文章编号:1000-5862(2015)03-0290-07

中文微博句子倾向性分类中特征抽取研究

徐雄飞,徐 凡,王明文*,左家莉,罗文兵
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:针对中文微博句子倾向性分类问题,在充分降低由于情感词典的扩充工作带来系统开销的基础上,抽取了中文微博句子中标点符号、情感词权重、词汇级和句法级等新型平面和结构化特征,探索了有效的特征选择方法. 在基准 COAE 和 NLP&CC 中文微博语料上进行双向交叉和独立实验,并研究了有效的不平衡性语料的处理方法. 实验结果表明:采用该文提出的特征后,中文微博句子倾向性分类的性能得到显著提升.

关键词:中文微博;句子倾向性;特征抽取;分类
中图分类号:TP 391 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2015.03.13

0 引言

随着 Web2.0 技术的逐渐普及,人们已经由信息获取者转变为网络内容的主要制造者,导致用户生成内容(User generated content,UGC)的数量以指数级速度增长,中文相关研究也越来越广泛[1-2]. 作为一种新型网络信息表达方式,微博具有短小精悍、使用便捷和传播迅速等特点,已经成为一种重要的信息获取和舆情传播的途径. 目前,国内外针对短文本倾向性分析评测已经开展,如国外有 TREC Blog Track 和 NTCIR 等,国内有 Chinese opinion analysis evaluation(COAE)和 NLP&CC(Natural language processing & Chinese computing)评测等,同时针对互联网产品的微博情感性评论将对产品的推广和改进工作具有非常重要的影响,这些均使得研究微博这种短文本有着重要的研究价值与现实意义.

目前,绝大多数方法将中文微博句子倾向性分析看成分类问题[4-13],其基本过程包括:(i)把标注好的中文微博句子倾向性语料分成训练语料和测试语料 2 个部分,从训练语料中提取出各种词汇、语法、句法和语义等特征,生成相应的微博句子倾向性训练实例集,并利用支持向量机(Support vector machine,SVM)、最大熵(Maximum entropy,ME)、朴素贝叶斯(Naive bayes,NB)等分类器训练得到分类器模型;(ii)利用分类器模型对微博句子测试样例进行预测,给出测试样例可能存在的倾向性.

然而,目前已有的方法主要存在 2 个问题:(i)情感词典过于复杂,如 COAE 2013 评测中取得较好分类性能的系统过于依赖于较庞大且复杂的情感词典,它们需要利用网络中的新词、同义词词林或 HowNet 等资源扩充原始情感词典,这些无疑增加了系统的复杂性;(ii)现有中文微博语料的交叉验证情况比较缺乏,如 COAE 2013 中的评测结果大部分采用 NLP&CC 2013 作为训练语料,而将 COAE 作为测试语料,然而两者并不一定兼容,因为 NLP&CC 2013 着重于细粒度的情感分类问题,属于情绪分析(兴奋、快乐、忧愁和悲伤);相反,COAE 2014 是粗粒度的情感分类(正性、负性和中性).

基于此,本文一方面在充分降低由于情感词典的扩充工作带来系统开销的基础上,提出简单有效的微博句子中的标点符号、情感词权重、词汇级和句法级等新型平面和结构化特征,并将这些特征指导微博句子倾向性分类;另一方面着重探索了 COAE 和 NLP&CC 微博语料的交叉验证情况,即分别将 COAE 和 NLP&CC 作为训练和测试语料,以便进一步验证本文提出的这些特征的有效性. 通过 COAE 和 NLP&CC 中文微博语料库的双向交叉和独立实验,结果表明采用上述特征后,中文微博句子倾向性分类的性能得到显著提升.

收稿日期:2015-02-19
基金项目:国家自然科学基金(61272212,61163006,61203313,61365002,61462045)资助项目.
通信作者:王明文(1964-),男,江西南康人,教授,博士生导师,主要从事信息检索、数据挖掘和并行计算的研究.

1　相关工作

文献 [3] 将中文微博文本情感分析分为 3 类任务: 文本预处理、情感信息抽取和情感分类,其中微博情感信息抽取包括情感词、主题和关系的抽取; 微博情感分类主要有基于语义词典的情感计算和基于机器学习的情感分类 2 大类. 基于此,根据最新的中文微博情感分析工作,将中文微博句子倾向性分析方法分为 3 大类: 基于语义词典的方法、基于机器学习的方法、语义词典和机器学习相结合方法.

针对基于语义词典的方法,文献 [4] 着重考虑了传统情感词典中缺乏微博中的新词和变形词等情况,结合同义词词林构建了新的情感词典(包括情感词词表、程度副词表和否定词表 3 大类) ,同时利用点间互信息技术对中文微博句子进行倾向性判断. 文献 [5] 利用中文情感词库,先对微博句子中的情感词评分,然后利用句法分析技术,抽取出微博句子中的词和父节点的词对信息,并对这些词对进行综合评分,最后按层次求和得到整个句子和整个文档的得分. 文献 [6] 利用微博中的表情符号信息扩充了传统的情感词典,通过对大量微博中与表情"共现"的文本的情感倾向性分析,确定表情的情感倾向,并以此构建面向情感倾向分析的表情情感词典. 文献 [7] 结合网络新词和基础情感词,分别构建了 4 大类词典: 基础情感词典、表情符号词典、否定词词典和双重否定词词典,同时探索并集成了否定和双重否定等汉语语言学特征和微博情感表达特征等信息.

针对基于机器学习的方法,文献 [8] 提出了基于 SVM 的层次结构的多策略方法,该层次策略将微博分别考虑为分句和不分句 2 种情况,同时结合 1 步 3 分类(直接进行正性、中性和负性判断) 和 2 步 3 分类(先进行主观和客观句的识别,然后再进行前面提到的 3 种极性判断) ,并且着重分析了微博中句子与句子之间涉及的主题相关和无关性对倾向性分类性能的影响. 文献 [9] 探索了微博话题间的切换对句子倾向性的分类影响,通过对分话题模型训练可以弥补训练数据的不足,从而提升微博观点句识别系统的性能. 文献 [10] 利用 NB 和 ME 分类器,采用微博句子中的 unigram(1 元文法) 和 bigram(2 元文法) 组合的字或词等特征进行"国产电影"和"高铁"2 个不同话题的微博情感分类.

针对语义词典和机器学习相结合方法,文献 [11] 首先以《知网》和《同义词林》为基础,构建了完备的微博情感词典,采用向量空间模型表示微博文本,抽取出微博语义和情感词及影响因子等特征,并采用情感强度作为主要特征权重,利用多特征融合方法构建中文微博情感分析计算模型. 文献 [12] 首先利用网络流行语扩充现有情感词典,然后结合此词典和提取的微博文本特有的特征向量,最后提出了由短句至长句极性的 2 步走的中文微博情感分析策略. 文献 [13] 着重研究了中文微博的依存句法分析方面的信息,抽取出微博句子中依赖关系的 nsubj(主语) 和状语修饰的 advmod(副词态) 关系,然后结合情感词典进行微博句子的倾向性识别.

2　中文微博情感语料库简介

目前,仅在 COAE 和 NLP&CC 中涉及到中文微博句子的倾向性分析的研究工作. 为清晰起见,表 1 针对微博的主题数、来源、规模、类别和实例等方面对中文微博语料库进行了对比,其中 NLP&CC 2013 和 COAE 2014 没有对主题进行分类,COAE 没有公开微博的具体来源. 由于未取得 COAE 2013 的测试语料,仅列出 COAE 2013 的训练语料的统计数据,后续的相关实验也在此基础上进行.

从表 1 可以看出 NLP&CC 2012 与 COAE 的任务比较相近,两者属于粗粒度的情感分类,即对中文微博句子进行正性、负性或中性分类; 相反,NLP&CC 2013 却属于细粒度的情感分类问题,它需要对中文微博句子进行情绪分析(兴奋、快乐、忧愁、悲伤) .

3　中文微博句子倾向性分类方法

本文采用情感词典和机器学习相结合的方法进行中文微博句子倾向性分类,利用文献 [14] 发布的原始情感词汇本体进行中文微博句子的情感特征方面的抽取,而没有采用文献 [4-7,10-13] 扩充情感词典等复杂工作,目的是为了最大限度地降低整个系统的复杂性.

3.1　特征抽取

表 2 例举了该文所采用的特征,主要分为 4 大类: (i) 标点符号特征,(ii) 情感词典特征,(iii) 词汇特征,(iv) 句法特征. 采用标点符号特征的原因在于人们通常采用标点符号(如问号、感叹号和省略号等) 来表达一定的情感; 采用情感词典特征(如最大正性、负性情感词权重等) 的原因在于这些情感词所具有的权重大小直接衡量了人们的情感表达强弱; 采用词汇特征(如形容词、名词等) 的原因在于

它们在一定程度上显示了情感的表达方式,如重复名词、突出形容词等. 以上这些特征主要是捕获了中文微博句子中的平面结构信息,属于平面特征. 相反,句法特征(如中文微博句子对应的短语成分句法树中从根节点到形容词的路径信息)可以使得本文试图挖掘出人们在表达情感时所采用的句式,如ROOT-IP-VP-VA 等. 这种句法路径信息体现了一定程度上的结构性,属于一种结构化特征. 为了清晰起见,表 2 给出了本文所采用的特征集以及例 1 所对应的特征值.

例 1 "三星手机很漂亮! 它包含了很多先进技术... @ mark @ sunshine"

表 1 中文微博情感语料库对比

对比项	NLP&CC		COAE	
	NLP&CC 2012	NLP&CC 2013	COAE 2013 (训练语料部分)	COAE 2014
主题数	20	未知	1	未知
来源	腾讯微博	新浪微博	未知	未知
已标注规模/条	3 416	300 000	1 838	5 000
类别	情感分析(正性、负性、中性)	情绪分析(兴奋、快乐、忧愁、悲伤)	情感分析(正性、负性)	情感分析(正性、负性)
实例	<weibo id="2"> <sentence id="1">#iPad3#这么麻烦的东西怎么还有那么多人在用,又是越狱又是破解.</sentence> </weibo> 此句针对 iPad 表明了负性情感.	<weibo id="2"> <sentence id="1">厦门午后的阳光很好,海边的风很舒服,可是即将要离去,有点淡淡的忧伤,但还是很开心.</sentence> </weibo> 此句中主要情绪是幸福和喜悦,次要情绪是悲伤.	<weibo id="2"> <sentence id="1">#成功学院饮食文化节#蒙牛真的很牛,太牛了,哈哈.</sentence> </weibo> 此句针对蒙牛产品表明了正性情感.	<weibo id="2"> <sentence id="1">强烈推荐北京银行信用卡. 人家北京银行不差钱儿,是偶用过的信用卡中活动最多、礼品最好滴.</sentence> </weibo> 此句针对北京银行信用卡表明了正性情感.

为清晰起见,图 1 给出了例 1 对应的句法树结构,可以分别抽取出从根节点到 VA(predicative adjective,表语形容词)和 JJ(adjective or numeral ordinal,形容词或序数词)的路径信息. 如针对图 1 左边的句法树,可以得到"ROOT-IP-VP-VP-VA(漂亮)"路径.

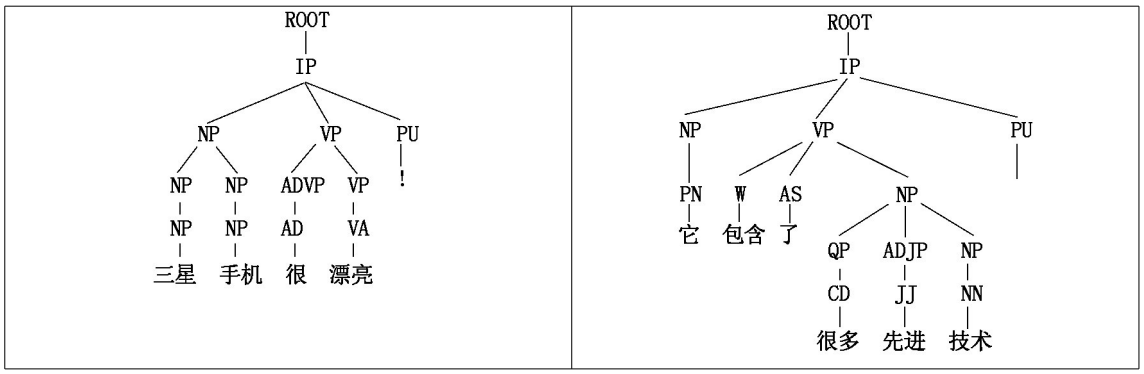

图 1 例 1 对应的句法分析树

3.2 特征选择

由于表 1 中的特征类型较多,它们的组合类型数目也相当多,所以采用传统的"爬山法"[15] 难以选择出有效的特征. 因此,利用在文本分类中广泛采用的信息增益(Information gain,IG)特征选择方法[16],能够有效地计算出这些特征的增益情况,从而能够选择出相对较优的特征集合. 其中,信息增益为

$$G(t) = -\sum_{i=1}^{m} p(c_i)\log P(c_i) + p(t)\sum_{i=1}^{m} p(c_i \mid t)\cdot \log p(c_i \mid t) + p(\bar{t})\sum_{i=1}^{M} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t}), \quad (1)$$

其中 $p(c_i \mid t)$ 为文本包含特征项 t 且属于类别 c_i 的概率,$p(c_i \mid \bar{t})$ 为文本不包含特征项 t 且属于类别 c_i

的概率,$p(c_i)$ 为文本类别的概率,$p(t)$ 为特征项 t 的概率.

表 2　中文微博句子倾向性分类特征

特征类别	特征序号	备注	特征值
标点符号特征	F1	微博中问号的个数	0
	F2	微博中感叹号的个数	1
	F3	微博中省略号的个数	1
	F4	微博中@ 符号的数目	2
情感词典特征	F5	分词后匹配情感词典中最大正性情感词权重	(漂亮 6)(先进 5)($MaxPos=6$, $AllPos=11$)
	F6	分词后匹配情感词典中最大负性情感词权重	($MaxNeg=0$, $AllNeg=0$)
	F7	微博中所有正性情感词权重之和与所有负性情感词权重之和的差值	$AllPos-AllNeg=11$
	F8	微博中所有正性情感词权重之积	(漂亮 6)(先进 5)　$6\times5=30$
	F9	否定词个数是基数还是偶数	(基数个 =0,偶数个 =1)
	F10	微博中所有负性情感词权重之积	0
词汇特征	F11	微博中形容词的个数	1　(AD 很)
	F12	微博中动词的个数	1　(VV 包含)
	F13	微博中名词的个数	2(NN 手机、技术)
	F14	微博中情态助动词的个数	0
	F15	微博中时间名词的个数	0
	F16	微博中基数词的个数	1(CD 很多)
	F17	微博中固有名词的个数	1(NR 三星)
句法特征	F18	微博短语句法分析树中从 ROOT 节点到 VA(predicative adjective,表语形容词) 的路径	图 1 句法树中的路径: ROOT-IP-VP-VP-VA(漂亮)
	F19	微博短语句法分析树中从 ROOT 节点到 JJ(adjective or numeral ordinal,形容词或序数词) 的路径	图 1 句法树中的路径: ROOT-VP-NP-ADJP-JJ(先进)

注: 表 2 中的特征值来自例 1,其中 $MaxPos$ 为最大正性情感词权重,$AllPos$ 为所有正性情感词权重之和,$MaxNeg$ 为最大负性情感词权重,$AllNeg$ 为所有负性情感词权重之和.

4　实验结果及分析

为了验证本文提出的特征的有效性,进行了相应的中文微博句子倾向性分类实验.

4.1　实验设置

本文采用 COAE 和 NLP&CC 中文微博语料进行相应实验,采用 Stanford 句法分析器对中文微博句子进行短语句法分析,采用 LibSVM [17] 作为分类器(利用 $T=0$ 号核函数). 为了与 COAE 的评测结果具有可比性,遵照他们的评测方法,采用 NLP&CC 2013 作为训练语料,COAE 2014 作为测试语料,针对语料中的正性和负性 2 类情感进行识别(由于 NLP&CC 2013 是一种情绪分析,与 COAE 的处理过程一致,将“兴奋”和“快乐”作为正性情感,将“忧愁”和“悲伤”作为负性情感),分别就微博的正性和负性情感采用 Precision、Recall、Macro_F1 和 Micro_F1 进行评测. 由于在 COAE 2014 的评测过程中需要提交约 10 000 条左右微博. 因此,分别按照包含测试集微博的 20%、50%、65%、80% 和 100% 共 5 种情况进行实验. 此外,对于单个独立 COAE 和 NLP&CC 语料采用 10 倍交叉验证方式. 由于 NLP&CC 2013 语料太大,选择太多的标注实例会给分类器带来噪音,于是从中随机选择 10 000 条用于训练和测试.

4.2 实验结果及分析

本节分别针对 NLP&CC 与 COAE 语料的双向交叉实验,以及单独语料实验进行详细的结果分析.

4.2.1 与现有最新系统(COAE 2014 评测前 3 名)性能对比实验 表 3 列出了系统与已有的 COAE 2014 最新评测前 3 名的系统性能对比情况. 其中,P^+为正性情感的 precision,R^+为正性情感的 recall,$F1^+$为正性情感的 F1 值,P^-为负性情感的 precision,R^-为负性情感的 recall,$F1^-$为正性情感的 F1 值,Macro 为宏平均,而 Micro 为微平均. 同时,为了最大限度地与 COAE 的评测过程一致,在评测时同样加入了"干扰"过程,如"50% 测试集"为有 50% 的微博数与标准测试集对应,另外的 50% 的微博由人为加入.

由表 3 可知:(i) 采用特征选择方法过后,系统检测性能具有一定程度的提升. 通过进一步实验分析,经过特征选择后的集合为 12 个:{F1,F2,F3,F5,F6,F7,F8,F11,F12,F13,F18,F19},明显少于原始的 19 个特征. 这些充分说明了当特征类型较多时,可以采用较为有效的信息增益技术选择相对较优的特征集合,从而避免复杂的"爬山法"特征选择技术;(ii) 该系统在包含 20% 的测试集时能够取得较高的 precision,而相对较低的 recall,原因在于本文提出的表 2 所示的这些特征属于细粒度的,它们涵盖了标点符号级、情感词典级、词汇级和句法级等特征,具有较高的区分度. 同时,由于只选择了 20% 的测试集,导致 recall 较低. 随着包含测试集比例的提高,系统在保证较高的 precision 前提下使得 recall 有较大幅度的提升,从而使得系统整体性能有明显提升. 其中,该系统在包含 65% 左右的测试集情况下,系统取得的性能与 COAE 2014 前 3 名的系统具有很高的可比性,在包含 100% 的测试集情况下,整个系统的检测性能有显著提升.

表 3 与现有最新系统性能对比

对比实验	系统	P^+	R^+	$F1^+$	P^-	R^-	$F1^-$	Macro_P	Macro_R	Macro_F1	Micro_P	Micro_R	Micro_F1
特征选择前	20% 测试集	0.890	0.168	0.280	0.912	0.195	0.321	0.901	0.181	0.302	0.901	0.180	0.300
	50% 测试集	0.891	0.419	0.570	0.866	0.462	0.603	0.879	0.441	0.587	0.879	0.439	0.586
	65% 测试集	0.889	0.544	0.675	0.810	0.562	0.664	0.850	0.553	0.670	0.850	0.552	0.670
	80% 测试集	0.866	0.652	0.744	0.779	0.664	0.717	0.822	0.658	0.731	0.822	0.658	0.731
	100% 测试集	0.817	0.762	0.788	0.749	0.806	0.777	0.783	0.784	0.783	0.783	0.783	0.783
特征选择后	20% 测试集	0.932	0.175	0.295	0.946	0.202	0.333	0.939	0.189	0.314	0.939	0.188	0.313
	50% 测试集	0.905	0.426	0.579	0.885	0.472	0.615	0.895	0.449	0.598	0.895	0.447	0.597
	65% 测试集	0.888	0.543	0.674	0.841	0.583	0.689	0.865	0.563	0.682	0.865	0.562	0.681
	80% 测试集	0.876	0.66	0.753	0.793	0.677	0.730	0.835	0.668	0.740	0.835	0.668	0.742
	100% 测试集	0.844	0.757	0.798	0.753	0.842	0.795	0.799	0.799	0.799	0.797	0.797	0.797
COAE 2014 现有系统	MI&TLAB_run2(第 1 名)	0.918	0.586	0.715	0.886	0.500	0.639	0.902	0.543	0.677	0.904	0.547	0.681
	hit_run5(第 2 名)	0.873	0.574	0.692	0.887	0.500	0.640	0.880	0.537	0.666	0.879	0.540	0.669
	COAE2014_sjtu(第 3 名)	0.954	0.429	0.592	0.885	0.630	0.736	0.919	0.530	0.664	0.914	0.522	0.664

注:采用 NLP&CC 2013 作为训练集,COAE 2014 作为测试集.

4.2.2 其它 NLP&CC 与 COAE 语料双向交叉实验 由于信息增益特征选择技术能够使得系统性能有所提升,于是,采用特征选择过后的 11 个特征集进行接下来的实验. 表 4 列举了采用 100% 测试集环境下的 COAE 和 NLP&CC 双向交叉实验结果. 限于篇幅,仅报告系统的 Accuracy,不分别报告正性和负性情感的 Precision、Recall、Marco_F1 和 Micro_F1. 由于未取得 COAE 2013 的测试数据集,于是仅利用 COAE 2013 的训练数据集作为双向交叉实验语料进行试验.

(i) 针对同类型的语料而言,检测性能相对较稳定,如采用 COAE 2013 和 COAE 2014 分别作为训练和测试语料下的检测性能相差不大. NLP&CC 上的实验结果有所差别的原因在于 NLP&CC 2012 的正性和负性情感严重不平衡(正负比例达 1:4),造成检测检测性能的偏斜,当将 NLP&CC 2012 进行过采样(将正性情感扩充至负性情感数目)和欠采样(将负性情感压缩至正性情感数目)技术后,并把采样过后的语料作为训练数据,此时系统的整体性能有较大幅度的提升,如利用 NLP&CC 2012 作为训练语料和 COAE 2013 作为测试语料时,系统的 Accuracy 为 52.47%;但将 NLP&CC 2012 语料进行欠采样和

过采样处理后,系统的 accuracy 分别达到 75.24% 和 74.10%.

(ii) 针对不同类型的语料而言,系统性能体现出较大的差异,例如: 采用 NLP&CC 2013 作为训练集,COAE 2013 和 COAE 2014 作为测试集时的性能较好,但相反却比较差,原因在于 NLP&CC 2013 语料相对较大,一般来说采用相对较大的语料作为训练集能够取得相对较好的检测性能. 同理,NLP&CC 2012 作为训练集,COAE 2014 作为测试集时的性能较差,但相反却比较好,原因在于 NLP&CC 2012 语料的规模比较小,而 COAE 2014 规模比较大.

(iii) 因为没有取得 COAE 2013 的测试数据集,所以没有与 COAE 2013 评测结果进行直接对比. 但是,根据 COAE 2013 的评测论文集的结果看来,前 3 名的评测性能的 F1 值在 33.00% 左右,而且他们均对原始的情感词典进行了不同程度的扩充工作,从这一点看来,本文的系统仅依赖于原始的情感词典,降低了系统的复杂程度,具有较高的可比性.

表 4　双向交叉语料实验结果

对比语料	测试语料	训练语料	Accuracy/%
未做语料平衡处理	COAE2014	NLP&CC2013	79.72
	NLP&CC2013	COAE2014	61.11
	COAE2013	NLP&CC2013	77.90
	NLP&CC2013	COAE2013	60.17
	COAE2014	NLP&CC2012	46.88
	NLP&CC2012	COAE2014	78.79
	COAE2013	NLP&CC2012	52.47
	NLP&CC2012	COAE2013	79.02
	COAE2013	COAE2014	77.70
	COAE2014	COAE2013	78.68
	NLP&CC2013	NLP&CC2012	50.00
	NLP&CC2012	NLP&CC2013	80.12
已做语料平衡处理	COAE2014	NLP&CC2012-欠采样	77.20
	NLP&CC2012-欠采样	COAE2014	69.90
	COAE2013	NLP&CC2012-欠采样	75.24
	NLP&CC2012-欠采样	COAE2013	70.27
	NLP&CC2013	NLP&CC2012-欠采样	60.91
	NLP&CC2012-欠采样	NLP&CC2013	70.63
	COAE2014	NLP&CC2012-过采样	75.76
	NLP&CC2012-过采样	COAE2014	67.58
	COAE2013	NLP&CC2012-过采样	74.10
	NLP&CC2012-过采样	COAE2013	68.23
	NLP&CC2013	NLP&CC2012-过采样	59.76
	NLP&CC2012-过采样	NLP&CC2013	68.12

4.2.3　NLP&CC 与 COAE 独立语料实验　表 5 列举了 NLP&CC 和 COAE 独立语料实验结果,可以明确中文微博句子倾向性分类在 COAE 语料上的实验结果相对稳定,原因在于 COAE 2014 和 COAE 2013 的任务比较相近,两者都有对中文微博进行粗粒度的情感分析,即正性、负性情感. 相比较而言,在 NLP&CC 语料上的实验结果相差较大,原因在于 NLP&CC 2013 属于细粒度的情感分析,是一种情绪分析(兴奋、快乐、忧愁、悲伤),经过细粒度至粗粒度的转换,即将“兴奋”和“快乐”作为正例,“忧愁”和“悲伤”作为负例,导致系统的检测性能有所降低. 相反,NLP&CC 2012 属于粗粒度的情感分析,可以不经过转换直接用于训练或测试. 此外,在 NLP&CC 2012、COAE 2013 和 COAE 2014 上的实验结果较相近,同样验证了任务和语料相近可以得到较一致的检测结果.

表 5　独立语料实验结果

语料	Accuracy/%
COAE 2014	79.92
COAE 2013	78.32
NLP&CC 2013	60.92
NLP&CC 2012	81.22

5　结语

本文系统地研究了中文微博句子倾向性分类问题,在充分降低情感词典扩充工作带来的系统开销基础上,利用原始(未扩展)的情感词典,抽取了中

文微博句子中标点符号、情感词权重、词汇级和句法级等新型平面和结构化特征,同时探索了有效的特征选择方法. 并且在基准 COAE 和 NLP&CC 微博语料上分别进行双向交叉和独立实验,结合有效的不平衡语料处理方法. 实验结果表明采用上述特征,中文微博句子倾向性分类的性能得到显著提升.

对于将来工作,一方面将利用网络新词、How-Net 和《同义词词林》对基础情感词典进行扩充,以验证扩充的情感词典对系统性能的影响; 另一方面将探索更为有效的语言学方面的特征及相应的特征选择方法.

6 参考文献

[1] 陈晨,王厚峰. 中文跨文本人名同名同指消解研究[J]. 江西师范大学学报: 自然科学版,2015,39(2): 111-116.

[2] 钱鹏,黄萱菁. 中国古诗统计建模与宏观分析[J]. 江西师范大学学报: 自然科学版,2015,39(2): 117-123.

[3] 周胜臣,瞿文婷,石英子,等. 中文微博情感分析研究综述[J]. 计算机应用与软件,2013,30(3): 161-164.

[4] 张艳辉,杜文韬,刘培玉,等. 基于词典的微博的倾向性分析[C]. 第 5 届中文倾向性分析评测研讨会,2013: 50-52.

[5] 李岩,徐蔚然,陈光. PRIS_COAE COAE2013 评测报告[C]. 第 5 届中文倾向性分析评测研讨会,2013: 53-69.

[6] 王文远,王大玲,冯时,等. 一种面向情感分析的微博表情情感词典构建及应用[J]. 计算机与数字工程,2012,40(11): 6-9.

[7] 王勇,吕学强,姬连春,等. 基于极性词典的中文微博客情感分类[J]. 计算机应用与软件,2014,31(1): 34-37.

[8] 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.

[9] 罗凌,陈毅东,曹茂元. 微博观点句识别的话题影响研究[J]. 电脑知识与技术,2014,10(1): 123-127.

[10] 戴敏,庞磊,李寿山. 基于机器学习方法的中文微博情感分类方法研究[J]. 语文研究与创作,2011(15): 1-13.

[11] 杜振雷,张仰森,李文坤,等. 基于多特征融合的中文微博情感分类方法研究[C]. 第 5 届中文倾向性分析评测研讨会,2013: 44-49.

[12] 刘志广,董喜双,关毅. 中文微博情感倾向性研究[C]. 第 5 届中文倾向性分析评测研讨会,2013: 81-87.

[13] 朱艳辉,杜锐,鲁琳,等. 中文文本情感分析与比较句的识别研究[C]. 第 5 届中文倾向性分析评测研讨会,2013: 34-43.

[14] 徐琳宏,林鸿飞,潘宇,等. 情感词汇本体的构造[J]. 情报学报,2008,27(2): 180-185.

[15] Caruana R,Freitag D. Greedy attribute selection[C]. Proceedings of the 11th International Conference on Machine Learning,1994: 28-36.

[16] Li Shoushan,Xia Rui,Zong Chengqing,et al. A framework of feature selection methods for text categorization[C]. Proceedings of the ACL-IJCNLP,2009: 692-700.

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

The Research on Feature Extraction of Polarity Classification of Chinese Micro Blogging

XU Xiongfei,XU Fan,WANG Mingwen*,ZUO Jiali,LUO Wenbing

(School of Computer Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: According to Chinese micro blogging sentence polarity identification problem,while fully reducing due to the emotional lexicon expansion work brought on the basis of system overhead,many novel flat and structural features,e. g. punctuation,sentiment word weighting,lexical and syntactic level information,from Chinese micro blogging,together with the effective feature selection method has been extracted. In-depth bidirectional and independent experiments on both COAE and NLP&CC,along with the effective imbalance corpus handling method has been conducted. Evaluation results show that the effectiveness of our novel features. Its also show that the model significantly outperforms existing model currently in the research field.

Key words: Chinese micro blogging; sentence polarity; feature extraction; classification

(责任编辑: 冉小晓)