彼此之间常有的问候“吃了吗？”，有“寒暄”、“唠叨”的意思。校内网创始人王兴对 Twitter 所代表的微博服务有着敏锐而独到的认识，他于 2007 年 5 月建立了“饭否网”。“饭否网”成立后，用户数目增长迅速，到 2009 年上半年，用户数已经增长到数百万之众。与此同时，同样擅长技术活的其它创业团队也瞄准了微博这个行业，很快，叽歪、嘀咕、做啥等一批模仿 Twitter 的微博服务也开始正式上线。遗憾的是，到 2009 年年中，这些第一批中文微博服务网站由于某些原因停止运营。2010 年 11 月 25 日，饭否重新开放，老用户可以登录，新用户可以通过邀请码注册。和后来兴起的门户微博网站相比，人们把这批探路者称为独立微博网站。

2010 年国内微博迎来春天，微博似雨后春笋般崛起。新浪、腾讯、网易、搜狐四大门户纷纷开设微博服务。

新浪凭借自身新闻门户运营和博客服务等成功经验，率先在四大门户中迅速跻身微博领域。2009 年 8 月，新浪微博开始内测；2009 年 11 月 2 日，距离对

[2] Available at twitter.com

外公测仅 66 天，新浪微博用户数达到 100 万；2010 年 4 月 28 日，新浪微博用户数目超过 1000 万；2010 年 10 月，新浪微博用户数增长至 5000 万；截止到 2010 年年底，新浪微博的用户数超过了 6000 万。这一发展趋势超乎人们想象，远远超过了传统媒体的普及速度。

随着新浪微博的蓬勃发展，其他几家门户网站也纷纷加入微博领域。2010 年 4 月，腾讯微博正式开始内测。截止到 2010 年底，腾讯微博的用户数超过 1 亿，其发展速度大有跟新浪微博正面竞争的势头。

几乎在同一时间，网易微博、搜狐微博等各大门户网站纷纷提供微博服务，搜索引擎百度推出的 i 贴吧也整合了微博因素，与此同时，社交网站人人网、开心网等也提供了类似微博的服务。

一时间，微博平台及微博营销成为国内互联网巨头们竞相追逐的最新阵地。微博，以摧枯拉朽的姿态扫荡世界，成为全世界最流行的词汇，成为时下最热门最流行的服务。

微博的商业价值巨大，据文章《中金公司发研究报告：新浪微博竞争优势可持续》[21] 分析，微博主要有三大类盈利模式：

（1）直接盈利：这主要通过关联广告及实时搜索实现。关联广告可以是在搜索时嵌入广告、根据兴趣向用户推荐广告及植入“软广告”；实时搜索可以通过与微博运营商自己开发搜索引擎，或同领先的第三方搜索引擎提供商合作以谋取收益。

（2）交叉销售：这主要通过捆绑销售、与其它业务结合进行交叉销售、同第三方网站分享流量。捆绑销售是针对注册了微博服务的企业而言，微博服务商对这些品牌广告主进行收费；同其他业务结合进行交叉销售，比如将传统门户、在线视频、在线网游等通过微博进行推广；与第三方网站分享流量，例如目前水木社区就为用户提供了“特殊功能”，用户可以在发帖后选择将该帖分享到搜狐微博、腾讯微博、新浪微博。同样微博也可以为这些网站带来流量。

（3）开放 API：同 APP 开发商和内容提供商收入分成。开放 API，可以给微博服务商带来成千上万的第三方应用，这些应用可以跟微博服务商分成。

微博的未来不可估量，微博在给运营商带来无限商机的同时，微博的发展也会让人们的生活更加便捷轻松，会给人们带来更多的乐趣和享受！

1.3 研究内容及挑战

1.3.1 新浪微博简介

新浪微博是针对中文用户提供的微博服务。与 Twitter 相比，新浪微博在保留通信、社交功能的基础上，极大地强化了微博的媒体及传播功能。

图 1.1 所示，新浪微博首页会展示正在用微博的人，并及时更新大家正在说的微博内容。

图1.1 新浪微博的首页

图 1.2 是新浪微博的用户页面。在新浪用户页面上，左上方不仅能看见各种应用，包括微博、广场、微群、活动等，还能在右上方看见用户的个人信息，包括该用户的地域信息、关注了多少人、发表了多少微博。而在中间上方位置，用户可以输入自己想写的微博，在新浪微博上，对于字符的限制是 140 个中文字符。这点与 Twitter 不同的是，140 个英文字符往往是 20——30 个英文单词，一两句话。而 140 个中文字符内容要丰富的多。在正中间我们能看见每条微博

都既显示了文字信息，也显示了图片信息，同时还能看见这条微博的转发数目、评论数目。用户可以对这些微博进行转发和评论，及在该页面查看相关信息。

图1.2 新浪微博的用户页面

图 1.3 是一条具体的新浪微博，由用户李开复发布。图中 V 代表新浪的 VIP 认证，一般带 V 的用户都是名人。

//@蔡文胜：代表这条博文是转发用户蔡文胜的微博消息。

@京东刘强东 V：代表该消息是对用户京东刘强东 V 的回复。

同时，新浪微博还有一个话题标记，比如#微软对联#，这两个#之间代表的是话题也是用户写微博消息时自己写的，用户点击这个的时候会跳到跟这个话题相关的页面。

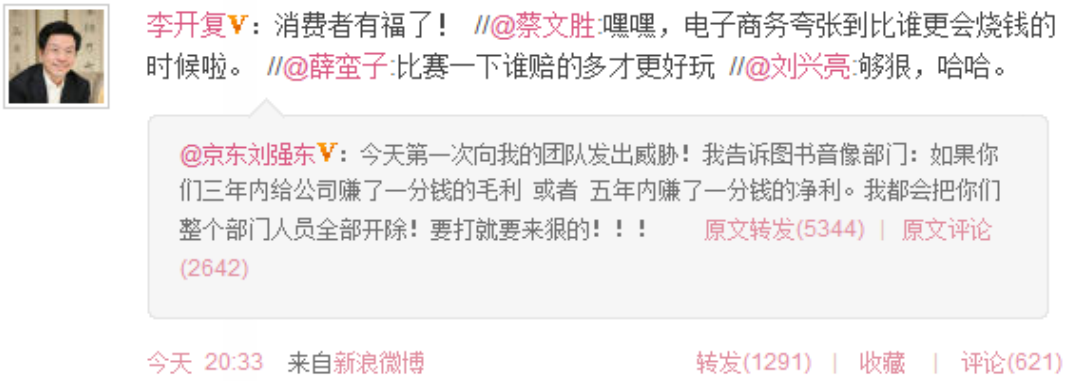

图1.3 新浪微博消息示例

1.3.2 研究内容

根据上文 1.1 节中提到的，本文主要是对新浪微博进行情感分析，分析用户的观点和情感，挖掘商业价值及了解用户意图。

通过观察，用户在微博平台发布的消息主要可分为两类：不指定主题的微博消息和指定主题的微博消息。

对于不指定主题的微博消息，在研究时主要存在两个问题：

（1）很难判断该消息的情感极性：表 1.1 所示。

表1.1 较难判断情感极性的中文微博

序号	微博消息
1	金鱼神马的都是浮云。
2	白茫茫的一片呀 这是苏州吗？
3	公司 D 人讲紧马利兄弟果只蘑菇
4	为啥不继续下雪呢

（2）微博消息本身不具有太大商业价值，对其它人影响不大：表 1.2 所示。

表1.2 商业价值不大的中文微博

序号	微博消息
1	我得瑟点什么好呢
2	Yummy
3	突然之間很想 show 下你…
4	小样，治不了你我还叫兽医!!!!

鉴于对不指定主题的微博消息研究的难度及有用程度的考虑，本文主要针对指定主题（主要涉及名人、电影、产品等）的微博消息进行情感分析，判断主题相关的微博消息的情感极性，系统输入输出如表 1.3 所示。

表1.3 系统的输入输出示例

输入：指定主题、与该主题相关的微博消息 示例： 主题：科比 微博消息：科比太酷了!!! [抓狂][爱你]
输出：每条微博消息对于指定主题的情感极性标签：正面情感（positive）、负面情感（negative）、中性（neutral） 示例：对于上面输入示例中的微博消息，该消息对于主题“科比”的情感极性为正面情感（positive）。

1.3.3 面临挑战

据我们所知，目前关于中文微博情感分析方面的研究工作尚处于起步阶段。在中文微博的情感分析方面主要存在以下几方面的挑战：

（1）中文情感词表较难构造：

由于中文词的多义性，导致了在构建情感词表时较难选择哪些词该加入正、负向情感词表，主要问题见表 1.4：

表1.4 构建中文词表面临的问题

序号	情况	示例
1	某种词性下具有情感	现实
2	代表不好事实的词	吵闹
3	本意无情感、寓意含情感	珍珠
4	某种音调时含情感	重用（zhong，四声）
5	本身无情感，加上否定词或肯定词带情感	意义、创造力
6	基于不同上下文情感不一样	好笑的
7	描述不同对象时情感不一样	高：高性能、债务高

（2）中文微博不同于英文微博及传统中文文本

中文微博允许 140 个中文字符，因此一条微博消息可以包含多个中文句子，表达的情感及意思较之英文微博要更为复杂。中文微博用语的不规范化较之传统中文文本要高很多，同时，中文微博字数较少，其句间关系没有中文传统文本紧密，经常出现省略主语等现象，加大了情感分析的难度；再次，中文微博表达的情感可能是发散的，针对多个主题的，这给情感分析造成了一些困难；

（3）中文分词、命名实体识别、句法分析工具的缺陷

由于中文分词、命名实体识别、句法分析工具的准确率较低，因此在做中文微博情感分析抽取特征时会影响特征的准确性，继而影响情感分析的正确性；

（4）新浪微博语料的关联信息无法获取：

本文通过新浪开放平台提供的 API[3]获取指定主题的微博消息，但是由于新浪未开放全部数据且抓取次数受限，因此本研究基本只用到微博正文本身，而用户之间的关联、微博消息之间的关联信息无法一一获取，而这些信息对于情感分析较为有用。

1.4 本文的章节结构

本文的章节结构如下：

第 1 章 引言

[3] Available at http://open.t.sina.com.cn/wiki/index.php/Trends/statuses

本部分分析了课题背景、微博发展现状，介绍了课题研究内容及面临的挑战，最后介绍了本文的章节结构。

第 2 章 情感分析综述

本部分主要对情感分析方面的研究工作进行介绍，包括不指定主题的情感分析、指定主题的情感分析、英文微博的情感分析、中文方面的情感分析工作。

第 3 章 中英文微博的特征比较

本部分对新浪微博和 Twitter 上的消息进行了相关特征的比较，包括标签、表情符号、链接等，阐述了中英文微博上的不同现象。

第 4 章 算法设计及实现

本部分提出了针对指定主题的新浪微博情感分析的具体算法，主要包括不指定主题的特征及指定主题的特征的研究。

第 5 章 实验结果及相关分析

本部分首先介绍了实验设置，然后展示了实验结果，并进行了一些分析。

第 6 章 结论与展望

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。

第2章 情感分析综述

情感分析，也被称为观点挖掘、观点分析、主客观分析等。情感分析的目标是从文本中挖掘用户表达的观点以及情感极性。挖掘用户观点意义重大，比如用户的观点能吸引潜在用户，帮助其它用户做决定[1]，同时用户的观点还包含了较有价值的反馈[2]。除此以外，情感分析的技术还有助于自然语言处理领域其它研究领域的发展，比如自动文本摘要[3]及问答系统[4]等。

近些年来，情感分析成为自然语言处理研究领域的热门话题。一系列研究工作就此展开。本章节主要从英文方面的情感分析和中文方面的情感分析两方面的研究工作进行介绍。

2.1 针对英文的情感分析

在这一节，本文主要从主题无关的英文情感分析、主题相关的英文情感分析、英文微博的情感分析三方面的研究工作来进行介绍。

2.1.1 主题无关的情感分析

目前大多数情感分析方面的研究工作都是主题无关的，即单纯判断一个文档或者一句话的情感极性，而不考虑这个文档是针对某个主题的情感。在这一研究领域，主要有三类方法：

（1）基于词典的方法[8]

该类方法主要是将情感词表与人工制定的规则相结合。这类方法通常面临无法解决未登录词 (Out Of Vocabulary) 的问题。

基于情感词典最简单的做法是，是用已有资源，如 WordNet 等构建情感词典，然后去看文本中包含正向情感词和负向情感词的个数，根据公式 2-1 判断文本的情感极性。

$$polarity = \begin{cases} Positive\ (if\ \ posCnt > negCnt) \\ Negative\ (if\ \ posCnt < negCnt) \\ Neutral\ (if\ \ posCnt = negCnt) \end{cases} \quad (2\text{-}1)$$

（2）有监督的机器学习方法[11][12]

这类方法主要采用的机器学习模型有朴素贝叶斯(Naive Bayes), 最大熵(Maximum Entropy) 和支持向量机 (SVM)。

以 Pang 等人[11]的工作为例：

Pang 等人的工作主要是使用机器学习的方法划分电影评论的情感极性，即正向情感和负向情感。Pang 等人首先对文本进行预处理，包括否定词提取、一元词提取、二元词提取、词性标注、提取位置信息等。然后将这些作为特征，再分别使用 Naive Bayes、Maximum Entropy 和 SVM 的方法来进行情感极性的分类。实验结果表明，SVM 在选取一元特征时的分类效果最好，准确率达到了 83%。

（3）无监督的方法[9][10]

这类方法主要通过指定基本的情感词，计算待挖掘观点的文本中的情感短语与基本情感词之间的分值来决定情感导向。

以 Turney 等人[10]的研究工作为例：

Turney 等人从 epinions.com 上选取了手机、银行、电影及旅游目的地相关的评论作为实验数据。他们采用了三步法进行情感分析：

①第一步：抽取情感短语

	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

图2.1 评论中的二词短语模版

先对每条评论进行词性标注；然后根据一些预先定义的模版（如图 2.1 所示）抽取每条评论中存在的两词短语，图中第一个模版表示该短语的第一个词是形容词 (JJ), 第二个词是名词 (NN 或 NNS)，限定条件是其后的第三个词可以是任意词。

②第二步：估计抽取出来的二词短语的语义导向

作者借助 PMI（公式 2-2）及预先定义的情感词来估计二词短语的语义导向 SO (Semantic orientation)(公式 2-3)。公式 1-2 中，“excellent”和“poor”是预先定义的两个情感此。作者使用著名搜索引擎 AltaVista 中的 near 操作返回的结果来计算二词短语及预定义的情感词之间的 PMI，来计算二词短语的语义导向 SO。

$$PMI(word_1, word_2) = \log_2(\frac{p(word_1 \wedge word_2)}{p(word_1)p(word_2)}) \tag{2-2}$$

$$SO(phrase) = PMI(phrase,"excellent") - PMI(phrase," poor") \tag{2-3}$$

③第三步：计算每条评论中所有二词短语的平均语义导向

在这一步，如果一条评论中所有二词短语的平均语义导向为正则推荐，否则不推荐。

Turney 等人的实验结果如表 2.1 所示。

表2.1 Turney等人的实验结果

领域相关的评论	准确率
手机	84%
银行	80%
电影	65.83%
旅游目的地	70.53%

2.1.2 主题相关的情感分析

在与主题相关的情感分析方面，目前主要有两类方法：

（1）基于规则的方法[13][14]

以 Nasuka-wa 和 Yi [13]的工作为例，他们先对文本进行预处理，包括词性标注（part-of-speech tagging）、依存关系句法分析（dependency parsing）。然后采取基于规则的方法，针对形容词、动词、名词等预定义一些规则。最后将这些规则应用到预处理后的文本上来进行情感分析。

（2）基于特征（或属性）的观点挖掘和文本摘要[15]

这类工作主要是针对用户对产品的评价进行情感分类。目的是发现用户喜欢或不喜欢产品的哪些属性。

Hu 和 Liu [15]针对用户对于在线产品的评论进行分析，提取产品属性的情感摘要，输出示例如图 2.2 所示。其中 Digital_camera_1 (数码相机 1) 是用户评论的产品，picture quality (相片质量)、size(尺寸)是数码相机的特征（即属性）。该任务主要是输出每个特征下对应的正负情感的评论语句。

*Digital_camera_*1:
Feature: **picture quality**
Positive: 253
<individual review sentences>
Negative: 6
<individual review sentences>
Feature: **size**
Positive: 134
<individual review sentences>
Negative: 10
<individual review sentences>
…

图2.2 基于特征的情感摘要的输出示例

他们的方法主要分为三步：

①第一步：识别出用户表达观点的产品属性；

他们先对评论进行词性标注，然后提取产品的两类属性：常用属性（即经常被用户评论的属性）和不常用属性（即不常被用户评论的属性）。他们使用关联挖掘 (Association mining) [16]的方法提取产品的常用属性；对于不常用属性，他们从被情感词修饰的名词及名词性短语中获取。

②第二步：针对已识别出的每个产品属性，找出用户评论中对其表达正向情感或负向情感的评论。

在判定每条评论的情感导向时，他们先识别出每条评论中所有语句的所有形容词，由于用户常用形容词表达观点，因此作者在文中将这些情感词称为观点词。然后他们采用自举（bootstrapping）的技术基于 WordNet 来进行情感词的正负极性判定。最后根据评论中所有语句中情感词的语义导向来决定该评论的情感导向。

③第三步：根据前两步的结果产生如图 2.2 所示的摘要。

2.1.3 英文微博的情感分析

这里所指的英文微博是研究者针对 Twitter 上的微博消息（即 Tweets）所做的研究工作。

```
RT @twUser: Obama  is the first U.S. president not to
have seen a new state added in his lifetime.
http://bit.ly/9K4n9p  #obama
```

图2.3 Tweets示例

图 2.2 是一条 Tweets 的示例。其中，RT 是 retweet 的简称，代表这条 tweet 是转发之前的消息；@twUser 代表该条 Tweet 是对 twUser 这个用户的回复；#obama 是一个 hashtag；http://bit.ly/9k4n9p 是一个网页链接。

在英文微博上的情感分析工作也可以分为两类：主题无关的情感分析与主题相关的情感分析。

（1）主题无关的情感分析 [17] [18][19][20]

在这方面较有代表性的是 Davidiv 等人 [17]的工作及 Barbosa 和 Feng [18]的工作。

Davidiv 等人使用 Tweets 中的标签（hashtag）和笑脸符号（smileys）作为标签，训练出了一个有监督的类似 K 近邻（KNN）的分类器，来对 Tweets 进行情感分类。

Barbosa 和 Feng 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，如去除不一致情感的 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，他们采用了二步法：

①第一步：采用抽象特征训练分类器进行主客观性分类

作者针对 Tweets 抽取了两类特征：元特征和 Tweets 相关的语法特征(表 2.2)。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：正面情感极性 > 链接 > 较强主观性 > 大写字母开头的词的个数 > 动词。

表2.2 元特征和Tweets相关的语法特征

元特征：
①词性信息：带有观点的消息通常包含形容词和感叹词；
②词本身的主观性：该词具有较弱或较强主观性；
③词本身的情感极性：该词表征正向情感还是负向情感或者中性；
④否定词：如 don’t, never 等，用在情感词前会扭转情感极性
Tweets 相关的语法特征：
①是否转发：即 retweet；
②标签：即 hashtag；
③是否回复：即 reply；
④Tweets 包含的 link；
⑤标点符号：感叹号和问号；
⑥情感符号：即表情符号
⑦Tweets 中以大写字母开头的词的个数

②第二步：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

在这一步，作者对上一步中分为主观性的 Tweets 进行情感极性分类。作者使用公式 2-4 和 2-5 来修正情感词的情感极性。

$$pol_{pos}(w) = count(w, pos) / count(w) \quad (2\text{-}4)$$

$$pol_{neg}(w) = 1 - pol_{pos}(w) \quad (2\text{-}5)$$

除此以外，仍然沿用与第一步中相同的特征来训练分类器。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 > 正面情感极性 > 动词 > 表示正面情感的表情符号 >大写字母开头的词的个数。

（2）主题相关的情感分析[5]

目前只有 Jiang 等人[5]针对 Tweets 进行主题相关的情感分析。他们使用 1939 条 Tweets 和五折交叉验证的方法来构建训练、测试集，方法大致可以分为三步：

①第一步：将 Tweets 进行主观性和中性分类

②第二步：承接第一步，将表征主观性的 Tweets 分为正面情感和负面情感

③第三步：考虑 Tweets 的转发关系，采用基于图的方法提升效果

在第一步和第二步中，作者主要采用了三类特征：内容特征、情感词典特征和主题相关的特征，如表 2.3 所示。

作者对三类特征的效果进行了测试，实验结果(表 2.4)表明：无论在主客观分类上还是在正、负向情感分类上，增加情感词典特征和主题相关特征都能提升结果。引入主题相关的特征后，系统取得的总体最高准确率是 66.0%。

表2.3 内容特征、情感词典特征和主题相关的特征

内容特征：

①词；

②标点符号；

③表情符号；

④标签。

情感词典特征：

①正面情感词的个数；

②负面情感词的个数。

主题相关的特征：

首先使用一些方法扩展主题词，如共指消解、计算其它词与主题词的 PMI 值，选取 PMI 较高的那些词、选取包含主题词的名词词组中的词；得到扩展词后，使用句法解析来提取 7 类预先定义特征：（w_i 表示词，T 表示主题词及扩展主题词）

① w_i 是一个及物动词，T 是宾语 –> w_i*_arg2*

② w_i 是一个及物动词，T 是主语 –> w_i*_arg1*

③ w_i 是一个不及物动词，T 是主语 –> w_i*_it_arg1*

④ w_i 是一个形容词或名词，T 是它的前置名词 –> w_i*_arg1*

⑤ w_i 是一个形容词或名词，它与 T 由一个连系动词连接 –> w_i*_cp_arg1*

⑥ w_i 是一个形容词或不及物动词且单独出现在一个句子中，T 出现在它的前一句 –> w_i*_arg*

⑦ w_i 是一个副词，它所修饰的动词的主语是 T –> *arg1_v_*w_i

表2.4 主客观分类、情感极性分类结果

特征	主客观分类准确率	正、负向情感分类准确率
内容特征	61.1%	78.8%
+情感词典特征	63.8%	84.2%
+主题相关的特征	68.2%	85.6%

在第三步中，考虑 Tweets 的转发关系主要思想是基于三点假设：

①转发的 Tweets 与原始消息表达的感情基本一致；

②由同一用户在较短时间内发布的关于同一主题的 Tweets 具有相同的情感极性；

③回复的 Tweets 或者被回复的 Tweets 也许具有不同的情感极性，但是它们所谈论的话题是一样的，因此会指向相同的情感主题词。

基于这三点发现，作者提出了图模型算法来优化情感分类结果。实验结果表明在不考虑内容特征单纯考虑主题相关特征时，系统分类准确率为 66%，考虑了基于图模型的优化算法后，性能有所提升，准确率达到了 68.3%。

2.2 针对中文的情感分析

中文情感分析的工作主要集中在 NTCIR[4]和 COAE[5]两个评测上。

（1）NTCIR (National Institute of Informatics)

NTCIR 是由日本情报信息研究所于 2002 年主办的针对亚洲语言的跨语言信息检索评测会议，主要关注日、韩、中等亚洲语种的相关信息处理。该评测主要包含六项任务：

①主客观判别（Opinionated judgment）

②相关性判别（Relevance judgment）

③观点持有对象抽取（Opinion holder detection）

④观点评价对象抽取（Opinion target detection）

⑤情感极性判别（Polarity judgment）

⑥问答系统（Questioning & Answering）

在最近一期的 NTCIR 评测中，即 NTCIR-8 中，针对情感极性判别这个任务，在繁体中文语料和简体中文语料的评测，最好结果见表 2.9。

[4] Available at http://research.nii.ac.jp/ntcir/
[5] Available at http://www.ir-china.org.cn/Information.html

表2.5 NTCIR-8中情感极性判别的最好结果

特征	主、客观情感判别		正、负向情感判别	
	繁体中文	简体中文	繁体中文	简体中文
精确率	56.37%	41.34%	76.48%	67.39%
召回率	85.71%	83.35%	53.03%	52.90%
F 值	68.01%	55.27%	62.63%	59.27%

(2) COAE (Chinese Opinion Analysis Evaluation)

COAE 由中国中文信息学会信息检索专业委员会从 2008 年开始举办。每届评测国内外大约有 20 多家科研单位参加。该评测主要包含五项任务：

①中文情感词的识别及分类 (Identification of emotion words)

②中文情感句的识别及分类 (Identification of emotion sentences)

③中文观点句子抽取 (Identification of opinion sentences)

④中文观点评价对象抽取 (Opinion target extraction and polarity analysis of opinion sentences)

⑤观点检索 (Opinion retrieval)

针对中文的情感分析从词表资源及方法上较英文仍有一定差距。中文情感分析的方法基本是在英文情感分析方法上引入中文自身的一些语法信息、搭配信息等，在中文评价对象抽取方面，目前使用较多的方法有两种：

（1）借助中文依存句法分析（Dependency Parsing）及语义角色标注（Semantic role labeling）识别评价对象[7]；

（2）基于 CRFs 的方法抽取评价对象[22][23]。

目前中文的情感分析主要存在以下问题：

①中文需要分词，分词错误会对情感分析产生影响，如“英雄难过美人关”中的“难过”；

②中文情感词典构建的难点：现在很多情感词典都仅为每个词条赋予一种情感极性，但是中文词较为复杂，在不同的语境下同样的词有不同的含义或情感色彩，如“黑马”，一般认为黑马是黑色的马，但在某些语境下比喻实力难测的竞争者或出人意料的优胜者，含褒义色彩，这使得如何构建一个较好的情感词典成为一个问题；

③中文存在一些难点目前尚无较好的解决方案，如“反讽”、“褒义贬用”和“贬义褒用”；

④中文情感分析主要使用句内特征进行分析，而句间特征，篇章特征尚未得到充分应用；

⑤受限于标注数据的规模大小，单纯使用机器学习的方法难以取得较好效果。

2.3 本章小结

分析目前的研究现状，我们可以知道：到目前为止，关于中文微博情感分析相关的研究工作尚处于起步阶段。然而中文微博与传统中文文本有较大差别，因此仍有很多地方值得研究，比如：

（1）由于中文微博允许输入 140 个中文字符，其信息量较英文微博要大很多，可以是几个英文句子。那么将中文微博消息看作一条消息或拆成若干分句是否会对情感分析造成影响？

（2）在中文微博上，主题无关的特征中哪些较为有用？

（3）主题相关的特征的引入是否有利于中文微博的情感分析？

第3章 中英文微博消息的区别[6]

微博是一个信息共享、信息开放的平台。在微博上，人们通常更为随意地去记录生活的点点滴滴，更简单地表达自己的观点。微博上的消息文本与传统文本也有较大的区别。例如在微博上，很多消息会包含网页链接，人们会更多地取自己书写表情符号或者使用微博平台提供的表情符号，也会使用标签来对自己发布的消息进行标记等等。因此，本章通过对新浪微博和 Tweets 上的消息文本进行一些统计来阐述中英文微博消息的不同。同时，本章涉及比较的方面也为后续的算法提供了简单的分析参考和估计。

本章使用的新浪微博的数据为第 5 章中涉及的 6000 条消息，使用的 Tweets 为 Jiang 等人[5]所用的 1939 条英文 Tweets。

微博上主要涉及四类比较明显的特征：

①网页链接（Link）：用户通常在分享新闻、图片或视频时常常会在消息文本末端附上网页链接，通常以 http 开头，如：http://t.cn/hr7fY9 。一条消息中可能包含一个或多个网页链接。

②表情符号（Emoticon）：英文微博上的表情符号通常是用户自己输入，如“:)”；在新浪微博上，微博平台提供了一些表情符号，用户可以自主选择，如  ，表情符号在抓取下来的文本中的表现形式为被中括号包含的文本，示例的这个表情对应的文本为“[哈哈]” 。一条消息中可能包含一个或多个表情符号。

③标签（Hashtag）：在 Twitter 上，用户的标签以#开头，标签可以是一个单词，也可以使多个单词连在一起，如“#ladygaga”、“#Ilikeladygaga”。在新浪微博上，用户的标签以“#”开头，以“#结尾”，如“#将爱情进行到底#”。一条消息中可能包含一个或多个标签。

本章会首先对中英文微博上这三类特征进行统计分析，随后会对新浪微博的单句数目进行统计分析，最后会对中英文微博上的情感极性进行统计分析。

[6] 此工作是作者在微软亚洲研究院实习期间完成

3.1 链接统计

（1） 中英文微博上包含链接的消息数目及比例见表 3.1。

表3.1 包含链接的微博消息的数目及比例

	包含链接的微博消息数	总数	比例
新浪微博	545	6000	9.08%
Tweets	388	1939	20.01%

分析：从这里可以看出，新浪微博上包含链接的消息比例为 9.06%，Tweets 上比例为 20.01%。这说明在英文微博上用户使用链接更多。

（2） 除了总体比例统计，本文还对包含链接的微博消息中不同链接数目的消息的数目及比例进行了进一步统计，见表 3.2。

表3.2 包含不同链接数目的微博消息的比例

	新浪微博		Tweets	
包含的链接数	数目	比例	数目	比例
1	536	98.35%	380	97.94%
2	8	1.47%	8	2.06%
3	1	0.18%	0	0

分析：从这里可以看出，中英文微博上用户使用不同数目的链接的比例大致一样。同时，中英文微博上包含链接的消息中大约有 98%的微博消息是使用 1 个链接；使用 2——3 个链接的微博消息非常少。

3.2 标签统计

（1） 中英文微博上包含标签的消息数目及比例见表 3.3。

表3.3 包含标签的微博消息的数目及比例

	包含标签的微博消息数	总数	比例
新浪微博	226	6000	3.77%
Tweets	471	1939	24.29%

分析：从这里可以看出，新浪微博上包含标签的消息比例非常低，仅为3.77%。而 Twitter 上的比例为 24.29%，约为中文的 6 倍多。这也说明标签在英文用户中更为流行。

（2） 除了总体比例统计，本文还对包含标签的微博消息中不同标签数目的消息的数目及比例进行了进一步统计，见图 3.1。

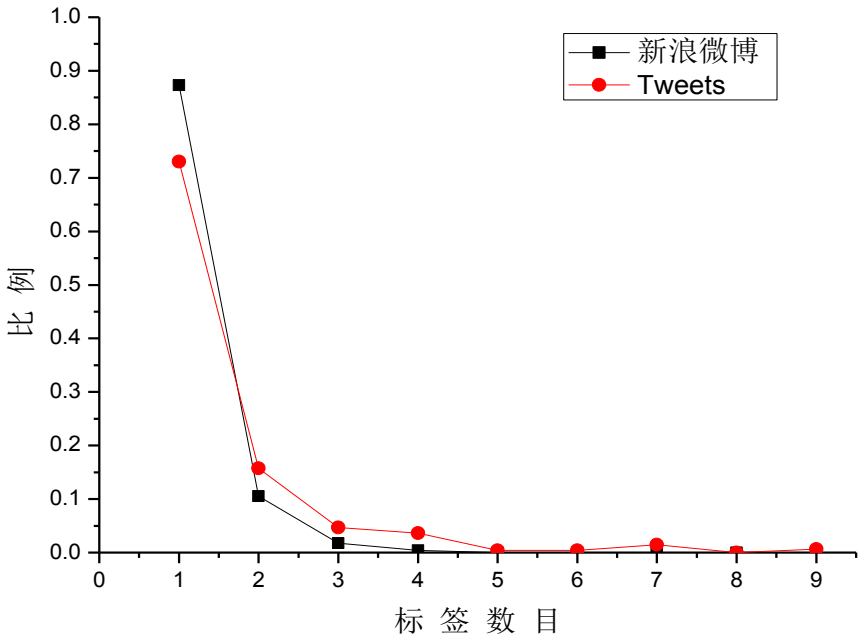

图3.1 包含不同标签数目的微博消息的比例

分析：从这里可以看出，中英文微博上用户使用不同数目的标签的情况还是有差别的。在新浪微博上，在使用标签的微博消息中，使用 1 个标签的用户接近 90%，而在 Twitter 上，使用 2——4 个标签的比例约为：这说明英文微博用户在使用标签时，使用多标签的倾向比中文微博用户高。

3.3　表情符号统计

在新浪微博上，微博平台给用户提供了多种表情符号，供用户选择，使得微博消息更为丰富多彩，如图 3.1 所示。

图3.2　新浪微博平台提供的表情符号

用户选择的表情符号通常反应了用户的心情，即表情符号暗含了情感色彩。如图 3.2 和图 3.3 所示。

图3.3　含有表情符号的表达正向情感的微博消息

图3.4 含有表情符号的表达负向情感的微博消息

（1） 本文在本节所作的统计是针对表 3.4 中的表情符号。

表3.4 中英文微博上的表情符号

		个数	内容
新浪微博	所有表情符号	--	如  ，对应[哈哈]；本处的表情符号为新浪消息文本中符合“[(.*?)]”的正则表达式的表情符号
Tweets	正向表情符号	9	":)",";)",":-)",";-)",":D",";]",":]",":p",";p"
	负向表情符号	3	":(", ":-(",":["

（2） 中英文微博上包含表情符号的消息数目及比例见表 3.5。

表3.5 包含表情符号的微博消息的数目及比例

	包含表情符号的微博消息数	总数	比例
新浪微博	977	6000	16.28%
Tweets	211	1939	10.88%

分析：从这里可以看出，新浪微博上包含表情符号的消息比例要比 Twitter 高 6%左右，这也许是因为新浪微博提供的表情选择比 Twitter 上需要用户手动输入表情更为便捷所致。

（3） 除了总体比例统计，本文还对包含表情符号的微博消息中不同表情符号数目的消息的数目及比例进行了进一步统计，见图 3.2。

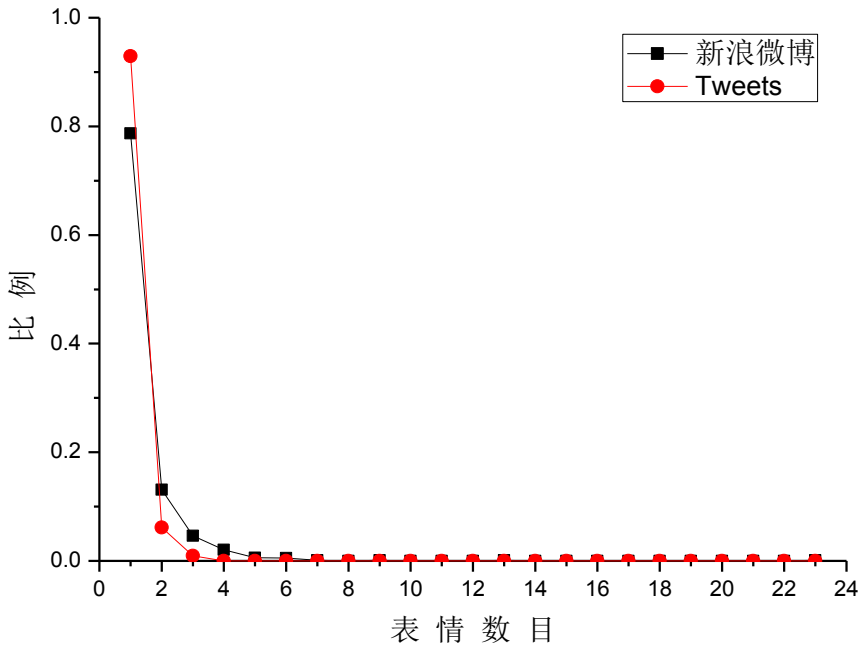

图3.5 包含不同表情符号数目的微博消息的比例

分析：从这里可以看出，针对包含表情符号的微博消息，在新浪微博上，用户使用 1 个表情符号的比例为 78%左右，使用 2——5 个表情符号的比例约为 20%。而 Twitter 上用户使用 1 个标签的比例约为 97%，使用多个表情符号的比例非常小。这说明中文微博用户在使用表情符号时，使用多表情符号的倾向比英文微博用户高。

3.4 句子情况统计

（1） 本节中所说的句子是指包含完整语义的单元，定义为以句号、感叹号、问号等标点符号结尾的句子（注：以逗号，分号等结尾的分句要小于句子的概念）。

在 Twitter 上，用户所发的消息被限定为 140 个英文字符，每个英文单词包含多个字母且英文单词之间需要用空格隔开，因此 Twitter 上的微博消息通常是一个句子或者一个句子外加标签的形式。因此在前人的研究中，在做英文微博情感极性时，都是将一条微博消息看作一个整体。

但是，在新浪微博上，用户所发的消息被限定为 140 个中文字符，而不是 140 个英文字符。这样用户所表达的语义更加丰富，很多微博消息包含的句子数目都不只一句，表 3.6 所示是新浪微博消息的句子情况统计。

表3.6 新浪微博消息的句子情况统计

	微博消息数	**句子总数**	**平均每条微博包含单句数**
新浪微博	6000	12940	2.157

（2） 除了总体比例统计，本文还对包含不同数目的句子的消息的数目及比例进行了进一步统计，见图 3.3。

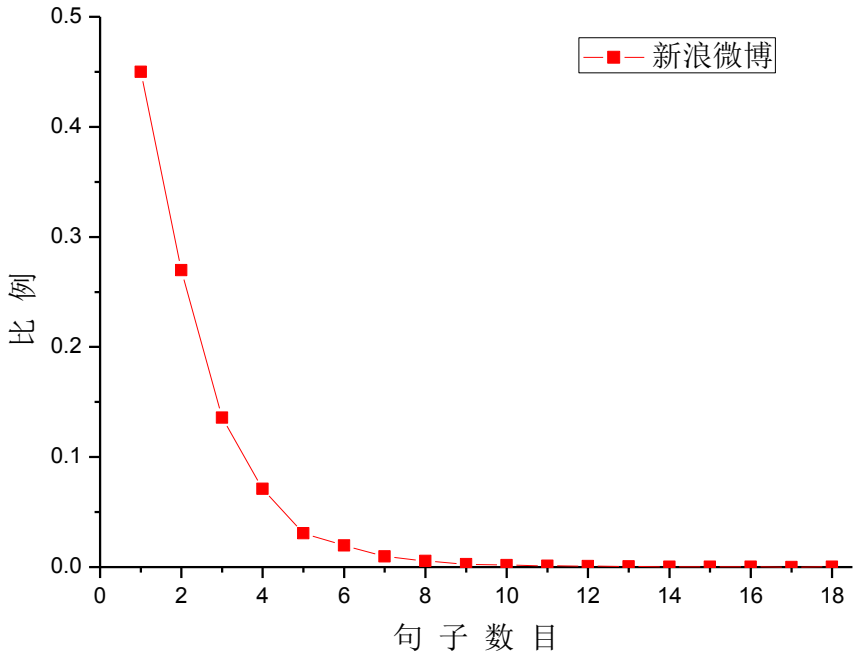

图3.6 包含不同句子数目的微博消息的比例

分析：从这里可以看出，仅包含 1 个句子的微博消息的比例为 45%，包含 2 个句子的微博消息数目为 27%左右，包含 3——7 个句子数目的微博消息的比例接近 26%。由此看来，新浪微博上一条微博的句子构成情况会比英文微博复杂。由此可以推测，考虑句子组成与不考虑句子组成情况对于中文微博的情感分析在效果上可能会有所差别，第 5 章会对此有更详细的讨论。

3.5 情感极性统计

这里所说的情感极性指正、负、中性情感极性，给出的新浪微博的示例如表 3.7 所示。

表3.7 不同情感极性的新浪微博消息示例

	微博消息示例
正向情感	<将爱情进行到底> 很不错 值得推荐 干嘛<将爱情进行到底>那么好笑的, 我以为是煽情的 谁知笑到我肚子疼 有些片段真的好经典。
负向情感	<将爱情进行到底>评价一般,看的只是回忆。 <将爱>电影版真的不好看,虽打着要将爱情进行到底的旗号,但是三段中,谁的爱情坚持到了底？！
中性情感	那首歌将爱情进行到底中,文惠在 CD 店里听的那首歌, 将爱情进行到底中的一首曲子 。 明天和学妹去看#将爱情进行到底#,大家都看了吗？

本节对于中英文微博的正、负、中性情感极性句子的比例情况进行了统计，见表 3.8。

表3.8 情感极性统计

		正向情感	负向情感	中性情感	总数
新浪微博	数目	1971	666	3363	6000
	比例	32.85%	11.10%	56.05%	100%
Tweets	数目	459	268	1212	1939
	比例	23.67%	13.82%	62.51%	100%

分析：从这里可以看出，中英文微博消息中各类情感的分布大致相似，都是中性情感比例最高，正向情感其次，负向情感最低，同时英文微博上中性情感比例相对高一些，中文微博上正向情感比例为 32%左右，负向情感比例为 11%左右，中性情感为 56%左右。从该比例构成中可以看出，表达情感的微博消息，即正、负向情感的微博消息的比例之和接近 44%，这也说明进行微博情感挖掘是可行且必要的。

3.6 本章小结

本章通过对中英文微博上的网页链接、标签、表情符号、情感极性分布情况进行统计及比较，从而分析了中英文微博上的区别。同时，本章还对新浪微博上的句子情况进行了统计及分析。得出的结论总结如下：

（1） 新浪微博中包含链接的消息比例（9.08%）低于 Tweets（20.01%），同时在包含链接的中英文微博消息中，接近98%的消息只包含1条链接；

（2） 新浪微博中包含标签的消息比例（3.77%）低于 Tweets（24.29%），同时中文微博中使用多标签的倾向比英文微博低；

（3） 新浪微博中包含表情符号的消息比例（16.28%）低于 Tweets（10.88%），同时中文微博中使用多个表情符号的倾向比英文微博高；

（4） 中英文微博的情感极性比例分布大致相同，英文微博中性情感比例偏高。中文微博上正向情感比例为32%左右，负向情感比例为11%左右，中性情感为56%左右。

（5） 新浪微博消息包含的平均句子数目超过 2 句，具体情况为：仅包含 1 个句子的微博消息的比例为 45%，包含 2 个句子的微博消息数目为 27%左右，包含 3——7 个句子数目的微博消息的比例接近 26%。

第4章 算法设计及实现[7]

4.1 算法设计流程图

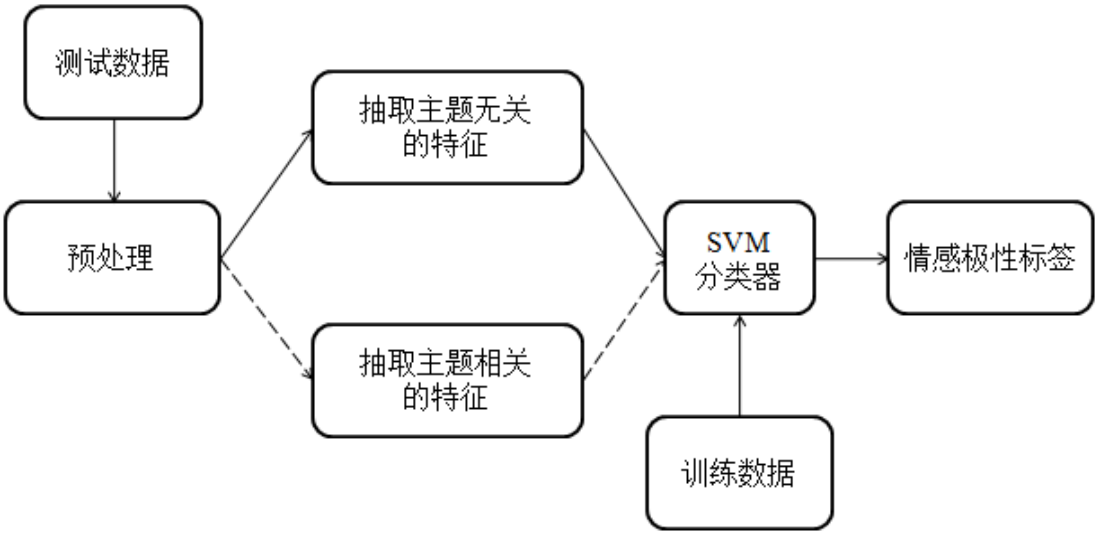

图4.1 算法设计流程图

新浪微博的情感分析问题在本文的工作中本质上而言是个分类问题。受到 Jiang 等人[5]的研究工作启发，针对中文微博，本文设计了图 4.1 所示的算法流程图。从图中可以看出，整个系统的大体思路是：

先将通过新浪 API 抓取的基于主题关键字的数据进行标注，然后按五折交叉验证的方法进行训练和测试。

首先对训练数据先进行预处理后，然后抽取主题无关的特征及主题相关的特征，训练 SVM 分类器来对测试数据进行情感极性分类，输出情感极性标签。

从图中可以看出，整个算法的核心在于抽取主题无关的特征及主题相关的特征，同时选用的训练方法也很关键，这些将在 4.2 节中详述。

[7] 此工作是作者在微软亚洲研究院实习期间完成

4.2 算法实现

4.2.1 基于表情符号的规则方法

基于情感词典的规则方法首先需要构造出一个正向表情符号表和一个负向表情符号表，然后提取每一条微博中出现的表情符号，统计每一条微博中出现的正向表情符号的个数、负向表情符号的个数。然后使用式 4-1 来计算整条微博的情感极性。

$$情感极性=\begin{cases}正向情感\ (如果正向表情符号数大于负向表情符号数)\\ 负向情感\ (如果正向表情符号数小于负向表情符号数)\\ 中性情感（如果正向表情符号数等于负向表情符号数）\end{cases} \quad (4\text{-}1)$$

该方法的优点是想法直观、简单，缺点是表情符号的覆盖率较低，对于不包含表情符号的微博消息不会做任何判定，直接被归类为中性情感。

4.2.2 基于情感词典的规则方法

基于情感词典的规则方法首先需要构造出一个正向情感词表和一个负向情感词表，然后对每一条微博进行分词，统计每一条微博中出现的正向情感词的个数、负向表情词的个数。在统计正、负向情感词的时候，由于中文里面存在否定转移的现象，例如句子“我很高兴。”，在“高兴”前面加个“不”，即“我很不高兴”，那么本身的正向情感词“高兴”前面出现了不，应该视作负向情感词，反之依然。在得到正、负向情感词后，使用式 4-2 计算整条微博的情感极性。

$$情感极性=\begin{cases}正向情感\ (如果正向情感词数大于负向情感词数)\\ 负向情感\ (如果正向情感词数小于负向情感词数)\\ 中性情感（如果正向情感词数等于负向情感词数）\end{cases} \quad (4\text{-}2)$$

该方法的优点是想法直观、简单，缺点是情感词典的覆盖率较低，对于不包含情感词的微博消息不会做任何判定，直接被归类为中性情感。

4.2.3 基于 SVM 的方法

（一）SVM 简介

在这一节，本文将介绍一下核心的分类器支持向量机（SVM）。本文使用的 SVM 分类工具 libsvm[8]是台湾大学林智仁 (Chih-Jen Lin) 博士等开发的一套支持向量机算法库。

SVM（Support Vector Machine），中文名为支持向量机，是 Vapnik 等人在对线性分类器提出了另一种设计最佳准则。

（1） SVM 的主要思想及特点

对于线性可分的数据，可以画出特点一条直线直接将元组分开。对于非线性不可分的数据，SVM 使用一种非线性映射，将原训练数据映射到较高的维。从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能。

在新的高维空间上，SVM 会基于结构风险最小化理论搜索线性最佳分离超平面，即将一类元组与其他类分离的决策边界。SVM 使用支持向量（即基本训练元组）和边缘（由支持向量定义）发现该超平面。

SVM 的学习可以表示为凸优化问题，因此能利用已知的有效算法发现目标函数的全局最小值。而其他分类方法，如基于规则的分类器和人工神经网络，大多采用一种基于贪心学习的策略来搜索假设空间，一般只能获得局部最优解。

SVM 不仅可以解决两类问题，而且可以处理多分类问题。

（2） SVM 的核函数

使用不同的核函数，会得到不同的 SVM，常用的有 4 种核函数：

①线性核函数（linear）：K(x,y) = x ·y；

②多项式核函数（polynomial）：K(x,y) = [(x ·y)+1]d；

③径向基函数（radial basis function）：K(x,y) = exp(-|x-y|^2/d^2)；

④二层神经网络核函数（sigmoid）：K(x,y) = tanh(a(x ·y)+b)。

[8] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

（二）方法介绍

本节我们对于 SVM 的训练方法进行探讨，首先依据是否对微博进行分句划分为两大类策略，然后再在每类策略下细分，共包括四种方法，详见图 4.2。后续我们将对图 4.2 中的方法及特征进行进一步阐述。

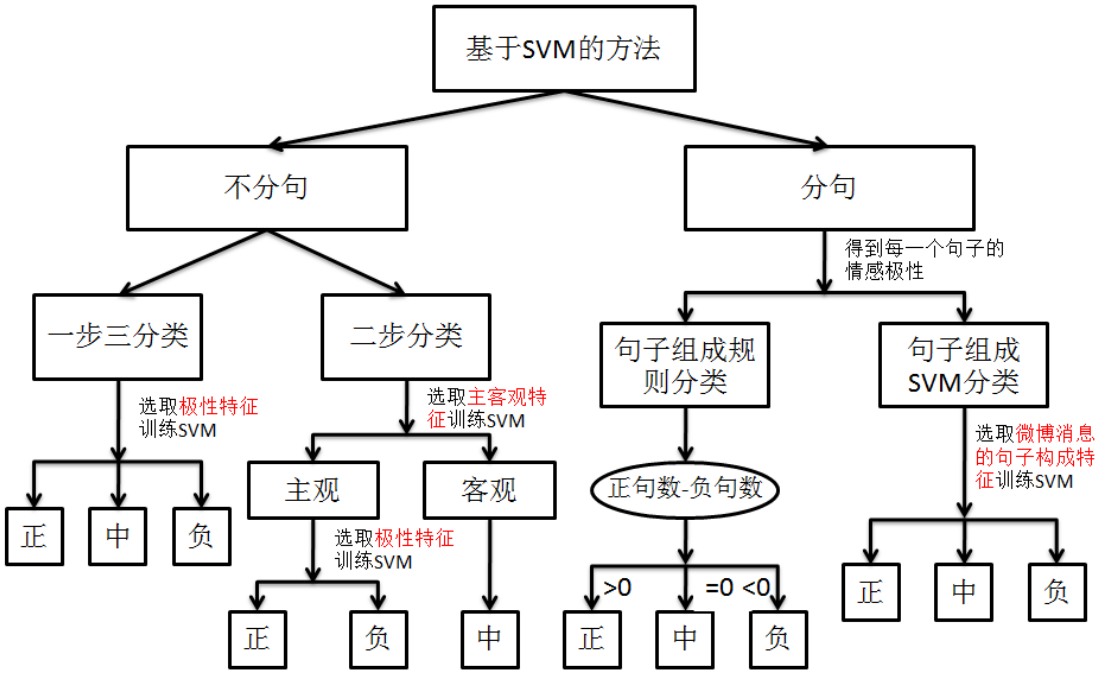

图4.2 基于SVM的情感分析的不同策略及方法

下面我们对这两大类策略、四种方法进行详细介绍。

第一类：不分句，将整个微博消息看作一个整体。在这一类方法中，我们考虑两种策略：

①一步三分类

此方法是最直接的想法，只需要得到训练数据的情感极性结果（即正、负、中性情感），同时每一条微博消息进行整体特征提取，直接训练一个三分类的 SVM 分类器，然后使用该分类器直接根据测试数据所提取的特征直接分为正、负、中性三类情感。

②二步分类

二步分类法是参考了 Barbosa 和 Feng [18]的工作。所谓二步分类是首先用训练数据的主客观情感标注结果和对应特征训练第一个 SVM 分类器来进行主客观分类。然后再使用带有主观情感的训练数据中的正、负向情感标注结果训练第二个 SVM 分类器来进行正、负向情感的分类。

对于测试数据，该方法会先提取特征，先使用第一个 SVM 分类器对测试数据进行主、客观分类；然后对于分为主观情感的测试数据再提取特征使用第二个 SVM 分类器将其分正、负向情感。

第二类：分句，将整个微博消息拆分为若干句子。这里的句子是依据句号、感叹号等进行切分的句子，不包括逗号。之所以会提出分句策略，是因为我们对新浪微博数据进行观察后，发现微博消息通常包含很多句。而句与句之间的情感极性可能不同。在这一类方法中，我们得到每一个句子的情感极性时所使用的方法是选择一步三分类和二步分类中效果最佳的方法。下面我们考虑两种方法：

①句子组成规则分类

该方法主要依据一条微博中各个句子情感极性情况，使用规则来进行分类。该方法具体步骤如下：

首先将训练数据中的每一条微博拆成若干个句子，使用每个句子的情感极性标注结果和对应提取的特征进行来训练 SVM 分类器，该分类器能对每个句子进行情感极性分类。

对于测试数据，将每一条微博消息拆分成若干个句子，提取特征，使用训练的 SVM 分类器分别得到每一句的情感极性，使用公式 4-1 得到该条微博的情感极性。

$$情感极性=\begin{cases}正向情感\ (如果正向情感句子数大于负向情感句子数)\\负向情感\ (如果正向情感句子数小于负向情感句子数)\\中性情感（如果正向情感句子数等于负向情感句子数）\end{cases}\qquad(4\text{-}3)$$

②句子组成 SVM 分类

该方法与句子组成规则分类的类似，首先也需要获得一条微博中各个句子情感极性情况。

该方法具体步骤如下：

首先将训练数据中的每一条微博拆成若干个句子，使用每个句子的情感极性标注结果和对应提取的特征进行来训练第一个 SVM 分类器，该分类器能对每个句子进行情感极性分类。

然后使用 4.2.3 节中的微博消息的句子构成特征训练第二个 SVM 分类器，来对微博消息进行情感分类。

对于测试数据，将每一条微博消息拆分成若干个句子，提取特征，使用训练好的第一个 SVM 分类器得到每一句的情感极性，根据该结果继续提取特征，使用第二个 SVM 分类器得到该微博消息的情感极性。

4.2.4 主题无关的特征抽取

本节将对主题无关的特征进行详细介绍，包括主客观分类特征、极性分类特征和微博消息的句子构成特征。

（1） 主客观分类特征

主客观分类特征主要是将一条微博消息分成主观消息和客观消息。主观消息即表达了观点或情感的消息。此处，我们参考了前人工作及第 3 章中的特征分析，主要考虑的特征除了包含微博消息中常见的链接、表情符号特征，还包含传统文本中所使用的情感词典和上下文特征，详见表 4.1。分词使用的程序为中国科学院计算技术研究所研制出的汉语词法分析系统 ICTCLAS[9]。

根据实际数据观察，发现标签在新浪微博上出现得极少（低于 4%），且在指定主题下的微博中的标签对于情感的贡献更是微乎其微，因此此处不考虑标签特征。

[9] Institute of Computing Technology, Chinese Lexical Analysis System. Available at http://ictclas.org/index.html.

表4.1 主客观分类特征

特征序号	特征类型	特征内容	描述
1	链接	是否含有 url 链接	链接通常以 http:开头
2	表情类型	表情符号个数	正向表情：34 个 负向表情：32 个
3	情感词典	情感词个数	正向情感词：4751 个 负向情感词：3651 个
4	情感短语	情感短语个数	正向情感短语词：103 个 负向情感短语词：6 个
5	上下文	中文词是否出现、形容词个数、动词个数、感叹号是否出现、问号是否出现	中文词采用 ICTCLAS 的中文词表统计，共 80224 个词； 分词采用 ICTCLAS 进行分词

注：①考虑到用户的“懒惰性”及使用表情的普遍情况，表 4.1 中涉及的正负向表情均只选用了新浪提供的默认表情类，具体如表 4.2。

②关于情感词：由于情感的主观性与不确定性，因此本文构建情感词典时，只挑选了绝对在任何情况下绝对表达正向情感的词和绝对表达负向情感的词，如“喜欢”和“讨厌”。由于篇幅所限，本文在附录中给出了本文构建的词表中 100 个常用的正向情感词和负向情感词。

③关于情感短语词：

汉语中，有些词本身没有情感偏向，但是如果前面被“有”、“没有”、“缺乏”等词修饰之后就会带有情感偏向。比如“文化”本身是中性词，前面被“有”修饰的话，“有文化”就变成了带有正向情感的短语；又比如“问题”本身是中性词，前面被“有”修饰的话，“有问题”就变成了带有负向情感的短语。

因此这里构建两个词表，一个是被“有”一类的词修饰能表征正向情感，被“无”一类的词修饰表征负向情感，如“文化”，这里简称为正向情感短语词；一个是被“有”一类的词修饰能表征负向情感，被“无”一类的词修饰表征正向情感，如“问题”，这里简称为负向情感短语词。

表4.2 正、负向表情符号

	个数	内容
正向表情	34	([爱心传递])，([蜡烛])，([绿丝带])，([植树节])，([兔子])，([熊猫])，([给力])，([威武])，([奥特曼])，([呵呵])，([嘻嘻])，([哈哈])，([爱你])，([可爱])，([偷笑])，([害羞])，([酷])，([鼓掌])，([亲亲])，([太开心])，([顶])，([握手])，([耶])，([good])，([ok])，([赞])，([蛋糕])，([心])，([咖啡])，([话筒])，([月亮])，([太阳])，([干杯])，([萌])
负向表情	32	([神马])，([浮云])，([晕])，([泪])，([抓狂])，([哼])，([怒])，([汗])，([困])，([睡觉])，([衰])，([吃惊])，([闭嘴])，([鄙视])，([挖鼻屎])，([失望])，([生病])，([怒骂])，([懒得理你])，([右哼哼])，([左哼哼])，([嘘])，([委屈])，([吐])，([可怜])，([打哈气])，([疑问])，([做鬼脸])，([弱])，([不要])，([伤心])，([猪头])

（2） 极性分类特征

极性分类特征适用于两种情况，一种是三分类，将微博消息分为正向情感、负向情感和中性情感；一种是二分类，即在主观消息中，继续将微博消息分为正向情感和负向情感。

表 4.3 所示是极性分类特征。这里极性分类特征涉及的组成部分大致与主、客观分类特征一致，略有不同。

表4.3 极性分类特征

特征序号	特征类型	特征内容	描述
1	链接	是否含有 url 链接	链接通常以 http:开头
2	表情类型	正向表情符号个数、 负向表情符号个数	正向表情：34 个 负向表情：32 个
3	情感词典	正向情感词个数、 负向情感词个数 （统计正负向情感词时使用否定词表进行了情感转换判定）	正向情感词：4751 个 负向情感词：3651 个 否定词表：18 个
4	情感短语	正向情感短语个数、 负向情感短语个数	正向情感短语词：103 个 负向情感短语词：6 个 短语修饰词：6 个
5	上下文	中文词是否出现、形容词个数、动词个数、感叹号是否出现、问号是否出现	中文词采用 ICTCLAS 的中文词表统计，共 80224 个词； 分词采用 ICTCLAS 进行分词

（3） 微博消息的句子构成特征

表4.4 微博消息的句子构成特征

特征序号	特征类型	特征内容	取值
1	首句的情感极性	首句的情感极性：正、中、负向情感	1、0、-1
2	尾句的情感极性	尾句的情感极性：正、中、负向情感	微博消息大于 2 个单句时，取值为 1、0、-1；只有一个单句时，设置为-2
3	正向情感句子数	正向情感句子的数目	整数
4	负向情感句子数	负向情感句子的数目	整数
5	中性情感句子数	中性情感句子的数目	整数

通常一条中文微博会包含多个中文句子，参见 3.4 节。因此在微博消息的句子构成特征中较易想到的特征是正向、负向、中性情感句子数目。除此以外，考虑到用户在书写传统文本时的习惯，用户经常在首句或首段、尾句或尾段表达自己的观点或情感，从而奠定文章的基调，即所谓的“开篇点题”和“首尾呼应”，故此处还考虑了一条微博消息中首句和尾句的情感极性，参见表 4.4。

4.2.5 主题相关的特征抽取

这一节，我们简单介绍主题相关的特征抽取。通过观察微博，我们发现微博中的主题较为发散，而且存在省略主题词的情况，如表 4.5 所示：

表4.5 新浪微博主题发散及省略主语示例

主题词：将爱情进行到底
一条微博消息：
好累啊 两天都没好好睡觉了。
昨天晚上和聒噪的老马一起去吃了西餐。
看了场电影。
将爱情进行到底。
很不错啊。

从表 4.5 中可以看出，这条微博由 5 个句子组成。首先该微博消息存在主题发散的情况：句子 1 和句子 2 表达的情感完全与主题词“将爱情进行到底”无关。其次，该微博存在省略主题词的情况：句子 3、4、5 是针对主题词“将爱情进行到底”的，除了句子 4 是明确包含主题词的，句子 3 中的“电影”指代的是“将爱情进行到底”，句子 5 中省略了主语“将爱情进行到底”。

基于上述观察，目前对于主题相关的特征仅做了初步的简单处理，包含两种情况：

①包含主题词的情况：仅把一条微博消息中包含主题词的若干个句子抽取出来，判断这些句子的极性，从而确定该条微博消息的情感极性；

②零指代的情况（此处仅考虑句子中缺乏名词性短语或者代词的情况）：除了抽取符合①中要求的句子以外，对于微博中的一个句子，如果它不包含任何名词性短语和代词，即认为它省略了情感表达对象，同时认为它表达的情感是针对上一句的对象，如果上一句包含主题词的话，则认为当前句也对主题词表达了情感，也应该考虑该句的情感极性。

除此以外，本文还对距离窗口的方法进行了尝试：

对于构成微博消息的每个句子，先识别出句子中的情感词或情感短语，记为位置 i,以词为举例,看在窗口 distance 个词范围内,即[i-distance,i+distance]中是否出现主题词，如出现主题词则认为该句是主题相关，在使用句子组成 SVM 方法时考虑该句子的情感极性，否则不考虑该句子的情感极性

4.3 本章小结

本章主要是对主题相关的新浪微博的情感分析算法部分的详细阐述，首先介绍了整个算法的大致流程，然后具体介绍了算法中使用的 SVM 分类器的简单介绍，具体分类方法的思路介绍（包括两类共四种方法），分类方法所使用的特征（主要包括主题无关的特征和主题相关的特征两大类）。

下一章，将针对本章中涉及的分类方法、各类特征等进行进一步探讨。

第5章 实验结果及相关分析[10]

5.1 实验数据、评测方法及指标

本文使用的实验数据是 1.3.3 节中提到的 API 获取的与话题相关的数据。这里，我们共选用了影视、名人、产品三个类别，共 6 个话题的数据。各个话题下数据的标注结果见表 5.1。

表5.1 实验数据情感极性分布

话题	文件	正向感情条数	负向感情条数	中性感情条数	总条数	中性感情比例
影视	将爱情进行到底	336	75	589	1000	58.90%
	青蜂侠	270	183	547	1000	54.70%
名人	科比	645	79	276	1000	27.60%
	乔布斯	329	33	638	1000	63.80%
产品	iphone	188	83	729	1000	72.90%
	诺基亚	203	213	584	1000	58.40%
共计		1971	666	3363	6000	56.05%

评测方法选用的是五折交叉确认（five-fold cross-validation）。在五折交叉中，将初始数据随机划分为 5 个互不相交的子集或“折” D_1, D_2, D_3, D_4, D_5。每个折的大小大致相等。训练和检验进行 5 次。在第 i 次迭代，划分 Di 作检验集，其余的部分一起用来训练模型。这里每个样本用于训练的次数相同，并且用于检验一次。

[10] 此工作是作者在微软亚洲研究院实习期间完成

评测指标选的是准确率（accuracy），即 5 次迭代正确分类的总数除以原始数据中的元组总数。

5.2 三种方法比较

在这一节，我们将针对使用表情符号的规则方法、情感词典的规则方法和使用 SVM 的方法来进行试验。这里为简便起见，SVM 的方法采用一步三分类方法，所选用的极性分类特征采用表 4.3 中所使用的全部特征。

表5.2 三种方法比较

	基于表情符号的规则方法	基于情感词典的规则方法	基于 SVM 的一步三分类方法
准确率	56.583%	55.583%	**65.400%**

分析：从表 5.2 中可以看出，基于 SVM 的方法效果最好，为 65.400%，基于表情符号的规则方法略好于基于情感词典的规则方法效果最差，均在 56%左右。

鉴于基于 SVM 的方法效果最好，因此从下一节开始，将对使用 SVM 的方法进行详细研究，涉及特征选择、方法比较、核函数比较等方面。

5.3 主题无关的相关实验

在这一节，我们将针对 SVM 的方法从方法、特征和核函数三个部分来进行试验。

（1） 方法比较

首先我们对于不拆分句子的方法，即一步三分类和二步分类的效果进行了比较，见表 5.3。此处选取的特征时各自对应的 4.2.4 节中的所有特征。

表5.3 一步三分类与二步分类的效果比较

方法		准确率
一步三分类		**65.400%**
二步分类	二步分类主、客观	69.800%
	二步分类极性	63.866%

分析：从表 5.3 中可以看出，一步三分类的效果好于二步分类结果。这也许是由于二步分类中主、客观分类的准确率仅为 69.800%，这说明在第一步分类中已经有 30.200%的数据分类错误，而极性分类本身准确率也不高，因此二步分类的效果要比一部三分类差。

鉴于表 5.3 的结果，我们对于拆分句子的方法，即句子组成规则分类和句子组成 SVM 分类中的得到单个句子极性的 SVM 分类器选用一步三分类方法，结果如表 5.4。此处选取的特征时各自对应的 4.2.4 节中的所有特征。

表5.4 句子组成规则分类与句子组成SVM分类的效果比较

方法	准确率
句子组成规则分类	63.517%
句子组成 SVM 分类	**66.267%**

分析：对照表 5.3 和 5.4 可以看出，对于拆分句子和不拆分句子这两类方法，拆分句子的效果要好于不拆分句子的效果。对于这四种方法的效果比较来看，句子组成 SVM 分类 > 句子组成规则分类 > 一步三分类 > 二步分类。

（2） 特征比较

这里我们对于主客观分类特征、极性分类特征、微博消息的句子构成特征进行了分析探讨。

①主客观分类特征的结果见表 5.5。

表5.5 主、客观分类特征的效果比较

特征	准确率
所有特征	**69.800%**
所有特征-链接特征	69.750%
所有特征-表情特征	69.267%
所有特征-情感词典特征	69.317%
所有特征-情感短语特征	69.717%
所有特征-上下文特征	59.133%

分析：从结果来看，对于主、客观分类，最好的特征组合是使用全部特征，即表情特征+链接特征+情感词典特征+情感短语特征+上下文特征。各个特征的有效性依次为：上下文特征 > 表情特征 > 情感词典特征 > 情感短语特征 > 链接特征。

②极性分类特征的结果见表 5.6。

表5.6 极性分类特征的效果比较

特征	准确率
所有特征	65.400%
所有特征-链接特征	**65.717%**
所有特征-表情特征	64.783%
所有特征-情感词典特征	64.767%
所有特征-情感短语特征	65.333%
所有特征-上下文特征	58.933%

分析：从结果来看，对于极性分类，最好的特征组合是使用表情特征+情感特征+情感短语特征+上下文特征。各个特征的有效性依次为：上下文特征 > 情感词典特征 > 表情特征 > 情感短语特征。链接特征对于极性分类起负面作用。

③微博消息的句子构成特征的结果见表 5.7。

表5.7 微博消息的句子构成特征的效果比较

特征	准确率
所有特征	**66.267%**
所有特征-首句极性特征	64.800%
所有特征-尾句极性特征	64.433%
所有特征-首、尾句极性特征	64.933%
所有特征-三种情感极性句子数目特征	55.800%

分析：从结果来看，使用首句极性特征+尾句极性特征+三种情感极性句子数目特征效果最好。从效果上来看，三种情感极性句子数目特征 > 尾句极性特征 > 首句极性特征。

（3） 核函数比较

这里我们主要对 4.2.3 节中提到的 SVM 常用的四种核函数的效果进行比较。这里的比较是在一步三分类的框架中进行的（因为一步三分类是拆分句子方法的基础），具体结果见表 5.8。

表5.8 核函数的效果比较

特征	准确率
线性核函数（linear）	**65.400%**
多项式核函数（polynomial）	56.050%
径向基函数（radial basis function）	56.050%
二层神经网络核函数（sigmoid）	56.050%

分析：从结果来看，核函数的效果性能为：线性核函数 > 多项式核函数 = 径向基函数 >二层神经网络核函数

根据本小节的三组实验，我们得出如下结论：

①在方法选择上：

一步三分类方法要好于二步分类法。整体选择句子组成 SVM 分类方法最佳。

②在特征选择上：

对于主客观分类，选择链接特征+表情特征+情感词典特征+情感短语特征+上下文特征最好；对于极性分类，选择表情特征+情感词典特征+情感短语特征+上下文特征最好，引入链接特征效果反而下降。从特征有效性来看，对于主客观分类，是上下文特征 > 表情特征 > 情感词典特征 > 情感短语特征 > 链接特征；对于极性分类，是上下文特征 > 表情特征 > 情感词典特征 > 链接特征。

对于微博消息的句子组成特征，选择首句极性特征+尾句极性特征+三种情感极性句子数目特征效果最好。从效果上来看，三种情感极性句子数目特征 > 尾句极性特征 > 首句极性特征。

③在核函数选择上：选择线性核函数效果最好。

因此，为达到最佳效果，“最佳组合”为：整体的方法选择句子组成 SVM 分类，句子组成特征选择首句极性特征+尾句极性特征+三种情感极性句子数目特征；在对单个句子进行分类时选择一步三分类方法，特征选择表情特征+情感词典特征+情感短语特征+上下文特征；核函数选择线性核函数。

在选定“最佳组合”后，我们首先对极性分类特征的覆盖率进行了统计，见表 5.9。

表5.9 最佳组合下特征覆盖率统计

	不拆分句子（总数：6000）		**拆分句子（总数：13004）**	
	数目	**比例**	**数目**	**比例**
表情特征	794	13.233%	861	6.669%
情感词典特征	3736	62.267%	5415	41.847%
情感短语特征	66	1.110%	77	0.595%
上下文特征	5934	98.900%	11981	92.589%
表情+情感词典+情感短语特征	4021	67.017%	6071	46.917%
所有特征	5948	99.133%	12758	95.703%

分析：在不拆分句子的情况下，表情特征、情感词典、情感短语特征的覆盖率为 67.017%，这说明另外大约 33%的微博消息的情感分类仅依靠上下文特征来进行；在拆分句子的情况下，链接特征、表情特征与情感词典的覆盖率仅为

46.917%，这说明另外大约 53%的句子的情感分类仅依靠上下文特征来进行。同时无论是拆分句子还是不拆分句子，四类特征总体的覆盖率都高于 95%。

针对“最佳组合”，我们对各个文件及总体进行了实验，并得到了结果见表 5.10。

表5.10 主题无关特征最佳组合的效果

主题	文件	微博消息数	主题无关准确率
影视	将爱情进行到底	1000	67.400%
	青蜂侠	1000	66.800%
名人	科比	1000	72.900%
	乔布斯	1000	69.400%
产品	Iphone	1000	69.800%
	诺基亚	1000	61.200%
共计		6000	66.467%

分析：从结果来看，主题无关的最佳方法的总体准确率达到了 66.467%。而细看各个主题，按准确率排序，名人 > 影视 > 产品。通过分析实验数据，我们发现，新浪微博上用户对于名人发表观点的时候使用的表达较为单一，例如对科比的评论很多是说“科比，当之无愧 MVP!”，“喜欢科比”。因此分类相对容易些。而对于影视的评论，用户关注的主题更为发散，如主角、画面等，因此分类时比名人要难。而对于产品，由于产品包含不同的型号，公司，而且产品的属性（如屏幕、信号等）非常多，因此分类问题更为复杂。由此，导致了各个主题下的分类准确率的差别。

为了进一步研究分类结果，本文输出了最佳组合下总体的分类结果矩阵，见表 5.11。

表5.11 主题无关特征最佳组合时情感极性判定结果

	标注人员标注结果			
句子组成 SVM 分类结果		正向情感	负向情感	中性情感
	正向情感	255	20	126
	负向情感	24	42	47
	中性情感	115	71	500

对表 5.11 进行进一步计算，可以得到表 5.12。

表5.12 主题无关特征最佳组合下各种情感的精确率和召回率

	精确度	**召回率**
正向情感消息	63.591%	64.721%
负向情感消息	37.168%	31.579%
中性情感消息	72.886%	74.294%

从表 5.12 中可以看出，中性情感的准确率和精确度最高，其次是正向情感，对于负向情感的分类效果最差。造成这一结果的原因与表 5.1 中的情感极性标注结果相吻合，在表 5.1 的标注数据中，中性情感消息的比例最高，其次是正向情感，最后是负向情感。

此处，我们给出一些分类正确的微博消息示例，如表 5.13 所示：

表5.13 分类正确的微博消息示例

	示例
正向情感消息	今日睇佐<将爱情进行到底>。好感动啊~ [太开心] 青蜂侠'哈哈[鼓掌] 我瞻仰乔布斯,我喜欢苹果!
负向情感消息	昨天我看了<将爱情进行到底>,真的没有想到,徐静蕾怎么拍了这么一部戏,白瞎我一晚上的时间 我想说的是,<青蜂侠>太烂了! 乔布斯! 你是傻子!
中性情感消息	将爱情进行到底看完之后,脑子里就一句话,等你爱我我我我我我我我。 周日一起去河西中影看周董演绎青蜂侠! 分享图片:乔布斯喝的是水啊

在情感分析中，我们除了关注分类准确率以外，还很关注误分的情况。将正向情感误分成中性情感和负向情感分为中性情感这两类错误不是很严重。但是如果将正向情感误分成负向情感或者将负向情感分为正向情感比较严重。在表 5.11 中，有 24 条微博消息正向情感被误分为负向情感，有 20 条负向微博消息被误分为正向情感。我们对于结果一一查看，发现主要有以下几类错误（表 5.13 和表 5.14，其中绿色表示系统中识别出的表情符号，红色表示正向情感词，紫色表示负向情感词）：

表5.14 正向情感消息被误分成负向情感消息的原因及示例

原因	例子
1、文中出现正向表情，但是有负向情感词	在做公交车 上面在放周杰伦 青蜂侠的采访 [呵呵] 手机照的不太清晰啦 我已经精简半天了[害羞] 信用卡升级新增了一个，旧的还没注销成功也没改号，暂时还不能购买 iphone 的相片处理软件
2、仅有正向情感词和上下文特征	但是单靠乔布斯一家做到这样真的是个奇迹。 乔布斯不是为了钱在战斗，而是为了荣誉在战斗 中国人民真的富强了，都快全民苹果了^_^今天看到保安 GG 和洗头 mm 也都用 iphone4 哎。
3、含负向表情	虽然不材，也真想和科比这睦巨星们打上一场蓝球赛[泪]
4、正向情感词明显多于负向情感词，仍分错	苹果的再度崛起，绝不仅仅是乔布斯一个人的功劳，一个骁勇的将军背后，一定要有一支善战的团队追随，众志成城，则无坚不摧！ 走进食堂，久违了的宾至如归的感觉啊，定睛一看，果然，黑漆漆的全是人头，大家都对阔别已久的食堂燃起了熊熊的思念之情，队伍浩浩，俨然让我想起了老版诺基亚手机里的经典游戏——贪食蛇撑满框框头尾相接的壮景。
5、仅包含上下文特征	第一次睇 3D 电影，献比之前觉得个戏名会磨灭我兴趣同有数个人同我讲吾好睇噶 青蜂侠，我觉得，比想像中好吾止少少 青蜂侠看得我笑了一整场~ 谁能灭得了科比，我还真服了。 果然是科比得全明星 mvp，太棒了 科比—我的神，我果然没看错你！ 班上的 iphone 持有量走上升趋势… 但发觉 iphone 葛 game 始终吾会好似 psp 甘打到吾停手。
6、表达了多种情感，出现的负向情感词不少于正向情感词	本以为春节档的电影就这么毁了，还好<青蜂侠>让人眼前一亮，算是继<让子弹飞>后，一部不错的电影，至于<致命伴旅>，倒还不如去看<危情谍战>了。 鄙视这些站着说话不腰疼的人们，#青蜂侠#很屌，赞。 我不知道 33 岁的科比职业生涯还有多久，但就是这样一个手

	指缠着厚厚的绷带'膝盖没有软骨的科比依然会为胜利和冠军去拼命。 科比果然不负众望，在家乡父老面前拿了 MVP，加油！ 今天与同学讨论乔布斯，虽然乔布斯以前真的不好，但人不是完美的，不能因为一些错误而断定一个人的性格，我当然不是盲目崇拜他，只是他的能力与魅力 。 （这个句子是 negation 算法的问题，不出现在盲目面前，反义为正向，此时 negation 未遇到标点符号，未消解，于是遇上崇拜，反义为不崇拜，为负向） 貌似乔布斯有点不舒服，您一定要挺住啊！ 谁告诉我 iphone5 什么时候出阿 我特想换手机阿 烂夏普 希望 iphone5 不要像 iphone4 一样变成街机了阿 上帝保佑 诺基亚 2630 让我第一次用手机上网，登录 QQ，下载歌曲，忘不了那时的兴奋心情。 连诺基亚都能玩儿坏，真有才啊！ 其实入手诺基亚也没有错，苹果价格偏贵。
7、名词本身无正负向情感，前面加有无就出现了情感，目前的算法中未考虑	今年贺岁档期电影全部看完，就觉得非二，子弹飞和青蜂侠有意思。

表5.15 负向情感消息被误分成正向情感消息原因及示例

原因	例子
1、文中出现正向表情	看了青蜂侠 也很 2[耶]
2、仅有负向情感词典和上下文特征，却未被分为负向情感	事实证明，3D 真的很野鸡，青蜂侠真的很孬种，加藤真的好 superman！ FoGsNoW: 看了乔布斯步履蹒跚的视频，内心不免难过，即便如此，我仿佛看到这个改变世界的人仍在微笑。 哈哈，最讨厌 iphone 的手写短信了！ 分享图片 C7，你到底祸害了多少无知少年，诺基亚，你骗了多少人的银子，还真是在你眨一眨眼间的时候，全球最起码卖了两台诺基亚，今年，你的销售业绩，智能手机领域，你不是老大了哈哈哈。
3、情感词表未覆盖到	看来乔布斯 真的病的很重 苹果危险了 搞错啊，又是科比？ 我还是买了 iphone，信用卡真不是个好东西
4、仅包含上下文特征	诺基亚终于死了，不过建议他们还是为了复活改一下操作系统。 诺基亚是手机，苹果是酷 哇，我白白扔了个 iphone 啊！ 千年道行催白发，塞班倒，诺基亚。
5、表达了多种情感，出现的正向情感词不少于负向情感词	伟大的乔布斯，看着酸酸的！ 诺基亚 N8 抗住了浸水测试，但是在常规跌落测试时出现重新，铝制机身在压力测试中得到了不错的成绩，但是磨损测试并没有多么理想，最后的得分仅有 53 分，总分 100。 骂得好，作为诺基亚的资深用户，这也是我想骂的。 长的还不错，可是诺基亚的机子小毛病好多。 真好笑啊，科比完胜詹姆斯，一个自私自利，一个带动全队力挽狂澜。

分析：从表 5.13 和表 5.14 中可以看出，情感消息误分主要包括几种情况：

①情感词典未覆盖到相关情感此；

②微博消息中表情与文本表达的情感不一致；

③仅包含上下文特征的消息情感较难识别；

④微博消息中表达了多种情感，较难识别；

⑤微博消息中情感是针对多个主题而言，甚至表达的情感与主题词无关；

⑥中文微博消息中存在着英文情感词，本系统尚未考虑。

5.4 主题相关的相关实验

在主题相关的实验部分，由于在主题无关的实验部分得到的最好效果是采用“句子组成 SVM 分类”方法，因此在做主题相关的实验时仍采用句子组成 SVM 分类方法，极性分类特征也选用主题无关的特征，只是选取的句子是满足 4.2.5 中提到的两种情况的句子，结果如下表：

表5.16 主题相关特征的效果

	准确率
仅考虑包含主题词的句子的情感极性	67.183%
+包含零指代主题词的情况的句子的情感极性	**67.283%**

分析：从表 5.15 中可以看出，加入“零指代”后，准确率约提高了 0.1 个百分点（由 67.183%提升到 67.283%）。这比主题无关的最好效果 66.467%提高了 1%左右。

距离窗口算法的结果见图 5. 1：

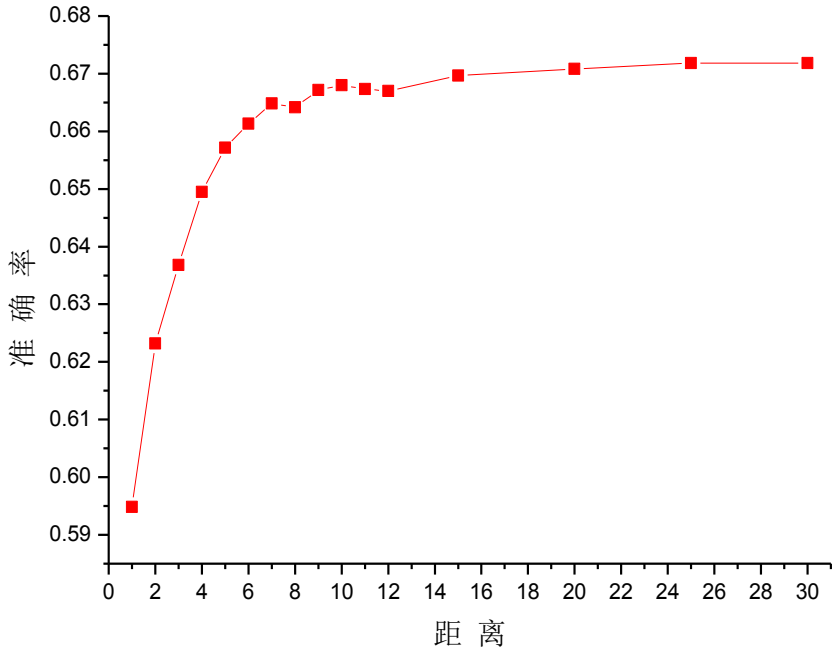

图5.1 不同距离范围内包含主题词的结果

分析：从图 5.1 中可以看出当距离从 1——25 时，准确率基本在持续上升，超过 25 以后，准确率基本不变，这时由于距离较大，退化到每个句子仅包含主题词的情况，因此准确率为 67.183%。由此可以看出，考虑距离窗口的效果较差。

引入主题相关的特征后，进行错误分析，观察原始数据发现，存在以下几个问题：

①包含主题词的句子表达的情感未必是针对该主题词的情感：例如，当主题词为“青蜂侠”时，微博消息“将爱情进行到底比青蜂侠好看多了！”这句话虽然包含了主题词“青蜂侠”，但是“好看”这一正向情感却是针对“将爱情进行到底”的，整条微博消息对于“青蜂侠”是负向情感；又比如主题仍是“青蜂侠”，微博消息“今天跟同学一起去看了青蜂侠，我同学人很好，跟他在一起很开心。”该消息表达的正向情感“好”与“开心”都是针对博主同学而非“青蜂侠”，该微博消息对青蜂侠是中性情感；

②在考虑主题相关的特征时，由于仅简单地考虑了 4.2.4 中定义的包含主题词的句子和零指代的情况，没有考虑其它情况，因此漏掉了其它对主题词表达情感的句子，简单总结，由以下几种情况：（比如主题词是“青蜂侠”）

a）省略宾语：微博消息是“青蜂侠。我觉得挺好看的。”这里“我觉得挺好看的”表达的正向情感也是针对“青蜂侠”；

b）首句省略主语：微博消息“恩恩，不错哦~~青蜂侠”，这里的“恩恩，不错哦~~”的正向情感也是针对“青蜂侠”。

c）微博中的句子未涉及主题词，但涉及主题词的属性，其表达的情感也应该视作对主题词表达情感：微博消息“今天去看了青蜂侠。画面挺炫。”，这里的“画面”是电影“青蜂侠”的一个相关属性，也在评论“青蜂侠”。

5.5 本章小结

本章首先对实验设置进行了说明，包括实验数据、评测方法及评测指标。然后对三种方法进行了评估，再进一步对 SVM 的方法进行了两部分的实验结果分析，即主题无关的相关实验和主体相关的相关实验。

在本章中，我们首先对于基于表情的规则方法、基于词典的规则方法与基于 SVM 方法进行了大致比较，发现基于 SVM 的方法效果最好。于是后续着重对基于 SVM 的方法进行了细致考察，主要进行了主题无关和主题相关的实验。

在主题无关的实验中，本文对方法、特征和核函数极性了比较。在方法选择上：一步三分类方法要好于二步分类法。整体选择句子组成 SVM 分类方法最佳。在特征选择上：对于极性分类，选择表情特征+情感词典特征+情感短语特征+上下文特征最好。对于微博消息的句子组成特征，选择首句极性特征+尾句极性特征+三种情感极性句子数目特征效果最好。在核函数选择上：选择线性核函数效果最好。

在选择最佳组合时，整体的准确率为 66.467%。在本文所选的影视、名人、产品三个相关主题的微博消息情感中，名人主题准确率最高，产品主题准确率最低，在 5.2 节中，本文对目前方法给出了一些错误分析。

在主题相关的实验中，本文对 4.2.5 节中采用的仅包含主题词和零指代的情况进行了实验，发现添加零指代后效果为 67.283%，比仅包含主题词的方法提高

了 0.1%，比主题无关的方法提升了 1%左右，本文在 5.3 小节中对于主题相关的方法给出了一些错误分析。

在下一章中，本文将进行全篇进行总结，指出目前系统存在的问题，并阐明下一步工作。

第6章 结论与展望

6.1 结论

近年来，微博在国内强势崛起，越来越多的公司开始进军微博领域，争相提供微博服务。越来越多的互联网用户在微博平台上注册，通过发布微博来分享信息，关注名人，拓展社交圈。每天有无以计数互联网用户通过微博服务来对热门事件、名人、产品等发表观点，表达情感，微博上的信息量巨大，其潜在价值无可估量。而目前没有任何对中文微博进行情感分析的研究，因此，对于新浪微博的情感分析极为迫切且非常重要。

本文通过从新浪微博开放平台抓取主题相关的数据，请标注人员进行标注，使用标注数据进行实验，采用五折交叉验证，着重从主题无关的特征和主题相关的特征两部分来进行分析探讨，最终使用主题无关特征的方法效果最好，即在做极性分类时，采用表情、情感词典、情感短语、上下文四种特征，方法采用句子组成 SVM 分类，句子组成特征选择选择首句极性特征+尾句极性特征+三种情感极性句子数目：准确率为 66.467%。

目前针对主题相关的特征做得非常初步，主题相关的特征引入后，效果由66.467%提升到 67.283%，但由于没有考虑句法信息和指代情况，未来需要更深入地对主题相关的特征进行研究。

6.2 存在的问题

本文涉及的工作框架主要存在以下问题：

1）目前本系统所采用的情感词典与情感短语的覆盖率有限，导致微博消息中的情感词未识别，从而对分类结果造成影响；

2）由于微博上存在很多网络流行语，如“给力”，“神马”“红人”之类的，这些词无法被现有的分词系统较好的支持，因此现有系统尚未考虑这些网络流行语的识别，这会对分类结果造成影响；

3）由于没有针对每一个句子进行句法分析，因此未能识别出每一句的情感表达的对象是否是针对主题词的，因而对句子的情感分类会出现错误；

4）由于用户在书写新浪微博时的随意性，新浪微博中经常出现主题词被省略的情况，这一现象被称为指代消解，指代消解是一个难度较高的研究课题，本文目前尚未深入研究，这对情感分类的结果造成一些影响。同时有时用户在谈及某部电影，是通过对其属性，例如剧情、画面等属性进行描述的，而不是直接针对主题词进行描述，这也会对结果造成影响。

5）同义现象：通过观察数据，我们发现无论是在判定句子是否主题相关以及采用上下文特征是，都存在词与词之间的同义现象。同义现象分两种，一种是中英文对译的，如(诺基亚，nokia),（iphone，苹果手机）等；还有一种是缩略语的关系，即全称与简称，如（将爱情进行到底，将爱），（北京大学，北大）。在判定主题的时候，如果将主题词进行同义扩展，能够更准确得得到对主题词表达情感的句子，从而提升情感分析的效果。目前本文尚未考虑引入同义扩展，但是本人在硕士期间曾做过缩略语识别的相关研究。主要思路是采用锚文字语料库，根据同义词通常会指向相同链接这一观察来进行了缩略语对的挖掘。实验结果表明：我们共得到了 109,288 对候选缩略语对，在前 500 对上得到的准确率为 91.4%，Bpref 值为 0.967。

6.3 下一步的工作

本文主要有下面的可改进和进一步的研究方向：

1）提高情感词典的覆盖率，从而能尽可能多地识别出情感词；

2）构建网络用语词典，针对这类型的词，由于无法借助现有的分词系统，需要采用新算法匹配识别，包括否定转移的处理；

3）更深入地研究主题相关的特征，主要包括两方面：

①使用句法解析工具对每个句子进行句法解析，识别真正跟主题相关的情感；

②考虑更丰富的指代情况，如省略宾语和主题词相关属性指代等；

4）考虑将缩略语识别的技术引入，来实现同义扩展，从而更好的进行情感分析；

5）考虑引入微博中人与人之间的社交网络关系或者消息与消息指间的关系来构建图模型进一步提升结果。

附录

本论文中构建情感词典时使用的 100 个常用的正向情感词和负向情感词示例：

编号	正向情感词	负向情感词
1	希望	差
2	喜欢	贵
3	支持	坏
4	不错	讨厌
5	快乐	垃圾
6	谢谢	反对
7	推荐	严重
8	开心	俗
9	民主	无聊
10	幸福	麻烦
11	成功	悲剧
12	欢迎	失败
13	重要	暴力
14	感谢	恐怖
15	期待	郁闷
16	才能	非法
17	可爱	错误
18	精神	恶心
19	机会	批评
20	邀请	贪
21	努力	流氓

22	解决	疯狂
23	美女	不满
24	方便	失望
25	肯定	浪费
26	好看	不对
27	雅	危机
28	和谐	崩溃
29	开放	低俗
30	理解	紧张
31	完成	颠覆
32	清楚	无耻
33	祝福	不爽
34	保护	邪恶
35	值得	威胁
36	漂亮	恐惧
37	帮忙	伤害
38	敏感	绝望
39	独立	后悔
40	稳定	难过
41	适合	不良
42	舒服	伤心
43	平安	生气
44	喜爱	困难
45	强大	难受
46	恢复	错过
47	感动	鄙视
48	纪念	糟糕
49	经典	遗憾
50	恭喜	变态
51	伟大	涉嫌

52	美好	破坏
53	有趣	黑暗
54	优惠	腐败
55	认真	污染
56	美丽	抱怨
57	顺利	谎言
58	和平	孤独
59	理想	不公
60	温暖	白痴
61	有效	指责
62	流行	不幸
63	完美	抄袭
64	精彩	煽动
65	英雄	故障
66	进步	敌人
67	亲爱	寒冷
68	正确	无视
69	天才	尴尬
70	满意	受害
71	文明	反动
72	创意	可恶
73	好听	诱惑
74	聪明	烦恼
75	好意	混乱
76	梦想	地狱
77	真理	小人
78	爱国	该死
79	创造	侵犯
80	正义	绑架
81	满足	吵架

82	怀念	奴才
83	完整	歧视
84	羡慕	迫害
85	赞赏	坏人
86	解放	讽刺
87	难得	末日
88	好人	谣言
89	尊重	悲伤
90	多谢	欺负
91	成长	难看
92	轻松	猥琐
93	兴奋	折磨
94	好事	罪恶
95	熟悉	罪名
96	优秀	落后
97	吸引	不和
98	勇气	损失
99	特色	暴露
100	创新	诈骗

插图索引

表格索引

參考文獻

[1] M.Q. Hu and B. Liu, 2004. Mining and Summarizing Customer Reviews, In ACM SIGKDD 2004, pp.168-177.

[2] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2): 1-135.

[3] M.Q. Hu and B. Liu. 2006. Opinion Extraction and Summarization on the Web. In AAAI06, Boston, pp. 1621-1624.

[4] H. Yu and V. Hatzivassiloglou. 2003. Towards Answering Opinion Question: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, In EMNLP'03, pp.129-136.

[5] Long Jiang, Mo Yu, Ming Zhou and Xiaohua Liu. 2011. Target-dependent Twitter Sentiment Classification. ACL 2011.

[6] Binyang Li, Lanjun Zhou, Shi Feng, Kam-Fai Wong. 2010. A Unified Graph Model for Sentence-based Opinion Retrieval. In ACL 2011, pp. 1367-1375.

[7] Lanjun Zhou, Yunqing Xia, Binyang Li and Kam-Fai Wong. 2010. WIA-Opinmine System in NTCIR-8 MOAT Evaluation. NTCIR-8 Workshop Meeting, June 15–18, 2010, Tokyo, Japan, pp 286-292.

[8] Lun-Wei Ku, Tung-Ho Wu, Li-Ying Lee, and Hsin-Hsi Chen. 2005. Construction of an Evaluation Corpus for Opinion Extraction, In NTCIR-5, pp.513-520, Japan, 2005.

[9] V. Hatzivassiloglou and J. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In COLING'00, pp. 299-305.

[10] P. Turney. 2002. Thumbs up or thumbs down? Semantic Orientate-on Applied to Unsupervised Classification of reviews. In ACL'02, pp. 417-424.

[11] B. Pang, L.L. Lee and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In EMNLP'02, pp.79-86.

[12] S. Dasgupta and V. Ng. Mine the Easy. 2009. Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification. In ACL'09, pp. 701-709.

[13] Tetsuya Nasukawa, Jeonghee Yi. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In Proceedings of the 2nd International Conference on Knowledge Capture, pp. 70-77.

[14] Xiaowen Ding and Bing Liu. 2007. The Utility of Linguistic Rules in Opinion Mining. In SIGIR-2007 (poster paper), e, pp. 811-812.

[15] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA, Aug 22-25, pp.168-177.

[16] Agrawal, R. & Srikant, R. 1994. Fast algorithm for mining association rules. In VLDB'94.

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

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

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

[20] Ravi Parikh and Matin Movassate. 2009. Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques. CS224N Final Report, pp.1-18.

[21] 中金公司发研究报告：新浪微博竞争优势可持续。Available at http://tech.sina.com.cn/i/2011-03-01/18305231638.shtml

[22] 张姝, 贾文杰, 夏迎炬, 孟遥, 于浩. 2008. 基于 CRF 的评价对象抽取技术研究. 第一届中文倾向性分析评测(COAE 2008)论文集, pp. 70-76.

[23] 徐冰, 王山雨. 2009. 句子级文本倾向性分析评测报告. 第二届中文倾向性分析评测(COAE 2009)论文集, pp. 69-73.

致　谢

首先感谢导师孙茂松教授和微软亚洲研究院周明博士对我的悉心指导，孙老师及周明博士严谨治学的态度与循循善诱的教导，使我受益匪浅。

其次，我要感谢人工智能实验室的马少平、朱小燕等老师的指导，他们在开题报告中给予了我宝贵的意见和建议，开阔了我的思路，让我更好的改进和完善毕业设计。

同时，我还要感谢清华实验室及微软亚洲研究院的师兄师姐，是他们勤奋好学的态度以及无微不至的关心感染了我、鼓励了我。他们营造了一个非常融洽的学术讨论氛围，并给予我宝贵意见，还给我提供了重要的资源，使我在毕业设计过程中能更快更好地解决问题。

最后我还要感谢父母对我的爱护和教育，感谢周围同学给予我的帮助、关心和支持。

本论文是基于作者在微软亚洲研究院期间的实习工作完成的。同时本课题承蒙国家自然科学基金资助，基金号为 No. 60873174，特此致谢。

此外，本人在硕士期间的汉语缩略语识别的研究工作受到清华——搜狐搜索技术联合实验室项目的支持，特此致谢。尤其感谢搜狐的研发人员佟子健、王灿辉、张杨、杨磊等的悉心指导及共同讨论。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签名：_____日期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1987 年 5 月 7 日出生于湖南省湘西州永顺县。

2005 年 8 月考入清华大学计算机科学与技术系计算机科学与技术系，2009 年 7 月本科毕业并获得工学学士学位。

2009年9月免试进入清华大学计算机科学与技术系攻读计算机科学与技术系，攻读工学硕士至今。

在学期间发表的发表的学术论文及研究成果

论文：

[1] **Lixing Xie**, Yabin Zheng, Zhiyuan Liu, Maosong Sun, Canhui Wang. Extracting Chinese Abbreviation-definition Pairs from Anchor Texts. Accepted by the 10th International Conference on Machine Learning and Cybernetics. ICMLC 2011. Regular paper. (EI 会议).

[2] **谢丽星**、孙茂松、佟子健、王灿辉。基于用户查询日志和锚文字的汉语缩略语识别。全国第十届计算语言学学术会议 (CNCCL-2009), pp. 551-556.。

[3] Yabin Zheng, Zhiyuan Liu, **Lixing Xie**. Growing Related Words from Seed via User Behaviors: A Re-ranking Based Approach. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (ACL-SRW 2010), 2010, 49-54. Poster Paper.

[4] Yabin Zheng, **Lixing Xie**, Zhiyuan Liu, Maosong Sun, Yang Zhang, Liyun Ru. Why Press Backspace? Understanding User Input Behaviors in Chinese Pinyin Input Method. Accepted by Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), 2011. Short Paper.

[5] Zhiyuan Liu, Yabin Zheng, **Lixing Xie**, Maosong Sun, Liyun Ru, Yang Zhang. User Behaviors in Related Word Retrieval and New Word Detection: A Collaborative Perspective. Accepted by the ACM Transactions on Asian Language Information Processing (Special Issue on Chinese Language Processing), April 2011. Regular Paper. (EI 刊源)

专利：

谢丽星、孙茂松、佟子健、王灿辉。汉语缩略语处理方法及装置。专利申请号：200910088377.6。公开号：CN101599075。公开日期：2009.12.09。

软件登记：

司宪策, 郑亚斌, 李景阳, 孙茂松, **谢丽星**。中英文文本分类软件, 软件登记号 2009SRBJ0174, 2009 (SEWM 大规模网页分类评测第一名)