基于词典的中文微博情绪识别

牛 耘 潘明慧 魏 欧 蔡昕烨
(南京航空航天大学计算机科学与技术学院 南京 210016)

摘 要 微博等社交媒体已成为表达个人情绪和感受的重要平台。自动分析微博文本表达的情绪对于迅速了解大众情绪走向以及调节个人情绪有着重要的意义。文中首次针对中文微博中的情绪进行自动分析,识别微博表达的喜、哀、怒、惧情绪。提出以词典为依据的基于规则的方法,通过实验详细分析了中文情绪词典在社交媒体文本分析中的现状,讨论了存在的主要问题。并深入讨论了微博中情绪表达的语言特点,为建立高精度的情绪分析系统提供了依据。

关键词 微博,情绪分析,情绪词典
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.09.048

Emotion Analysis of Chinese Microblogs Using Lexicon-based Approach

NIU Yun PAN Ming-hui WEI Ou CAI Xin-ye
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract The proliferation of microblogs has created a digital platform where people are able to express themselves through a variety of means. Automatic analysis of the emotional content in microblogs plays an important role in capturing popular feelings and adjusting personal mood. In this paper, a lexicon-based approach was proposed to automatically determine whether a microblog expresses one of the four basic emotions: joy, sadness, anger, and fear. We performed an extensive analysis of current Chinese emotion lexicons to understand their roles in analyzing social media text. The experimental results show that lexicon is a crucial resource in emotion analysis. The results also reveal limitations of current Chinese emotion lexicon. The characteristics of emotion in microblgs are identified for building advanced emotion analysis system.

Keywords Microblog, Emotion analysis, Emotion lexicon

1 引言

随着互联网的迅速发展,微博、博客、论坛等社交媒体成为大众沟通情感和信息的重要途径。其中微博更新快、信息量大、传播广、有巨大影响力,吸引了越来越多的用户。通过微博即时地抒发自己对生活的感受已经成为互联网上的时尚。这些微博文本蕴含了巨大的商业价值,分析其中蕴含的情绪色彩能够帮助预测电影票房、进行舆情监控、了解用户体验等。

本文对微博中所表达的情绪进行自动分析判别。情绪(emotion)是人的各种感觉、思想和行为的一种综合的心理和生理状态,是对外界刺激所产生的心理反应,以及附带的生理反应,如:高兴、生气、伤心等。情绪分析(emotion analysis)即自动判别文本所传达的作者自身的心情、心理反应和情绪状态。由于情绪的复杂性和敏感性,不同的研究对其类别的划分也有很大差异。其中 Ekman[1] 通过研究人的面部表情,提出了6种基本情绪状态:喜(joy)、哀(sadness)、怒(anger)、惧(fear)、恶(disgust)、惊(surprise)。其中"喜"表达了积极、正面的情绪;"惊"大部分时候表达了惊喜之情,然而也有惊慌失措的负面情绪。而其余4种情绪则是消极、负面的。这6种基本情绪分类被自然语言处理领域的自动情绪识别研究广泛采纳[2-4]。本文也将以这6类为目标,分析微博文本中表达的情绪。

本文研究的情绪分析问题(emotion analysis)不同于传统的情感分析问题(sentiment analysis 或 opinion analysis)。二者虽然都属于情感计算[5](affective computing),但是存在着很大的差异。情感分析是对带有情感色彩的文本进行自动分析并预测其情感极性,即判断文本表达的意见或态度是正面、还是负面的。情感分析领域的研究工作集中于对事物所作评论的自动分析,主要的研究对象是产品评论,即自动判别一条评论是对产品的肯定或是否定。虽然情绪分析与情感分析都有极性(正面、负面)的判断,但情绪和评论无论在内容上还是

到稿日期:2013-09-06 返修日期:2013-12-14 本文受教育部高等学校博士学科点专项基金(20103218120024,20123218120041),国家自然科学基金青年科学基金(61202132),校青年科创基金(NS2012073)资助。

牛 耘(1974-),女,博士,副教授,主要研究方向为自然语言处理;潘明慧(1988-),女,硕士生,主要研究方向为自然语言处理,E-mail:panminghui@nuaa.edu.cn;魏 欧(1974-),男,博士,副教授,主要研究方向为软件工程;蔡昕烨(1983-),博士,讲师,主要研究方向为智能计算与生物信息。

表达方式上都存在着很大的差异。首先，人们抒发情绪与发表评论的表达是不同的。情绪的表达并不总伴随着对事物的评论，二者是各自单独出现的。比如，“今天好开心！”表达了作者强烈的高兴的主观情绪，但并不针对某个具体的事物。同样，“看央视春晚到泪流满面，却孤身漂泊在外，待回国后一定要好好陪伴家人。”中表达了作者因思念亲人而悲伤难过的心情，但并未进行任何评论。类似地，很多对事物的评论并不包含作者的情绪。比如，“摩托罗拉 XT 610 白色款很有质感。”中“质感”是对手机的肯定评价，但这句话并没有反应作者的个人情绪。更重要的是，即使是由于某事物引发的情绪，对事物的评论和作者自身的情绪也经常表现出不一致甚至完全相反的极性。比如，“演得太好了，我几次泣不成声。”对于电影本身给予了积极正面的评价，然而作者的情绪则是悲伤的。又如，“电影成功再现了 1942 年的饥荒场景，随着情节的发展心情越来越沉重。”肯定了影片的拍摄，然而作者的心情是悲哀的。另外，从文本特点来看，本文所研究的情绪载体为来源广泛、种类繁多、内容琐碎、表达方式极为口语化的微博文本，而产品评论的主题相对比较集中。在研究目标方面，目前情感分析一般把评论分为正面、负面或者中性。而对于复杂多变的情绪状态，需要更加细致的分类才能更接近作者的真实情绪。

以微博为典型例子的社交媒体已成为抒发个人情绪的重要平台。对微博所表达情绪的自动分析对于迅速掌握大众的情绪走向，预测民众需求有着重要的现实意义。情绪分析还能够帮助用户分析自身情绪以及情绪波动变化原因，来调整个人状态。自动情绪分析也是与情感计算相关的很多应用中重要的子任务。这些应用包括自然语言界面[6]、电子学习环境[7]、安全信息学[8]等。然而目前尚缺乏针对中文微博进行情绪分析的独立研究。虽然英文微博的情绪研究已取得了一定的成果，但中、英文微博在分词、网络用语等方面存在的诸多差异使得英语情绪分析技术难以直接应用于中文。

有鉴于此，本文针对中文微博进行了专门的情绪研究，分析了一条微博是否表达了某种情绪。本文提出基于规则的方法，以词典为依据对微博文本表达的情绪种类进行判断，重点在于讨论作为情绪分析重要资源的中文情绪词典的现状。基于规则的方法的优势在于情绪识别的结果不依赖于标注数据的数量。更为有利的是，基于规则的方法对所取得的结果能够给予清晰的解释。这对于认识情绪表达的特点，进一步建立高精度的情绪分析系统是至关重要的。本文设计了粗颗粒度和细颗粒度两种不同粒度的情绪识别的实验，从词典的作用和微博中情绪表达的语言特点两个方面进行了深入探讨，主要贡献在于：

(1)建立了第一个中文微博的情绪人工标注数据集。

(2)提出以词典为依据的基于规则的方法，详细分析了中文情绪词典在社交媒体文本分析中的现状，讨论了存在的主要问题，为情绪词典的建设提供了参考。

(3)深入分析了微博在情绪表达方式上的语言特点，为建立高精度的情绪自动分析系统提供了重要依据。

本文第 2 节介绍情绪分析相关工作；第 3 节提出基于词典的中文微博情绪分析的方法；第 4 节讨论微博情绪分析数据集的建立过程；第 5 节详述实验过程及结果分析；最后进行总结。

2 相关工作

2.1 英文社交媒体中的情绪分析

近来，情绪分析问题由于其广泛的应用价值引起越来越多的关注。针对英文的情绪分析已取得了一些初步的研究成果。计算语言学领域著名的语义评估会议 SemEval 在 2007 年设立了一个评测任务来对新闻标题进行情绪分析，所用的数据集包含 1250 个句子。为了更好地理解情绪分析问题，该任务强调对情绪进行词法语义分析[9]。从方法上看，机器学习的算法被用于对情绪进行分类，但其中大部分未考虑分类目标的层次性[10,11]。而 Ghazi 等[3]在判断情绪类别时采用了分层次的方式，不同的层次上分类的目标不同，上层的分类依赖于其下层的分类结果，实验结果表明这种分类策略优于无层次结构的分类方式。基于规则的方法是情绪分析研究采用的另一种主要策略。Golder 和 Macy 采用基于词典的方法对上百万的不同地域、不同文化背景的博主发表的 Twitter 微博进行了自动情绪分析，清晰地识别出人们情绪随时间呈周期性变化的模式[12]。Paltoglou 和 Thelwall[13]利用基于情绪词表的规则方法对 Twitter 等微博进行情绪分类以及情绪强度的分析。实验中使用了 3 个社交媒体数据集，一个包含 472 篇来自社交新闻网站的文章，一个包含 658 篇来自社交网站上的博文，另一个包含 216 篇微博。实验结果表明多数情况下这种方法优于监督的机器学习方法。

2.2 中文社交媒体中的情绪分析

目前还没有针对中文社交媒体中的情绪进行的专门研究。在对社交媒体的情感分析中，谢丽星等[14]考察了微博中的评论是正面或负面的。该文重点研究了基于 SVM 层次结构多策略方法，提取的特征包括主题无关和主题相关的特征。杨亮等[15]则根据微博文本中情感词数量以及所表达情感的变化来发现热点事件。

如前文所述，情绪在表达方式上以及所属类别上都与评论有着很大的不同。事实上，由于情绪和评论表达的差异，已经有中文词典明确将情绪词和评论词划分开来。比如被广泛使用的中文概念知识库 HowNet(知网)的“情感分析用语集”中分别收录了评价词 6846 个和表达情绪的词 2090 个，这为情绪分析提供了重要的词典资源。

综上所述，目前尚未针对中文微博所表达的情绪进行研究，也缺乏对作为情绪分析重要资源的中文情绪词典的现状分析，而且中、英文微博在分词、网络用语等方面存在的诸多差异使得英语情绪分析技术难以直接应用于中文。比如，中文自动分词时可能把一个情绪词分散开。例如，偷笑是明显的正向情绪，但被分成了“偷”和“笑”两个词，对情绪分析造成困扰。中文微博中包含大量谐音字(词)，如稀饭/喜欢，果酱/过奖，这给自动情绪分析带来困难。因此，针对中文微博中表达的情绪开展研究有着重要的现实意义。

3 基于词典的中文微博情绪分析

本文提出以词典为依据的基于规则的方法来对微博表达的情绪进行自动分析。无论是在英语的情绪分析或是在传统评论分析研究中，词典都是非常重要的资源[12,13]。由于中文微博情绪分析尚在起步阶段，深入讨论中文情绪词典在社交媒体文本分析中的现状对于发展情绪词典，建立高精度的情

绪分析系统有着重要意义。本文选取了 HowNet 和 C-LIWC 这两个被广泛应用的情感词典，其共同点是二者都明确地将表达情绪的词汇独立于其它类别的词汇。本节首先详细分析了两个词典的构成，然后给出了利用词典对微博所表达情绪进行判断的基于规则的方法。

3.1 词典介绍

(1)C-LIWC(Chinese Linguistic Inquiry and Word Count)词典是在 LIWC 基础上形成的词典。Pennebaker 等人研究建立的 LIWC 软件(Linguistic Inquiry and Word Count, 2007)[16]用于对文本描述中的单词进行分析。其核心为包含约 4500 个从社会学、健康学以及心理学方面挖掘的情绪和认知的单词的词典。LIWC 是英文情绪分析研究应用的重要词典[12,13]。台湾科技大学人文社会学科研究人员根据中文特性将 LIWC 词典翻译改编为中文版本(C-LIWC)[17]。其中包含语言特性 30 类(如：副词、介词等)、心理特性 42 类(如：正向情绪词、负向情绪词等)，共有 72 个类别，总计 6862 个词。其中和情绪相关的类别有 positive emotion、negative emotion、anxiousness、anger 和 sadness。词典中每个词都有一个或多个类别属性，如担忧属于 negative emotion 和 anxiousness。

(2)HowNet(知网)是中科院计算机语言信息中心创建的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础内容的常识知识库[18]。HowNet 明确将表达情绪的词(分为正向情绪词，如高兴，和负向情绪词，如害怕、生气)和发表评论的词(如短小精悍、坚固、鄙俗、枯燥)分开[1)]。与 C-LIWC 不同，HowNet 不包含更具体的情绪种类。表 1 详细列出了两个词典中情绪词的交集和并集包含的单词个数以及单词示例。

表 1 词典正/负向情绪词示例

词典	类别	词数	示例
C-LIWC	正向情绪词	476	呵呵、容光焕发、开心
	负向情绪词	693	低迷、狂躁、泪流满面
HowNet	正向情绪词	836	朝思暮想、称心、欢乐
	负向情绪词	1254	担惊受怕、忿忿不平
重合部分	正向情绪词	68	爱慕、开心、崇拜
	负向情绪词	142	担忧、发怒、焦急、惧怕
合并词典	正向情绪词	1244	轻快、陶醉、爽、赞
	负向情绪词	1805	操心、内疚、闷闷不乐

如表 1 所列，正、负向情绪词并不均衡，在两个词典中正向情绪词都少于负向情绪词。仅从词典的收录上看，负向情绪比正向情绪的用词更加多样化。还可以看到，C-LIWC 包含的词条数少于 HowNet，其中正向情绪词为 HowNet 的 56.9%，负向情绪词为 HowNet 的 55.3%。值得注意的是两个词典的重合部分非常少，共有的正向情绪词只占两个词典正向情绪词总数的 5.47%，而共有的负向情绪词占两个词典负向情绪词总数的 7.87%。这表明不同的分析角度、不同的评价标准可能得到差异很大的情绪词集合。这从一个侧面体现出情绪分析问题的复杂性。

3.2 基于规则的情绪分析方法

本文提出基于规则的方法，利用词典进行微博情绪自动分析，目的在于研究中文情绪词典在社交媒体文本分析中的现状。为了更直接地反映词典的贡献，我们建立了简单直观的规则来对一条微博表达的情绪进行判断。找到微博所包含的情绪词，情绪词数量最多的那种情绪为该微博的主要情绪。下面给出了规则的详细说明。

给定一条微博文本 t，假设待判断的情绪种类集合为 $E=\{e_1,\cdots,e_i,\cdots,e_m\}$，其中 m 为情绪类别的总数，那么对 t 的情绪判断过程如下：

1. 使用中文分词系统对 t 进行分词处理得到单词序列 q。

2. 对 q 中的单词与情绪词典中的情绪词进行匹配，对于每种情绪类别 e_i，统计匹配到的该类的情绪词个数 n_i。

3. 对于每种情绪类别 e_i，计算 t 所对应的情绪值 v_i，公式如下：

$$v_i=\begin{cases}1, & \max\limits_{1\leqslant j\leqslant m\cap j\neq i}\{n_j\}<n_i\\ 0, & \text{otherwise}\end{cases}$$

即当一条微博中包含了多类情绪词时，该微博主要表达的情绪为情绪词数量最多的那个类。此时 v_i 的值为 1。

4. 判断 t 所对应的情绪类别：如果存在 $1\leqslant i\leqslant m$，使得 $v_i=1$，那么 t 属于情绪类 e_i；否则，t 的情绪类别无法判别。

根据上述过程可以看出，对于微博文本 t，如果情绪词个数集合 $\{n_1,\cdots,n_i,\cdots,n_m\}$ 中存在唯一的最大值，那么可以判断 t 的情绪。否则，$\{v_1,\cdots,v_i,\cdots,v_m\}$ 中所有的值均为 0，无法判别 t 的情绪类别。相应地，如果 t 未匹配到词典中任何的情绪词，t 的情绪也无法判别。对于后一种情况我们说 t 未被词典覆盖。我们定义词典的覆盖率为被词典所覆盖的微博的条数占微博总条数的百分比。

4 微博情绪分析数据集

目前对于中文微博的情绪分析还在起步阶段，尚无标注了情绪类别的微博数据集发表。为此，我们首先对新浪微博进行了人工标注，建立了第一个中文微博情绪标注数据集。

4.1 数据获取

我们从新浪微博中选取了 4 个主题“科比”、“燃油涨价”、“我国留学生澳洲遇袭”、“武广高铁”，使用新浪提供的 API 抓取微博文本。每个主题下载了 400 条微博消息，由两名标注人员各自独立进行标注。在标注微博文本时删除了广告和转发内容相同的文本。每条微博标注为喜、哀、怒、惧、恶、惊和其它共 7 类中的一类，被划为其它类的微博文本有以下几个特点：

①微博文本的情绪不属于喜、哀、怒、惧、恶、惊 6 类。如，不在自己国家好好呆着，出去出事了吧……

②微博表达的情绪不明显。如，不知说什么才好！

③微博中多种情绪混杂，并且缺少占主导地位的情绪。如，在高原，缺氧；回平原，醉氧。我的躯体真无聊  ！好在肩负粉丝 600 万的公益活动做完了 

4.2 实验数据集

我们将两名标注员标注结果一致的微博文本提取出来作为实验所依赖的人工标注集，以保证数据的可靠性。将标注微博以图 1 的方式进行存储：

1) HowNet 中表达情绪的词被称为正面情感词和负面情感词。为了避免多个概念造成混淆，且便于与 C-LIWC 进行比较，本文分别将正面情感词和负面情感词称为正向情绪词和负向情绪词。

〈weibo id="393" emotion-type="Joy:〉
〈content〉#史诗对决#支持科学，支持湖人，科比勇夺六冠[奥特曼][奥特曼][奥特曼][帅][帅][帅]〈;
〈/weibo〉
〈weibo id="394" emotion-type="Joy"〉
〈content〉姐今天也坐了会儿武广高铁，真滴很快！打了会儿小嗑睡就到鸟！[哈哈][兔子]〈/content〉
〈/weibo〉
〈weibo id="395" emotion-type="Sdd"〉
〈content〉[可怜]悲剧喽这下，这次不知道损失多少米米〈/contnet〉
〈/weibo〉

图 1　微博情绪标注数据集

其中 id 是对微博的编号，emotion-type 为人工标注的微博文本的情绪类别，content 是微博正文。

根据标注统计结果，4 个主题的微博文本中包含的恶和惊的数据非常少（分别为 4 条和 31 条），并且在 C-LIWC 和 HowNet 中都未包含这两种情绪，因而在下面的实验中只对被标注为喜、哀、怒、惧 4 种情绪的 807 条微博文本进行分析。表 2 显示了这些微博在 4 个主题及 4 种情绪类别中的分布。

表 2　微博在不同主题及不同情绪中的分布

	情绪类别	喜	哀	怒	惧	合计
主题	科比	210	12	33	1	256
	燃油涨价	19	14	99	7	139
	我国留学生澳洲遇袭	13	20	76	101	210
	武广高铁	152	23	24	3	202
	合计	394	69	232	112	807

如表 2 所列，一个事件往往存在占主流的情绪。比如，主题为“科比”的微博文本中喜的情绪占大多数，达到 82.0%；主题为“燃油涨价”的微博文本中怒的情绪占大多数，为 71.2%；主题为“我国留学生澳洲遇袭”的微博文本中则有两种主要情绪即怒和惧，分别为 21.2%和 28.4%；主题为“武广高铁”的微博文本中喜的情绪占大多数，达到 75.2%。识别主流情绪有助于迅速了解大众情绪走向，而分析非主流情绪则能够全面掌握事件对不同群体产生的影响。

5　实验结果与分析

本文采用基于规则的方法，利用词典进行微博情绪分析。为了深入了解词典的作用及其存在的问题，我们进行了粗颗粒度和细颗粒度两组实验。粗颗粒度情绪识别将一条微博分为正向或负向情绪。在我们的实验数据中，喜类是唯一的一种正向情绪，而哀、怒、惧都属于负向情绪。细颗粒度分析则识别更具体的情绪，判断一条微博表达喜、哀、怒、惧 4 种情绪中的哪一种。

由于 HowNet 只有正向情绪词和负向情绪词两类，并未作更细致的类别划分，因此只将其用于粗颗粒度分类实验。在 C-LIWC 中则既包含正向情绪词（positive emotion）、负向情绪词（negative emotion），同时还包含 sadness 词、anger 词和 anxiousness 词。因而将其用于粗、细颗粒度两组实验中。其中，正向情绪词、负向情绪词用于粗颗粒度情绪分类。在细颗粒度实验中，sadness 词由于与哀情绪存在直接的对应关系，因此用于识别哀。类似地，anger 词用于识别怒。词典中虽然并没有直接与喜和惧对应的分组，但是 positive emotion 中包含快乐、高兴、开心、狂喜、呵呵、幸福、知足等词，因此我们将它用于识别喜类情绪。类似地，anxiousness 中包含不安、可怕、怕、风险、冒险、害怕、恐怖、恐惧等词，因此将它用于识别惧类情绪。表 3 列出了进行粗、细颗粒度情绪判别所利用的词典中相应的词分组所包含的词条数。词典中缺乏与 6 种基本情绪对应的词的分组本身就是目前情绪词典存在的不足。由于 positive emotion 包含的词并不仅限于表达喜类情绪，将它们用来识别喜类会带来噪音，然而这是词典中唯一能够帮助识别喜类情绪的分组。

表 3　词典中情绪词分组包含的词条数

词的分组	词典	
	C-LIWC	HowNet
positive emotion 正向情绪/喜	476	836
negative emotion 负向情绪	693	1254
sadness/哀	128	—
anger/怒	249	—
anxiousness/惧	111	—

5.1　正/负向情绪判别

本节的实验采用 3.2 节中阐述的基于规则的方法对 4.2 节中描述的数据集中的每条微博进行分析并判断其表达的是正向情绪还是负向情绪。该方法中的步骤 1 使用了清华大学中文分词系统[19]，对情绪的分析分别使用 C-LIWC 和 HowNet 进行判断。由于两个词典的重合率很低（详见表 1），进而将两个词典进行了合并，测试了合并词典对情绪识别的影响，结果见表 4。

表 4　两个词典正、负向情绪判定结果

	C-LIWC 的规则方法	HowNet 的规则方法	合并词典的规则方法
覆盖率/%	66.7	63.8	**78.2**
准确率/%	**70.2**	58.5	63.8

总体来说两个词典的覆盖率都不太高。C-LIWC 的覆盖率为 66.7%，略高于 HowNet。两个词典合并后尽管覆盖率得到很大提升，但仍只达到 78.2%，表明微博中尚存在大量的表达情绪的词汇未被收录进来，包括一些常用情绪词如嘻嘻等，还有很多网络词汇，比如 v5、哈尼、果酱等。导致覆盖率低的原因还包括繁体中文的匹配问题以及英文情绪词的匹配问题。有些微博夹杂了表达情绪的英语单词如 good、OMG、COMEON。因为 C-LIWC 和 HowNet 只包括中文词，在判断情绪时这些英语单词就未能发挥作用。另外有些微博中虽然并不包含情绪词但整体上却表达了某种情绪，比如，“让他们看看什么叫做中国功夫!:)”这条微博中没有明确表达情绪的单词，因而不会被词典覆盖。

在情绪判断的准确率（accuracy）方面，可以看到 C-LIWC 比 HowNet 高了约 12 个百分点，表明它收录的词与微博文本中情绪的表达更吻合。两个词典合并后准确率并没有相应提高，而是相比 C-LIWC 下降了约 6 个百分点。可见，仅靠合并词典并不一定能提高情绪识别的准确率。

总的说来，C-LIWC 取得了更高的覆盖率和准确率。为进一步了解它在情绪判断中的作用，我们计算了 C-LIWC 词典在正负向情绪分类中的精确率（precision）、召回率（recall）和 F 值（F-score），结果如表 5 所列。

表5 C-LIWC词典正/负向情绪判定结果

类别	精确度/%	召回率/%	F值/%
正向情绪	65.5	**87.9**	75.1
负向情绪	**80.8**	52.1	63.4

可以看出，C-LIWC对正向情绪识别的总的结果优于对负向情绪的识别，其表现在F值高了约12个百分点。具体来说，C-LIWC对正向情绪识别的召回率较高，精确度比较低，而对负向情绪判断则是精确度较高，召回率比较低。统计表明近一半的负向情绪被误判为了正向情绪，导致对正向情绪的识别取得了较高的召回率。我们对抽取的120条人工标注为负向情绪但被误判的微博进行了详细分析，从词典和微博文本两个方面探讨了导致错误的主要因素。

(1)词典的主要问题

·词典收录不全面。在这120条微博中，50.8%的微博包含了至少一个明确表达负向情绪的词且该词未被词典收录，比如无良、危险、心惊胆战、唉、泪奔、悲剧。因此在判断情绪时这些词未能起到标识负向情绪的作用。

·词典中没有针对情绪给出相应的词性信息。一些情绪词只在具有某种词性时才体现出情绪上的倾向性，而作为其它词性出现时并不具有这种倾向性。例如，“好”作为形容词时通常表达正向情绪，而作为副词时则通常不具有情绪上的倾向性。类似地，“希望”作为名词表达正向情绪，而作为动词则一般只表示愿望；“安全”作为形容词或副词时表达正向情绪，而作为名词则通常不具有情绪上的倾向性。在下面的例子中，“好”、“希望”和“安全”并不带有情绪色彩。

好悲哀。

阿暴在悉尼，希望只是个别现象，地球太危险了。

出国在外须小心谨慎注意出行安全，希望此事得到顺利解决。

由于词典中没有与情绪对应的词性信息，因而在与微博进行匹配时无法区分上述情况而导致误判。

(2)微博文本情绪表达的特点

·一条微博虽然主要表达的是某种情绪但也出现了其它的情绪。比如，

感觉还是在自己的国家比较安全啊，因为到一个陌生的地方去都会遭到别人的歧视，中国人在国外受到的特殊待遇还少吗？唉

这条微博表达的主要情绪是对在国外的中国人所受的不公正而伤感。在表达这个主题的同时采用了对比方式来叙述在国内的安全，因而出现了正向情绪的表达(比较安全)，对识别整条微博的情绪造成干扰。

·一些情绪词在某些上下文中并不表现出情绪倾向，或表现出不同于词典标注的情绪倾向。

当前新一轮物价开涨，都是油价惹的祸。就纳闷了？为什么一边在极力抑制通货膨胀，可还一边在鼓励牵动全局的可恨的燃油涨价呢？

这条微博表达了强烈的生气、愤怒的情绪，其中出现的鼓励在词典中被标注为正向情绪词，然而在这里的上下文中它并不表达任何正向的情绪。再比如，

今天湖人与雷霆的比赛解说员说的是不是有些过分了啊，什么科比在场上没有什么用啊，就你有用，你这素质不会这样吧，是不是走后门才赶上这份工作的啊…你行你上啊

这条微博中有用在词典中属于正向情绪词，然而被用在非常口语化的表达中，与上下文语气一致，表达了生气这种负向情绪。在这些情况下仅仅按照词典的标注而忽略了上下文信息则容易造成误判。

·否定用法导致语义反转。比如，不/太平、不/尊重、绝非/好意、从未/开心。这些表达中虽然包含了正向情绪的词，如太平、尊重、好意、开心，但由于否定词如不、绝非、从未的出现而使得其真正体现的是担心、害怕、生气等负向情绪。由于规则中未对否定用法进行处理，导致误判。

·在表达负向情绪的反讽语句中一般只出现正向情绪词，造成负向情绪被判断为正向情绪。

应该热情赞美中石油和中石化为推动国家物价攀升所做的不懈的努力，持之以恒，乐此不疲！

·分词错误导致词典匹配失败也会造成误判，比如真心痛、真得小心。心痛、小心等提示情绪的重要线索由于分词的错误而未发挥作用。

从以上的分析可见，提高正向和负向情绪判断的精度应综合考虑词典和情绪表达特点两个方面的因素。

词典方面，大量表达情绪的词尚未被词典收录，因而补充情绪词是提高精度的重要方面。其中很多词为口语化的网络用语，很难从传统的词典中获得。从微博文本中自动获取网络情绪词将是扩充情绪词典的重要途径。另外，在建立词典时对情绪词标注体现其情绪色彩的具体的词性将有助于准确判断文本的情绪倾向。

从微博文本中情绪的表达方式上看，应充分考虑情绪词所处的上下文的影响。首先，对整条微博文本进行词性标注分析将有助于明确词语在该条具体微博中的含义和作用，从而更准确地判断它在当前微博中是否带有情绪色彩。制定规则时应根据词性来判断微博的情绪倾向。其次，否定用法可起到改变情绪倾向的作用，是判断微博情绪的重要线索。由于微博极为口语化的表达方式，出现了大量新的否定词和否定用法。收集各种否定词、否定用法和否定句式将对情绪自动判别发挥重要作用。第三，一些情绪词在某些特定上下文中出现时不表现出情绪倾向或呈现与词典中不同的情绪倾向(上下文相关情绪词)。通过对微博文本进行句法分析或依赖关系分析来找到与这些情绪词相互作用的句子成分，将有助于确定具体语境所体现的情绪倾向。

5.2 四类情绪判别

本节中将目标情绪细分为4种，即喜、哀、怒、惧。如前所述，C-LIWC词典中的positive emotion(476个)用来判断喜类情绪，sadness(130个)用来判断哀类情绪，anger(129个)用来判断怒类情绪，anxiousness(111个)用来判断惧类情绪。在HowNet中，正向情绪词和负向情绪词未被具体地划分，故在本节实验中未采用。

利用C-LIWC词典采用基于规则的方法判断一条微博表达的具体情绪，取得了62.0%的覆盖率和63.4%的准确率。与正、负两类情绪识别结果类似，基于C-LIWC词典的规则方法的覆盖率不是很高。对于各种情绪识别的精确度、召回率和F值见表6。

如表6所列，喜类取得了最高的F值，而其它几类的F值都比较低。怒类微博识别的精确度达到83.0%，说明C-LI-

WC 词典收录的 anger 词情绪色彩鲜明且在情绪表达上较少歧义，然而由于其召回率低，导致 F 值不高。哀类和惧类的精确度和召回率都比较低，特别是召回率与喜类相比差别非常大。

表 6 C-LIWC 类情绪识别结果

类别	精确度/%	召回率/%	F 值/%
喜	63.5	**97.5**	**76.9**
哀	33.3	25.0	28.6
怒	**83.0**	34.2	48.4
惧	37.5	8.6	14.0

表 6 中的一个突出的现象是喜类的召回率非常高，而 3 种负向情绪的召回率都很低。相对而言 3 种负向情绪的召回率从高到低依次为怒、哀、惧。注意到微博与词典中各类情绪词相匹配的个数由高到低分别为：positive emotion词数为 144 个(占 positive emotion 词总数的 30.3%)，anger 词数为 49 个(占 anger 词总数的 38.0%)，sadness 词数为 32 个(占 sadness 词总数的 24.6%)，anxiousness 词数为 22 个(占 anxiousness 词总数的 19.8%)。可见各类情绪与词典匹配词的个数减少的趋势与相应类所取得的召回率降低的趋势是一致的。这从一个侧面表明微博中表达怒、哀、惧的很多词在词典中未收录，这是导致其召回率低的直接原因。

为了进一步分析导致怒、哀、惧召回率低的因素，我们对 4 类情绪分别统计了平均每条微博匹配到词典中各类情绪词的个数，见表 7。可以看出，每类情绪的微博匹配 positive emotion词的个数都很高。例如，对于喜类，平均每条微博匹配 positive emotion 词的个数为 1.28，而匹配其它 3 类情绪词的个数相对较低(0.06)；哀类微博匹配 positive emotion 词的个数平均达到 0.43 个，约为其匹配的 sadness 词的 2 倍；对于怒和惧类微博也存在着类似的情况，平均每条怒和惧类微博匹配到的 positive emotion 词的个数也分别高于匹配到 anger 和 anxiousness 词的个数，特别是对于惧类，平均每条微博匹配 positive emotion 词的个数为 1.05，而匹配 anxiousness 词的个数仅为 0.16。positive emotion 词在各类微博中都出现得较频繁是造成怒、哀、惧 3 种情绪的微博被误判为喜类的另一主要原因。

表 7 4 类情绪的微博中匹配词典中各类词的均值

	Positiveemotion	sadness	anger	anxiousness
喜微博	**1.28**	0.06	0.06	0.06
哀微博	**0.43**	0.22	0.13	0.09
怒微博	**0.57**	0.13	0.47	0.10
惧微博	**1.05**	0.03	0.11	0.16

为了寻找能够帮助识别惧、哀、怒 3 种情绪的线索，我们对这 3 类微博各随机抽取了 40 条进行分析，发现每种情绪在表达时都有各自突出的特点。这些特点为进一步设计有效的分类算法提供了重要的依据。

(1)否定用法常被用来表达惧情绪。对 40 条惧类微博的分析表明，22.5%含有否定用法，如下例所示。

唉！在哪都不太平！在国外更不安全！

哎，其实在国内也不一定能保证安全。

上面句子中太平、安全、保护在词典中属于 positive emotion，然而由于否定用法如不、不一定的应用，这些微博其实表达的是惧的情绪，而非喜的情绪。因而对于惧类情绪的分析应充分了解微博情绪中否定用法的表述方式。注意到对 positive emotion 的否定可能表达不同种负向情绪，如惧、哀、怒，因而应根据具体的表达方式有针对性地制定规则来提高识别的精度。

(2)表情符常常出现在哀类微博中表达情绪，如下例所示。

艾走了，麦替补了，科比老了，我觉得自己的青春随他们而去了 

燃油涨价了！ 

这些微博中没有出现任何明确表示伤心的词语，然而表情符的使用强烈地表达了哀的情绪。对 40 条哀类微博的分析表明，其中 27.5%由表情符表达情绪。因此如何有效地利用表情符是情绪判断中的重要问题。

(3)怒情绪的表达经常使用反问句式。对 40 条怒类微博的分析表明，其中 20%都包含有反问句式。比如，

放手燃油涨价，却约谈欲涨价的食用油企业。都是产油的，待遇的差别咋就这么大呢？

宣布燃油涨价，这本无可厚非，成本涨了，销售涨价也是合情合理的，但是一个关键的问题是，你既然有宣布燃油涨价的权利，那么为啥不宣布全民涨工资？

因而，除了对情绪表达文本进行句法和词法分析，对于句型和句式的定义和描述也是判断情绪类别的重要方面。

综合以上分析，情绪词典是进行情绪分析的重要资源。而目前的中文情绪词典无论从规模或是针对性方面都存在着很大的改进空间，尤其是远远不能满足识别更加细致的情绪类别的需要。对于以微博为代表的社交媒体文本，其中大量的表达情绪的网络词汇的收录是建设高适应性情绪词典亟待解决的问题。同时，不同情绪类别的微博文本在表达方式上也呈现出不同的特点，为建立高精度的自动情绪识别系统提供了重要的依据。

结束语 以微博为典型例子的社交媒体已成为大众表达心情感受的重要渠道。自动分析微博的情绪倾向将有助于预测事件走向、大众需求进而作出迅速及时的反应，也能够帮助微博用户分析自身的情绪变化以调整个人状态。本文建立了首个包含喜、哀、怒、惧 4 种情绪的人工标注微博数据集，选取了两个被普遍使用的情绪词典 C-LIWC 和 HowNet，采用基于词典的规则的方法对两个词典进行了深入的比较分析。对实验结果的分析表明了情绪词典的重要性，同时发现目前中文情绪词典存在着对网络用词收录严重不足、对细致情绪种类针对性不强的问题。另外，从语言描述方面详细讨论了微博不同种情绪各自的语言表达特点，为建立更有效的情绪自动判定系统提供了重要的依据。后续的工作将以此为基础，从 3 个方面展开进一步的研究：针对基本情绪类别扩充情绪词典以更好地服务于社交媒体的自动情绪分析；建立高精度的自动情绪分类系统；收集表达 6 种基本情绪中的惊、恶情绪的微博，更全面地进行情绪分析。

参 考 文 献

[1] Ekman D. Facial Expression and Emotion [J]. American Psychologist, 1993, 48(4): 384-392

(下转第 289 页)

结束语 分析了分布估计算法的优缺点，对基本的分布估计算法进行改进并将它用于求解软硬件划分问题。在分布估计算法中加入了精英克隆选择操作，加强了局部搜索能力，并对概率模型进行修正改善了多样性易丧失的问题，改进后的算法进行软硬件划分问题的求解，以时间和功耗为约束条件、以成本为优化目标进行优化。将结果与基本 EDA 以及文献中算法相比，结果表明 IEDA 算法可以获得更优结果，可以有效求解软硬件划分问题。

参 考 文 献

[1] Abdelhalim M B, Habib S E-D. An integrated high-level hardware/software partitioning methodology[J]. Design Automation for Embedded Systems, 2011, 15(1): 19-50

[2] Koudil M, Benatchba K, Tarabet A, et al. Using artificial bees to solve partitioning and scheduling problems in codesign[J]. Applied Mathematics and Computation, 2007(186): 1710-1722

[3] 邹谊，庄镇泉，杨俊安. 基于遗传算法的嵌入式系统软硬件划分算法[J]. 中国科学技术大学学报，2004，34(6)：724-731

[4] Wu Yue, Zhang Hao, Yang Hong-bin. Research on parallel HW/SW partitioning based on hybrid PSO algorithm[J]. Algorithms and Architectures for Parallel Processing, 2009, 5574: 449-459

[5] Wang P, Wu J G. Efficient heuristic and tabu search for hardivare/software pattitioning[J]. Computer, Science, 2012, 39(1): 290-294

[6] Huang Yue, Kim Y. Applying hybrid neural fuzzy system to embedded system hardware/software partitioning[C] // Third International Conference on Intelligent Computing, ICIC 2007. 2007, 4682: 660-669

[7] Wu Jigang, Srikanthan T. Algorithmic aspects of area-efficient hardware/software partitioning[J]. The Journal of Supercomputing, 2006, 38(3): 223-235

[8] Larranaga P, Lozano J A. Estimation of distribution algorithms: A new tool for evolutionary computation[M]. Netherlands: Springer, 2002

[9] Chen S-H, Chen M-C, Chang Pei-chann, et al. Guidelines for developing effective Estimation of distribution algorithms in solving single machine scheduling problems[J]. Expert Systems with Applications, 2010, 37(9): 6441-6451

[10] Izquierdo C E, Velarde J L G, Melián-Batista B, et al. Hybrid Estimation of distribution algorithm for the quay crane scheduling problem[J]. Applied SoftComputing, 2013, 13: 4063-4076

[11] Wang Ling, Wang Sheng-yao, Xu Ye. An effective hybrid EDA-based algorithm for solving multidimensional knapsack problem [J]. Expert Systems with Applications, 2012, 39: 5593-5599

[12] Ceberio J, Irurozki E, Mendiburu A, et al. A review on estimation of distribution algorithms in permutation-based combinatorial optimization problems [J]. Progress in Artificial Intelligence, 2012, 1(1): 103-117

[13] Santana R, Larrañaga P, Lozano J A. Combining variable neighborhood search and Estimation of Distribution Algorithms in the protein side chain Placement problem[J]. Journal of Heuristics, 2008, 14(5): 519-547

[14] Branke J, Lode C, Shapiro J L. Addressing sampling errors and diversity loss in UMDA[C]// Proceedings of the 9th annual conference on Genetic and evolutionary computation. London, England, UnitedKingdom, 2007: 508-515

(上接第 258 页)

[2] SemEval2007[OL]. http://nlp.cs.swarthmore.edu/semeval/

[3] Ghazi D, Inkpen D, Szpakowicz S. Hierarchical versus Flat Classification of Emotions in Text[C]// Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Los Angeles, California, 2010: 140-146

[4] Bellegarda J R. Emotion Analysis Using Latent Affective Folding and Embedding[C]// Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Los Angeles, California, 2010: 1-9

[5] Picard R W. Affective Computing[M]. Cambridge: MIT Press, 1997

[6] Cosatto E, Ostermann J, Graf H P. Lifelike talking faces for interactive ser-vices[J]. Proc. IEEE, 2003, 91(9): 1406-1429

[7] Ryan S, Scott B, Freeman H, et al. The Virtual University: The Internet and Resource-based Learning [M]. London, UK: Kogan, 2000

[8] Abbasi A. Affect Intensity Analysis of Dark Web Forums[C]// Proc. IEEE Int. Conf. Intelligence and Security Informatics (ISI). New Brunswick, NJ, 2007: 282-288

[9] Strapparava C, Mihalcea R. Learning to Identify Emotions in Text[C]// Proc. ACM Symposium on Applied computing. Fortaleza, Brazil, 2008: 1556-1560

[10] Alm C. Affect in text and speech[M]. University of Illinois at Urbana-Champaign, Department of Linguistics, 2008

[11] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis [J]. Computational Linguistics, 2009, 35(3): 399-433

[12] Golder S A, Macy M W. Diurnal and Seasonal Mood Vary with Work, Sleep, and Day length Across Diverse Cultures [J]. Science, 2011, 333(6051): 1878-1881

[13] Paltoglou G, Thelwall M. Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media[J]. ACM Transactions on Intelligent Systems and Technology(TIST), 2012, 3(4): 66

[14] 谢丽星，周明，孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报，2012，26(1)：73-83

[15] 杨亮，林原，等. 基于情感分布的微博热点事件发现[J]. 中文信息学报，2012，26(1)：84-90

[16] LIWC[OL]. http://www.liwc.net/

[17] C-LIWC[OL]. http://c-liwc.blogbus.com/

[18] HowNet[OL]. http://www.keenage.com/

[19] 清华大学中文分词演示系统[OL]. http://nlp.csai.tsinghua.edu.cn/app/wordSegment/