基于分类的微博新情感词抽取方法和特征分析

刘德喜[1)] 聂建云[2)] 万常选[1)] 刘喜平[1)] 廖述梅[1)]
廖国琼[1)] 钟敏娟[1)] 江腾蛟[1)]

[1)](江西财经大学信息管理学院 南昌 330013)

[2)](蒙特利尔大学计算机科学与运筹学系 蒙特利尔 H3C 3J7 加拿大)

摘 要 情感或情绪分析在舆情分析、商品评论分析、商品推荐等领域应用广泛,而文本中的情感或情绪分析通常以情感词典为基础.人工情感词典虽然准确但构建代价大、难以及时更新,很难适应微博这类新情感词快速更迭的数据.微博平台为新情感词的发布和传播提供了便捷的途径,是新情感词的重要来源.考虑到已有规模较大的人工情感词典及大量包含新情感词的微博数据,在统计、分析、对比中、英两种语言微博中情感词分布差异的基础上,提出了与特定语言无关的基于分类思想的微博新情感词抽取方法 cNSEm. cNSEm 根据微博数据集和情感词典自动构建训练数据、训练分类器并判别候选词的情感极性,最后采用投票机制确定候选词的情感极性.通过大量而细致的实验,分析了 cNSEm 在中、英文两种语言的微博数据上的表现、六类特征的作用和用法以及抽取的新情感词对微博情感分类任务的帮助作用.实验结果表明,cNSEm 比经典的基于共现和极性传播的方法要好,特别是当考虑中文微博数据集中的名词类情感词时.对 cNSEm 抽取的新情感词进行了直接和间接两种方法评测,前者利用人工情感词典作参照,后者考察抽取的新情感词对情感分类的帮助作用,从评测指标上看,cNSEm 抽取的新情感词与人工情感词典的质量相当,并且 cNSEm 能适应有较大差异的中、英两个语种.

关键词 微博;新情感词抽取;cNSEm 方法;特征分析

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2018.01574

A Classification Based Sentiment Words Extracting Method from Microblogs and Its Feature Engineering

LIU De-Xi[1)] NIE Jian-Yun[2)] WAN Chang-Xuan[1)] LIU Xi-Ping[1)] LIAO Shu-Mei[1)]
LIAO Guo-Qiong[1)] ZHONG Min-Juan[1)] JIANG Teng-Jiao[1)]

[1)](*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang* 330013)

[2)](*Department of Computer Science and Operations Research, University of Montreal, Montreal H3C 3J7, Canada*)

Abstract Text sentiment analysis tries to get the orientation (attitude, point of view, or emotion) of information publishers, which is widely used in the field of public opinion supervision, product reviews analysis, et al., and has become one of the hottest topics in natural language processing, social media processing, data mining, etc. Sentiment analysis or emotion analysis on text is always based on a sentiment dictionary. Manually-built sentiment dictionary may produces high accuracy however with limited coverage and updating difficulty, which is hard to cope with situation under Web 2.0, where new sentiment words are created more frequently and spread more quickly.

收稿日期:2016-06-25;在线出版日期:2017-05-26.本课题得到国家自然科学基金(61762042,61363039,61562032)、江西省落地计划项目(KJLD14035)、江西省自然科学基金(20171BAB202021, 20152ACB20003)资助.**刘德喜**,男,1975 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为社会媒体处理、信息检索、自然语言处理. E-mail: dexi.liu@163.com. **聂建云**,男,1963 年生,博士,教授,博士生导师,主要研究领域为信息检索. **万常选**,男,1962 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为 Web 数据管理、数据挖掘. **刘喜平**,男,1981 年生,博士,副教授,主要研究方向为 Web 数据管理、数据挖掘. **廖述梅**,女,1976 年生,博士,副教授,主要研究方向为信息管理与信息系统. **廖国琼**,男,1969 年生,博士,教授,博士生导师,主要研究领域为社会计算. **钟敏娟**,女,1976 年生,博士,副教授,主要研究方向为 Web 数据管理、数据挖掘. **江腾蛟**,女,1976 年生,博士,讲师,主要研究方向为情感分析.

Microblog platforms, such as Twitter and Sina Weibo, allow users to publish and transmit information freely, and become important sources of new sentiment words. By using large manually-built sentiment dictionaries and microblog data with mass sentiment words online, this paper analyzes distribution difference of Chinese and English sentiment words, and cNSEm is proposed to extract new sentiment words from microblogs, based on classification principle. cNSEm automatically generates candidate samples, which are classified by a trained classifier, and then sorted and extracted according to a voting strategy. The classification based methods have been used to extract new sentiment words in some related works. However, most of them extracted sentiment words from web pages, Wordnet, or product reviews, and candidate words are usually constrained on adjectives. cNSEm has to deal with not only the informal expression of microblogs but also the expanded POS candidates, especially when nouns are included. Based on some carefully designed experiments, we analyze the performance of cNSEm on both Chinese and English microblogs. We also analyze and compare the impacts of six categories of features used in cNSEm, including context, POS, language mode, modify relationship, sentence feature and co-occurrence with other sentiment words. Experimental results show that six categories of features employed by cNSEm play different roles in sentiment words extraction and polarity setting in different languages. Experimental results on Chinese microblogs also show that the classical co-occurrence besed methods are effective when candidates are adjectives, but their performance degraded when nouns are included. However, cNSEm performs better than co-occurrence based methods, especially when nouns are considered as candidate sentiment words on Chinese microblogs. To evaluate cNSEm performance, we also test the impacts of extracted sentiment words on sentiment classification tasks. Experimental results on Chinese microblogs show that the performance of microblog subjectivity classification and polarity classification has been improved significantly after sentiment dictionary expanded by cNSEm, and cNSEm performs better than benchmark method. As for classifying subjective terms on English microblogs, the benchmark method and cNSEm perform closely, while cNSEm perform better than benchmark method for polarity classification task. Surprisingly, the sentiment words extracted by cNSEm are more helpful for sentiment classification tasks than manual sentiment dictionaries. In conclusion, both the direct evaluation results by ideal sentiment dictionaries and the indirect evaluation results by sentiment classification tasks show that the new sentiment word extracted by cNSEm are competitive with manual sentiment words. Moreover, cNSEm is adaptive to both Chinese and English microblogs, which have great difference between two languages.

Keywords　microblogs; new sentiment words extraction; cNSEm method; feature engineering

1　引　言

文本情感分析(以下简称“情感分析”)是利用自然语言处理、机器学习、数据挖掘等技术,通过文本内容分析其作者的观点、态度、情感或情绪,分析的文本对象包括新闻、评论、微博等.情感词典在文本情感分析任务中扮有重要角色,是很多情感分析方法中情感倾向性和情感极性判断的重要依据,情感词典的质量甚至会直接或间接决定着情感分析的效果[1].因此,包括大连理工大学、清华大学、台湾大学等很多研究机构或团队,花费了巨大代价通过人工方式构建高质量的情感词典.

Web 2.0 的发展和手持终端设备的普及让数以亿计的用户通过博客、微博、微信、Twitter 等平台走进自媒体时代,普通民众的参与使得各种网络用语或网络新词快速更迭并迅速传播,这其中不乏大量带有情感的新词或不规范用词,如中文新情感词“弱爆”、“傻 X”、“逆天”等,英文新情感词如“goooood(good 的不规范使用)”“dobe(逗比,形容

人逗、傻得可爱)""obamacize(像奥巴马那样努力奋斗)"等.

新情感词有两种:一是带有情感极性(又称情感倾向性)的未登录词;二是未被已有情感词典收录的登录词,但在某些特定时期、特定领域或特定上下文中表现出了情感极性. COAE2014[①] 的任务 3 定义新情感词为前者,即新情感词是指那些未在通用词典中出现的且带有情感极性的词. 由于本文的实验中部分情感词被随机选择出来用于评测,因此综合考虑这两种新情感词. 由于这些新词并未被通用词典收录,也不会被常用的语义词典如 WordNet、同义词林、HowNet 等词典资源收录,很难用基于词典的方法获取其情感极性.

Twitter、新浪微博、人人网等社会网络平台是普通民众参与自媒体的重要平台,包含大量新情感词,同时也是很多基于情感分析的舆情分析、商品评论分析、商品推荐等工作的数据源. 因此,基于微博的情感分析或情感词抽取受到学者的普遍关注并取得了大量成果. 但由于微博数据主题复杂、语法不规范等特点,使得一些相对成熟的、用于商品评论等特定领域的、或以语法分析为基础的方法无法适应. 已有的新情感词抽取方法通常具有以下三个共同假设:假定候选词仅为形容词、动词、副词;假定情感词之间、相同极性情感词之间的共现程度更高;假定已知的种子情感词典规模较小,通常为数十条或上百条,可以在共现的基础上通过极性传播和传递(Propagation)逐步扩展得到新情感词. 然而,第 3 小节的实验分析发现,与英文情感词分布不同的是,在中文微博中,除形容词、动词、副词外,还有大量的情感词以名词词性出现,而名词类情感词与其它情感词之间的共现并不比名词类非情感词与其它情感词之间的共现更频繁,这导致点互信息等共现特征无法有效地区分名词类情感词与非情感词. 另外,在出现两个以上情感词的中、英文微博中,有近一半的微博中两个或多个情感词的极性并不完全一致.

目前已有的中、英文情感词典非常丰富,例如大连理工大学的 DUTSD[②]、清华大学的 THUSD[③]、知网的 HNSD[④]、台湾大学的 NTUSD[⑤] 等中文情感词典,SentiWordNet[⑥]、MPQA[⑦] 等英文情感词典,这些情感词典收录了少者数千、多者上万条情感词,可以作为训练样本. 此外,类似 Twitter、微博等平台每天数以亿计的微博量,可以用作情感词的上下文或特征. 因此,基于机器学习的新情感词提取方法是比较适合微博数据的,而特征选择则成为该方法成败的关键. 然而,目前还没发现有相关文献系统地讨论什么类型的特征对于微博新情感词抽取是有效的和必要的,也没有文献对从不同语种的微博上抽取新情感词进行对比分析.

本文的工作是在参加 COAE2014 微博新情感词抽取任务基础上的进一步扩展,主要工作包括:

(1)充分利用人工情感词典和微博数据,提出基于分类的微博新情感词抽取算法 cNSEm(classification based New Sentimental words Extracting from microblog)(第 4 节). 已有基于分类的情感词抽取方法中,一般限定候选词为形容词,情感词来源也以 WordNet 语义词典、网页或商品评论数据为主. cNSEm 针对微博数据,并且将中、英文中的名词也纳入候选词. cNSEm 不需要语义词典(微博中的新词通常没有收录进现有的语义词典中)或带有情感极性标注的数据集(收集困难、不适合微博数据),只需要一些种子情感词、通用词典和大量微博数据,获取方便.

(2)通过在中、英文两类微博数据上的大量实验,分析了上下文、词性、语言学模式、修饰关系、句子特征、与情感词的共现等 6 类特征(第 5 节)对 cNSEm 的影响,结果显示,对于不同语种,这 6 类特征对新情感词的抽取及极性判断作用不尽相同(第 6 节). 本文在特征选择时借鉴了已有的研究成果,包括那些并非基于分类方法抽取情感词时用到的线索,但更多地是对这些线索或特征进行抽象,使得特征不再依赖于具体的语言,具有与语言无关的特点.

(3)在中文微博数据集上的实验结果显示(第 6 节),基于共现和极性传播的方法 GPC 的性能对形容词类型的情感词抽取是有效的,但增加名词等更多候选词后,GPC 性能下降严重,cNSEm 方法则表现出良好的性能,抽取得到的新情感词词典与理想(人工)情感词典之间的 Rprec 值与多个人工情感词典之间的 Rprec 值相当.

(4)实验分析了新抽取的情感词对微博情感分类的影响(第 6 节),结果显示,对于中文微博,利用 cNSEm 扩展得到的新情感词能显著提高微博主观性分类和极性分类的效果,且较 GPC 方法要好;而对于英文微博的主观性分类,GPC 与 cNSEm 方法

① http://www.liip.cn/CCIR2014/pc.html
② http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx
③ http://www.datatang.com/data/44522
④ http://www.keenage.com/html/c_bulletin_2007.htm
⑤ http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html
⑥ http://sentiwordnet.isti.cnr.it/
⑦ http://mpqa.cs.pitt.edu/

表现相当，但在极性分类方面，cNSEm 要显著好于 GPC. 并且，扩展后的词典在对微博情感分类的帮助上超过实验中选择的人工词典.

2　相关工作

尽管有不少关于新词识别的文献，但本文的焦点是判断新词的情感极性，因此，对于中文采用能够识别新词的分词系统同时完成分词与新词识别工作，对于英文则视不在给定词典中的词为新词.

商品评论分析、舆情分析、商品推荐等以文本情感分析为基础的应用具有重大的商业价值和社会意义，吸引着越来越多的企业和科研院所参与研究，在文本情感分析的基础研究和应用上取得了丰硕成果[2-6]. 由于情感词典在文本情感分析中的重要作用[7-8]，很多研究者，特别是情感分析工作开展较早的研究者，花费大量人力物力通过人工筛选与标注的方式构建人工情感词典，如前文提到的 DUTSD、THUSD、HNSD、NTUSD、SentiWordNet、MPQA 等，这些高质量的人工情感词典为本领域的研究作出了卓越的贡献. 然而，这些人工情感词典在情感分析时并不能完全满足需要，有大量相关工作是先扩展人工情感词典或种子情感词，再在此基础上开展情感分析工作.

情感词抽取或情感词典扩展方法可以分为两大类，一是基于共现或相似度的方法，二是基于分类的方法. 基于共现或相似度的方法通常利用种子情感词、语义词典以及包含情感词的文本等资源，以候选词与种子情感词的共现、相似性等信息为依据，抽取情感词并判断其极性. 例如，文献[9]认为，与种子情感词共现且用“and”连接的形容词是情感词，且极性与种子情感词相同，而用“but”连接时极性相反. 文献[10-12]则将点互信息 PMI(Point-Wise Mutual Information)作为考察共现强度的指标：与正向种子情感词共现越强、与负向种子情感词共现越弱，该候选词是正向情感词的可能性越大，反之则其是负向情感词的可能性越大. 同理，候选词与种子情感词越“相似”，其是情感词的可能性越大，其极性的判断与基于共现类似，而相似性计算方法不尽相同，包括基于上下文的相似度[13]、基于词在 WordNet 中的语义距离[14]、基于 WordNet 中词的释义[15]等等. 而文献[16]则基于词干来考察这种相似性.

考虑到种子情感词的规模较小，与种子情感词的共现分析或相似性分析不够全面，有学者通过构建词汇图并让情感极性在图中传播的方式，将那些与种子情感词无直接共现或共现较弱的候选词也纳入考察范围[17-21]，或者采用多次迭代的策略，每次抽取少量情感词，多次迭代，不断扩展[22]. 还有一类基于共现和极性传播的方法不仅考虑情感词之间的共现，还考虑情感词与情感对象(例如商品评论中的评论对象)之间的共现，认为在商品评论这类数据中，情感词及评论对象不会孤立地出现，因此当发现评论中有情感词时，附近应该有评论对象，同理，有评论对象的上下文本中应该有情感词. 因此，情感词和评论对象互为证据、协同抽取[23-26].

基于分类的方法将情感词抽取和情感极性判断视为分类问题，通过分类模型，将候选词划分到正极性、负极性和无极性三个类别中. 特征的选择和训练数据的选择是这类方法之间的主要区别. 例如文献[27]以 WordNet 的同义词集为特征、文献[28]以 WordNet 中词的释义(gloss)为特征，它们的训练数据都来 WordNet 中的种子情感词. 文献[29]的训练数据则来自用户在评论中对产品的打分情况，希望抽取的情感词对产品自动打分有帮助.

尽管在新情感词抽取上已有大量研究成果，但在针对不同语种的微博数据时，新情感词抽取效果仍有很大的提升空间，有些核心问题还有待探索，具体体现在：

(1) 候选词词性的限定有局限性. 目前大部分新情感词抽取工作都将候选词限定在形容词、副词等词性上[28-30-32]，抽取方法的有效性也得到了充分的证明，但实验发现，在中文语种的微博中，这种仅以形容词词性为抽取对象是有局限性的[33]. 例如，文献[28]显示，形容词和副词是情感词的概率(0.3966 和 0.3570)远大于动词和名词是情感词的概率(0.1104 和 0.0998). 在英文微博数据上的统计显示，仅以形容词、副词、动词为候选词时，可以覆盖 85% 以上的情感词，因此，权衡准确率和召回率，很多经典方法只将形容词和副词等作为候选词. 然而，在中文微博中，形容词和副词仅覆盖了 21% 的情感词，而仅以名词词性出现在数据集中的情感词就占到已知情感词典的 40%. 因此，对于中文微博，考虑名词类情感词是必要的，但大量以名词词性出现的词并非情感词，这给新情感词的抽取带来巨大挑战.

文献[21]在基于中文微博数据构建情感词典时，将候选词从形容词扩大到成语、习惯用语等，但依然没考虑名词. 文献[34]在构建跨领域的中文情感词典时，除了形容词、副词外，考虑了形容词-名词

短语,仍没有离开“形容词”.文献[35]分析了以名词词性出现的商品可能蕴含的情感极性,但在构建通用的情感词典时,商品并不能纳入情感词典中.例如,微博中关于“中石油”的评论大都是负向,但“中石油”不是情感词.文献[36]在完成文本情感分类任务时,以 400 个情感词为种子,从约 40 亿个 Web 页面里抽取了近 18 万条“新”情感词,尽管扩展后得到的情感词有利于文本情感分类任务,但这些词明显不适合都视为新情感词,因为与种子情感词共现的上下文信息对文本情感分类的确能起到帮助作用.对这近 18 万条“新”情感词的评测结果也证实了这一点:规模上它是 WordNet LP(利用 Label Propagation 方法在 WordNet 上构建的情感词典)的 30 余倍,但对 WordNet LP 的覆盖不到一半.

(2)基于分类的方法依赖语义词典或标注数据集.3.3 小节以及文献[33]中的实验分析显示,将名词作为候选词后,基于共现的情感词抽取方法已不适合微博数据,而目前基于分类的方法或者依赖于语义词典[27-28],或者需要直接[37]或间接标注的数据集[29],这在利用微博数据集抽取新情感词时是不现实的.一方面微博中的“新”情感词在语义词典中不存在,另一方面对微博进行情感标注工作量太大.

(3)特征的选择和使用有待进一步分析.本文提出的基于分类的微博新情感词抽取方法 cNSEm 是以前期参加 COAE2014 评测为基础的[33],由于时间限制,参加评测时所采用的特征仅为种子情感词和候选词上下文中的 n_gram 特征,新的特征及用法还有待进一步挖掘.

(4)多数经典的方法将新情感词的抽取视为中间过程,并利用扩展后的情感词典对文本情感分类的改善作为评价新情感词质量的标准[36].实验发现,如果任务本身就是抽取新情感词而不是文本情感分类,这种间接评价的方式存在误导.例如,在 COAE2014 提供的微博数据中,大量关于“蒙牛”的微博都是负极性的,因此,如果将“蒙牛”视为负极性的情感词是有利于微博情感分类的,但将“蒙牛”纳入到情感词典中不太适合.

3 数据分析

文献[33]统计了情感词在中文微博中的词性分布和情感词之间的共现情况,指出对于中文微博,候选词限定在形容词上以及基于共现的新情感词抽取方法是不合适的.本节将对中、英两种语言的微博数据中情感词的分布情况和共现情况进行对比分析,旨在进一步明确将名词作为候选词的必要性及所带来的挑战.

3.1 数据准备

中文微博数据集 $\mathcal{D}^C$. COAE 2014 任务 3 提供了约 1 千万条中文微博,从中随机选择 50 万条,删除用户名和超级链接等预处理后,用 ICTCLAS2013① 分词并标注词性,用 Stanford 的 CoreNLP② 进行依存句法分析,得到本文实验用的中文微博数据集 $\mathcal{D}^C$. ICTCLAS2013 考虑了中文微博的特点,可以发现新词,如“百菜价”、“套现”、“毒舌”、“钜惠”、“帅哥”等.

中文情感词词典 $\mathcal{S}^C$. 常用的四部中文情感词典 DUTSD、THUSD、HNSD 及 NTUSD,去掉极性为“0”或者在词典内部存在极性歧义的情感词条后,各情感词典的词条数如表 1 所示.

表 1 常用中文情感词典收录的情感词数量

	DUTSD	THUSD	HNSD	NTUSD
正向情感词数量	11174	5566	4431	1810
负向情感词数量	10740	4467	4231	6537

如果将其中一部情感词典视为理想的情感词典(表 2 中的各列),其它词典(表 2 中的各行)相对该理想情感词典的召回率如表 2 所示.其中“P±”表示考虑情感极性时的召回率,即要求情感词的极性判断也正确,而“P”表示不考虑极性时的召回率.反之,如果将各行视为理想情感词典,表中的值则表示其它词典(各列)对理想情感词典的准确率.

表 2 常用中文情感词典之间的召回率
(以各列为理想情感词典)

	DUTSD		THUSD		HNSD		NTUSD	
	P	P±	P	P±	P	P±	P	P±
DUTSD			0.6252	**0.6196**	0.4950	**0.4747**	0.2293	**0.2243**
THUSD	0.2863	**0.2837**			0.3182	**0.3065**	0.2476	0.2416
HNSD	0.1957	**0.1876**	0.2747	**0.2646**			0.1918	0.1816
NTUSD	0.0873	**0.0854**	0.2060	0.2010	0.1848	0.1750		

相比较而言,DUTSD 收录的情感词最丰富,如果将其它三部情感词典视作理想情感词典,DUTSD 对其它三部词典的召回率也最高.然而,尽管 DUTSD 的规模分别是 THUSD、HNSD 及 NTUSD 的 2.2、2.5 和 2.6 倍,但其召回率只有 0.6196、0.4747 和 0.2243,准确率分别为 0.2837,0.1876,0.0854.HNSD 与 THUSD 的规模相当,二者之间的召回率

① http://ictclas.nlpir.org
② http://nlp.stanford.edu/software/corenlp.shtml

和准确率仅为 0.31 和 0.26.各部词典对共同收录的情感词的极性判断上是比较一致的，平均一致率达 96.90%.以上这些值可以作为评测新情感词的参照.

情感词典$\mathcal{S}^{C}$以 DUTSD 为基础，并根据中文微博数据的特点和任务需要做了如下补充和过滤：(1)补充微博中带情感极性的表情符；(2)补充 COAE2014 任务 3 评测时所用的新情感词；(3)过滤掉 DUTSD 中有不同极性的词条(部分词条在不同词性时被标注了不同的极性)以及极性标注为“0”的词条；(4)过滤掉没有出现在数据集$\mathcal{D}^{C}$中或者未被 ICTCLAS2013 正确分词的词条；(5)过滤掉长度超过 4 个汉字或字符的词条.得到的$\mathcal{S}^{C}$共包含情感词 7565 条，其中正极性 3964 条、负极性 3601 条.尽管$\mathcal{S}^{C}$无法覆盖全部情感词，但包含了大多数常用的且出现在$\mathcal{D}^{C}$中的情感词，其规模也与 THUSD、HNSD 及 NTUSD 相当，因此$\mathcal{S}^{C}$中情感词的分布具有一定代表性.

中文非情感词词典$\mathcal{O}^{C}$. 人工情感词典通常是在通用词典上筛选得到的，因此那些被多个人工情感词典过滤掉的词可以被视为没有情感极性，如式(1)所示：

$$\begin{aligned}&\mathcal{O}=\text{CommonDict}-\text{MixedSD},\\&\text{MixedSD}=\text{DUTSD}\cup\text{HNSD}\cup\text{THUSD}\cup\text{NTUSD}\cup\\&\qquad\text{FACIAL}\cup\text{COAESD}\end{aligned}\tag{1}$$

其中 CommonDict 是在 COAE2014 任务 3 的通用词典基础上，补充了人名、地名等，扩展后的规模为 2836.74 K；MixedSD 是多个情感词典的并集，共含 38906 个情感词(其中正极性 16272 条、负极性 18898 条、无极性 3124 条、有极性歧义 612 条).除前文提到的多个人工情感词典外，补充了 FACIAL 和 COAESD 两个情感词典，其中 FACIAL 为标注带有情感极性的表情符，COAESD 为 COAE2014 任务 3 评测时所用的新情感词词典，并去除未被 ICTCLAS2013 正确分词的词条.

英文微博数据集$\mathcal{D}^{E}$. $\mathcal{D}^{E}$来自文献[22]的作者所提供的一百万 tweet id 号，由于部分 tweet 在下载时已被删除，$\mathcal{D}^{E}$中的 tweet 仅有 991248 条.用 Stanford 的 CoreNLP 标注词性并分析依存关系后，得到$\mathcal{D}^{E}$.

英文情感词词典$\mathcal{S}^{E}$. 选用 MPQA 中那些在多个相关文献中用作种子情感词典的、主观性强的、且在$\mathcal{D}^{E}$中出现的词，再补充它们的曲折变化形式(inflectional forms)(对于动词、形容词或副词)或复数形式(对于名词)，得到包含正极性 2416 条、负极性 3218 条的情感词典$\mathcal{S}^{E}$.

英文非情感词词典$\mathcal{O}^{C}$. SentiWordNet 对 WordNet 中全部词条都计算了其具有正向极性和负向极性的概率，那些正向极性和负向极性概率都为 0 的词条可视为非情感词.对名词补充其复数形式，对动词、形容词和副词，补充其曲折变化形式，得到包含 214162 条非情感词的$\mathcal{O}^{C}$.

为表述方便，下文称中文数据集和词典为$\mathcal{R}^{C}=(\mathcal{D}^{C},\mathcal{S}^{C},\mathcal{O}^{C})$，英文为$\mathcal{R}^{E}=(\mathcal{D}^{E},\mathcal{S}^{E},\mathcal{O}^{E})$.

3.2　情感词在微博数据集中的词性分布

表 3 是情感词在$\mathcal{R}^{C}$和$\mathcal{R}^{E}$中的词性分布，其中n_s表示以相应列中的词性呈现在微博数据集中的情感词数量，n_t表示以该词性呈现的全部词数(以 K 为单位).“n_new”为 ICTCLAS2013 标注的未登录的名词，单独列出的原因在于它是未登录词，是不在已知词典中的“新”情感词的重要来源.由于情感词在数据集中会以不同的词性多次出现，因此，表 3 中各词性的情感词占情感词总数的比例$n_s/|\mathcal{S}^{C}|$(或者$n_s/|\mathcal{S}^{E}|$)之和大于 1.给定一个词，特别是英文中的形容词或副词，尽管其词性可以通过规则辅助判断，但本文直接采用词性标注的结果，原因在于，一是词的词性需要放在实际环境中才能确定，二是规则等辅助手段较难适应新词或不规范的词.

表 3　情感词在数据集中的词性分布

	中文情感词					
词性	noun	verb	adj	adv	n_new	others
n_s(K)	2.95	2.77	1.46	0.16	0.21	0.37
$n_s/\|\mathcal{S}^{C}\|$	0.39	0.37	0.19	0.02	0.03	0.05
n_t(K)	67.74	16.88	3.33	1.39	22.86	66.83
n_s/n_t	0.04	0.16	0.44	0.11	0.01	0.01
	英文情感词					
n_s(K)	4.00	3.56	2.95	1.13		1.00
$n_s/\|\mathcal{S}^{E}\|$	0.71	0.63	0.52	0.20		0.18
n_t(K)	205.14	66.93	64.13	14.08		18.43
n_s/n_t	0.02	0.05	0.05	0.08		0.05

表 3 显示，从中文微博中抽取新情感词时，有必要将标注为名词词性的词视为候选词.在中、英文微博数据中，大量的情感词以名词形式出现过，名词类情感词在中、英文微博中分别占到 42%(含 noun 和 n_new)和 71%.由于部分情感词在不同上下文本环境下会呈现不同的词性，进一步统计显示，在$\mathcal{R}^{E}$中，不考虑名词会遗漏 15%的情感词，但在$\mathcal{R}^{C}$中遗漏高达 40%.此外，在$\mathcal{R}^{E}$中，形容词可覆盖情感词的 52%，再考虑动词和副词后，可覆盖 84%的情感词；

但在$\mathcal{R}^C$中,形容词仅覆盖19%,考虑动词和副词后也仅能覆盖56%.因此,英文新情感词抽取时,只考虑形容词、动词和副词是比较恰当的,但中文新情感词抽取时有必要考虑名词.

表3还显示,考虑名词会给新情感词的抽取带来大量噪声,特别是对于中文微博新情感词的抽取.在英文微博数据集$\mathcal{R}^E$中,标注为形容词、动词和副词的全部词条中,分别有5%、5%和8%是情感词,高于名词的2%.而在中文微博数据集$\mathcal{R}^C$中,标注为名词和未登录名词("n_new")的词条分别有67.74K和22.86K个,但其中仅有2.95K(4%)和0.21K(1%)是情感词,远低于形容词的44%,动词的16%和副词的11%.本文中标为"n_new"的情感词较少的原因是我们使用的情感词典$\mathcal{S}^C$主要源自DUTSD,其中未登录词的数量非常有限,这也是我们在构建$\mathcal{S}^C$时补充一些情感词的主要原因.尽管在$\mathcal{S}^C$中纳入更多的未登录词可以提高这一比例,但名词给中文新情感词抽取带来的挑战依然非常严峻.

3.3 情感词之间的共现分析

在基于共现的情感词抽取方法中,一个基本的假设是情感词或者同极性的情感词之间有较强的共现.然而,在$\mathcal{R}^E$和$\mathcal{R}^C$上的统计显示,几乎所有的情感词都有与其它情感词共现过,情感词在同一微博中的共现现象比较普遍,但共现的情感词之间极性冲突也比较显著,如表4所示.

表4 情感词在微博中的共现统计

	中文情感词		英文情感词	
情感词数量	>=2	有极性冲突	>=2	有极性冲突
微博条数(K)	238.15	93.64	228.12	106.21
比例(%)	47.63	18.73	25.03	11.66

表4显示,对于中文微博,包含两个以上情感词的微博数占总微博数的47.63%,但其中有近40%(93.64/238.15)的微博中多个情感词的极性并不一致.而对于英文微博,有近一半(106.21/228.12)微博中的多个情感词极性不一致.此外,统计发现,$\mathcal{S}^C$的7565条情感词中有7534条存在共现,其中,7376条与同极性的词存在共现,7131条与不同极性的词存在共现;$\mathcal{S}^E$的5634条情感词中有5282条存在共现,其中4804(4776)条与极性相同(相异)的情感词共现过.

为了更深入地分析情感词与非情感词共现的统计特性,并考虑到大量相关工作中PMI被用于分析共现强度,本文统计了各种词性的候选词与已知情感词之间的PMI平均值,如表5所示.中文候选词为文档频率大于等于2,词长2~4个汉字;英文候选词为文档频率大于等于5、词长2~30个字符且包含字母.其中,英文候选词文档频率阈值5来自文献[22],而$\mathcal{D}^C$的规模只有$\mathcal{D}^E$的一半,因此,中文候选词文档频率阈值设置为2.表5中,"±"代表情感词,"+"代表正向情感词,"−"代表负向情感词,"0"代表非情感词,"+,+"表示正向情感词之间的PMI值,以此类推.由于中文非情感词词典$\mathcal{O}^C$收录的大都是登录词,因此大量被ICTCLAS2013标注为"n_new"词性的非情感词并未出现在$\mathcal{O}^C$中,为了去除"n_new"带来的偏差,表5中"noun"词性指除"n_new"外的其它名词.

表5显示,英文微博中,情感词与情感词之间的PMI("±,±"列)比非情感词与情感词之间的PMI("0,±"列)要高(表5中加粗的数值),这一规律适用于形容词、动词、副词类的情感词,特别是形容词类的情感词.这说明,如果以形容词、动词、副词三类词为候选词,选择与已知情感词共现较高的候选词作为新情感词是合适的.加之形容词、动词、副词已能覆盖84%的英文情感词(表3所示),因此经典方法只考虑形容词、动词、副词是合适的.但共现分析在形容词等词性上得到的规律并不适用于名词,相比名词类的情感词,名词类的非情感词与已知情感词的共现更强(表5中灰色底纹的数值).出现这种情况的主要原因是,在大量以名词词性出现的词条中,情感词所占的比例太少,而名词类的非情感词比例高达96%(中文)和98%(英文)(如表3所示),这给共现分析方法带来严重噪声.例如在COAE2014的中文微博数据集中,对"蒙牛"产品较多的负面评论使得"蒙牛"这一名词类非情感词与负极性的情感词共现更频繁.

表5 不同词性的情感词与已知情感词之间的PMI值

词性	中文情感词							英文情感词						
	+,+	0,+	−,+	−,−	0,−	±,±	0,±	+,+	0,+	−,+	−,−	0,−	±,±	0,±
noun	77.69	126.51	39.95	49.01	82.17	99.70	191.16	48.80	48.21	40.47	59.04	60.97	86.85	101.23
verb	192.04	177.99	106.92	130.59	125.78	256.06	290.87	33.66	16.04	33.35	49.01	28.39	**66.86**	**34.79**
adj	324.17	151.14	161.38	189.10	103.98	**424.27**	**241.38**	50.36	10.93	39.41	62.90	19.28	**85.55**	**22.24**
adv	231.94	294.76	164.09	176.83	259.25	366.92	534.96	16.59	13.25	22.46	32.59	34.16	**37.75**	**32.11**

表 5 还显示，在与已知情感词的共现方面，中文微博中的形容词与英文微博中的形容词有相似的规律. 因此，可以推测，如果只以形容词为候选词，基于 PMI 方法抽取中文新情感词也是比较准确的. 然而，仅考虑形容词会遗漏 81% 的情感词. 当考虑动词、副词和名词时，共现情况完全相反了，非情感词与已知情感词的 PMI 更高，高出相应类别的情感词与已知情感词的 PMI 值，各词类高出的比例分别为：动词 14%(290.87 vs 256.06)、副词 46%(534.96 vs 366.92)、名词 92%(191.16 vs 99.70). 由于与已知情感词共现更高的名词不再是情感词(当然，共现低的更不可能是情感词)，因此传统基于共现的新情感词抽取方法不再适合中文微博数据.

通过表 5，还可以分析 PMI 是否适合作为情感词极性判断的依据. 在英文微博中，除了正极性的副词外，极性相同的情感词较极性相异的情感词共现更强烈. 加上仅以副词词性出现的情感词数量较少，因此，对于英文微博数据，PMI 适合作为情感词极性判断的依据. 对于中文微博，全部词性都满足“极性相同的情感词较极性相异的情感词共现更强烈”，因此，可以推断 PMI 也适合用来判断中文微博中情感词的极性.

4　基于分类的微博情感词抽取算法 cNSEm

通过第 3 小节的分析发现，基于 PMI 共现的方法显然不适合从中文微博中发现新情感词，需要考察更多的线索. 直观上，可以用来判断一个候选词是否是情感词的线索包括：候选词与情感词的共现信息、候选词所在的上下文、上下文中的用词规律、与其它词的修饰关系，甚至是所在句子的情感极性等等. 如果将这些线索视为候选词的特征，并给定一部情感词典和一部非情感词典，则通过一个分类器来判断候选词是否有极性以及有何极性则是比较自然的想法. 本节以中文微博新情感词抽取为例，介绍基于分类的微博情感词抽取算法 cNSEm，而 cNSEm 中使用的特征将在下一节详细描述. 文献[33]介绍了该算法及其参加 COAE2014 任务 3 的评测情况，此处只对 cNSEm 的基本思想作简要描述.

cNSEm 算法包括 6 个步骤：

S1：构建数据集. 采集微博数据并进行预处理，经过分词(对于中文微博)、词性标注、依存句法分析等过程，得到微博数据集. 本文采用 3.1 节中的数据集 $\mathcal{D}^{C}$.

S2：构建情感词词典和非情感词词典. 本文用 3.1 节中的 $\mathcal{S}^{C}$，$\mathcal{O}^{C}$.

S3：构建用于分类的训练样本. 训练样本是微博数据集中的情感词(正、负极性的标签分别为“+1”和“−1”)和非情感词(标签为“0”)，情感词的特征在第 5 节中详细分析.

S4：选择候选词. 未作为训练样本，且词长、文档频率和词性满足设定的要求，则被视为候选词.

S5：训练并分类. 利用训练样本训练分类器，并对候选词分类. 本文用 libleaner 1.94[①] 作分类器，参数为“−s 4 −e 0.1”.

S6：候选词排序. 由于同一候选词在不同的上下文中会有不同的情感极性，加之分类器分类准确率的限制，在数据集中多次出现的候选词会被贴上不一致的标签. 本文考虑候选词在大部分环境下的情感极性，并假设出现分类错误的情况相对较少，因此采用基于投票的策略确定候选词的情感极性，采用候选词被分类为情感词和非情感词次数的比例对候选词排序，如式(2)所示.

$$polarity(t)=\mathrm{Sgn}(C_t^+-\alpha\cdot C_t^-)\frac{C_t^++C_t^-}{C_t^0+1} \quad (2)$$

其中 C_t^+、C_t^-、C_t^0 表示候选词 t 被分类为正、负极性情感词和非情感词的次数. 参数 α 用于平衡训练样本不均问题，本文设置 $\alpha=1$，并在第 6 节讨论 α 的影响. $polarity>0$ 则 t 为正极性，$polarity<0$ 则 t 为负极性，$polarity$ 绝对值越大则 t 作为情感词的可信度越高.

在分析第 3 节的表 5 时发现，从总体上看，PMI 适合中文微博数据中情感词的极性判断，而在 cNSEm 算法中，通过式(2)的投票机制，也可以对情感极性作出判断. 为了对比两种方案，本文设计了 cNSEm 算法的一个变种，称为 cNSEm-PMI. 与 cNSEm 的不同之处在于，cNSEm-PMI 的步骤 S3 中，只将候选词贴上有情感和无情感两类标签；步骤 S6 中 $polarity$ 的正负由候选词与正向和向负情感词的 PMI 之差决定，即

$$polarity(t)=\mathrm{Sgn}(pmi_t^+-\alpha\cdot pmi_t^-)\frac{C_t^{\pm}}{C_t^0+1} \quad (3)$$

其中 pmi_t^+(或 pmi_t^-)分别为 t 与已知正向(或负向)情感词的 PMI 之和，$C_t^{\pm}$ 为 t 被分类为情感词的

① http://www.csie.ntu.edu.tw/~cjlin/liblinear/

次数.

由于已有的文献或者针对英语语种,或者针对WordNet、商品评论等数据集,或者仅考虑形容词等个别词性、亦或是将情感词扩展视为情感分类的中间步骤.为了说明 cNSEm 的效果,本文基于经典的情感词抽取或识别算法,设计了作为参照的 GPC 算法,该算法基于共现(Co-occurrence),同时考虑了情感极性在图(Graph)中的传递(Propagation),其主要思想与经典的情感词抽取方法一致.

GPC 算法包括 5 个步骤:

S1:构建图 G. 以种子情感词和候选词为结点 V,以词间的共现、相似性等关系为边 E,构建带权无向图 $G=(V,E,P)$,$p^+(v_t)$ 和 $p^-(v_t)$ 是结点 v_t 为正向情感词和负向情感词的概率(用于排序的分值,不一定满足概率要求),$\langle p^+(v_t),p^-(v_t)\rangle\in P$;边 $e_{i,j}\in E$ 的权重 $w_{i,j}$ 表示结点 v_i 和 v_j 之间的关系强度.

S2:初始化图 G 中各结点的情感概率 P. 给定种子情感词词典 $S=S^+\cup S^-$,其中 S^+ 为正极性情感词,S^- 为负极性情感词,结点 v_t 属于正、负情感词的概率如式(4):

$$p^+(v_t)=\begin{cases}1,\text{如果 } v_t\in S^+\\0,\text{其它情况}\end{cases},\quad p^-(v_t)=\begin{cases}1,\text{如果 } v_t\in S^-\\0,\text{其它情况}\end{cases} \tag{4}$$

S3:情感极性传播. 根据结点 v_t 到达种子情感词的路径更新 v_t 为情感词的概率 $p^+(v_t)=F^+(v_t,G)$,$p^-(v_t)=F^-(v_t,G)$. 此处 F^+ 和 F^- 是情感极性传播方法的抽象,在不同的文献中情感极性传播方法不尽相同. 重复该传播过程直到收敛.

S4:抽取新情感词. 根据情感概率 P 抽取 δ 个或者满足阈值条件 ε 的结点为新情感词.

S5:将新情感词加入情感词典 S,重复步骤 S2 到 S4,直到没有新情感词产生,或者抽取的新情感词数量达到上限.

经典的基于共现和极性传播的方法可以看成是 GPC 方法的特例. 下面给出 3 个例子予以说明.

文献[36]从 Web 页面抽取新情感词,可视为在 GPC 中,结点 V 中候选词设定为满足频率高于指定阈值等统计特性的 n-gram,边的权重为词或 n-gram 的上下文之间的余弦相似度,情感极性传播的迭代次数为 1,候选词 v_t 的正(或负)向极性强度为 v_t 与 S^+(或 S^-)中各情感词的"相似性"之和,即

$$p^+(v_t)=\sum_{v_i\in V,i\neq t}\max_{\text{path}(v_t,v_i)}\left(p^+(v_i)\cdot\prod_{e_{j,k}\in\text{path}(v_t,v_i)}w_{j,k}\right),$$

$$p^-(v_t)=\sum_{v_i\in V,i\neq t}\max_{\text{path}(v_t,v_i)}\left(p^-(v_i)\cdot\prod_{e_{j,k}\in\text{path}(v_t,v_i)}w_{j,k}\right),$$

$$polarity(t)=p^+(v_t)+p^-(v_t) \tag{5}$$

其中 path(v_t,v_i) 为 v_t 到 v_i 的路径,$e_{j,k}$ 为构成路径的一条边,路径长度不超过设定的阈值.

文献[22]从 tweets 中抽取新情感词,以文档频率大于指定阈值的词为候选词,候选词 v_t 的情感极性概率是 v_t 与已知情感词在 tweets 集合中共现的概率.此时,可将 GPC 做如下设置:$w_{t,i}$ 为 v_t 与 v_i 共现的概率,候选词 v_t 的极性及概率计算如下:

$$p^+(v_t)=\sum_{v_i\in V,i\neq t}p^+(v_i)\cdot w_{t,i},$$

$$p^-(v_t)=\sum_{v_i\in V,i\neq t}p^-(v_i)\cdot w_{t,i},$$

$$polarity(t)=p^+(v_t)-p^-(v_t) \tag{6}$$

抽取 $|polarity(t)|>\varepsilon$ 的 δ 个词加入情感词集合. 如果有新情感词加入,重新初始化图 G 中的 P,并重复上述过程直到没有发现新情感词为止.

文献[38]基于候选词与有情感倾向的 tweets 之间的 PMI 来计算候选词的情感极性,但通常 tweet 的情感极性是未知的. 如果假设 tweet 的情感极性是由 tweet 中包含的已知情感词来决定的,则其基本思路与文献[22]一致,不同的只是文献[38]基于 PMI 计算 $p^+(v_t)$ 和 $p^-(v_t)$.

本文依据相关文献,选用 PMI 作为两词的关系强度,并设计有迭代和无迭代两个参照系统:GPC1 参照文献[36],种子情感词的极性只传递给距离它最近的候选词,无间接传递,无步骤 S5 的迭代,新情感词一次返回;GPC2 参照文献[22],每轮迭代扩充 50 个"新"情感词. 由于 GPC2 每轮迭代有新情感词加入,相当于考虑了情感极性的间接传递. 候选词的情感概率(仅用于排序,不满足概率要求)及极性判断参照文献[36],如式(7)所示. 按 $|polarity(v_t)|$ 从大到小选择候选词,$polarity$ 符号的正负即是候选词情感极性的正负,β 用于平衡种子情感词典中正负极性情感词不平衡、或者数据集中正负极性情感词不平衡的问题.

$$polarity(t)=\sum_{v_i\in V,i\neq t}p^+(v_i)\cdot PMI(v_i,v_t)-\beta\cdot\sum_{v_i\in V,i\neq t}p^-(v_i)\cdot PMI(v_i,v_t),$$

$$\beta=\sum_t\sum_{v_i\in V,i\neq t}p^+(v_i)\cdot PMI(v_i,v_t)\Big/\sum_t\sum_{v_i\in V,i\neq t}p^-(v_i)\cdot PMI(v_i,v_t) \tag{7}$$

5　分类特征

有大量关于评论或微博情感分析的文献采用了分类方法，但分类特征却不能直接用于本文的任务. 这不难理解，相邻的两个候选词的上下文几乎完全一样，但它们有无情感及情感极性却可能大不相同. 尽管如此，其中的部分特征可以借鉴. 例如，如果微博中出现了叹号“!”，则其中存在情感词的可能性会增加. 本节通过对相关文献中所用的特征或线索进行归纳总结并抽象，提出潜在有用的、与特定语言无关的候选特征集合，并在第 6 节对候选特征的作用进行实验验证和分析.

本文选择的候选特征包括如下 6 类：

F1：共现信息(Co-OCcurance，COC). 该特征假设正/负向情感词与其它正/负向情感词(或集合)间共现频率更高. 点互信息 PMI 是目前新情感词抽取时最常用的特征[11,39]，有些文献甚至将其视为唯一的特征[11,16]，它也经常被用作基于图的新情感词抽取方法中边及其权重设置的重要依据.

F2：上下文(ConTeXt，CTX). 该特征假设不同微博中上下文相似的词具有相似的情感极性. 文献[36]提取候选词周围长度为 6 的窗口内的词作为上下文，计算候选词与已知情感词的相似性，并构建图. 该特征还假设，已知极性(prior polarity)的情感词在与上下文组合在一起时，该词的情感极性或者组合得到的短语(phrase)的情感极性(contextual polarity)会发生变化[40-41]，而文献[42]认为将词组合在一起的二元组和三元组更有利于产品品论的情感分类.

F3：语言学模式(Linguistic PaTterns，LPT). 该特征假设作者在表达多个极性相关的情感词时会使用类似的语言学规则，例如通过“和(and)”联系的两个形容词极性相同，而用“但是(but)”“然而(however)”联系的两个形容词极性相反[9,30]. 文献[16]在抽取主观名词如“感觉(feeling)”、“拒绝(repudiation)”时，定义了一组包括候选词的词干、线索词例如代名词、情感词(如“will”)等特征. 文献[39]在对评论进行情感分类时，认为尽管形容词对情感的判断有重要的指示作用，但还要看它与其它词或者词性是按什么规则组合在一起的.

F4：修饰关系(MoDiFication，MDF). 该特征假设情感词在句法树中所扮演的角色，以及与其它词之间的修饰关系具有一定的规律. 文献[41]在短语级别上分析一个词是否具有情感倾向性时，所采用的特征除了包括与该词相邻的词是否是形容词、是否是副词、是否是程度词等信息外，还利用依存分析的结果，考察该词是否被主观词通过特定的依存关系(adj，mod 或 vmod)修饰. 文献[31]在计算形容词之间的相似性时，利用表达修饰关系的三元组($w1$；*relation*；$w2$)作为特征.

F5：文档(微博或句子)特征(DOCument features，DOC). 该特征假设具有情感倾向性的文档中存在情感词的可能性更大. 尽管该类特征不能细粒度地确定具体的情感词，但可以用于指示文档中存在情感词的可能性，以及该情感词的可能极性[41].

F6：词性(Part Of Speech，POS). 该特征的依据是不同词性的词被分类为情感词的概率是不同的[28]. 尽管这些概率在不同语种的数据集上存在差别，但它依然是判断情感词的重要特征. 另外，词性经常与其它特征结合使用，例如，将上下文中的标记换成相应的词性，从而形成一些语言学模式.

除上述 6 类特征外，基于词典的词义相似度(Semantic Similarity)假设语义相近的词具有相似的情感极性，而 WordNet、HowNet 等语义资源通常是计算语义相似度的重要依据[14]. 然而，由于新情感词大多并未被这类语义资源收录，所以本文不考虑该类特征. 此外，在新情感词抽取过程中，通常会将以上一种或多种特征混合使用，以提高分类效果[31,41].

本文依据上述 6 类特征，设计如表 6 所示的 5 大类特征. 词性特征并未在表 6 中显式地列出，它通过两个方面体现，一是利用词性进行候选词的过滤，二是结合上下文形成语言学模式特征. 对于语言学模式，尽管人工构建的更精准，但覆盖有限，因此本文采用通配符或者词性替换部分上下文等形式，以模拟实际的语言学模式. 表 6 中，候选词的上下文指该词前后窗口长度各为 l 的“标记”序列(词或标点符号). 如果没有特殊说明，cNSEm 使用各类特征中标有“*”号的特征，而其它特征将在第 6 节的实验中用作参照.

表 7 给出部分候选特征取值的一个例子，其中假设已知“脑残”和“游手好闲”的情感极性为负，“富二代”和“才怪”的情感极性未知，其它词为非情感词，上下文窗口长度 $l=4$.

表 6 cNSEm 中的候选特征集

类别	特征	说明
与情感词共现 COC	**coc_pmi**(*)	候选词与正向、负向情感词以及非情感词之间的 PMI,用最大值进行归一化,共现窗口为整条微博.
	coc_dice	将 coc_pmi 中的 PMI 换为 Dice 系数.
上下文 CTX	**ctx_nglr**(*)	候选词上下文所形成的 n_gram.
	ctx_u	候选词上下文中的标记,即 unigram.
	ctx_ulr	候选词上下文中的标记及该标记与候选词的相对位置(在候选词前或者后).
修饰关系 MDF	**mdf_all**(*)	与候选词有修饰关系的词及其修饰类型形成的二元组. 修饰关系来自 Standford parser 的依存分析.
	mdf_near	只考虑 mdf_all 中距离候选词最近的修饰或被修饰关系.
语言学模式 LPT	**lpt_ctx_nglr_?p**(*)	保留 ctx_nglr 中距离候选词最远的标记,其余全部用通配符"?"替换,并附上候选词的词性.
	lpt2_ctx_nglr_?p	与 lpt_ctx_nglr_?p 类似,但保留 ctx_nglr 中距离候选词最远的 2 个标记.
	lpt_ctx_nglr_±p	与 lpt_ctx_nglr_?p 类似,但不用通配符"?",而是将其中的已知情感词替换为其情感极性.
	lpt_ctx_nglr_?	与 lpt_ctx_nglr_?p 类似,但不附上候选词的词性.
	lpt_ctx_nglr_pp	与 lpt_ctx_nglr_?p 类似,但不用通配符"?",而是用相应词所标记的词性.
	lpt_mdf_all_±	将 mdf_all 中修饰候选词的情感词替换为其情感极性.
文档特征 DOC	**doc**(*)	候选词所在微博中正向、负向情感词、叹词(!,词性被标注为 wt,e,o,y 的词)及其它标记的个数,并用数据集中最长文档长度归一化.

表 7 候选特征举例

DOC2499139	*//@高雷雷:这样的爹脑残!这孩子将来不是游手好闲的富二代才怪!…*
预处理及分词结果	*这样*/rzv *的*/udel *爹*/n *脑残*/n *!*/wt *这*/rzv *孩子*/n *将来*/t *不*/d *是*/vshi *游手好闲*/vl *的*/udel *富二代*/n_new *才怪*/n *!*/wt
依存分析结果	句子 1:assmod(*爹*-3,*这样*-1),assm(*这样*-1,*的*-2),nsubj(*脑残*-4,*爹*-3),root(ROOT-0,*脑残*-4) 句子 2:det(*孩子*-2,*这*-1),nsubj(*是*-5,*孩子*-2),dep(*是*-5,*将来*-3),neg(*是*-5,*不*-4),root(ROOT-0,*是*-5),assmod(*才怪*-9,*游手好闲*-6),assm(*游手好闲*-6,*的*-7),nummod(*才怪*-9,*富二代*-8),attr(*是*-5,*才怪*-9)
候选词"富二代"的部分特征值:	
ctx_nglr	{*的*/&,*游手好闲*/*的*/&,*是*/*游手好闲*/*的*/&,*不*/*是*/*游手好闲*/*的*/&,&/*才怪*,&/*才怪*/*!*}
ctx_u	{*不*,*是*,*游手好闲*,*的*,*才怪*,*!*}
ctx_ulr	{*不*/&,*是*/&,*游手好闲*/&,*的*/&,&/*才怪*,&/*!*}
lpt_ctx_nglr_?p	{*的*/n_new,*游手好闲*/?/n_new,*是*/?/?/n_new,*不*/?/?/?/n_new,n_new/*才怪*,n_new/?/*!*}
lpt2_ctx_nglr_?p	{*的*/n_new,*游手好闲*/*的*/n_new,*是*/*游手好闲*/?/n_new,*不*/*是*/?/?/n_new,n_new/*才怪*,n_new/*才怪*/*!*}
lpt_ctx_nglr_±p	{*的*/n_new,*游手好闲*/*的*/n_new,*是*/-/*的*/n_new,*不*/*是*/-/*的*/n_new,n_new/*才怪*,n_new/*才怪*/*!*}
lpt_ctx_nglr_?	{*的*/&,*游手好闲*/?/&,*是*/?/?/&,*不*/?/?/?/&,&/*才怪*,&/?/*!*}
lpt_ctx_nglr_pp	{*的*/n_new,*游手好闲*/udel/n_new,*是*/vl/udel/n_new,*不*/vshi/vl/udel/n_new,n_new/*才怪*,n_new/n/*!*}
doc	{0,2,2,11}(归一化前)
候选词"才怪"的部分特征值:	
mdf_alr	{&/*ASSMOD*/*游手好闲*,&/*NUMMOD*/*富二代*,*是*/*ATTR*/&}
mdf_nlr	{&/*NUMMOD*/*富二代*,*是*/*ATTR*/&}
lpt_mdf_all_±	{&/*ASSMOD*/-,&/*NUMMOD*/*富二代*,*是*/*ATTR*/&}

6 实验分析

新情感词抽取时,中、英文微博数据集、情感词典和非情感词典分别采用 3.1 小节中的$\mathcal{R}^{C}$和$\mathcal{R}^{E}$. 将其中的情感词典$\mathcal{S}$随机均分为两部分$\mathcal{S}_{tr}$和$\mathcal{S}_{ts}$,其中$\mathcal{S}_{tr}$用于 cNSEm 的训练或用作 GPC 中的种子情感词集合,$\mathcal{S}_{ts}$用于测试. 另外,抽取新情感词的数量上限设定为 10 K 条. 需要说明的是,由于仅有文档频率高于指定阈值(英文 5,中文 2)、长度介于[2,4](中文)或[2,30](英文)之间的词才可能成为候选词,因此,去掉$\mathcal{S}_{ts}$中不满足该阈值条件的情感词. 考虑到英文测试情感词典中词汇数量较少,本实验用 Liu 的情感词典[43]进行补充.

6.1 评测方法

以$\mathcal{S}_{ts}$为理想结果,选择多组评测指标,包括:

(1) Bpref[44](简称 Bp). 用于评测的新情感词$\mathcal{S}_{ts}$规模有限,未被$\mathcal{S}_{ts}$和$\mathcal{O}$收录的词也可能是情感词,Bpref 将这部分词排除在外. 同时,Bpref 也考虑返回结果排序的问题.

(2) 平均精确率 AP. 这是信息检索结果评测的权威方法,适合本文对候选词抽取结果的评测.

(3) Rprec(简称 Rp). 当返回规模与$\mathcal{S}_{ts}$相同时的召回率(或准确率).

(4) Rprec2(简称 Rp2)：当返回规模为$\mathcal{S}_{ts}$两倍时的召回率.

如果在评测时要求候选词的极性也正确，则相应的指标为 Bp±、AP±和 Rp±、Rp2±.

6.2　实验结果分析

(1) cNSEm 与 GPC 在不同候选词性上的性能

按照第 3 小节的分析，对于中文微博，名词作为候选词是必要的，但对英文则不然. 实验考察了以下不同的候选词词性集合：

PosAll：所有词性，也即不考虑候选词词性.

PosAVDN：形容词、动词、副词和名词. 根据各个词性的词是情感词的概率大小，依次选取词性，直到 98%的情感词被覆盖. 最终依次选择的中文候选词性为{a,an,al,ad,d,dl,b,bl,vl,v,vn,vi,vd,z,n,nl}，英文候选词性为{JJ,JJR,JJS,VB,VBZ,VBG,VBN,VBD,VBP,RB,RBR,RBS}.

PosAVDN′：仅用于中文，在 PosAVDN 中添加 n_new 词性. 关于词性 n_new，我们将在实验中单独讨论.

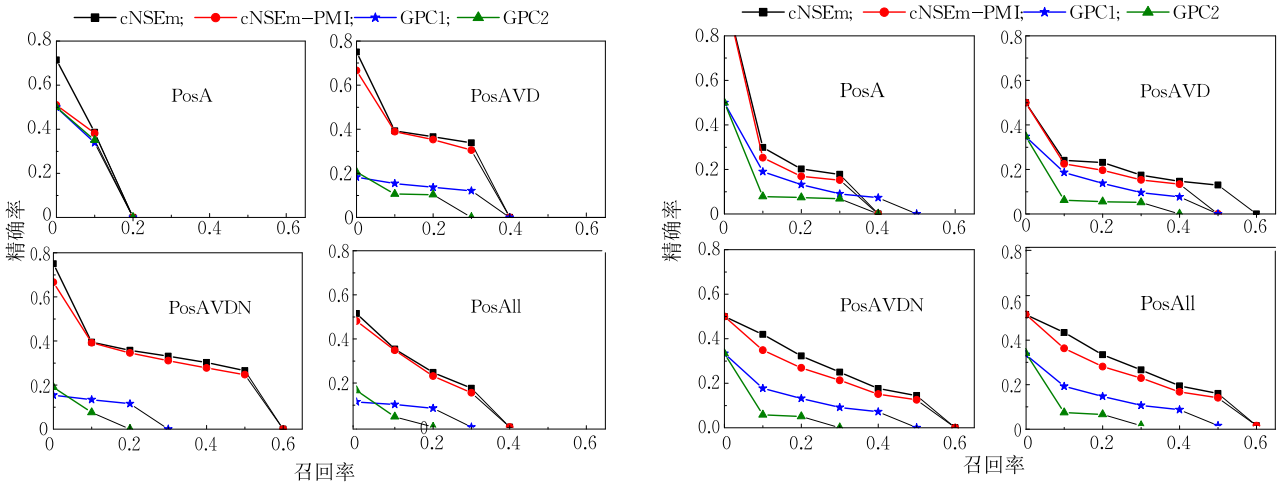
PosAVD：为考察名词作为候选词带来的影响，将 PosAVDN 候选词性集合中的名词去掉.

PosA：大量相关文献中只考虑形容词词性的情感词，作为对比，此处也仅以形容词作为候选词.

cNSEm 与 GPC 在不同候选词性上的新情感词抽取评测结果如表 8 所示，图 1 则展示了它们 11 点插值的 PR 曲线，以方便在更细粒度上观察 cNSEm 与 GPC 在不同候选词性上的表现. 该组评测中，cNSEm 利用了全部 5 类特征，特征组合为：coc_pmi+ctx_nglr+mdf_all+lpt_ctx_nglr_?p+doc，本文将该组特征组合作为后续实验中的默认特征组合.

表 8　情感词抽取结果评测

	(a) 中文数据集$\mathcal{R}^C$($\alpha=1,l=3$)								(b) 英文数据集$\mathcal{R}^E$($\alpha=1,l=5$)							
	Pos	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±
GPC1	A	0.1833	0.1361	0.0798	0.0512	0.1882	0.1417	0.1882	0.1417	0.2659	0.1878	0.1066	0.0596	0.1931	0.1538	0.3027	0.2303
	ADV	0.1647	0.1342	0.0686	0.0451	0.1686	0.1459	0.3146	0.2527	0.2520	0.1861	0.0987	0.0573	0.1864	0.1535	0.3030	0.2370
	AVDN	0.1210	0.1054	0.0442	0.0338	0.1406	0.1248	0.2511	0.2213	0.2391	0.1802	0.0905	0.0538	0.1814	0.1503	0.2986	0.2347
	All	0.1064	0.0913	0.0272	0.0209	0.1155	0.1012	0.2022	0.1784	0.2392	0.1804	0.0905	0.0538	0.1809	0.1500	0.2992	0.2352
GPC2	A	0.1849	0.1373	0.0812	0.0526	0.1901	0.1427	0.1901	0.1427	0.2006	0.1039	0.0831	0.0359	0.1338	0.0826	0.2669	0.1521
	ADV	0.1182	0.0918	0.0513	0.0314	0.1277	0.1026	0.2657	0.2030	0.1606	0.0789	0.0693	0.0239	0.1064	0.068	0.2416	0.1204
	AVDN	0.0744	0.0626	0.0246	0.0170	0.0923	0.0780	0.1784	0.1393	0.1446	0.0728	0.0602	0.0207	0.0989	0.0654	0.2271	0.1137
	All	0.0561	0.0468	0.0107	0.0073	0.0595	0.0515	0.1200	0.0933	0.1445	0.0731	0.0599	0.0206	0.0983	0.0643	0.2268	0.1140
cNSEm	A	0.1864	0.1418	0.0920	0.0595	0.1864	0.1443	0.1864	0.1443	0.5326	0.3500	0.1533	0.0807	0.2626	0.1940	0.4586	0.3292
	ADV	0.5109	0.3628	0.2386	0.1432	0.4272	0.323	0.5112	0.3854	0.7382	0.4546	0.2035	0.1035	0.2879	0.2099	0.4655	0.3324
	AVDN	0.7109	0.4874	0.3138	0.1846	0.4280	0.3238	0.6868	0.5141	0.8534	0.5211	0.3001	0.1559	0.3719	0.2687	0.5295	0.3766
	All	0.5152	0.3667	0.1707	0.1042	0.2908	0.2239	0.4356	0.3307	0.8555	0.5220	0.3004	0.1561	0.3734	0.2693	0.5292	0.3760
cNSEm-PMI	A	0.1872	0.1390	0.0923	0.0577	0.1872	0.1417	0.1872	0.1417	0.5472	0.327	0.1589	0.0689	0.2448	0.1576	0.4533	0.3030
	ADV	0.5159	0.3437	0.2405	0.1323	0.4261	0.3032	0.5165	0.3693	0.7516	0.4321	0.2060	0.0887	0.2931	0.1937	0.4539	0.3053
	AVDN	0.7204	0.4724	0.3165	0.1745	0.4296	0.3087	0.6865	0.4980	0.8662	0.4853	0.3018	0.1306	0.3623	0.2399	0.5254	0.3516
	All	0.4983	0.3326	0.1664	0.0944	0.2982	0.2199	0.4264	0.3056	0.8685	0.4863	0.3025	0.1308	0.3632	0.2408	0.5275	0.3536



(a) 中文数据集$\mathcal{R}^C$　　(b) 英文数据集$\mathcal{R}^E$

图 1　不同候选词性上新情感词抽取结果的 PR 曲线(要求极性判断正确)

表 8 和图 1 展示了丰富的信息,对于中、英文微博新情感词抽取的结论和具体分析如下:

结论 1. GPC 方法对中文微博中形容词类的新情感词抽取是有效的. 表 8(a)显示,对于中文微博 $\mathcal{R}^C$,除 Rprec2 和 Rprec2±外,GPC1 和 GPC2 在候选词性 PosA 上的各项指标值都高于在其它候选词性上的指标值. 这说明,文献[10,18,29]中采用的基于共现和极性传播的新情感词提取方法对中、英两种语言的微博数据都有效,但前提是仅以形容词为候选词. 需要说明的是,GPC 方法的 Rprec2 和 Rprec2± 指标在候选词性 PosA 上偏低的原因并不是因为抽取的准确率低,而是因为形容词的数量非常有限,导致返回的新情感词数量也非常有限,通过对比图 1(a)中 GPC 方法在候选词性 PosA 和 PosADV 上的 PR 曲线不难发现这一点,在 PosA 上,各方法在召回率为 0.2 时准确率已经为 0.

结论 2. 对于中文微博 $\mathcal{R}^C$,GPC 方法和 cNSEm 方法在形容词类的新情感词抽取上无显著差别,但更多候选词性加入后,cNSEm 方法远好于 GPC 方法. 表 8(a)和图 1(a)均显示,在 PosA 上,GPC 的两种变形和 cNSEm 的两种变形在性能上没有显著差别. 需要说明的是,cNSEm 在召回率为 0 时的准确率明显高于其它三种方法,其原因是 11 点准确率的计算采用插值,因此返回结果的第 1 个或前几个如果正确,就会使召回率为 0 的点上有较高的准确率. 表 8(a)和图 1(a)还显示,在动词、副词、名词等更多候选词性加入后,单靠共现的 GPC 方法已无法适应中文微博,并且候选词性越多,各项评测指标越低(结论 1 中分析了 Rprec2 和 Rprec2± 在 PosADV 上较在 PosA 上更高的原因). 这与第 3 小节的统计分析结果一致,即在中文新情感词抽取时,名词词性的加入会带来大量噪音,给新情感词的识别带来极大挑战,这也是 GPC 方法在加入名词后表现不佳的主要原因. 除 PosAll 外,cNSEm 方法在加入更多候选词性时,各项指标反而呈增长趋势,这表示 cNSEm 能够应对不同词性的情感词.

结论 3. 对于英文微博 $\mathcal{R}^E$,GPC1 在各类候选词性上的表现无差异,而 cNSEm 方法在各类候选词性上的表现都远好于 GPC1 方法. 3.1 小节的统计显示,英文微博中形容词性的情感词与已知情感词有较强的共现,动词、副词次之,名词最低,因此,增加动词、副词、名词等候选词性,几乎不会影响 GPC1 方法对候选词的排序,依然是形容词性的情感词排在最前,而名词性的情感词排在最后. 加之形容词、动词对情感词有很高的覆盖率,因此表 8(b)显示 GPC1 方法在各类候选词性上表现几乎完全一致,这与 3.1 小节的统计相符. 表 8(b)和图 1(b)还显示,cNSEm 方法在各个词性上的表现都较 GPC1 方法要好,说明 cNSEm 方法在不同语种上有很强的适应性. 另外,增加动词、副词、名词等候选词性后,cNSEm 的各项指标稳步上升,一方面因为相同词条以多种词性出现,更多候选词性使得 cNSEm 最后的投票统计更合理,另一方面极性较强的动词或名词等词性有机会被抽取,从而提高了抽取效果.

结论 4. 基于传递的方法仅适用于抽取准确率较高的情况. 表 8(a)和图 1(a)显示,在中文微博上,对于形容词性的情感词,尽管 GPC1 和 GPC2 差异不明显,但各项指标显示 GPC2 略优于 GPC1. 更多候选词性加入后,GPC2 的各项指标显著低于 GPC1. 表 8(b)和图 1(b)则显示,在英文微博上,GPC1 在各类候选词性上表现稳定,但 GPC2 随着更多词性的加入,各项评价指标显著下降. 这不难理解,当更多候选词性加入后,基于共现的方法抽取性能下降,导致新加入的情感词质量下降,进一步恶化后续的抽取质量.

结论 5. cNSEm 能适应不同词性的新情感词. 对于中文微博,PosA 对新情感词的覆盖率太低,因此返回结果达不到测试集的规模. 而在 PosADV 和 PosAVDN 上的返回结果都能达到测试集的规模,因此,各方法的 Rprec 和 Rprec± 指标在这两种候选词性上具有可比性. 表 8(a)中,cNSEm 方法的 Rprec 和 Rprec± 指标在 PosADV 和 PosAVDN 两种候选词性上保持稳定,这说明,随着名词性情感词的加入,cNSEm 方法抽取新情感词效果并没有降低,说明该方法能很好地适应多种词性的情感词. 意料之外的是,当 cNSEm 在 PosAVDN 词性集合上抽取与理想情感词典规模相同的新情感词时,抽取结果对理想情感词典的召回率达到 0.42(不考虑极性),当抽取规模再扩大一倍时,对理想情感词典召回率达到 0.6727,这已达到或接近 3.1 小节中人工情感词典 DUTSD 对 THUSD、HNSD 及 NTUSD 的召回率及 HNSD 与 THUSD 之间的召回率. 然而,相比人工方式,cNSEm 对情感词正负极性判断的准确性只有 77%,与人工之间的平均一致率 97% 还有较大的距离. 对于英文微博,随着动词、副词、名词的加入,cNSEm 的各项指标也稳步上升.

结论 6. cNSEm 在判断情感词极性上较 PMI 更有优势. 表 8(a)和表 8(b)均显示,如果不考虑情感词的极性是否划分正确(指标 Bpref、AP、Rprec、Rprec2),cNSEm-PMI(优化式(3)中参数 α 后)在候

选词性 PosAVDN 上抽取结果与 cNSEm 不相上下，在中文微博上大部分指标的分值还较 cNSEm 略高. 但如果考虑情感词的极性，情况刚好相反，cNSEm 的表现要明显超过 cNSEm-PMI. 这说明，情感词极性的判断不能仅依赖于它们与已知情感词的共现，cNSEm 中用到的其它特征也起到帮助作用.

“评测时不考虑极性”是指在评测新情感词时，不判断其极性是否被贴上正确的标签. 此时的情感词抽取可视为一个二分类问题，即将候选词分为“有极性”和“无极性”两类. cNSEm 是将情感词的抽取视为一个三分类问题，把候选词分为“正极性”、“负极性”和“无极性”三类，然后将“正极性”和“负极性”的情感词合并为“有极性”；而 cNSEm-PMI 则将情感词的抽取视为一个二分类问题. 因此，如果在评测时不考虑新情感词的极性是否正确，在分类方法、特征都相同时，经过优化，cNSEm-PMI 的分类性能高于 cNSEm 是容易理解的，划分为三类带来错误的可能性更大.

如果评测时考虑新情感词极性是否被贴上了正确的标签，仅用与正、负情感词共现的 PMI 方法 cNSEm-PMI 则不如 cNSEm，后者直接在分类时将候选词分为“正极性”、“负极性”和“无极性”三类. 尽管如此，将 PMI 用于极性判断也是比较有效的，cNSEm-PMI 在 Bp±、AP±、Rp±和 Rp2±评测指标上只比 cNSEm 分别低 3.08%、5.47%、4.66%和 3.13%.

结论 7. 根据词性进行候选词过滤是必要的. 对于中文微博，4 种方法在候选词性 PosAll 上的各项评测指标显示，如果不对情感词的词性加以约束，而将所有词性都作为候选，会严重影响新情感词的抽取准确率. 对于英文微博，要视所选择的方法确定是否有必要通过词性进行过滤：如果采用 GPC1 方法和 cNSEm，则没有必要，但如果采用传递方法，最好只抽取形容词性的情感词.

(2) 上下文窗口长度对 cNSEm 的影响

表 9 显示了不同上下文窗口长度时 cNSEm 的评测结果. 通常认为，上下文窗口长度越大，特征中的信息就越丰富，对新情感词抽取越有利. 但表 9 的结果显示，上下文窗口长度 l 从 1 增加到 3(中文)和 5(英文)时，cNSEm 的各项评测指标也逐渐增加，但 l 继续增大时，cNSEm 的性能反而下降. 这说明，词的情感及情感极性更多地反映在其前后 3 个词(中文)或 5 个词(英文)的用词规律上，而与距离更远的词已没有太大关系，纳入这些词反而会干扰新情感词的抽取. 另外，从评测分值上看，英文微博对上下文窗口长度不如中文微博敏感，窗口长度为 2 和 5 时各项评测指标值比较接近. 因此如果考虑效率，可以直接将 l 设置为 2.

表 9　窗口长度 l 对 cNSEm 的影响

(a) 中文数据集 $\mathcal{R}^{C}$(PosAVDN, $\alpha=1$)

l	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±
1	0.5918	0.3919	0.2798	0.1571	0.3955	0.2939	0.6011	0.4401
2	0.7164	0.4817	0.3143	0.1791	0.4222	0.3114	0.6767	0.5017
3	0.7109	0.4874	0.3138	0.1846	0.4280	0.3238	0.6868	0.5141
4	0.6628	0.4512	0.2997	0.1688	0.4348	0.3251	0.6630	0.4925
5	0.6235	0.4128	0.2813	0.1508	0.4269	0.3085	0.6236	0.4512
6	0.6000	0.3999	0.2728	0.1455	0.4259	0.3066	0.6001	0.4364

(b) 英文数据集 $\mathcal{R}^{E}$(PosAVDN, $\alpha=1$)

l	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±
1	0.7739	0.4764	0.2656	0.1417	0.3359	0.2457	0.4940	0.3548
2	0.8524	0.5251	0.2834	0.1500	0.3548	0.2594	0.5132	0.3716
3	0.8815	0.5309	0.2911	0.1502	0.3568	0.2556	0.5263	0.3725
4	0.8688	0.5281	0.2973	0.1527	0.3681	0.2626	0.5237	0.3737
5	0.8534	0.5211	0.3001	0.1559	0.3719	0.2687	0.5295	0.3766
6	0.8454	0.5120	0.2950	0.1491	0.3670	0.2643	0.5190	0.3664
7	0.8329	0.5026	0.2948	0.1466	0.3696	0.2629	0.5188	0.3641

(3) 参数 α 对正、负向情感词抽取的影响

表 10 中，Bp＋、AP＋(或 Bp－、AP－)为仅考虑正(或负)极性情感词时的 Bpref 值和 AP 值. 式(2)中参数 α 用于调节正向与负向情感词样本不均衡导致的分类偏差，提高 α 有助于提高负向情感词的召回率. 表 10(a)显示，对于中文微博，尽管提高 α 值对负向情感词的抽取结果有改善，但这种改善在 α 大于 2 以后不再明显，其中 AP－在 α 大于 2 后反而下降. 而对于英文微博，$\alpha=1$ 是最佳的选择. 本组实验中，中文微博上 α 设置 1.5 或 1 时在 Bpref±和 AP±各有优劣，为统一中、英文，本文实验中默认 $\alpha=1$.

表 10　参数 α 对 cNSEm 的影响

(a) 中文数据集 $\mathcal{R}^{C}$(PosAVDN, $l=3$)

α	Bpref＋	Bpref－	Bpref±	AP＋	AP－	AP±
0.5	0.5383	0.3546	0.4561	0.2019	0.1209	0.1626
1.0	0.5187	0.4516	0.4874	0.2101	0.1537	0.1846
1.5	0.5026	0.4645	0.4852	0.2120	0.1542	0.1854
2.0	0.4724	0.4770	0.4765	0.2070	0.1564	0.1825
2.5	0.4620	0.4796	0.4729	0.2055	0.1538	0.1786
3.0	0.4311	0.4726	0.4550	0.1917	0.1480	0.1669

(b) 英文数据集 $\mathcal{R}^{E}$(PosAVDN, $l=5$)

α	Bpref＋	Bpref－	Bpref±	AP＋	AP－	AP±
0.5	0.4535	0.5087	0.4934	0.1156	0.1731	0.1486
1.0	0.3625	0.6057	0.5211	0.0826	0.1984	0.1559
1.5	0.3048	0.6146	0.5103	0.0583	0.1996	0.1481
2.0	0.2521	0.6228	0.5017	0.0482	0.1991	0.1444
2.5	0.2312	0.6252	0.4978	0.0444	0.1989	0.1427
3.0	0.1997	0.6272	0.4914	0.0356	0.1994	0.1397

(4)各类特征的作用

为了考察各类特征在 cNSEm 中的作用,本文在 cNSEm 默认特征组合的基础上,每次去掉其中一种类型的特征,并重新优化上下文窗口长度.相对于默认特征组合,去掉某一特征后,cNSEm 各项评测指标值增减的百分比如表 11 所示,其中第 2 行中 cNSEm 采用默认特征组合,上下文窗口长度 $l=3$(中文)或 5(英文),候选词性集合为 PosAVDN."－COC"表示在默认特征组合中去掉"共现特征 COC",窗口长度优化后仍为 3,"－CTX($l=4$)"表示去掉上下文特征,窗口长度优化后为 4.对表 11 的分析结论如下.

表 11 各类特征对 cNSEm 的影响

(a)中文数据集 $\mathcal{R}^{C}$(PosAVDN,$\alpha=1$,$l=3$)

	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±
cNSEm	0.7109	0.4874	0.3138	0.1846	0.4280	0.3238	0.6868	0.5141
－COC	－8.37	－14.20	－8.57	－21.13	0.37	－5.87	－4.97	－9.96
－CTX	0.17	－3.90	1.21	－4.71	－1.50	－4.82	－1.70	－4.01
－CTX($l=4$)	3.64	0.16	4.43	－0.76	－2.73	－4.82	－0.66	－2.57
－MDF	－2.11	－3.28	－1.37	－2.98	－1.05	－2.29	－1.35	－2.41
－LPT	－5.36	－3.49	－11.57	－10.46	－5.70	－4.57	－4.89	－3.54
－DOC	1.36	－4.64	1.82	－6.99	－0.63	－5.87	－0.47	－4.77

(b)英文数据集 $\mathcal{R}^{E}$(PosAVDN,$\alpha=1$,$l=5$)

	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±
cNSEm	0.8534	0.5211	0.3001	0.1559	0.3719	0.2687	0.5295	0.3766
－COC	－2.48	－6.33	－4.40	－9.49	－0.94	－2.83	－2.42	－5.79
－CTX	2.03	3.38	－7.70	－5.71	－7.34	－5.95	－2.47	－0.48
－MDF	－0.02	0.79	－2.50	－0.58	－1.96	－1.41	－1.32	0.13
－LPT	2.05	2.46	－14.73	－15.33	－8.52	－8.34	－5.93	－4.41
－DOC	4.07	2.69	4.03	2.63	－0.46	－1.75	1.38	0.37
－DOC($l=3$)	4.39	5.66	5.33	6.29	0.00	－0.11	3.74	4.01

结论 8. 共现特征对情感词抽取及情感词的极性判断是非常重要的.表 11 显示,去掉 COC 特征后,除 Rprec 基本不变外,各项评测指标下降显著,特别是 AP±指标,下降了 21.13%(中文)和 9.49%(英文).这说明,尽管仅用 PMI 对于微博中新情感词抽取是不够的,但基于 PMI 等共现的经典方法是合理的,共现信息在情感词的识别和情感极性的判断上有重要的指示作用.对比考虑极性和不考虑极性时的各项指标,发现去掉 COC 特征后,考虑极性的各项指标下降更严重.这说明,COC 是判断情感词极性的重要特征之一,原因是,尽管一条微博中可能会使用多个有倾向性冲突的情感词,但多数情况下一条微博表达的情感是单一的,因此所用的多个情感词的倾向性也是一致的.这与文献[11,16,39]中"正/负向情感词与其它正/负向情感词(或集合)间共现频率更高"的假设是一致的,也符合 3.3 小节中的观察结果,即"极性相同的情感词比极性相异的情感词共现更强烈".再者,去掉 COC 后,Rprec 指标基本不变,但 Rprec2 指标明显下降,这说明仅使用 COC 以外的其它特征,情感倾向性比较明显的词(排名靠前)已经能够较好地被 cNSEm 识别出来.但对于情感倾向不明显的词(排名靠后),COC 能起到辅助作用.

结论 9. 中文微博中语言学模式特征可以弥补上下文特征的缺失.表 11(a)中,对于中文微博,在去掉上下文词汇特征 CTX 后,如果上下文窗口仍然为 3(注意到语言学模式特征 LPT 也与上下文窗口有关),则 cNSEm 的 Bpref±、AP±和 Rprec±指标明显下降.但优化(增大)窗口后,去掉 CTX 对 Bpref±、AP±两项指标并无明显影响,反而使得 Bbref 和 AP 指标略有上升.造成这种结果的主要原因是,在 cNSEm 默认的 LPT 特征 lpt_ctx_nglr_?p 中,保留了 CTX 特征 ctx_nglr 中距离候选词最远的词,从而保留了部分上下文特征.因此,当 CTX 特征缺失后,增大窗口的长度,使得 LPT 成了 CTX 的一个补充,减少了 CTX 缺失的影响.然而,在英文微博上我们却不能下同样的结论.表 11(b)中,优化上下文窗口后(仍为 5),去掉 CTX 特征仍然导致除 Bpref 和 Bpref±以外的各项评测指标明显下降,说明对于英文微博,上下文特征和语言学模式特征起到了不同的作用.另外,对于中、英文微博,去掉语言学模式特征 LPT 会严重影响情感词抽取的质量,这说明中、英文情感词的使用都有一些隐式的规律,因此 LPT 特征对中、英两种语言的微博数据中情感词抽取都至关重要.相比中文,LPT 特征对从英文微博中抽取情感词的影响更大,这也反映出英文情感词的用词模式更突出或者更容易被刻画,对新情感词的抽取更有帮助,因此文献[9,30]等直接采用一些显示模式来抽取英文情感词.

结论 10. 修饰特征对新情感词的抽取影响较小,但依然有帮助.表 11 显示,去掉修饰特征 MDF 后,中、英文微博上的各项评测指标都会下降,但降幅不大.降幅不大有两个可能的原因,一是句法分析不够准确,导致长距离依赖分析不准确;二是长距离依赖相对较少,而短距离依赖又可以由上下文特征或语言模式特征来弥补.而去掉 MDF 特征后性能下降的原因是,尽管上下文中包含了大部分与候选词形成修饰关系的词,但却没有修饰类型,因此可以认为修饰类型对情感词及其极性的

判断有帮助.

结论 11. 文档特征对中、英文微博情感词抽取的影响差别很大. 对于中文微博，文档特征有利于情感词的极性判断. 表 11(a)中，去掉 DOC 特征对新情感词的抽取影响不大，但会显著降低情感极性判断的正确率，使得各项考虑极性的评测指标显著下降. 这说明，尽管文档特征对新情感词的抽取没有帮助，但文档包含的正负情感词的数量对文档中新情感词的极性判断是有帮助的，这类似于 COC 特征的作用. 然而，表 11(b)却显示，从英文微博中抽取新情感词时，增加 DOC 特征反而严重损害了 cNSEm 的性能. 更仔细的观察发现，这是由于中、英文不同语种的微博中情感词和非情感词的分布差异导致的. 相比中文微博 140 个汉字，英文 tweet 140 个字符包含的词条要少很多，因此，出现情感词共现的可能性也小很多. 表 4 显示，$\mathcal{R}^{C}$中含两条以上情感词的微博占 48%，而$\mathcal{R}^{E}$中只有 25%. 也就是说，英文微博中已知情感词和新情感词大都是独立出现的. 对于已知情感词，它们的 DOC 特征中已知的“正(或负)情感词个数”通常是 1，而由于新情感词通常也是独立出现的，所以它们的 DOC 特征中已知的“正(或负)情感词个数”通常是 0，这对 cNSEm 中的分类是非常不利的.

(5)各类特征的用法

表 11 显示了第 5 节定义的各类特征对 cNSEm 的影响，而表 12 则展示了各类特征的不同用法对 cNSEm 的影响. 表 12 的第 2 行 cNSEm 采用默认特征，其它各行则表示替换默认特征并优化上下文窗口长度后各评测指标增减的百分比. 例如，第 3 行中 dice 表示将默认特征组合中共现特征 COC 由点互信息 PMI 替换为 Dice 系数. 分析表 12 可得出以下结论.

表 12　各类候选特征的不同用法对 cNSEm 的影响

(a)中文数据集$\mathcal{R}^{C}$(PosAVDN，$\alpha=1$，$l=3$)

类	候选特征	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±
	cNSEm	0.7109	0.4874	0.3138	0.1846	0.428	0.3238	0.6868	0.5141
COC	dice	3.25	2.93	2.36	2.60	−1.54	−0.65	1.03	1.03
CTX	ctx_u	−0.52	−3.02	0.29	−4.01	−2.59	−5.37	−1.89	−3.23
	ctx_u($l=4$)	3.11	0.57	3.03	−2.00	−2.17	−4.42	−0.47	−1.38
	ctx_ulr	0.28	−2.75	0.38	−4.77	−1.87	−3.92	−1.89	−3.33
	ctx_ulr($l=4$)	3.87	−0.45	4.11	−3.20	−1.99	−5.81	−0.66	−2.92
	−CTX($l=4$)	3.64	0.16	4.43	−0.76	−2.73	−4.82	−0.66	−2.57
MDF	mdf_nlr	−0.48	−0.90	−0.41	−0.65	−1.05	−1.20	−0.77	−1.28
LPT	lpt2_ctx_nglr_?p	−11.87	−12.02	−10.01	−15.28	−0.33	−1.45	−8.71	−9.04
	lpt_ctx_nglr_±p	−5.28	−5.01	−4.56	−6.12	0.91	0.65	−2.78	−2.76
	lpt_ctx_nglr_?	−2.98	−2.81	−9.05	−9.70	−4.70	−3.83	−3.86	−3.29
	lpt_ctx_nglr_pp	−2.45	−2.71	−2.10	−3.20	0.05	−0.80	−1.54	−1.48
	lpt_mdf_all_±	0.11	−1.46	0.67	−1.14	−0.14	−0.99	−0.09	−1.32

(b)英文数据集$\mathcal{R}^{E}$(PosAVDN，$\alpha=1$，$l=5$)

类	候选特征	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±
	cNSEm	0.8534	0.5211	0.3001	0.1559	0.3719	0.2687	0.5295	0.3766
COC	dice	2.04	3.63	−3.67	−1.99	−1.56	0.41	0.38	2.68
CTX	ctx_u	0.47	−1.44	−11.53	−13.86	−9.30	−10.27	−5.93	−7.20
	ctx_u($l=2$)	−4.53	−2.38	−10.43	−5.84	−9.68	−6.81	−5.44	−3.56
	ctx_ulr	0.50	1.38	−10.50	−8.72	−10.54	−8.22	−5.59	−3.40
	−CTX	**2.03**	**3.38**	**−7.70**	**−5.71**	**−7.34**	**−5.95**	**−2.47**	**−0.48**
MDF	mdf_nlr	0.27	0.58	−0.97	0.96	−1.80	−0.33	−0.66	0.77
LPT	lpt2_ctx_nglr_?p	−1.39	−1.40	−4.53	−4.75	−1.56	−0.67	−3.46	−3.88
	lpt_ctx_nglr_±p	−0.47	−1.65	−5.36	−8.40	−3.98	−4.32	−2.36	−2.42
	lpt_ctx_nglr_±p($l=2$)	0.68	2.71	−5.80	−1.99	−4.92	−2.49	−1.59	0.45
	lpt_ctx_nglr_?	−0.41	−0.13	−13.73	−16.10	−5.46	−7.26	−4.06	−4.25
	lpt_ctx_nglr_?($l=4$)	2.30	1.29	−13.53	−15.78	−6.48	−7.48	−4.06	−3.80
	lpt_ctx_nglr_pp	0.05	0.48	−2.80	−3.34	−2.42	−2.05	−0.49	0.45
	lpt_ctx_nglr_pp($l=3$)	3.01	3.32	−2.80	−1.99	−3.60	−2.05	−0.38	1.30
	lpt_mdf_all_±	−0.57	−0.02	−1.57	−0.90	−0.24	0.52	−1.53	−0.40

结论 12. 对于中文微博，共现特征选择 Dice 系数较 PMI 更好. 已有文献中通常用 PMI 作为主要特征抽取新情感词，但对于中文微博数据，尽管二者的表现差距不大，但总体上 Dice 系数表现更佳. 将

pmi 替换为 dice 后，除了 Rprec 和 Rprec±两项指标外，其它指标都略有提升.我们做了类似于表 5 的统计分析，发现在中文微博上，Dice 系数较 PMI 对情感词有更好的区分能力.

结论 13. 上下文特征中的 unigram 特征不利于新情感词的抽取.为了便于对比，表 12 中将去掉 CTX 特征后(－CTX 行)各项指标值再次罗列出来.对于英文微博，采用 CTX 特征的 ungram 形式(ctx_u 和 ctx_ulr)还不如去掉 CTX 特征(－CTX 行)，这在全部评测指标上都有不同程度的体现.而对于中文微博，这一评测结果尽管不如英文微博上明显，但使用 unigram 要么在部分指标上与去掉 CTX 相当，要么更差.上下文的 unigram 形式不利于新情感词抽取的原因可以通过一个简单的示例来说明：设微博$\langle t_1,t_2,s,t_4,t_5\rangle$中 s 为情感词，其它为非情感词，相邻的情感词 s 和非情感词 t_4 的 ctx_u 特征分别为$\{t_1,t_2,t_4,t_5\}$和$\{t_1,t_2,s,t_5\}$，二者有 3/4 是相同的，区别非常小，不利于分类.而 s 和 t_4 的 ctx_nglr 特征分别为$\{t_2/\&,t_1/t_2/\&,\&/t_4,\&/t_1/t_5\}$和$\{s/\&,t_2/s/\&,t_1/t_2/s/\&,\&/t_5\}$，没有重复特征，因此有较好的区分能力.这也同时也说明，在文本或微博情感分类上有效的上下文特征不能直接用于新情感词的抽取.

结论 14. 远距离的修饰特征对中文微博新情感词的抽取有帮助.与考虑全部修饰关系的 mdf_all 相比，只考虑距离候选词最近的修饰关系(mdf_nlr)时，cNSEm 的各项指标在中文微博上均有下降.但与结论 10 中不考虑修饰特征相比，这种下降幅度非常小.这说明，对于中文微博，远距离的修饰关系有用，但作用非常有限，这与结论 10 中的分析是一致的.对于英文微博，mdf_all 与 mdf_nlr 各有千秋，没有显著性的差异.一个可能的原因是，在英文微博上，上下文窗口设置为 5，更长距离的依赖关系已经比较少或者不准确了.

结论 15. lpt_ctx_nglr_?p 特征能较好地刻画情感词的用词模式.表 12 中，语言学模式的 5 种替换特征中，除给 cNSEm 的极少数指标带来 3%左右的提高外，大部分指标在替换后都明显下降，这在中、英文微博上是相似的.特征 lpt_ctx_nglr_±p 不用通配符而采用情感词的极性，用以模拟文献[9,36,42]中的模式，例如通过“和(and)”联系的两个形容词极性相同等.实验结果显示，用特征 lpt_ctx_nglr_±p 替换掉默认特征 lpt_ctx_nglr_?p 后，除中文微博上 Rprec 和 Rprec±基本不变外，各项评测指标均下降，其中 AP±指标下降高达 6.12%(中文)和 8.40%(英文).这说明中、英文微博中两个情感词之间的语言学模式不明显或者不易描述.此外，当观察各替换特征在 Rprec 和 Rprec±(英文的 Rprec2 和 Rprec2±)指标上的表现时，发现 lpt_ctx_nglr_±p 和 lpt_ctx_nglr_pp 特征的表现与 lpt_ctx_nglr_?p 特征相当.由于 Rprec 和 Rprec±(或 Rprec2、Rprec2±)指标是返回规模与测试词典相同(或 2 倍)时的召回率，可以看作是对排名靠前或情感倾向比较明显的情感词的评测.结合表 11 中去掉 LPT 特征导致 cNSEm 的 Rprec、Rprec±(或 Rprec2、Rprec2±)指标下降 5%左右，我们可以大胆地做出结论：包含了上下文词汇和候选词词性的语言学模式对微博中情感倾向比较明显的情感词的用词规律有较好的刻画能力.再者，lpt2_ctx_nglr_?p 试图在 lpt_ctx_nglr_?p 的基础上通过增加一个词(或者减少一个通配符)来增强准确率，但实验显示这种调整是失败的，优化窗口长度后，各项评测指标依然有大幅下降，其中 AP±指标下降高达 15%(中文)和 5%(英文)左右.

结论 16. 候选词的词性有助于刻画情感词的语言学模式.除了将词性用于候选词的过滤外，本文还将词性引入语言学模式特征 LPT 中.表 12 中，特征 lpt_ctx_nglr_?中不包含候选词的词性，用该特征替换 lpt_ctx_nglr_?p 后，全部指标都大幅下降，其中 AP 和 AP±下降约 10%(中文)和 15%(英文)，说明词性不仅在过滤候选词时非常重要，而且在描述情感词的用词模式中也扮演着重要的角色，这与我们的常识是一致的.此外，当我们试图将 LPT 中除候情感词以外的通配符“?”全部替换为相应词的词性时(特征 lpt_ctx_nglr_pp)，结果也变糟了，各项评测指标均有不同程度的下降.由此可见，情感词自身的词性有助于描述情感词的用词规律，但这个用词规律与上下文中其它词的词性关系不大.

(6)未登录名词的影响

在中文微博数据上，在通用词典中不存在但被 ICTCLAS2013 分词及词性标注系统识别出来并且标注为“n_new”的词，即未登录的名词.尽管本文在构建情感词典时考虑到该问题，增加了部分未登录的情感词，但与被标注为 n_new 的总词数 22.86K 相比，情感词典中包含的 n_new 类型的情感词仍然非常有限，仅有 0.21K.从表 13 的评测结果来看，在候选词性集合中加入了 n_new 后(PosAVDN′)，各项评测指标都显著下降，说明 n_new 这一类词依然是

cNSEm 的一大挑战. 解决该问题至少存在两种途径，一是扩大用于训练的情感词典中未登录词的比例，这可以在 cNSEm 方法的帮助下，通过人工筛选来完成；二是专门针对 n_new 这一类词选择特征组合、训练分类器，以及增加更为老练的过滤规则，我们将在未来工作中继续探讨.

表 13　候选词性 n_new 对 cNSEm 的影响($\alpha=1, l=3$)

POS	Bp	Bp±	AP	AP±	Rp	Rp±	Rp2	Rp2±
PosAVDN	0.7109	0.4874	0.3138	0.1846	0.4280	0.3238	0.6868	0.5141
PosAVDN′	0.5856	0.4358	0.2146	0.1289	0.3360	0.2564	0.5102	0.3849

(7) 参加 COAE2014 评测结果

cNSEm 参加了 COAE2014 的中文微博新情感词抽取任务. 该任务要求从 1000 万条中文微博中抽取不超过 1 万条新情感词. 由于参加比赛时间紧迫，待处理的数据量大，因此没有对 cNSEm 方法进行参数优化，并且在用词性进行过滤后，只选择了语言学模式 LPT 一个特征. 尽管只用了一个特征，cNSEm 仍在 26 支参赛系统中取得了排名第二的成绩. 表 14 列出了排名前 5 的参赛系统，其中 UdeM-t3-2 是基于 cNSEm 的.

表 14　cNSEm 在 COAE2014 上的评测结果

参赛系统	P	R	F1	P±	R±	F1±
SXU	0.20420	0.20961	0.207	0.16400	0.16834	0.166141662
UdeM-t3-2	0.16590	0.17029	0.168	0.14850	0.15243	0.150439338
KMUST_LIIP	0.20492	0.17460	0.189	0.16167	0.13775	0.148754542
ICT_WDSE	0.17700	0.17142	0.174	0.14630	0.15017	0.148209741
iip−2	0.26177	0.12441	0.169	0.21793	0.10357	0.140410638

(8) 新情感词对微博情感分类的帮助(间接评测)

为进一步评估 cNSEm 方法的性能，本文用 cNSEm 抽取的新情感词扩展情感词典，利用扩展前后的情感词典对微博进行情感分类，考察所抽取的新情感词对微博情感分类是否有帮助. 考虑到情感词典之间的差异性(如 3.1 节表 2 所示)，为充分考察 cNSEm 对不同情感词典的适应性，对中文微博，本组实验除采用基于 DUTSD 的 $\mathcal{S}^{C}$ 外，还选择了与 DUTSD 差异最大的 NTUSD 作为 cNSEm 的训练词典；对于英文，除采用 $\mathcal{S}^{E}$ 外，另外选择 SentiWordNet(SWN)中情感倾向较强的情感词(简称 SWNHQ)用作训练. 作为参照，实验中还评测了中文情感词典 MixedSD(多个人工情感词典的并集)和英文情感词典 MPQA($\mathcal{S}^{E}$ 只是 MPQA 中极性强度较高的情感词)，它们可以认为是情感词的理想或人工扩展结果. 各情感词典的规模如表 15 所示.

表 15　情感分类中使用的情感词典

	情感词典	词条数	正向词条数	负向词条数
中文	MixedSD	35170	16272	18898
	$\mathcal{S}^{C}$	7565	3964	3601
	NTUSD	8347	1810	6537
英文	MPQA	12296	4183	8113
	$\mathcal{S}^{E}$	5634	2416	3218
	SWNHQ	11806	3786	8020
	SWN	65994	30017	35977

用于扩展新情感词的中文微博数据集是从 COAE2014 微博集合中随机抽取的 100 万条微博，英文微博数据集为 $\mathcal{R}^{E}$ 中的 $\mathcal{D}^{E}$. 根据前面的实验结果，本组实验中，对中文微博，cNSEm 的特征为默认特征组合，上下文窗口长度 l 为 3，候选词性集合为 PosAVDN′，即包含 n_new 词性(真实环境下需要抽取未登录的新情感词)；对于英文微博，cNSEm 的特征不包含 DOC 特征，其它类型的特征仍为默认特征，候选词性集合为 PosAVDN，上下文窗口长度 l 为 5.

为考察 cNSEm 所抽取的情感词的通用性，中文微博情感分类任务选用两个数据集，分别来自 COAE2014 的情感分类子任务(称 COAE2014 测试集)和来自 CC&NLP2013 的情绪分类子任务(称 CC&NLP2013 测试集). COAE2014 情感分类子任务提供了 5000 条微博，其中 2656 条被标注正向，2344 条被标注为负向. CC&NLP2013 情绪分类子任务中提供了 1 万条微博，其中 2674 条被标注为“高兴”、“喜好”，2453 被标注为“厌恶”、“恐惧”、“悲伤”、“愤怒”、“惊讶”等情绪，4873 条被标注为“无情绪”. 本实验将“高兴”和“喜好”情绪视为正向，其它情绪视为负向. 英文测试数据集为文献[22]作者提供的英文 tweets 测试集(原文 2K)，下载得到 1766 条，其中“正”、“负”和“无极性”的 tweets 数量分别为 531 条、298 条和 764 条，另有 173 条正负极性兼有.

情感词扩展的数量为测试数据中原始(扩展前)情感词数量的 λ 倍. 本实验中，λ 分别取值 0.1 和 1，前者考察少量扩展时的影响，后者考察情感词典规模扩大到原始情感词典 2 倍时的影响.

由于本任务主要考察新情感词的抽取质量，因此微博情感分类方法采用基于情感词典的 Naive 情感分类方法[22]. 该方法通过微博中包含的情感词的数量来判断其情感倾向性，即，如果有 k 条及以上情感词出现在微博中，则该微博有情感倾向，其情感极性为微博中正、负情感词的数量中的多者；如果微博中正、负情感词的数量相同，则只认为该微博有情

感,不对其极性作判断.尽管有文献在实验中尝试不同的 k 值,但本文在多个情感词典和多个测试集上的结果都显示,对于中文微博,$k=1$ 是最佳的.对于英文微博,$k=2$ 在降低召回率的代价下,提高了精确率,但扩展情感词前后所表现出来的特点和规律与 $k=1$ 是一致的.因此,本文呈现的实验结果都是 $k=1$ 时的评测结果.

表 16 和表 17 是微博情感分类的评测结果,其中 P、R、F1 表示微博主观性(只考虑微博是否有情感,而不考虑其情感极性是否判断正确)分类的准确率、召回率和二者的调和平均值;而 P±、R±、F1± 表示微博极性分类的准确率、召回率及二者的调和平均值,极性分类的评测只在测试集中正、负两种极性的微博上进行.由于 COAE2014 测试集中只有带情感的微博,没有无情感的微博,因此不对其主观性分类进行评测,只在该数据集上作极性分类的评测.观察表 16 中文微博的主观性分类结果,可以看出,即使选择 $\mathcal{S}^{C}$ 这种情感倾向性较高的(或者称高质量的)情感词典,如果采用基于情感词典的 Naive 主观性分类方法,分类的召回率高达 85% 以上,而准确率却只有 54%,如果采用人工词典 NTUSD,召回率超过 92%,而准确率只有 52%,相当于几乎将测试集中全部的微博都贴上"有情感"的标签,这显然是没有意义的.尽管如此,与 GPC 方法相比,用 cNSEm 方法扩展的情感词典在微博主观性分类上还是获得了较高的准确率.

表 16 基于中文微博情感分类任务的间接评测结果

情感词典	扩展方法	λ	CC&NLP2013 测试集						COAE2014 测试集		
			P	R	F1	P±	R±	F1±	P±	R±	F1±
MixedSD			0.5133	0.9980	0.6779	0.5988	0.5976	0.5982	0.6760	0.6738	0.6749
$\mathcal{S}^{C}$	GPC1	0	0.5488	0.9578	0.6694	0.6598	0.5660	0.6093	0.8229	0.6516	0.7273
		0.1	**0.5213**	**0.9817**	**0.6810**	0.6126	0.6013	0.6069	0.7322	0.6848	0.7077
		1	**0.5134**	**0.9957**	**0.6775**	0.5773	0.5748	**0.5760**	0.7590	0.7552	**0.7571**
	cNSEm	0.1	**0.5458**	**0.8902**	**0.6767**	0.6661	0.5929	0.6274	0.8225	0.6736	0.7406
		1	**0.5225**	**0.9729**	**0.6799**	**0.6628**	**0.6448**	**0.6537**	0.8059	0.7550	**0.7796**
NTUSD	GPC1	0	0.5208	0.9214	0.6655	0.6556	0.6041	0.6288	0.7946	0.6508	0.7155
		0.1	0.5179	0.9834	0.6785	0.5732	0.5637	0.5684	0.6990	0.6474	0.6722
		1	0.5135	0.9957	0.6776	0.5318	0.5295	**0.5306**	0.6567	0.6420	**0.6493**
	cNSEm	0.1	0.5210	0.9308	0.6681	0.6614	0.6156	0.6377	0.8019	0.6816	0.7369
		1	0.5204	0.9739	0.6783	0.6566	0.6394	**0.6478**	0.8000	0.7542	**0.7764**

表 17 基于英文微博情感分类任务的间接评测结果

情感词典	扩展方法	λ	PosAVDN						PosA					
			P	R	F1	P±	R±	F1±	P	R	F1	P±	R±	F1±
MPQA		0	0.6356	0.8723	0.7354	0.6267	0.5489	0.5852	0.6356	0.8723	0.7354	0.6267	0.5489	0.5852
$\mathcal{S}^{E}$	GPC1	0	0.6806	0.7146	0.6972	0.6684	0.4789	0.5580	0.6806	0.7146	0.6972	0.6684	0.4789	0.5580
		0.1	**0.6491**	**0.8343**	**0.7301**	0.5690	0.4777	0.5194	0.6661	0.8184	0.7344	0.5962	0.4934	0.5400
		1	**0.5964**	**0.9571**	**0.7349**	**0.4540**	**0.4343**	**0.4439**	0.6106	0.9391	**0.7400**	0.5058	0.4765	**0.4907**
	cNSEm	0.1	**0.6706**	**0.7435**	**0.7052**	0.6742	0.5018	0.5754	0.6769	0.7695	0.7202	0.6421	0.4934	0.5580
		1	**0.6286**	**0.9002**	**0.7403**	**0.6148**	**0.5525**	**0.5820**	0.6432	0.8922	**0.7475**	0.6201	0.5573	**0.5870**
SWNHQ	GPC1	0	0.6839	0.6068	0.6430	0.6647	0.4017	0.5008	0.6839	0.6068	0.6430	0.6647	0.4017	0.5008
		0.1	0.6724	0.7046	0.6881	0.5873	0.4138	0.4855	0.6737	0.7046	0.6888	0.5955	0.4174	0.4908
		1	0.6043	0.8962	0.7219	0.4257	0.3800	0.4016	0.6220	0.8673	0.7244	0.4651	0.4017	0.4311
	cNSEm	0.1	0.6840	0.6417	0.6622	0.6829	0.4391	0.5345	0.6837	0.6427	0.6626	0.6779	0.4367	0.5312
		1	0.6434	0.8174	0.7200	**0.6696**	**0.5452**	**0.6010**	0.6634	0.8084	0.7288	0.6667	0.5428	0.5984
SWN	GPC1	0	0.5764	0.9860	0.7275	0.5655	0.5573	0.5614	0.5764	0.9860	0.7275	0.5655	0.5573	0.5614
		0.1	0.5727	0.9910	0.7259	0.5189	0.5139	0.5164	0.5745	0.9930	0.7279	0.5188	0.5151	0.5169
		1	0.5685	0.9980	0.7244	0.5036	0.5030	0.5033	0.5692	0.9980	0.7249	0.5036	0.5030	0.5033
	cNSEm	0.1	0.5756	0.9920	0.7285	0.5645	0.5597	0.5621	0.5751	0.9940	0.7286	0.5728	0.5694	0.5711
		1	0.5682	0.9980	0.7241	**0.5894**	**0.5887**	**0.5890**	0.5714	0.9980	0.7267	0.5749	0.5742	0.5745

表 16 和表 17 显示,当情感词典的规模仅增加 10%时,相比 cNSEm,GPC 方法扩展的情感词大幅度地提高了主观性分类的召回率,但同时使其准确率严重下降.此时,单从 F1 值上看,GPC 方法扩展的情感词典对微博主观性分类的帮助要比 cNSEm 方法更大.但进一步观察扩展词典中排名靠前的"新情感词"发现,GPC 方法扩展的新情感词中,有大量文档频率较高的非情感词,从而使得主观性分类的

召回率大大提升，但准确率却严重受损. 当情感词典规模扩大 1 倍时，cNSEm 和 GPC 两种方法扩展的情感词典在微博主观性分类上获得了相近的 F1 值，但在准确率方面，cNSEm 方法扩展的情感词典明显更好.

对于情感极性分类问题，表 16 和表 17 显示，GPC 方法扩展的新情感词导致微博的极性分类准确率和 F1 值大幅度下降，这与文献[22]的结果一致. 而用 cNSEm 方法将情感词典扩大 1 倍时，显著提升了极性分类的召回率和 F1 值. 意料之外的，对于中、英文上的多部情感词典，用 cNSEm 方法扩展的情感词典不仅没使微博极性分类的准确率下降，反而略有提升. 例如，对中文情感词典 $\mathcal{S}^{C}$ 和 NTUSD 的扩展，使 CC&NLP2013 测试集上微博极性分类的准确率分别从 0.6598 和 0.6556 提升到 0.6628 和 0.6565；对英文情感词典 SWNHQ 和 SWN 的扩展，使得 tweets 上极性分类的准确率分别从 0.6647 和 0.5655 提升到 0.6696 和 0.5894. 另外，利用 cNSEm 将 $\mathcal{S}^{C}$ 和 NTUSD 扩展一倍时，得到的情感词典在微博主观性分类的准确率和 F1 值上、在极性分类的准确率、召回率和 F1 值上均超过混合人工情感词典 MixedSD，特别是在极性分类的准确率上，高出 10%（CC&NLP2013 测试集）和 19%（COAE2014 测试集）. cNSEm 对英文情感词典 SWNHQ 的扩展有类似的结论，将 SWNHQ 扩展一倍时，微博情感极性的召回率达到 MPQA 的效果，但准确率却较 MPQA 高出近 7%. 从这组数据可以看出，用 cNSEm 扩展出的新情感词在质量上几乎可以与已知（或人工）的情感词典竞争.

表 16 和表 17 还显示出 cNSEm 方法对种子情感词典（用于训练）、扩展情感词用的微博数据集、语种等均有较强的适应能力，通过 cNSEm 方法扩展的新情感词也具有通用性. 首先，在 cNSEm 对种子情感词典的适应性方面，中文种子情感典选择了词条及词典规模相差较大的 $\mathcal{S}^{C}$ 和 NTUSD 作为对照，英文种子情感词典选择了 $\mathcal{S}^{E}$ 和 SWNHQ 作为对照. 实验结果显示，用 cNSEm 方法扩展后，对微博主观性分类和情感极性分类都有改善，而对情感极性分类的改善幅度更大. 其次，在 cNSEm 对扩展情感词用的微博数据集的适应性方面，我们在评估候选特征、上下文窗口长度等因素时所选用的微博数据集是从 COAE2014 中随机采样的 0.5M 条微博，而在本小节扩展情感词时，重新采样了 1M 条微博，情感分类结果显示，用 0.5M 条微博上分析得到的 cNSEm 特征及参数，在 1M 条微博数据集上抽取得到的新情感词仍然有较高的质量. 第三，语种适应性方面，我们选择了中、英文两种语种的微博，尽管各类特征（如文档特征 DOC）或参数（如上下文窗口长度）在不同语种的微博上表现稍有差异，但 cNSEm 仍能很好地适用于中、英两种语言的微博数据集. 最后，在扩展得到的情感词的通用性方面，COAE2014 测试集与用于新情感词扩展的 1M 条微博都来自 COAE2014 评测任务，是同源的，但 CC&NLP2013 测试集用于情绪分类任务，与新情感词扩展所用的数据集差异较大. 表 16 显示，基于 $\mathcal{S}^{C}$ 训练的 cNSEm 并在 COAE2014 测试集上所抽取的新情感词对两个测试集上微博情感极性分类效果都有显著的改善：对于 CC&NLP2013 测试集，当扩展规模到 100%时，F1±从 0.6093 提高到 0.6537，提高了 7.29%；在 COAE2014 测试集上的表现类似，情感词典扩展后，微博情感极性分类的 F1±从 0.7273 提高到 0.7796，提高了 7.19%. 这说明，cNSEm 方法扩展的新情感词具有通用性.

表 17 中还列出了将候选词限定在形容词上扩展得到的新情感词对情感分类任务的影响. 结果显示，在 $\mathcal{S}^{E}$ 和 SWNHQ 两个种子情感词典及多项评测指标上看，对于英文微博情感分类任务，如果选择 GPC 方法扩展情感词，将候选词限定在形容词上更好. 而对于 cNSEm 方法，在候选词性 PosA 或候选词性 PosAVDN 上扩展的新情感词对英文微博情感分类任务并未表现出明显差别. 结合新情感词抽取的其它评测方法及表 18 中的新情感词示例，我们一方面更加确信了 cNSEm 方法可以适应不同的候选词性，同时也更加质疑通过情感分类的任务是否足以说明或对比不同方法所抽取的新情感词的质量. 下面通过一个例子来说明这一质疑：设微博$\langle t_1, t_2, s, t_4, s^e\rangle$中 s 为种子情感词，s^e 为扩展得到的新情感词. 情感词扩展前，对该微博的分类仅依据 s，而扩展后，对该微博的分类依据 s 和 s^e，相当于利用了微博中的更多信息. 在这种情况下，即便 s^e 不是情感词，也可能对情感分类起到有益的作用.

关于表 18 中新情感词样例的进一步说明：为了能够评测 cNSEm 的效果并且不增加主观性，实验设计时并没有通过人工来评测抽取到的“新情感词”，而是利用被普遍使用的人工情感词典进行评测（主要基于大连理工大学的人工情感词典）. 具体而言，非情感词是指那些在通用词典中出现但没在任何人工情感词典中出现过的词（本文的人工情感词

典包括六部分,分别是来自大连理工大学、知网、清华大学、台湾大学的情感词典、标注带有情感的新浪微博表情符号以及 COAE2014 提供的新情感词典,如式(1)所示),而用于训练的情感词典和用于测试的情感词典都基于大连理工大学的人工情感词典,训练和测试各用一半.这样,评测实验中的候选词包括三部分:(1)没被通用情感词典收录的情感词;(2)用作测试的情感词;(3)包含在其它人工情感词典中的词.例如:表 18 中"不正确"不在通用词典中,"悲哀"来自大连理工大学的情感词典,而"水灵"则包含在其它情感词典中.因此,就出现了表 18 中新情感词看上去不"新"的现象.在实际提交到 COAE2013 的参赛系统中,训练用的情感词典是基于大连理工大学的情感词典并进行了适当扩展,而候选词也只是那些不在通用词典中的词.例如,UdeM-t3-2 参赛系统提交的前 20 个词是:很好、给力、白皙、尼玛、达人、伤不起、吃货、和美、佳品、高端、发飙、大礼、坑爹、柔润、你妹、的真、无语、柔滑、淡定、傻逼.这些词都未在给定的通用词典中出现.

表 18 新情感词样例

("+"、"−"分别表示该新情感词极性为"正"或"负")

	中文微博		英文 tweets	
Rank	cNSEm	GPC	cNSEm	GPC
1	水灵/−	自己/−	elixer/−	people/−
2	频繁/−	没有/−	able/+	feel/−
3	深沉/−	觉得/−	guilty/−	tcot/−
4	可悲/−	结果/−	unheard/−	obama/−
5	矛盾/−	不要/−	painful/−	even/−
6	矿泉喷雾/−	事情/−	sensitive/−	stop/−
7	不正确/−	人家/−	optimistic/−	think/−
8	悲哀/−	东西/−	good-looking/+	shit/−
9	邪说/−	不知道/−	unable/−	poor/−
10	吃力/+	政府/−	feel/−	ca/−
11	甜香/+	事件/−	unaware/−	bit/−
12	恍恍惚惚/−	还是/−	needless/−	still/−
13	陷入/−	知道/−	victorious/−	left/−
14	苍白/−	司机/−	liable/−	away/−
15	阳刚/+	可能/−	analytical/+	get/−
16	别扭/−	其实/−	productive/+	never/−
17	合适/+	就是/−	indicative/−	someone/−
18	惭愧/+	行为/−	loveeee/+	really/−
19	残忍/−	时候/−	carpal/−	trying/−
20	可怕/−	社会/−	unfocused/−	way/−

7 结论与展望

情感词典对文本情感分析任务具有重要意义.人工情感词典虽然准确但构建的代价很大,难以适应微博这类新情感词快速更迭的数据集.

本文针对中英文微博数据中情感词的词性分布、情感词共现等特点,提出了基于分类的微博新情感词抽取方法 cNSEm,并且将名词等形容词外的其它词性纳入候选词集合中.cNSEm 利用人工情感词典和微博数据集构建训练数据,训练分类器并对候选词进行极性分类.实验结果显示,与基于共现和极性传播的 GPC 方法相比,仅考虑形容词类型的情感词时,cNSEm 与 GPC 性能相当,但扩大候选词的词性集合后,cNSEm 在多项评测指标上都远好于 GPC,其 Rprec 指标达到了人工情感词典的性能.

实验还发现,在情感词的极性判断方面,尽管从统计上看,通过与已知正、负极性情感词的共现可以有效地判别候选词的极性,但其准确率仍然明显低于 cNSEm 中的极性判断方法.

本文还通过大量的实验来分析上下文、词性、语言学模式、修饰关系、文档特征、与情感词的共现等各类特征,以及上下文窗口长度等参数对 cNSEm 的影响,发现文档特征不利于 cNSEm 从英文微博中抽取新情感词,其它各类特征对 cNSEm 性能的提高都有帮助,特别是语言学模式和与情感词的共现特征.尽管各类特征对情感词的抽取和极性判断有不同程度的帮助,但不同的使用方法会带来不同的效果,例如:用 n_gram 表示的上下文特征要比 unigram 更好,在语言学模式中引入候选情感的词性能提高 cNSEm 的性能.

除了利用理想情感词典对 cNSEm 进行评测之外,本文还通过考察扩展的新情感词对微博情感分类的影响对 cNSEm 进行间接评测.评测结果显示,cNSEm 方法对种子情感词典、扩展情感词用的微博数据集、语种等均有较强的适应能力,通过 cNSEm 方法扩展的新情感也具有良好的通用性.

cNSEm 方法还参加了 COAE2014 的微博新情感词抽取子任务,尽管当时只用到了词性和语言学模式两类特征,并且未进行参数优化,cNSEm 仍在 26 支参赛系统中排名第二,显示了较强的竞争力.

下一步的主要工作:

(1)在确定新情感词及其极性时,cNSEm 是对出现在不同场景下候选词的分类结果进行了简单的投票统计,因此,得到的新情感词只能说明在大部分情况下该词可能是情感词,以及其可能的情感极性.然而,在不同的上下文中,词的情感极性可能会发生变化,因此,有必要在 cNSEm 的基础上,分析带有极性歧义的情感词,挖掘这些情感词在不同极性时的用词规律,以增强情感分析的准确性.

(2)由于 cNSEm 方法中用到的特征利用了句

法分析的结果，如词性、依存关系等，而句法分析的效率将成为 cNSEm 效率的瓶颈，因此，如果要进行新情感词的在线抽取，还需要探索可替代句法分析结果的其它特征.

(3) 在从微博中抽取新情感词时，发现很多分词系统对微博的分词效果较差，这也是有待克服的一大障碍.

(4) 如何利用 cNSEm 方法在抽取新情感词的同时，抽取情感或评论的对象，也是我们感兴趣的工作.

(5) cNSEm 方法的成败很大程度上依赖于特征选择，把特征选择问题交给深度学习来完成，将是我们未来的工作之一. 也相信本文所进行的特征分析工作对用深度学习的方法抽取新情感词是有帮助的，例如，在确定是否要对微博进行句法分析、是否将词性等作为深度学习的输入、如何更恰当地选择上下文等方面.

致　谢　感谢加拿大蒙特利尔大学刘晓华博士对本文工作的建议和帮助，感谢审稿专家提出的宝贵意见!

参　考　文　献

[1] Liu De-Xi. Effect of sentimental word expansion on the performance of microblog sentimental classification task. Journal of Chinese Computer Systems, 2016, 37(5): 957-965(in Chinese)
(刘德喜. 情感词扩展对微博情感分类性能影响的实验分析. 小型微型计算机系统, 2016, 37(5): 957-965)

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

[3] Zhao Yan-Yan, Qin Bing, Liu Ting. Sentiment analysis. Journal of Software, 2010, 21(8): 1834-1848(in Chinese)
(赵妍妍, 秦兵, 刘挺. 文本情感分析. 软件学报, 2010, 21(8): 1834-1848)

[4] Liu B. Sentiment Analysis and Opinion Mining. USA: Morgan & Claypool, 2012: 1-165

[5] He Yan-Xiang, Sun Song-Tao, Niu Fei-Fei, Li Fei. A deep learning model enhanced with emotion semantics for microblog sentiment analysis. Chinese Journal of Computers, 2017, 40(4): 773-790(in Chinese)
(何炎祥, 孙松涛, 牛菲菲, 李飞. 用于微博情感分析的一种情感语义增强的深度学习模型. 计算机学报, 2017, 40(4): 773-790)

[6] Huang Fa-Liang, Feng Shi, Wang Da-Ling, Yu Ge. Mining topic sentiment in microblogging based on multi-feature fusion. Chinese Journal of Computers, 2017, 40(4): 872-888(in Chinese)
(黄发良, 冯时, 王大玲, 于戈. 基于多特征融合的微博主题情感挖掘, 计算机学报, 2017, 40(4): 872-888)

[7] Jiang L, Yu M, Zhou M, et al. Target-dependent Twitter sentiment classification//Proceedings of the ACL 2011. Oregon, USA, 2011: 151-160

[8] Bravo-Marquez F, Mendoza M, Poblete B. Combining strengths, emotions and polarities for boosting Twitter sentiment analysis//Proceedings of the WISDOM 2013. Chicago, USA, 2013: 1-9

[9] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives//Proceedings of the ACL 1997. Madrid, Spain, 1997: 174-181

[10] Turney P D, Littman M L. Measuring praise and criticism: inference of semantic orientation from association. ACM Transactions on Information Systems, 2003, 21(4): 315-346

[11] Kaji N, Kitsuregawa M. Building lexicon for sentiment analysis from massive collection of HTML documents//Proceedings of the EMNLP-CoNLL 2007. Prague, Czech Republic, 2007: 1075-1083

[12] Feng S, Zhang L, Li B, et al. Is Twitter a better corpus for measuring sentiment similarity?//Proceedings of the EMNLP2013. Washington, USA, 2013: 897-902

[13] Yu H, Deng Z H, Li S. Identifying sentiment words using an optimization-based model without seed words//Proceedings of the ACL 2013. Sofia, Bulgaria, 2013: 855-859

[14] Kamps J, Marx M, Mokken R, et al. Using WordNet to measure semantic orientations of adjectives//Proceedings of the LREC 2004. Lisbon, Portugal, 2004: 1115-1118

[15] Andreevskaia A, Bergler S. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses//Proceedings of the EACL 2006. Trento, Italy, 2006: 209-215

[16] Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping//Proceedings of the HLT-NAACL 2003 - Volume 4. Edmonton, Canada, 2003: 25-32

[17] Rao D, Ravichandran D. Semi-supervised polarity lexicon induction//Proceedings of the EACL 2009. Athens, Greece, 2009: 675-682

[18] Esuli A, Sebastiani F. Pageranking WordNet synsets: An application to opinion mining//Proceedings of the ACL 2007. Prague, Czech Republic, 2007: 442-431

[19] Awadallah A, Radev D. Identifying text polarity using random walks//Proceedings of the ACL 2010. Uppsala, Sweden, 2010: 395-403

[20] Awadallah A, Abu-Jbara A, Jha R, et al. Identifying the s emantic orientation of foreign words//Proceedings of the ACL 2011. Oregon, USA, 2011: 592-597

[21] Xu G, Meng X, Wang H. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources//Proceedings of the COLING 2010. Beijing, China. 2010: 1209-1217

[22] Volkova S, Wilson T, Yarowsky D. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams//Proceedings of the ACL 2013. Sofia, Bulgaria, 2013: 505-510

[23] Qiu G, Liu B, Bu J, et al. Expanding domain sentiment lexicon through double propagation//Proceedings of the IJCAI 2009. California, USA, 2009: 1199-1204

[24] Zhao W X, Jiang J, Yan H, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid//Proceedings of the EMNLP 2010. Massachusetts, USA, 2010: 56-65

[25] Lazaridou A, Titov I, Sporleder C. A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations//Proceedings of the ACL 2013. Sofia, Bulgaria, 2013: 1630-1639

[26] Xu L, Liu K, Lai S, et al. Walk and learn: A two-stage approach for opinion words and opinion targets co-extraction//Proceedings of the WWW 2013. Rio de Janeiro, Brazil, 2013: 95-96

[27] Kim S M, Hovy E. Determining the sentiment of opinions//Proceedings of the COLING 2004. Geneva, Switzerland, 2004: 1367-1373

[28] Esuli A, Sebastiani F. SentiWordNet: A publicly available lexical resource for opinion mining//Proceedings of the LREC 2006. Genoa, Italy, 2006: 417-422

[29] Mohtarami M, Lan M, Tan C L. Probabilistic sense sentiment similarity through hidden emotions//Proceedings of the ACL 2013. Sofia, Bulgaria, 2013: 983-992

[30] Peng W, Park D H. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization//Proceedings of the ICWSM 2011. Barcelona, Spain, 2011: 273-280

[31] Wiebe J. Learning subjective adjectives from corpora//Proceedings of the in AAAI/IAAI 2000. Texas, USA, 2000: 735-740

[32] Hatzivassiloglou V, Wiebe J. Effects of adjective orientation and gradability on sentence subjectivity//Proceedings of the COLING 2000. Saarbrucken, Germany, 2000: 299-305

[33] Liu De-Xi, Nie Jian-Yun, Zhang Jing, et al. Extracting sentimental lexicons from Chinese microblog: A classification method using N-Gram feature. Journal of Chinese Information Processing, 2016, 30(4): 193-205(in Chinese)
(刘德喜, 聂建云, 张晶等. 中文微博情感词提取: N-Gram为特征的分类方法. 中文信息学报, 2016, 30(4): 193-205)

[34] Du W, Tan S, Cheng X, et al. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon//Proceedings of the WSDM 2010. New York, USA, 2010: 111-120

[35] Zhang L, Liu B. Identifying noun product features that imply opinions//Proceedings of the ACL-HLT 2011. Portland, USA, 2011: 575-580

[36] Velikovich L, Blair-Goldensohn S, Hannan K, et al. The viability of web-derived polarity lexicons//Proceedings of the HLT-NAACL 2010. Los Angeles, USA, 2010: 777-785

[37] Rao Y H, Lei J S, Liu W Y, et al. Building emotional dictionary for sentiment analysis of online news. World Wide Web, 2014, 17(4): 723-742

[38] Becker L, Erhart G, Skiba D, et al. AVAYA: Sentiment analysis on Twitter with self-training and polarity lexicon expansion// Proceedings of the SemEval 2013. Atlanta, USA, 2013: 333-340

[39] Turney P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews//Proceedings of the ACL 2002. Philadelphia, USA, 2002: 417-424

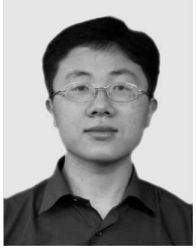
[40] Ding X, Liu B, Yu P S. A holistic lexicon-based approach to opinion mining//Proceedings of the WSDM 2008. California, USA, 2008: 231-240

[41] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis//Proceedings of the HLT/EMNLP 2005. Vancouver, Canada, 2005: 347-354

[42] Dave K, Lawrence S, Pennock D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews//Proceedings of the WWW 2003. Budapest, Hungary, 2003: 519-528

[43] Hu M, Liu B. Mining and summarizing customer reviews//Proceedings of the SIGKDD 2004. Washington, USA, 2004: 168-177

[44] Buckley C, Voorhees E M. Retrieval evaluation with incomplete information//Proceedings of the SIGIR 2004. Sheffield, UK, 2004: 25-32



LIU De-Xi, born in 1975, Ph. D., professor, Ph. D. supervisor. His research interests include social media processing, information retrieval, and natural language processing.

NIE Jian-Yun, born in 1963, Ph. D., professor, Ph. D. supervisor. His research interests focus on information retrieval.

WAN Chang-Xuan, born in 1962, Ph. D., professor, Ph. D. supervisor. His research interests include Web data management and data mining.

LIU Xi-Ping, born in 1981, Ph. D., associate professor. His research interests include Web data management and data mining.

LIAO Shu-Mei, born in 1976, Ph. D., associate professor. Her research interests include information management

and information system.

LIAO Guo-Qiong, born in 1969, Ph. D., professor, Ph. D. supervisor. His research interest is social computing.

ZHONG Min-Juan, born in 1976, Ph. D., associate professor. Her research interests include Web data management and data mining.

JIANG Teng-Jiao, born in 1976, Ph. D., lecturer. Her research interests focus on sentiment analysis.

Background

Sentiment analysis has wide and important applications in the field of public opinion analysis and product reviews analysis, which has attracted wide attention from academic and enterprise in recent years. As an important resource for text sentiment analysis, the sentiment dictionary should have full coverage, be updated frequently and labelled precisely. A sentiment dictionary collected and labeled manually is more accurate than an auto-generated one, but its disadvantages of limited coverage and updating difficulty are magnified in Web 2.0 era, where the new sentiment emerged frequently and spread rapidly. Therefore, the automatic or semi-automatic methods should be explored to extracted new sentiment words from web data, especially from microblogs where users express their sentiments conveniently.

In this paper, we analyzed and compared sentiment words distribution in Chinese and English microblogs respectively, including the distribution of POSes and co-occurrence of sentiment words. Compared with English microblogs, there are more challenges when extracting new sentiment words from Chinese microblogs, especially when nouns are taking into consideration. Basing on analysis, we proposed a classification based sentiment words extraction method cNSEm. cNSEm makes fully use of existing resources such as manual sentiment dictionaries and huge amount of microblogs, and generates training data automatically.

Experimental results show that cNSEm performs as good as classical method GPC after the candidates are constrained to the adjectives. Additionally, cNSEm perform significantly better than GPC if more POSes are taken into consideration, and Rprec score shows that sentiment words extracted by cNSEm are competitively compared with manual sentiment dictionary. cNSEm is robust on different sentiment seeds (for training), different microblog datasets (where new sentiment words extracted from), and different languages. Moreover, the impacts of different categories of features employed by cNSEm are analyzed by carefully designed experiments.

This team has done some works about sentiment analysis on financial text, which were published on *Chinese Journal of Computers*.

This work is supported by the Natural Science Foundation of China (Nos. 61762042, 61363039 and 61562032), the Transformation Project of Scientific and Technological Achievements from Universities in Jiangxi Province (No. KJLD14035), and the Natural Science Foundation of Jiangxi Province (Nos. 20171BAB202021 and 20152ACB20003). These projects are focus on summarizing content or sentiment of multi-microblogs, and extracting the opinions of products and their attributes in microblogs. Sentiment words are the key resource for these projects.