文章编号：1003-0077(2015)04-0134-10

基于多样化特征的中文微博情感分类方法研究

张志琳，宗成庆

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190）

摘 要：随着 Web 2.0 时代的兴起，微博作为一个新的信息分享平台已经成为人们生活中一个重要的信息来源和传播渠道。近年来针对微博的情感分类问题研究也越来越多地引起人们的关注。该文深入分析了传统的情感文本分类和微博情感分类在特征表示和特征筛选上存在的差异，针对目前微博情感分类在特征选择和使用上存在的缺陷，提出了三种简单但十分有效的特征选取和加入方法，包括词汇化主题特征、情感词内容特征和概率化的情感词倾向性特征。实验结果表明，通过使用该文提出的特征选择和特征加入方法，微博情感分类准确率由传统方法的 73.17%提高到了 84.17%，显著改善了微博情感分析的性能。

关键词：中文微博；情感分类；机器学习；特征选择

中图分类号：TP391 **文献标识码**：A

Sentiment Analysis of Chinese Micro Blog Based on Rich-features

ZHANG Zhilin, ZONG Chengqing

(National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China)

Abstract: Micro blog, a new information-sharing platform, is now playing an important role in people's daily live with the rise of Web 2.0. And micro blog sentiment analysis research also attracts more attention in recent years. This paper provides an in-depth analysis on the difference of feature representation and feature selection between the traditional sentiment classification and micro blog sentiment analysis. To avoid the drawbacks of feature selection of existing methods, we propose three simple but effective approaches for feature representation and selection, including the lexicalization hashtag feature, the sentiment word feature, and the probabilistic sentiment lexicon feature. Experimental results show that our proposed methods significantly boost the micro blog sentiment classification accuracy from 73.17% to 84.17%, outperforming the state-of-the-art method significantly.

Key words: Chinese micro blog; sentiment analysis; machine learning; feature selection

1 引言

微博是微博客（Micro Blog）的简称，是一种基于用户关系的信息分享、传播和获取的平台，用户可以通过 WEB、WAP 以及各种客户端组建个人社区，以 140 字左右的短文本更新信息，并实现及时分享[①]。目前微博已经从各个方面渗透到了人们的日常生活和工作当中。以国内新浪微博为例，截止到 2012 年 12 月 31 日，用户数目已经超过了 5 亿人。

微博的快速发展引发了研究人员对于微博处理的兴趣。其中针对微博的情感分析研究是目前微博研究中最热，关注度最高的研究领域之一。情感分类是情感分析研究中的基本任务，该任务旨在对文本按照情感极性进行褒贬分类。与普通文本相比，微博由于其本身所具有的特点，如句子短，用词口语化，网络词汇较多等，使得对微博进行情感分类研究更具挑战性。

目前，微博情感分类的方法主要有基于规则的方法[1-3]和机器学习方法[4-6]两类。规则方法中主要采用了表情符号和情感词作为规则的统计特征。机器学习方法都是将情感分类作为一个普通的分类问

① http://baike.baidu.com/view/1567099.htm

收稿日期：2013-08-26 **定稿日期**：2013-11-15

基金项目：国家自然科学基金(61402478)

题来对待。微博情感分类的机器学习方法开始主要沿用了文本分类的方法，一般采用一元语言模型和二元语言模型等特征。之后，结合微博本身的特点，开始陆续提出了一些新的解决方法，例如，利用 Twitter 的标签(hashtag)和笑脸符号(smileys)等进行情感分类。目前，基于机器学习方法的情感分类基本流程都是对预处理后的微博数据进行特征的获取和加工。这些特征主要包括：主题词、链接、标点符号是否存在，正负极性表情符号的个数和正负极性情感词的个数等作为特征进行分类器的训练，取得了一定的成效。传统的方法要么只侧重于直接从训练语料中提取特征，要么只依赖于情感词典，而大量的工作表明，情感词典和从训练语料中抽取的特征对情感分类都非常重要。考虑到两者在某种程度上互为补充、互相关联，我们相信，如果能够很好地将两者结合起来，发挥各自所长，必将对情感分类有很大的帮助。

正是基于这种动机，我们研究了情感词典与从训练语料中抽取的知识的结合方式。其基本思路是：1)对于有关主题的特征，我们不仅考虑主题是否出现，而且考虑主题词的特定内容；2)对于情感词，不仅分析情感词加入的方法，而且研究情感词加入的数量对于整体分类效果的影响；3)考虑到通用的情感词典首先不能及时覆盖和添加日新月异的网络用语，同时针对微博数据也没有权重区分，我们提取了微博用语来丰富拓宽通用情感词典，并使用微博数据对该词典进行倾向性概率打分，将概率打分作为特征取代原始的布尔特征，从而更加真实地反映微博情感倾向。

针对上述分析，本文经过词汇化主题特征的选取、情感词特征的加入和概率化情感词倾向性特征的加入，逐渐丰富特征，既结合了外部词典资源，又充分利用了微博数据本身，使得中文微博情感分析的准确率从 73.17%上升到了 84.17%。

本文其余部分的结构组织如下：第二节介绍相关的工作；第三节阐述了本文特征设计的主要内容；第四节给出了实验结果以及分析；第五节主要阐述本文的结论，并展望下一步的工作。

2　相关工作

这一部分我们分别按照传统情感分类和微博情感分类的相关工作进行陈述。

传统情感文本分类研究主要有两类：一类是基于词典的方法，另一类是基于统计机器学习的方法。基于词典的方法代表工作有 Lu[7] 和 Turney[8] 等。Lu[7] 等使用通用情感词典，比如 WordNet 中的同义词、反义词信息以及一些语法规则，来判断微博的情感极性。它的缺陷在于过于依赖外部词典。Turney[8] 利用 PMI-IR 方法计算出现在文本中符合规则的短语的情感倾向，通过这些情感倾向的平均值来判断文本极性。基于词典的方法过于依赖相关的知识库(词典、规则库等)支撑，这些知识库一般是由语言专家总结出来的，但是，这些规则难以描述不确定性事件，且规则与规则之间的相容性难以得到有效的控制。

基于统计机器学习的方法，代表工作有 Pang[9] 和 Mullen[10] 等。Pang[9] 将统计机器学习方法引入到电影评论的褒贬分类任务中。文章中使用了包括一元词、二元词、词性标注等若干特征，选用了朴素贝叶斯、最大熵、支持向量机训练模型。实验结果表明，支持向量机的效果最理想，且选用一元词特征，特征值采用布尔值时取得了最好的准确率。Pang 的分析是在英文语料中进行的，对于中文是否仍然有效还有待考证。除此之外，Pang 的特征是对训练语料进行统计得到的，没有使用外部词典，对于训练语料的依赖性太强，所以泛化能力相对较差。Mullen[10] 等使用 SVM 分类器，将不同来源的各个特征信息进行综合，提升了分类效果。

对于微博情感分析，英文的微博情感分析相对中文微博情感分析来说效果较好。Go 等[11] 首次提出对微博文本进行情感分析的思想。文章中将表情符号加入到了选取的特征中，取得了很好的效果。Pak 和 Paroubek 等[12] 利用表情符号组织标注了一个 Twitter 微博文本情感极性数据集，并且使用 N 元词汇(N-grams)作为特征进行分类，没有使用任何情感词典，与 Pang[3] 的工作类似，过于依赖训练语料，泛化能力较差。除此之外，Davidov 等[13] 使用了 Tweets 中的标签(hashtag)和笑脸符号(smileys)作为特征，训练出了一个有监督的类似 K 近邻(KNN)的分类器，用于对 Tweets 进行情感分类。针对中文微博的研究仍处于起步阶段，已采用的方法包括基于表情符号的规则方法、基于情感词典的规则方法以及机器学习的方法[4, 14]。在这些方法中，对于微博情感分类的特征选取比较单一[6]，主要还是借助于外部资源对微博表情符号、情感词的统计信息上，或者是直接沿用传统的情感分析的方法，采用 N-grams 作为特征，忽略了情感词典的作用，

缺乏对于两者结合的探讨和研究。比较有代表性的文章是,谢丽星的基于层次结构的多策略中文微博情感分析和特征提取[14],文章中采用了基于层次结构的多策略分析框架,并且引入了一些新的特征,实验证明了基于SVM的一步三分类来解决情感分类取得了比较好的结果。

3 特征分析

像处理其他分类问题一样,情感分类的两项关键任务是设计有效的分类器和选取有效的特征。对于分类器设计而言,很多分类器模型已较为成熟,那么,特征的选择与使用方式无疑成了被重点关注的焦点问题。为此,本文也是将重点放在特征的获取、选择和组合方法研究上。分类器使用基于支持向量机(SVM)的分类模型。

首先我们给出目前研究工作中常用的效果不错的特征作为我们不同特征组合实验对比的基本特征。通过对已有工作的总结,我们引入以下六大类特征,细化为14个小类特征作为基本特征,记作BaseSet(表1)。

表1 基本特征(BaseSet)

类型编号	类型	特征标号	特征内容	描述
1	主题	1	是否含有hashtag	以"#"包围的部分替换为hashtag
2	链接	2	是否含有url	以"http://"开头的链接
3	标点符号	3	是否含有问号	句子中含有?
		4	是否含有感叹号	句子中含有!
4	表情符号	5	正向表情符号个数	正向表情符号:34个
		6	负向表情符号个数	负向表情符号:30个
5	情感词典	7	正向情感词个数	正向情感词:10 353个
		8	负向情感词个数	负向情感词:14 980个
6	词性	9	副词(AD)个数	采用中科院自动化所Urheen① 工具进行分词和词性标注
		10	感叹词(IJ)个数	
		11	普通名词(NN)个数	
		12	代词(PN)个数	
		13	表语形容词(VA)个数	
		14	其他动词(VV)个数	

由表1所示,我们可以更加清晰地了解到,在已有的工作中,主要是将主题、链接以及标点符号是否出现作为特征,缺乏对主题特征的进一步挖掘。另外,对于情感词典,也仅仅利用正负向情感词的个数作为特征,而没有涉及情感词本身的内容,且缺少不同情感词对于分类影响程度不同的区分性,这无疑会大大影响情感分类的性能。

所以,我们有必要从微博的特点出发,逐步引入词汇化主题特征、情感词特征以及概率化情感词倾向性特征作为基础特征的扩展,以提高分类精度。

3.1 词汇化主题特征

通过对微博数据的观察,我们发现很多的微博都含有主题词,例如,"#天主教#那些假借信仰而误导世人者必下地狱……",其中由"#"包围的"天主教"就是主题词。对于某一主题下的微博,情感极性往往会有一定的倾向性。例如,在谈到"富二代"和"官二代"的主题中,负向情感极性就比较多。但是,在谈及一个产品的时候,正负极性的比例会和产品本身有很大的关系。因此,微博的主题词内容能够给情感分类带来一定的先验知识。

① http://www.openpr.org.cn/

基于以上分析，我们在已有的方法仅考虑一条微博是否有主题的基础上，更进一步探讨了主题内容对于情感分类的影响。我们将 Hashtag 的内容作为特征加入到了分类中。例如，“奖状植入广告＃满天飞的广告，就不能留点净土！”中，我们将“奖状植入广告”作为一个词汇化主题特征直接加入到分类中。特征权重，采用 0/1 二值化权重，出现为 1，不出现为 0。

由于某些微博的主题词出现的概率非常小，对分类提供的帮助不大，所以，我们只选取那些出现频次大于某一阈值的主题词作为特征，特征描述见表 2。

表 2　分类特征-主题词内容特征

类型编号	类型	特征标号	特征内容	描述
1	主题	15	主题词内容	微博训练集中出现频率较多的主题词

3.2　情感词特征

如表 1 所示，在 BaseSet 中对于情感词特征只是简单统计了正向情感词数和负向情感词数。这样处理只用到了情感词语极性对于句子分类的部分信息，而没有考虑到情感词语本身对于微博情感的贡献。有很多的情感词，可以直接表征句子的情感极性，例如，在训练数据中，凡是出现“坑爹”这个负向情感词的句子均为负极性。回顾前人的工作，对于“坑爹”这个词只是作为负极性词，仅仅增加了负向情感词的个数，这样势必导致它的作用很容易被其他情感词的正负极特征所湮没，因此，我们将情感词本身作为一元词汇（unigram）特征，加强了情感词本身在分类器中的作用。

其次，如果将情感词典中的情感词全部作为情感词特征的话，会造成特征空间膨胀并加重数据稀疏，分类结果会很差。所以，引入多少情感词也是一个值得我们思考和研究的问题。

为了找到合适的情感词特征空间，我们采用了两种方法对情感词进行筛选。

方法 1　直接按照情感词典中的词语在训练集合中出现的频次进行排序，选取其中的前 N 个进行了测验。特征的描述如表 3 所示。

对于情感词的特征权重我们采用的仍然是二值化方法，出现为 1，不出现则为 0。

方法 2　在第 1 种方法中，情感词的频次并不能够严格地说明情感词的重要程度。所以，为了解决这个问题，我们采用了 CHI[15-16] 的打分方法，而不是简单地通过频次进行排序。

表 3　分类特征-情感词特征 1

类型编号	类型	特征标号	特征内容	描述
5	情感词典	16	训练集中出现的频次较高的情感词	通过计算频次选取训练集中出现较多的情感词

χ^2 统计量（CHI）是特征项 t_i 和类别 C_j 之间的相关联程度，并假设 t_i 和 C_j 之间符合具有一阶自由度的 χ^2 分布。特征对于某类的 χ^2 统计值越高，它与该类之间的相关性就越大，携带的类别信息就越多。公式（1）给出了 χ^2 的计算方法。

$$\chi^2(t_i,C_j)=\frac{N\times(A\times D-C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{1}$$

其中，N 表示训练语料中文档的总数；A 表示属于 C_j 类且包含 t_i 的文档的频数；B 表示不属于 C_j 类但包含 t_i 的文档的频数；C 表示属于 C_j 类但不包含 t_i 的文档的频数；D 是既不属于 C_j 类也不包含 t_i 的文档的频数。

对于多类问题，基于 χ^2 统计量的特征提取方法采用下面的方法提取，分别计算 t_i 对于每个类别的 CHI 值，然后，在整个训练语料上计算。如式（2）所示。

$$\chi^2_{\mathrm{MAX}}(t_i)=\max_{j=1..M}\{\chi^2(t_i,C_j)\} \tag{2}$$

其中，M 表示的是类别的总数。

通过打分以后，我们仍然对于词语作为特征的数量进行了分析。特征描述见表 4。

表 4　分类特征-情感词特征 2

类型编号	类型	特征标号	特征内容	描述
5	情感词	17	训练集中出现的 CHI 打分较高的情感词	通过计算 CHI 选取训练集中出现较高的情感词

3.3　概率化的情感词倾向性特征

在前面的特征中，对于情感词我们通过查阅情感词典进行了频次计数。但是，这样的做法忽略了情感词本身的倾向性概率问题。对于不同的情感词来说，所带有的情感倾向性概率是不一样的。例如，在微博中表达负面情绪时，“坑爹”要比“猥琐”的倾

向性要深的多。因此，我们认为有必要引入概率化的情感词倾向性特征。

现在已开始有研究人员关注到带情感倾向性概率的情感词典的重要性，并根据自己的理解和定义对传统的情感词典添加情感倾向性信息。但是，他们往往直接将这些情感词典搬来使用，这样就会有很大的问题。第一，面对新的微博领域，微博构词灵活，新词出现的频率较高，情感词典本身不能覆盖微博中的情感词。第二，微博情感词的倾向性分布与情感词典的分布未必一致，如果强制把情感词典的倾向性强加于微博分类未必起到很好的作用。

为了克服上面的问题，我们做了如下的处理：对于情感词典，尽可能地从各个领域收集情感词，也从微博中加入微博常用语，扩大覆盖范围。之后，为了适应微博情感词用语环境，我们在标注数据中对微博情感词典的倾向性概率进行打分。打分的准则如式(3)所示。

$$\text{Score}(word_i) = \frac{\text{Counts}_i(word_i)}{\text{Sum}(word_i)} \quad (3)$$

其中，$i \in \{pos, neg\}$，pos 和 neg 分别表示正面和负面两个极性。

$\text{Counts}_i(word_i)$ 表示某一极性的词语在已标注情感极性的样本中对应的极性出现的次数。具体情况是，如果 $word_i$ 是一个正极性的情感词，我们在正极性的样本中统计其出现的次数；如果是负极性情感词，则在负极性的样本中进行统计。

$\text{Sum}(word_i)$ 表示该词语在整个语料中出现的次数。

利用上述方法，我们构建了一个适应微博分类的带有倾向性概率的情感词典。利用这个情感词典，我们对原来的简单采用频次叠加的方法替换为进行概率叠加的方法，从而反应整条句子的情感程度。特征的描述如表5所示。

表5 分类特征-情感词典特征

类型编号	类型	特征标号	特征内容	描述
5	情感词典	18	情感词的倾向性概率叠加	微博中出现情感词，按照情感词语进行的正负极性的各自叠加

4 实验结果及分析

4.1 实验设置

实验使用的语料来自两个方面，一个是第一届自然语言处理与中文计算会议(NLP&CC)评测中的微博语料①。另一部分使用由新浪API抓取的电影、名人和热点事件的微博。我们一共标注了客观句以及正、负极性的微博各1 200句。最后，从这三个类别中各随机抽取1 000句作为训练集，200句作为测试集，进行实验。

情感词典一部分来自HowNet[17]的情感词词典，一部分来自NTUSD情感词集合，另一部分来源于我们自己人工标注的网络常用语。最后，情感词典包含正向情感词10 350个，负向情感词14 980个。

除此之外，我们还从新浪API获取了官方的表情符号，并且选取了常用的64个表情符号，其中，包括34个正向表情符和30个负向表情符号，组成正负极性表情符号词典。

在对微博处理的过程中，我们还用到了否定词列表，否定词列表包括15个词，包括："不"，"不是"，"不然"，"不行"，"不要"，"没"，"没有"，"无"，"否"，"非"，"不够"，"不可"，"未"，"绝非"，"并非"。处理否定词的算法比较简单，对于每一个情感词，我们对其开了一个长度为4的前驱窗口，判断该情感词语的前面4个词内，出现的否定词的个数，如果出现的是偶数次，则情感词的极性不改变，如果是奇数，则翻转情感词的极性。另外，为了避免微博分类中停用词带来的影响，我们收集了一个停用词表，包含50个停用词和无用符号。例如，"的"，"了"，"在"，"有"，"和"，"就"等。

整个实验采用的分类器是libsvm[18]②，该分类器的设置都使用默认参数设置。

4.2 数据预处理

在执行分类之前我们对数据进行了预处理。预处理的主要工作包括以下几个方面。

1) 数据中所有的半角符号和全角符号的统一化；

2) 所有主题词用"Hashtag"标签进行替换；

3) 所有@信息用"AT"标签进行替换；

4) 所有链接用"URL"标签进行替换；

5) 所有数字用"NUM"进行替换。

经过上述处理之后，对于微博内容进行分词处

① http://tcci.ccf.org.cn/conference/2012/pages/page04_evares.html

② http://www.csie.ntu.edu.tw/~cjlin/libsvm/

理，我们使用 Urheen[19] 进行分词和词性标注①。为了减少分词错误，我们人工筛选了 152 个微博常用语组成了一个微博词典以辅助分词。

4.3 实验结果

根据上述分析，我们对本文提出的新的特征加入和使用方法进行了一系列的实验验证。

在实验结果中，我们对三个类别的每个类别计算了其准确率、召回率和 F1 值得分，并计算整体的正确率。

4.3.1 基础特征分类性能

使用 BaseSet 作为特征集合的实验结果见表 6。

表 6 基准系统的分类效果

类别	准确率/%	召回率/%	F1 值/%	正确率/%
负极性	60.58	83.00	70.03	NA
正极性	75.80	72.00	73.85	NA
中性	94.85	64.50	76.78	NA
总体	NA	NA	NA	73.17

从该组实验结果，我们可以看出，对于客观句（中性句）的识别准确率很高，达到了 94.85%，但是召回率却偏低。这个现象产生的原因可能是由于我们的情感词典规模较大，覆盖面较广，而客观句中也出现了这些情感词。但是，BaseSet 中对于情感词典只是简单地对频次进行叠加，使得出现在客观句中的情感词与出现在主观句中的情感词统一对待，将客观句误判为主观句。同时，由于被判为客观句的句子，含有极少的情感词，所以，基本上全部属于客观句，因此，准确率非常高。同时对负极性句子判断的准确率偏低。这个现象在很多的微博情感分类中经常见到。因为在微博中人们习惯于采用“否定＋褒义词”的说法，去表达贬义的概念，但是，却很少用“否定＋贬义”的说法来表达褒义的概念。例如，“这样的做法不是很好，以后有待改善！”，在这句话中，作者就使用了“否定＋褒义”的做法来表达贬义的概念。而微博贬义的识别也是一个难点，在文章中我们只是简单地使用前驱窗口（大小为 4）来识别贬义，这样就造成了很多贬义句的识别错误。而误把非贬义的表达按照模板强行规约到贬义中，例如，“这有什么不好的”，句子本意是正极性的句子，但是，我们却由于找到“好”，以及否定词“不”而误判成了负极性。中文中否定形式非常多样化，这可能是造成否定类别准确率较低的原因。

4.3.2 加入词汇化主题特征分类性能

在“BaseSet＋特征 15 词汇化主题特征”的实验中，我们选取了出现频次大于 4 的主题作为特征加入进来，结果如表 7 所示。

表 7 主题词加入的结果

类别	准确率/%	召回率/%	F1 值/%	正确率/%
负极性	67.54	77.00	71.96	NA
正极性	81.82	85.50	83.62	NA
中性	88.34	72.00	79.34	NA
总体	NA	NA	NA	78.17

从表 7 的结果中可以看出，加入词汇化主题特征以后，引入了对于同一类主题的先验概率，三个类别的 F 值都有了明显的提升，总体的准确率也有了显著的提高，提高了五个百分点。分析其原因，是因为加入了词汇化主题特征之后，对于特定主题引入了先验概率，将之前这一主题下误判的句子进行了纠正。

整体来看，实验结果说明词汇化主题特征在微博分类中有着非常重要的作用。

4.3.3 加入情感词特征分类性能

• 根据情感词频次选词

为了比较情感词特征的作用，我们对 3.2 节中提到的方法进行对比。方法 1 通过频次选取情感词的方法，我们分别选取了 25、50、75、100、150、200 个情感词作为特征加入到基本的特征集合中进行实验比较，使用的特征集合为 BaseSet＋特征 15＋特征 16。实验结果如表 8 所示。

表 8 方法 1 情感词选择结果比较

数目	类别	准确率/%	召回率/%	F1 值/%	正确率/%
0	负极性	67.54	77.00	71.96	78.17
	正极性	81.82	85.50	83.62	
	中性	88.34	72.00	79.34	
25	负极性	68.67	80.00	73.90	79.50
	正极性	83.50	86.00	84.73	
	中性	90.06	72.50	80.33	
50	负极性	69.13	79.50	73.95	79.50
	正极性	82.13	85.00	83.54	
	中性	90.80	74.00	81.54	

① http://www.openpr.org.cn/

续表

数目	类别	准确率/%	召回率/%	F1值/%	正确率/%
75	负极性	69.16	78.50	73.53	79.33
	正极性	81.52	86.00	83.70	
	中性	90.75	73.50	81.22	
100	负极性	69.70	80.50	74.71	79.83
	正极性	82.52	85.00	83.74	
	中性	90.80	74.00	81.54	
150	负极性	70.63	81.00	75.46	**80.33**
	正极性	83.66	84.50	84.08	
	中性	90.42	75.55	82.32	
200	负极性	69.40	80.50	74.54	79.67
	正极性	83.50	83.50	83.50	
	中性	89.30	75.00	81.53	

从上面可以看出加入情感词本身作为特征，比直接叠加而忽略词语本身作用的特征更有效。基本上所有类别的准确率和召回率都有所提升。这一结果验证了我们之前曾经说过的，很多情感词能够直接说明句子的极性，如“坑爹”，“伤不起”等词，这些情感词作为特征的有效性非常明显。

在实验中，加入150个词语作为特征的性能最好，准确率达到了80.33%，在基线系统上提高了2.16个百分点。但是从表8中，我们也注意到，准确率的提高并非与加入的词的数量成正比。以下我们对加入词语之后，各个类别的准确率变化情况做简要分析。如图1所示。

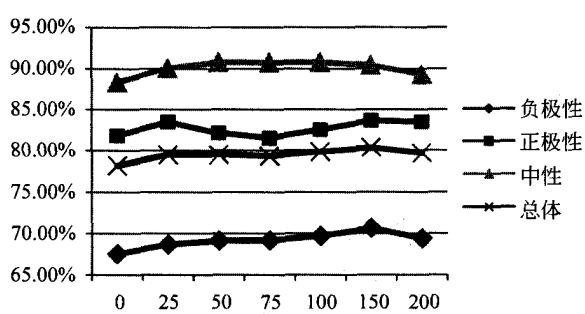

图1 方法1情感词趋势分析

从图1可以看出，除了正极性类别之外，其他的准确率出现先上升后下降的波动。在刚开始增加情感词时，分类器从中学习到了知识，准确率提升比较明显。当达到一定程度之后，继续增加情感词的话，会使得空间维度变大，数据稀疏，从而造成了准确率下降。

结果表明，加入情感词对于分类效果的提升有很大帮助，但是，盲目地加入情感词特征，反而会使效果下降。所以，情感词特征并不是越多越好。而是需要寻找一个比较合适的值。

从结果中我们可以看出，情感词特征的最大值大概在150到200之间。

• 根据CHI方法选情感词

我们按照3.4中的方法2，利用CHI进行词语的选取。同样也选取了25、50、75、100、150、200个词语作为特征进行分类比较。结果如表9所示。

表9 方法2情感词选择结果比较

数目	类别	准确率/%	召回率/%	F1值/%	正确率/%
0	负极性	67.54	77.00	71.96	**78.17**
	正极性	81.82	85.50	83.62	
	中性	88.34	72.00	79.34	
25	负极性	74.66	82.50	78.38	82.33
	正极性	84.39	86.50	85.43	
	中性	89.66	78.00	83.42	
50	负极性	75.23	80.50	77.78	82.83
	正极性	85.99	89.00	87.47	
	中性	88.27	79.00	83.38	
75	负极性	74.31	81.00	77.51	82.67
	正极性	85.92	88.50	87.19	
	中性	89.20	78.50	83.51	
100	负极性	75.23	82.00	78.47	**83.33**
	正极性	86.70	88.00	87.34	
	中性	89.38	80.00	84.43	
150	负极性	72.52	80.50	76.30	81.83
	正极性	85.51	88.50	86.98	
	中性	89.47	76.50	82.48	
200	负极性	75.61	77.50	76.54	82.50
	正极性	84.88	87.00	85.93	
	中性	87.37	83.00	85.13	

从上面可以看出利用方法2加入的词语在相同数量上，都要比之方法1加入的情感词特征有效。CHI更加合理地区分了情感词对于分类的重要性。对于负极性类别，在加入情感词时变化比较明显。究其原因，可能是使用频次选取情感词的方法只是

简单地计算该情感词整体的频次，而忽略了情感词对于不同类别的贡献度。CHI 方法能够比较合理地估计词语对于不同类别的贡献度。这样能够区分出经常出现在负极性类别中的“否定＋褒义”的表达方式中的情感词，在一定程度上增加了对于这种表达方式的识别度。从整体来看，方法 2 提高了五个百分点，在方法 1 的基础上又提高了三个百分点。这说明采用不同的方法来对情感词进行情感程度的区分是很重要的。

以下我们对加入词语之后各个类别的准确率变化情况以及与方法 1 的比较进行分析(图 2)。

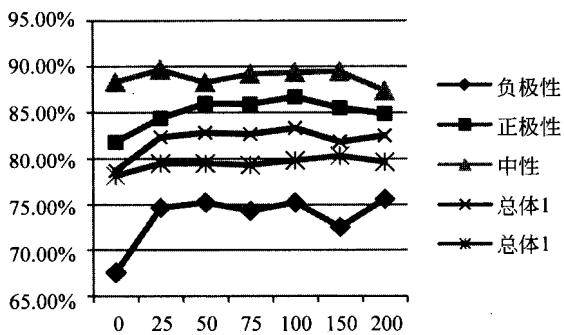

图 2　方法 2 情感词变化趋势

从图 2 可以看出，与方法 1 的变化趋势基本相同，所有类别在加入特征之后准确率都有很大的提升，但仍然有不同之处。首先，达到最高值之后两种方法都开始趋于平稳，而方法 1 之后下降比较慢，方法 2 却下降较快。从方法 2 的走向趋势我们可以看出，情感词特征的最大值大概在 100 到 150 之间。达到最大值的速度最多只需要 100 个词。所以，方法 2 比方法 1 在达到最好效果的速度上有优势，而且只需要较少的词就可以达到比方法 1 更好的结果。但是，相对而言在维数增加时下降的趋势也比较明显。在方法 2 中后续加入的情感词对于分类的作用并不明显，反而由于特征维数的增加带来了过多的噪声，使得方法 2 的下降趋势更加明显。

为了验证方法 2 在情感词维数增加时准确率的变化趋势，我们做了进一步实验。

从图 3 可以看出，在利用 CHI 增加特征维数的过程中，准确率开始提升，达到一个峰值后准确率在波动中逐步下降。这一结果再次验证了并不是维数越高越好的结论。

4.3.4　加入概率化情感词倾向性特征的分类性能

本实验以情感词特征中方法 2 达到最好结果的特征为基础，分析加上词典打分之后的情况，结果如表 10 所示。

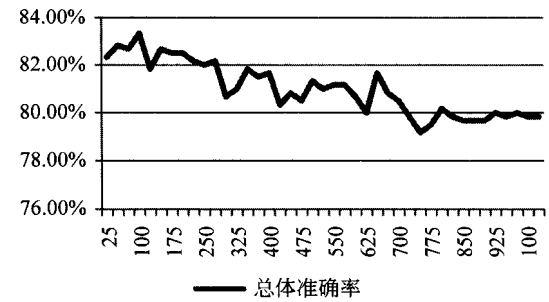

图 3　方法 2 情感词数量变化形势图

表 10　词典打分后的结果

类别	准确率/%	召回率/%	F1 值/%	正确率/%
负极性	76.23	81.00	78.54	NA
正极性	87.70	88.45	88.07	NA
中性	90.38	80.00	84.87	NA
总体	NA	NA	NA	84.17

可以从结果中看出，加入词典打分之后正极性和客观的分类效果有了明显的提高。总体的准确率也上升了 0.84 个百分点。分析词典的打分对于客观分类有提升的原因，可能是由于之前我们简单地对情感词进行累加频次，没有考虑各个情感词的倾向性概率对于整体句子情感分值的影响。使得本来非主观性的句子被误分为主观句。加入倾向性概率以后，有效地表达了句子中整体的情感极性。

总体来看，加入情感词的倾向性概率之后整体分类效果有了比较明显的提高，这验证了概率化情感词倾向性特征的有效性。

4.3.5　整体对比分类性能

最后我们将每一类特征中最好的结果放到一起进行比较，实验结果如图 4 所示。

从图 4 中可以直观地看出，在特征加入的过程中整体准确率在稳步上升。

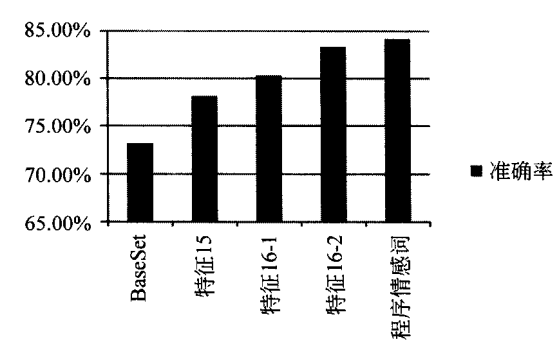

图 4　整体对比分类性能图

为了进一步验证我们方法的有效性，我们将本文提出的多样化特征分类方法与谢立星提出的一步三分类的方法进行比较，实验结果如表11所示。

表11 两种方法的对比

方法	正确率/%
多样化特征分类方法	84.17
一步三分类方法	79.53

通过对比可以看出，我们的方法比之谢丽星的方法提高了4.64个百分点，由此，更进一步说明了我们方法的有效性。

5 结论及下一步工作

本文在充分研究微博情感分类的基础上，结合传统方法，主要做出了以下贡献：1)对于有关主题特征，不仅考虑主题是否出现，而且考虑了主题词的特定内容；2)对于情感词，不仅深入地分析和探讨了情感词的加入方法，而且详细研究了情感词加入的数量对于整体分类的效果影响；3)考虑到通用的情感词典首先不能及时覆盖和添加日新月异的网络用语，同时针对微博数据也没有权重区分，我们提取了微博用语来丰富和拓宽通用情感词典，并使用微博数据对该词典倾向性概率进行打分，将概率打分作为特征取代原始的布尔特征，从而更加真实地反应微博情感倾向。实验表明，这种方法使得微博情感分类准确率达到了84.17%。

在下一步工作中，我们将研究对表述方式基本相似的主题进行聚类的方法，以减少领域不同带来的问题，并缓解数据稀疏问题。同时，探究情感词加入数目的规律，进一步提升待分类问题的分类效果。另外，针对中文否定形式的表达多样性，我们将提出中文微博否定形式的解决办法。

参考文献

[1] A Das, S Bandyopadhyay. Dr Sentiment knows everything! [C]//Proceedings of the ACL-HLT, 2011: 50-55.

[2] A Joshi, A Balamurali, P Bhattacharyya, et al. C-feel-it: A sentiment analyzer for micro-blogs[C]//Proceedings of the ACL-HLT, 2011 :127-132.

[3] P Chesley, B Vincent, L Xu, et al. Using verbs and adjectives to automatically classify blog sentiment[J]. Training, 2006, 580(263).

[4] 刘鲁，刘志明. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用，2012,48(1):1-4.

[5] L Jiang, M Yu, M Zhou, et al. Target -dependent twitter sentiment classification [C]//Proceedings of ACL-HLT, 2011:151-160.

[6] S Prasad. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods [N]. Technical Report, Stanford University, 2010, Available at http://www-nlp. stanford. edu/courses/

[7] Y Lu, M Castellanos, U Dayal, et al. Automatic construction of a context-aware sentiment lexicon: an optimization approach[C]//Proceedings of the 20th international conference on World wide web, 2011: 347-356.

[8] P D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002 :417-424.

[9] B Pang, L Lee, S Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques [C]//Proceedings of EMNLP, 2002:79-86.

[10] T Mullen, N Collier. Sentiment Analysis using Support Vector Machines with Diverse Information Sources [C]//Proceedings of EMNLP, 2004: 412-418.

[11] A Go, R Bhayani, L Huang. Twitter sentiment classification using distant supervision[J]. CS224N Project Report, Stanford University, 2009: 1-12.

[12] A Pak, P Paroubek. Twitter as a corpus for sentiment analysis and opinion mining[C]//Proceedings of LREC, 2010:1320-1326.

[13] D Davidov, O Tsur, A Rappoport. Enhanced sentiment learning using twitter hashtags and smileys [C]//Proceedings of the 23rd International Conference on Computational Linguistics,2010:241-249.

[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):73-82.

[15] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008.

[16] T Dunning. Accurate methods for the statistics of surprise and coincidence[J]. Computational linguistics, 1993, 19(1): 61-74.

[17] Dong Z, Dong Q. HowNet [EB/OL]. Available at http://www. keenage. com/ 2000

[18] C C Chang, C J Lin. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST),2011,2(3):1-27.

[19] K Wang, C Zong, K Y Su. A character-based joint

model for Chinese word segmentation[C]//Proceedings of the 23rd International Conference on Computational Linguistics, 2010:1173-1181.

张志琳(1988—)，硕士，助理研究员，主要研究领域为情感分类方法研究。
E-mail：zhilin.zhang@ia.ac.cn

宗成庆(1963—)，博士，研究员，主要研究领域为机器翻译、情感分类和自然语言处理等相关领域的研究。
E-mail：cqzong@nlpr.ia.ac.cn

(上接第 79 页)

[12] Meng Jiana, Lin Hongfei, Li Yanpeng. Knowledge transfer based on feature representation mapping for text classification [J], Expert Systems with Applications, 2011, 38(8): 10562-10567

[13] Andrew Arnold, Ramesh Nallapati, William W. Cohen. A comparative study of methods for transductive transfer learning[C]//Proceedings of the 7th IEEE International Conference on Data Mining Workshops. Omaha, Nebraska, USA: IEEE Computer Society, 2007: 77-82.

[14] Pengcheng Wu, Thomas G. Dietterich. Improving svm accuracy by training on auxiliary data sources [C]//Proceedings of the 21st International Conference on Machine Learning, Morgan Kaufmann, 2004: 871-878.

[15] Vikas C. Raykar, Balaji Krishnapuram, Jinbo Bi, et al. Bayesian multiple instance learning: automatic feature selection and inductive transfer[C]//Proceedings of the 25th International Conference on Machine learning. 2008: 808-815.

[16] Lawrence Page, Sergey Brin, Rajeev Motwani, et al. The PageRank citation ranking: bringing order to the web, Technical Report [R], Stanford University, Stanford, CA, 1998.

[17] 郑伟，王朝坤，刘璋等，一种基于随机游走模型的多标签分类算法[J]，计算机学报，2010，33(8)：1418-1425

[18] Thorsten Joachims. Text Categorization with Support Vector Machines: Leaning with Many Relevant Features[C]//Proceedings of the 10th European Conference on Machine Learning, 1998: 137-142.

孟佳娜(1972—)，博士，教授，主要研究领域为自然语言处理及文本挖掘。
E-mail：mengjn@dlnu.edu.cn

赵丹丹(1975—)，硕士，讲师，主要研究领域为自然语言处理及机器学习。
E-mail：zhaodd@dlnu.edu.cn

于玉海(1980—)，硕士，讲师，主要研究领域为深度学习及情感计算。
E-mail：yuyh@dlnu.edu.cn