文章编号：1003-0077(2012)01-0073-11

基于层次结构的多策略中文微博情感分析和特征抽取

谢丽星[1]，周 明[2]，孙茂松[1]

（1. 智能技术与系统国家重点实验室；清华信息科学与技术国家实验室(筹)；清华大学 计算机系，北京 100084；
2. 微软亚洲研究院，北京 100084）

摘 要：随着 Web2.0 时代的兴起，与微博相关的研究得到了学术界和工业界的广泛关注。该文使用新浪 API 获取数据，针对中文微博消息展开了情感分析方面的研究。我们对于三种情感分析的方法进行了深入研究，包括表情符号的规则方法、情感词典的规则方法、基于 SVM 的层次结构的多策略方法，实验表明基于 SVM 的层次结构多策略方法效果最好。其次，针对层次结构的多策略方法的特征选择进行了详细分析，包括主题无关、主题相关的特征。实验表明使用主题无关的特征时获得的准确率为 66.467%。引入主题相关的特征后，准确率提升至 67.283%。

关键词：新浪微博；情感分析；SVM

中图分类号：TP391 **文献标识码**：A

Hierarchical Structure Based Hybrid Approach to Sentiment Analysis of Chinese Micro Blog and Its Feature Extraction

XIE Lixing[1], ZHOU Ming[2], SUN Maosong[1]

(1. State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China; 2. Microsoft Research Asia, Beijing 100084, China)

Abstract: With the development of Web 2.0, micro blog has drawn substantial attention from both academia and industry communities. This paper utilizes micro blog API from Sina and carries out sentiment analysis on Chinese micro blog. We compare performances of three method, based on the emoticon, the sentiment lexicon and the hybrid approach over hierarchical structure using SVM, respectively. Through the experiments, we find that SVM based hybrid approach achieves the best performance. Furthermore, we analyze the contribution of various features in this model, including target-independent features and target-dependent features. Experimental results show that SVM based method can gain an accuracy of 66.467% with target-independent features, and an improved accuracy of 67.283% with the addition of target-dependent features.

Key words: sina micro blog; sentiment analysis; SVM

1 引言

微博，是一种新的信息发布及社交网络平台。用户注册微博服务后，可以关注名人、结交朋友，还能随意发表、回复及评论消息，来记录生活、分享心情、表达观点等。微博自问世以来，迅速吸引了大众的眼光，蓬勃发展。以国内的新浪微博[①]为例，截止

① Available at http://weibo.com/

收稿日期：2011-06-10 定稿日期：2011-09-18

基金项目：国家自然科学基金资助项目(60873174)

作者简介：谢丽星(1987—)，女，硕士，主要研究方向为缩略语识别、输入法和中文微博的情感分析；孙茂松(1962—)，男，博士，清华大学计算机系教授，博士生导师，主要研究方向为自然语言处理、信息检索和社会计算；周明(1964—)，男，博士，微软亚洲研究院主任研究员，博士生导师，主要研究方向为自然语言处理、机器翻译、搜索引擎和社会关系网络。

到2011年4月底，用户数超过了1.4亿。微博正在从各个方面渗透并影响人们的生活，包括大量的信息传播、更快的信息发现、与世界的连接等。

微博消息数量大，更新快，吸引了一大批学者对其进行研究。针对微博的自然语言处理研究已成为当前一个新的研究热点和前沿课题，而情感分析就是其中一个热点话题。情感分析，也被称为观点挖掘、观点分析、主客观分析等。情感分析的目的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，既能吸引潜在用户，帮助用户做决策[1]，又能得到产品反馈[2]，还能对政治选举等重大事件进行预测。除此以外，情感分析的技术还有助于自然语言处理领域其他研究方面的发展，例如，自动文本摘要[3]及问答系统[4]等。目前已经有许多针对英文的(即新闻、博客、微博等)情感分析系统相继问世[5-11]。在情感分析方面，主要使用的技术分两大类：一类是采用情感词典与规则相结合的方法，根据文本中所包含的正向情感词和负向情感词的个数来进行情感分类；另一类是采用机器学习的方法，选择文本中的一些特征，标注训练集和测试集，使用朴素贝叶斯(Naïve Bayes)、最大熵(Max Entropy)、支持向量机(Support Vector Machine)等分类器来进行情感分类。

微博作为一种新型的消息传递方式，写作简洁，与传统文本差异较大，主要表现在五个方面：①主题发散性。在传统文本中，文本所涉及的内容通常在同一主题下，主题较为集中，通篇会不断的提到主题词。但在微博中，主题较为发散。例如，这条微博“今天看了青蜂侠。我们后来还一起唱了KTV，真是开心的一天。”该微博中第一句的主题词是“青蜂侠”，第二句涉及的内容与“青蜂侠”无关。这是主题发散的表现；②省略成分的(主谓宾)。在传统文本中，用户使用的语法较为规范，而在微博中经常会出现省略主语、宾语的现象，例如，“青蜂侠不错！真好看啊！”该微博的第二句省略了主语“青蜂侠”；③省略上下文的(上下文隐含在其他的微博信息或者交互的信息中)。在微博中一条微博如果是回复某一条微博的内容，由于是对话的形式，通常会省略掉主题。例如，一条微博是“凤姐真讨厌！”另一条回复该微博的微博是“确实如此。”这两条微博讨论的都是“凤姐”，但是第二条微博省略了“凤姐”相关的上下文；④口语化：在微博中用户经常会使用时下流行的热门词，口语化的词等，例如，“坑爹”，“神马”，“不咋地”等；⑤包含链接、表情符号及标签等信息。微博中用户经常分享网页链接，使用表情符号表征情感，也经常给文本打上标签，标签可以反映主题。由于上述差别，使得微博情感分析任务更为复杂，也使得传统文本的分析方法无法适用，例如，传统文本中通常将一段文本内容表达的情感视为针对同一主题，但是微博中可能是多主题，使用单一主题的情感分析技术可能会造成失误；同时微博中口语化的词较多，如果口语词表达了情感，而传统文本的分析方法无法识别这类情感，也会对情感分析造成影响。因此针对中文微博的情感分析方面的研究显得尤为迫切和重要。

在微博情感分析中，目前英文微博相关的研究已经有了一些进展，例如，针对英文微博自身包含的属性如表情符号，标签(hashtag)等作为特征对微博进行情感分类，而针对微博的主题发散性，也有学者从主题无关和主题相关两方面进行分析。但到目前为止，针对中文微博的研究仍处于起步阶段，而中文微博与英文微博有很大不同：英文微博限制用户的输入文本不超过140个字符，这通常是一个句子，包含7～10个英文单词，涉及的主题和情感相对一致；而中文微博限制用户的输入文本不超过140个中文字符，这可以包含多个句子，每个句子涉及的主题可能不同，表达情感也可能不同。例如，在如下微博中“今天看了青蜂侠，很一般。场面一般，剧情一般。不过杰伦还是那样帅。”这条微博前两句是对电影“青蜂侠”的负面评论，第三句是对青蜂侠的扮演者的正面评论，表达的主题和情感均不同。

本文主要研究中文微博的情感分析。由于此前相关研究并不多，我们在研究中借鉴了普通文本情感分析的方法。在普通文本的情感分析方面，主要有两类任务：主题无关的情感分析和主题相关的情感分析。主题无关的情感分析不需要考虑待分析文本的评价对象，给出一个情感极性即可；主题相关的情感分析需要考虑待分析文本的评价对象，给出待分析文本针对该评价对象的情感极性。受此启发，本文将从主题无关和主题相关两个方面抽取特征，并应用于基于层次策略的中文微博情感分析。本文通过从新浪微博开放平台提供的API[①]抓取一定规模的数据，对中文微博的情感分析进行了研究。本文研究的输入为给定主题词及中文微博消息，如主题词：“科比”，中文微波消息：“科比太酷了!!!

① Available at http://open.t.sina.com.cn/wiki/index.php/Trends/statuses

[抓狂][爱你]”；输出为该条微博消息针对该主题词的情感，包括正向情感、负向情感、中性情感三种。对于该例，系统输出为正向情感。针对中文微博的情感分析，本文采用了二步法，首先引入主题无关特征，即使用链接、表情符号、情感词典、情感短语、上下文等特征训练 SVM 对中文微博进行情感分类；然后进一步引入主题相关的特征，即筛选微博中与主题相关的句子来进一步提升效果。本文主要的贡献有：①提出了基于层次结构的多策略分析框架；②中文微博特征的研究：研究了链接、表情符号等特征对于中文微博的有效性；同时提出了微博消息的句子构成特征。

本文的结构组织如下：第二章简单介绍相关工作；第三章阐述算法设计；第四章展示实验结果及相关分析；第五章简单地讨论本文与以往工作的区别；第六章是结论及下一步工作。

2　相关工作

微博，实时提供短消息的播放，作为数以亿计用户的日常信息沟通和人际交流的重要工具，已经成为互联网的新形态。关于微博的情感分析已成为时下热门话题，一系列研究就此展开。本章节将从针对英文的情感分析和针对中文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

本节将从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个方面的研究工作来进行介绍。

2.1.1　主题无关的情感

主题无关的情感分析是对指定文本给出情感极性，而不关心该情感极性所描述的对象。目前大多数情感分析方面的研究都是主题无关的，主要有三种方法：基于词典的方法、有监督的机器学习方法、无监督的方法。

基于词典的方法[12]。这类方法首先需要构建一个情感词典，主要包括正向情感词和负向情感词。然后利用情感词典统计待分析的文本中的正向情感词的数目和负向情感词的数目。最后依据它们的差值来进行情感极性的判定，即差值为正即为正向情感，差值为负为负向情感，差值为零为中性情感。情感词典的方法的局限性在于无法解决未登录词的问题。

有监督的机器学习方法[5,13-14]。这类方法主要是使用机器学习的模型，包括朴素贝叶斯（Naïve Bayes）、最大熵（Max Entropy）、支持向量机（Support Vector Machine）等来对文本进行情感分析。Pang 等人[5]的研究工作主要是对电影评论进行情感极性的分类，分为正向情感和负向情感。该工作首先对待分析的文本进行预处理，提取出若干特征，包含一元词特征（unigram）、二元词特征（bigram）、词性标注、词的位置信息等，然后使用这些特征来训练模型，选用的方法有朴素贝叶斯、最大熵、支持向量机。实验结果表明，支持向量机的效果最为理想，且在选用一元词特征时取得了最好的准确率为 83%。Li 等人[14]对于评论数据，首先提出了用于情感二分类的 Dependency-Sentiment-LDA 模型，它在情感分类的时候不仅考虑了情感词所表达的话题语境，而且还考虑了情感词的局部依赖关系。然后进一步探讨了情感多分类问题，提出了一种基于 Tensor 的评论分值预测方法。基于有监督学习方法精度较高，缺陷是依赖于人工标注语料库，语料库标注存在不一致性问题。

无监督的学习方法[15-16]。Turney 等人[16]对于手机、银行、电影及旅游目的地相关的评论的情感分析工作。他们选定了两个基本情感词（正向词：excellent，负向词：poor），然后他们制定了一些模版来提取短语，使用 PMI 分别计算待分析的文本中这些短语与基本正向情感词的关联度（记为正向关联度）和负向情感词的关联度（记为负向关联度），根据正向关联度与负向关联度的差值来判定该文本的情感极性。无监督的方法依赖于处理语料的领域范围，存在着对基准情感词的依赖性问题，正确率较低。

2.2.2　主题相关的情感分析

主题相关的情感分析主要包括两种方法，基于规则的方法和基于特征（或属性）的方法。

基于规则的方法[17-18]。这类方法主要是对文本进行预处理，包括词性标注、依存句法分析等，然后针对形容词、动词、名词等制定一些规则来对该文本判定情感极性。代表工作为 Nasukawa 和 Yi[17]的工作。

基于特征（属性）的方法[19]。这类工作除了需要对文本进行情感极性的判定，还需要按照产品的属性进行归类。代表工作为 Hu 和 Liu[19]针对用户对在线产品的评论进行的情感分析工作。他们的方法主要包含三步。首先识别出用户评论中涉及的产品属性；其次针对每个属性，得出评论中包含的正向

情感和负向情感的内容;最后将属性与对应情感极性的内容按某种形式输出。

2.2.3 针对英文微博的情感分析

对于微博的情感分析的研究主要是针对 Twitter① 上的消息 Tweets 而言的,本节将从主题无关和主题相关两方面进行介绍。

主题无关的情感分析[10,20-22]。Davidiv 等人[20]使用 Tweets 中的标签、表情符号等作为特征,训练了一个类似 KNN 的分类器来进行情感极性的分类;Barbosa 和 Feng[10]利用一些网站(即 Twendz、Twitter Sentiment、TweetFeel)对于 Tweets 所提供的情感分析的结果作为训练数据,然后选用一些特征,采用二步分类法来对 Tweets 进行分类,即先对 Tweets 进行主客观分类,然后再在被分为主观的 Tweets 中进行正、负向情感分类。

主题相关的情感分析。Jiang 等人[11]对 Tweets 的情感分类采用二步分类法,首先对 Tweets 进行主、客观分类,然后再对被分为主观的 Tweets 进行正、负向情感分类。与其他工作不同的是,Jiang 等人在分类时除了考虑了主题词,还对主题词进行了扩展,引入了主题相关的特征,此外还考虑 Tweets 间的转发关系,采用图模型的方法提升效果。未使用图模型之前,系统取得的最好准确率为 66%,引入图模型之后,系统的准确率提升到 68.3%。

2.2 针对中文的情感分析

目前针对中文的情感分析主要集中在 NTCIR② 和 COAE③ 两个评测上。

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。该评测主要包括六项任务,主客观判别、相关性判别、观点持有对象抽取、观点评价对象抽取、情感极性判别、问答系统。在 NTCIR-08 中,针对繁、简体中文,在主客观判别及情感极性判别这两项任务,评测的最好结果见表 1。

表 1 NTCIR-08 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%

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,情感词的识别及分类、情感句的识别及分类、观点句抽取、观点评价对象抽取、观点检索。在 COAE-09 中,极性判别的最好结果见表 2。

表 2 COAE-09 中情感极性判别的最好结果

裁判员	P@1000	Precision	Recall	F1	Accuracy-1000
1	0.662	0.662	0.158 033	0.255 155	0.158 033
2	0.612	0.612	0.153 268	0.245 143	0.153 268
3	0.544	0.544	0.149 986	0.235 142	0.149 986

总的来说,中文的情感分析方法与英文类似,大致有两种。

① 有监督的机器学习方法[23]。Zhao 等人[23]基于 CRF 模型引入“冗余特征”来研究情感分类问题,Zhou 等人。基于 SVM 模型来进行主客观及情感极性分类;

② 组合方法[25]。Li 等人[25]研究了该文具体研究四种不同的分类方法在中文情感分类上的应用,同时考虑到不同领域需要选择不同基分类方法才能获得更好的分类结果,采用一种基于 Stacking 的组合分类方法,用以组合不同的分类方法。

目前针对中文的情感分析较针对英文的情感分析无论从资源还是方法上来说都要相对初步一些。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,例如,“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;

① Available at twitter.com

② Available at http://research.nii.ac.jp/ntcir/

③ Available at http://www.ir-china.org.cn/Information.html

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。

3　算法设计

这章我们将介绍三种方法，分别是基于表情符号、情感词典、SVM 的层次结构的多策略方法。

3.1　基于表情符号的规则方法

在新浪微博上，微博平台提供了一些默认的表情符号，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多个表情符号。

本文针对新浪微博提供的表情符号进行了正、负向表情符号的分类，然后对于待分析的文本，从中提取中正、负向表情符号，依据公式(1)进行情感极性的分类：

情感极性＝

$$\begin{cases} \text{正向情感(如果正向表情符号数} > \text{负向表情符号数)} \\ \text{负向情感(如果正向表情符号数} < \text{负向表情符号数)} \\ \text{中性情感(如果正向表情符号数} = \text{负向表情符号数)} \end{cases} \quad (1)$$

3.2　基于情感词典的规则方法

情感词是情感极性判定中较为重要的考量依据。本文借鉴了传统的情感分析的方法，选取了常用的正、负向情感词构建了情感词典。在构建词典时，我们只选用了在任何情况下都绝对表征正、负向情感的词，如正向情感词“喜欢”、负向情感词“憎恨”。然后将待分析的文本进行了分词处理，依据该情感词典从中提取出正、负向情感词，依据公式(2)进行情感极性的分类。

情感极性＝

$$\begin{cases} \text{正向情感(如果正向情感词数} > \text{负向情感词数)} \\ \text{负向情感(如果正向情感词数} < \text{负向情感词数)} \\ \text{中性情感(如果正向情感词数} = \text{负向情感词数)} \end{cases} \quad (2)$$

3.3　基于层次结构的多策略分析框架

3.3.1　方法介绍

本文提出了基于层次结构的多策略分析框架，见图 1(下页)。使用的分类工具是 SVM(Support Vector Machine)，中文名为支持向量机，是由 Vapnik 等人提出的一种非常有潜力的学习技术，是一种基于统计机器学习理论的模式识别方法，主要用于模式识别领域。本文使用的 SVM 工具是由台湾大学林智仁 (Chih-Jen Lin) 博士等开发的一套支持向量机算法库 libsvm[①]。

表 3　中文微博消息包含不同情感极性句子的示例

主题词：青蜂侠
微博消息： 好失望啊！ 今天没买到致命伴旅的票。 不过还好，看了青蜂侠，杰伦好帅！

在前人针对英文微博的研究工作中，由于英文微博消息文本长度被限制在 140 个英文字符，这通常是一个句子，包含 7～10 个英文单词，因此之前的所有工作都是将一条微博消息当做一个整体来进行训练和测试。通过观察中文微博数据数据，我们发现中文微博文本长度被限制在 140 个中文字符，它可以包含多个句子，与英文微博相比语义更丰富，句与句之间的情感极性不尽相同。如表 3 所示，针对主题词“青蜂侠”的微博消息共包含三句，首句的情感极性是负向情感，第二句是中性情感，第三句是正向情感。如果将不同极性的句子作为一个整体赋予一种极性，也许会影响训练效果。因此在使用 SVM 对微博消息进行文本分类时，基于分句与不分句的考量分为两大类策略，共四种方法，见图 1(下页)。

下面我们对这两大类策略、四种方法进行简单介绍。

第一类：不分句，将一条中文微博消息看做一个整体，该微博的情感极性被视为一致，有两种方法。

① 一步三分类。从微博消息中提取极性分类特征，然后根据每条微博的正、负、中性情感标签，直接训练一个三分类的 SVM 分类器，对微博消息三分类；

② 二步分类。先从微博消息中提取主、客观分类特征，根据每条微博的主、客观标注情况，训练主、客观分类的 SVM 分类器，先对微博消息进行主、客观分类；然后对于分为主观的微博消息，再从微博消息中提取极性分类特征，根据每条微博的正、负向情感标签训练正、负向情感的 SVM 分类器，进一步将

① C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. Available at http://www.csie.ntu.edu.tw/～cjlin/libsvm/.

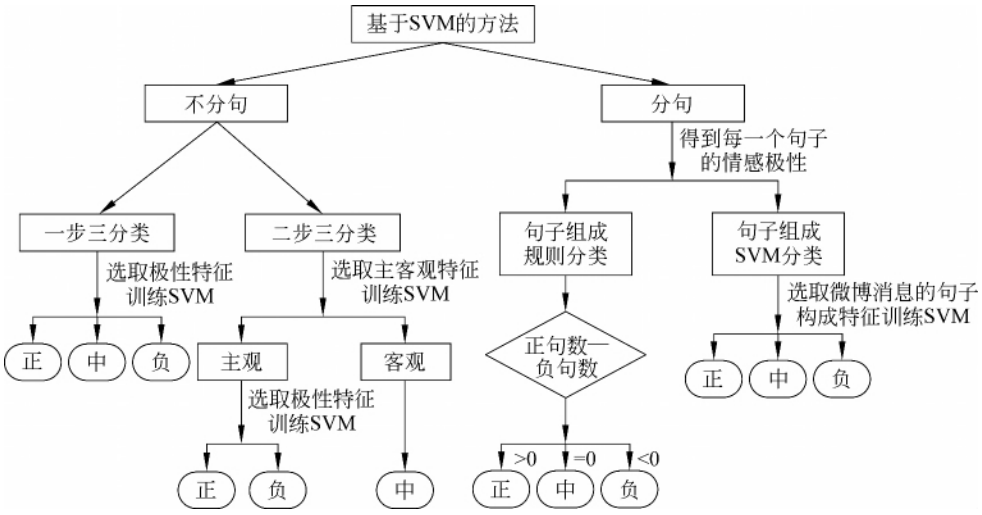

图1 基于层次结构的多策略情感分析框架

主观的微博消息分为正、负向情感；

第二类：分句，将一条中文微博消息使用程序自动拆分成若干个句子，然后将针对每个句子进行训练，有两类方法。

① 句子组成规则分类。首先使用上述两种方法中的最佳方法训练SVM分类器得到每一条微博消息中每个句子的情感极性，然后根据正、负向句子的个数，依据公式(3)对微博消息进行三分类；

$$\text{情感极性}=\begin{cases}\text{正向情感(如果正向情感句子数}>\text{负向情感句子数)}\\ \text{负向情感(如果正向情感句子数}<\text{负向情感句子数)}\\ \text{中性情感(如果正向情感句子数}=\text{负向情感句子数)}\end{cases} \tag{3}$$

② 句子组成SVM分类。首先使用上述两种方法中的最佳方法训练SVM分类器得到每一条微博消息中每个句子的情感极性，然后选取微博消息的句子构成特征，结合每条微博的情感极性再次训练SVM，将每条微博进行三分类。

3.3.2 主题无关的特征抽取

在上一节中，涉及三类特征，主、客观分类特征、极性分类特征、微博消息的句子构成特征，这里将一一详述。

(1) 主、客观分类特征(表4)。情感短语特征指的是“有意思”，“没文化”这类短语，它们的特点是中心词“意思”和“文化”本身是中性，但是前面出现了“有”、“无”这一类修饰词就含有极性色彩了。

(2) 极性分类特征(表5)。

(3) 微博消息的句子构成特征(表6)。考虑到中国人书写文章时“开门见山”的习惯，以及“首尾呼

表4 主、客观分类特征

序号	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	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：4 751个 负向情感词：3 651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80 224个词；分词采用ICTCLAS进行分词

表5 极性分类特征

序号	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	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(使用否定词表进行了情感转换判定)	正向情感词：4 751个 负向情感词：3 651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80 224个词；分词采用ICTCLAS进行分词

应”的句式，因此除了考虑正、负、中性情感句子数目以外，这里还考虑了首句、尾句的情感极性。

表 6　微博消息的句子构成特征

序号	类型	特征内容
1	首句的情感极性	首句的情感极性：正、中、负向情感
2	尾句的情感极性	尾句的情感极性：正、中、负向情感
3	正向情感句子数	正向情感句子的数目
4	负向情感句子数	负向情感句子的数目
5	中性情感句子数	中性情感句子的数目

3.3.3　主题相关的特征抽取

通过观察，我们发现中文微博消息不像电影、产品评论那样集中一个主题讨论，微博消息中存在着大量的主题发散及省略现象。如表 7 所示，针对主题词“将爱情进行到底”的微博消息共包含 5 个句子。首先句子 1 和句子 2 涉及的主题与主题词“将爱情进行到底”无关，这说明该消息存在主题发散的情况。其次句子 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”，这说明该消息存在省略主题词的情况。

表 7　中文微博消息中的主题发散及省略示例

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。

因此，在使用微博消息的句子构成特征时，在筛选句子的时候，需要进行主题相关的句子的筛选，具体考虑以下三种情况。

① 仅考虑包含主题词的句子的情感极性；

② 零指代的情况。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，即认为它表达的情感是针对上一句的对象，如上一句包含主题词，则也应该考虑该句的情感极性；

③ 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，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与主题相关。

4　实验结果及相关分析

4.1　实验设置

本文使用新浪提供的 API 抓取了影视、名人、产品三个领域，共六个主题的数据，最后每个主题选取了 1 000 条微博消息进行标注，结果见表 8。

评测方法选用的是五折交叉验证（five-fold cross-validation）。评测指标主要使用的是准确率，即 5 次迭代正确分类的总数除以原始数据中的元组总数。

表 8　中文微博消息中的主题发散及省略示例

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	1 000	58.90%
	青蜂侠	270	183	547	1 000	54.70%
名人	科比	645	79	276	1 000	27.60%
	乔布斯	329	33	638	1 000	63.80%
产品	iphone	188	83	729	1 000	72.90%
	诺基亚	203	213	584	1 000	58.40%
共计		1 971	666	3 363	6 000	56.05%

4.2　实验结果及分析

（1）我们首先对于三种方法进行比较，实验结果见表 9。

表 9　中文微博消息中的主题发散及省略示例

	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 8 可以看出，基于 SVM 的方法效果最好，基于表情符号的规则方法略好于基于情感词典的方法，准确率均在 56%左右。

（2）SVM 相关的实验

① 方法比较：这里我们对 3.3.1 节中提到的四种方法进行比较。从表 10 可以看出，一步三分类方法要比二步分类方法高出 1.5 个百分点。因此在后续实验得到句子级别的情感极性时采用一步三分类法进行训练和测试。从表 11 可以看出，采用句子组成 SVM 分类法效果好于句子组成规则分类方法。

表10 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	63.866%

表11 句子组成规则和句子组成SVM分类方法的效果比较

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

② 主题无关的特征比较。这里我们仅对极性特征、微博消息的句子构成特征进行比较分析。对于极性特征，从表12可以得出两点结论：(a)从有效性来看，上下文>情感词典>表情>情感短语，引入链接特征后效果反而变差；(b)最佳特征组合：表情+情感+情感短语+上下文。

表12 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%

对于微博消息的句子构成特征，从表13中可以得出两点结论：(a)从有效性来看，三种情感极性句子数目>尾句情感极性>首句情感极性；(b)最佳特征组合：首句极性+尾句极性+三种情感极性句子数目。

表13 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%

综上所述，在仅考虑主题无关的特征时，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。

我们采用“最佳组合”对数据进行了训练测试，得到结果见表14。从表14中可以看出，主题无关的最佳方法的总体准确率达到了66.467%。根据准确率从高到低排序，各个领域依次为：名人>影视>产品。通过对数据进行分析，我们得到了以下三点原因。

① 名人。用户对名人发表观点的时候使用的表达较为单一；

② 影视。用户关注的主题更为发散，如主角、画面等，分类时比名人要难；

③ 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，分类问题更为复杂。

表14 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	1 000	67.400%
	青蜂侠	1 000	66.800%
名人	科比	1 000	72.900%
	乔布斯	1 000	69.400%
产品	iphone	1 000	69.800%
	诺基亚	1 000	61.200%
共计		6 000	66.467%

同时，针对主题无关的实验，我们还进行了正、负、中向情感的精确率和召回率的统计，见表15。表中显示从评测指标来看，中性情感>正向情感>负向情感。这跟标注数据中各种情感极性的消息数目比例成正比。

表15 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%

进一步，我们对于主题无关特征最佳组合的错误类型进行了分析，结果见表16。

表 16　主题无关特征最佳组合下的错误类型

序号	类型	示例
1	情感词表未覆盖	将爱情进行到底真纠结。 青蜂侠看的反胃。
2	使用的词是口语化的词或网络热词	将爱，不咋地。 要不是为了看 JAY，青蜂侠真的是挺坑爹的。
3	误识别	今天去看了将爱，看的心情好差。
4	不含情感词	＃青蜂侠＃只有一个念头，让编剧回家吧。 第二次看青蜂侠看不下去＝＝！ 看了＜青蜂侠＞，周杰伦你的英文敢再闹太套点么！
5	反讽	作为一部动作片能拍到让我看着想睡着青蜂侠在某个意义上也是成功了…
6	情感词类型较多	青蜂侠，一堆不知叫什么的什么，很无聊，本来以为很好看，现在好啦，浪费时间，浪费期望值
7	英文情感词	青蜂侠 istupid[汗]
8	有正向词，但是语境是相反意思	但是，随着 iphone 和 Android 的冲击，无论是高端还是中低端市场，诺基亚的优势正在全面失去。

③ 主题相关的特征比较。首先我们考虑仅包含主题词的句子及零指代的情况，见表 17，从中可以看出引入零指代后，效果提升了约 0.1%，比主题无关的最好效果(66.467%)提升了 1%左右。

表 17　主题相关特征的效果比较

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

其次，考虑距离窗口方法，得到图 2。由图可知，考虑距离窗口在距离为 30 的时候最佳，此时退化到仅包含主题词的情况。

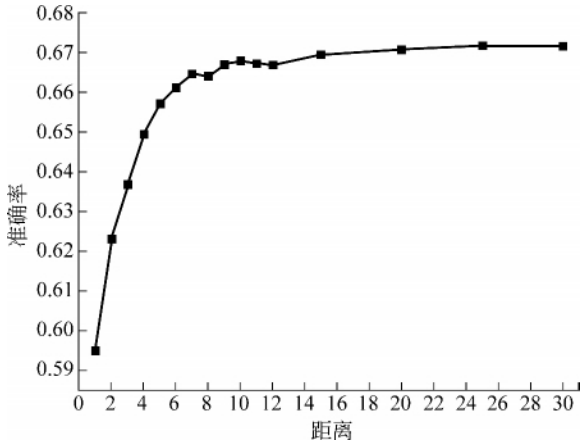

图 2　不同距离范围内包含主题词的结果

最后，我们对于引入主题相关特征后的系统进行了错误分析，发现主要有两类错误。

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，如当主题词为“青蜂侠”时，微博消息“致命伴旅比青蜂侠好看多了！”中的“好看”指的是“致命伴旅”；

② 指代情况较丰富，包括省略宾语，如“青蜂侠。我觉得挺好看的。”；首句省略主语，如“恩恩，不错哦！青蜂侠。”；句子未涉及主题词，但涉及主题词的属性，如“今天去看了青蜂侠。画面挺炫。”，“画面”是电影“青蜂侠”的一个属性。

5　讨论

目前针对中文微博的研究仍处于起步阶段，尚未有关于中文微博情感分析方面的研究工作。本文针对中文微博的情感分析进行了初步探讨。与中文传统文本情感分析工作相比，本文针对中文微博自身特点，相对于传统文本考虑将微博的特有属性如连接、表情符号等作为特征，同时考虑了微博中简单的省略现象，从而更好地进行情感分析；与英文微博的情感分析工作相比，由于英文微博通常只有一句，已有工作均将一条微博作为一个整体赋予情感极性来进行训练和测试，本文考虑到中文微博比英文微博语义更丰富，包含的句子数目更多，且句与句之间涉及的主题及情感可能不同，分别从微博级别和句子级别两方面来探讨了情感分析的效果，实验结果证明句子级别的效果更佳。

6 结论及下一步工作

近年来，微博在国内外强势崛起，成为时下焦点。本文通过使用新浪提供的API抓取一定规模的微博数据，并根据中文微博的特点，提出了基于层次结构的多策略情感分析框架，包括考虑分句与不分句的策略，并对微博的属性，如链接、表情符号、情感词典等进行了特征选择。此外，本文还采用基于表情符号的规则方法和情感词典方法进行分类。通过比较实验，我们发现与后两种方法相比，基于层次结构的多策略情感分析框架可以取得更好的分类效果。其中，在主题无关特征下取得的最好效果是66.467%，考虑主题相关特征后取得的最好效果为67.283%。

目前，本系统仍有很大的提升空间。后续，我们将考虑如下工作来进一步提升实验效果。

① 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；

② 更深入地研究主题相关的特征。例如，考虑引入句法分析及更好的指代消解技术来处理复杂的指代情况；

③ 考虑引入社交网络关系或者消息与消息之间的关系来构建图模型提升结果。

参考文献

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

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

[3] M. Q. Hu, B. Liu. Opinion Extraction and Summarization on the Web[C]//AAAI06, Boston: 1621-1624.

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

[5] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques[C]//ACL'02: 79-86.

[6] Bo Pang, Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]//ACL'04: 271-278.

[7] E. Riloff, J. Wiebe. 2003. Learning extraction patterns for subjective expressions[C]//EMNLP'03: 105-112.

[8] Glance, N., M. Hurst, K. Nigam, et al. 2005. Deriving marketing intelligence from online discussion [C]//SIGKDD'05: 419-428.

[9] Wilson, T., J. Wiebe, P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis[C]//HLT-EMNLP'05: 347-354.

[10] Luciano Barbosa, Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data[C]//Coling 2010 (poster paper): 36-44.

[11] Long Jiang, Mo Yu, Ming Zhou, et al. Target-dependent Twitter Sentiment Classification[C]//ACL 2011.

[12] Lun-Wei Ku, Tung-Ho Wu, Li-Ying Lee, et al. 2005. Construction of an Evaluation Corpus for Opinion Extraction[C]//In NTCIR-5 Japan, 2005: 513-520.

[13] S. Dasgupta, V. Ng. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification[C]//ACL'09: 701-709.

[14] Fangtao Li, Nathan Liu, Hongwei Jin, et al. Incorporating Reviewer and Product Information for Review Rating Prediction[C]//IJCAI 2011.

[15] V. Hatzivassiloglou, J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity [C]//COLING'00: 299-305.

[16] P. Turney. 2002. Thumbs up or thumbs down Semantic Orientate-on Applied to Unsupervised Classification of reviews[C]//ACL'02: 417-424.

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

[18] Xiaowen Ding, Bing Liu. 2007. The Utility of Linguistic Rules in Opinion Mining[C]//SIGIR-2007 (poster paper), 811-812.

[19] Minqing Hu, Bing Liu. Mining and summarizing customer reviews [C]//KDD-2004 (full paper), Seattle, Washington, USA, Aug 22-25, pp. 168-177.

[20] Dmitry Davidiv, Oren Tsur, Ari Rappoport. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys[C]//Coling 2010 (poster paper), pp. 241-249.

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

Technologies Project.

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

[23] Zhao Jun, Liu Kang, Wang Gen. Adding Redundant Features for CRFs-based Sentence Sentiment Classification[C]//EMNLP 2008: 117-126.

[24] Lanjun Zhou, Yunqing Xia, Binyang Li et al. WIA-Opinmine System in NTCIR-8 MOAT Evaluation [C]//NTCIR-8 Workshop Meeting, 2010.

[25] 李寿山，黄居仁. 基于 Stacking 组合分类方法的中文情感分类研究[J]. 中文信息学报，2010，24(5)：56-61.

《中文信息学报》征稿简则

一、《中文信息学报》主要刊登中文信息的基础理论、应用技术、中文信息处理系统及设备、中文信息的自动输入和人工编码输入、汉字字形信息、自然语言处理、计算语言学及民族语言文字信息处理及网上信息处理等方面的研究论文、技术报告、综述、通讯、简报、国内外学术活动等。

二、来稿要求和注意事项

1. 来稿内容力求正确，论点明确，文字简练，数据可靠，图表清晰，字数不超过 8000 字。

2. 文章题目不超过 20 个字，须有 200 字中文摘要和英文摘要。英文文摘应符合英文语法，概括论文内容，包括研究目的、方法、结果和结论。中英文摘要均应包括题目、作者姓名、单位名称、城市名、邮编、摘要、关键词。写明中图分类号。

有基金项目支持的写明基金名称、编号。

给出前三作者信息，包括姓名，出生年，性别，学位或职称，主要研究方向。

3. 文中图、表放在文稿中相应位置，并注明图号、图注。图中文字用六号宋体。

4. 文中外文字母、符号要分清大小写、正斜体；上下角标的位置高低应区别明显；全文计量单位要一致，或中文，或符号。

5. 参考文献只列最主要的，必须是已公开发行的书刊才能列入，最少不得少于 5 条。文献按文中出现先后次序编排，书写格式为：

专著：[序号] 作者. 题名[M]. 出版地：出版者，出版年 ：起止页码.

期刊：[序号] 作者(多作者用逗号分开，超过 3 个者用“，等”代替). 文章题目[J]. 刊物名称，年代，卷数(期数)：起止页码.

论文集：[序号] 作者. 题名[C]//编者. 论文集名. 出版地：出版者，出版年：起止页码.

学位论文：[序号] 作者. 题名[D]. 保存地点：保存单位，年份.

报告：[序号]作者. 题名[R]. 保存地点：保存单位，年份.

报纸文章：[序号] 作者. 题名[N]. 报纸名，出版日期(版次).

标准：[序号]制定单位. 标准编号，标准名称[S]. 出版地：出版者，出版年.

专利：[序号]专利所有者. 专利题名：专利国别，专利号[P]，公开日期.

电子文献：主要责任者. 电子文献题名[电子文献标识/载体类型]. [发表或更新日期]. 电子文献的出处或可获得地址.

电子文献标识：[DB]-数据库 [CP]-计算机程序 [EB]-电子公告

电子文献载体类型：[OL]-联机网络 [MT]—磁带 [DK]-磁盘 [CD]-光盘

6. 来稿请勿一稿二投，文责自负。不录用稿件概不退还，请自留底稿。来稿一经发表，按规定付给稿酬，并赠送单行本 2 册。

通信地址：北京 8718 信箱《中文信息学报》编辑部收，邮政编码 100190，电话：010-62562916。

本刊接收电子投稿，请以附件方式，将 WORD 文档发送至：cips@iscas.ac.cn。请写明作者工作单位、通信地址(邮政编码)、电话(手机)、E-mail。