

基于极性词典的中文微博客情感分类

王 勇¹ 吕学强¹ 姬连春² 肖诗斌¹

¹(北京信息科技大学网络文化与数字传播北京市重点实验室 北京 100101)

²(新华网络股份有限公司 北京 100101)

摘 要 微博客是近年来自然语言处理领域研究的热点。主要针对中文微博客中的情感分类展开研究。结合网络新词和基础情感词,同时考虑了情感词的极性情感强弱,构建四个词典,分别是基础情感词典、表情符号词典、否定词词典和双重否定词词典;在情感词典的基础上,融合汉语语言学特征和微博情感表达特征,提出一种新的基于极性词典的情感分类方法。实验准确率达到 82.2%。实验结果表明,提出的方法可以对中文微博进行较好的情感分类,有一定的应用价值。

关键词 微博客 情感分类 词典 语言学特征

中图分类号 TP391

文献标识码 A

DOI:10.3969/j.issn.1000-386x.2014.01.010

SENTIMENT CLASSIFICATION FOR CHINESE MICROBLOGGING BASED ON POLARITY LEXICONS

Wang Yong¹ Lü Xueqiang¹ Ji Lianchun² Xiao Shibin¹

¹(Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China)

²(Xinhua Net Co., Ltd., Beijing 100101, China)

Abstract Microblogging is the focus in research field of natural language processing recently. Our study in this paper is mainly in regard to the sentiment classification of Chinese microblog. In combination with new Internet words and basic emotional words and taking into account the strength of the polarity of emotions, we construct four lexicons, they are: the basic sentiments lexicon, emotional signs lexicon, negative words lexicon and double negative words lexicon respectively. On the basis of sentiments lexicon and fused in Chinese linguistic features and the sentiment expression features in microblogging, we propose a new sentiment classification method based on polarity lexicons. The precision in the experiments reaches 82.2%. Experimental result indicates that the method proposed in the paper can conduct the sentiment classification on Chinese microblog well, and has certain applied value.

Keywords Microblogging Sentiment classification Lexicons Linguistics features

0 引 言

互联网的兴起,特别是 Web2.0 时代的到来,使网民不再只是互联网的“消费者”,也成为了互联网的“生产者”。随着 Web 应用的增多,用户产生内容也呈爆炸式的增长,人们越来越多的在论坛、BBS、博客和微博等应用上表达自己的情感。其中,微博作为新生代应用的佼佼者,在近几年来取得了巨大的发展。

微博客,是一种通过关注机制分享简短实时信息的广播式的社交网络平台。用户注册微博服务后,可以关注自己感兴趣的人而不需对方的权限验证,同时,用户也可以实时地更新自己的状态或发表自己的观点。由于便捷性和草根性,微博的用户遍布世界各地和各个阶层,通过微博平台,用户可以对某种商品、某个电影或电视剧、某个热点事件等进行评论,发表自己的看法。其中关于商品的评论分析,对于生产厂商和观望该商品的潜在用户具有重要的参考价值;关于电影或者电视剧的评论分析,对于制片商和观众有重要的实际价值;关于某个热点事件的评论分析,对于国家对舆论的监督也具有重大的意义。因此,针对微博评论分析的研究,尤其是微博评论的情感分类研究,是

当前研究的一个热点。

文本的情感分类,就是按照文本表达的情感倾向性进行分类^[1]。当前文本情感分类的研究主要将微博情感分为两类(正向情感类、负向情感类)和三类(正向情感类、负向情感类、中性情感类)^[2]。微博的情感分类属于文本情感分类的一个分支,本文主要针对微博的情感分类进行研究,将微博情感分为三类。

1 相关研究

目前,国内外已经有很多关于文本情感分类的研究。总的来说,情感分类方法可以分为基于机器学习的方法和基于情感词典或者知识系统的方法。其中,基于机器学习的方法主要有朴素贝叶斯方法、最大熵方法和支持向量机方法等^[3-7]。文献[3]针对新闻文本的分类进行研究,分别利用朴素贝叶斯方法

收稿日期:2012-09-28。国家自然科学基金项目(61271304);国家科技支撑计划课题(2011BAH11B03);北京市教委科技发展计划项目(KM201211232023)。王勇,硕士生,主研领域:自然语言处理。吕学强,教授。姬连春,高工。肖诗斌,副教授。

和最大熵方法将新闻文本分为正面情感类和负面情感类,并采用词频和二值作为特征项权重,最终取得了较好的分类效果,最高分类准确率达到 90% 以上。文献[4]认为不同的领域或者不同的特征需要不一样的分类方法才能取得最好的分类效果,组合朴素贝叶斯、最大熵、支持向量机和随机梯度下降现行分类方法进行文本情感分类,结果证明组合后的分类结果优于单个分类的最优结果。基于情感词典或知识系统的方法利用已有的语义词典或知识系统建立初始词典,采用一定的方法来扩展词典^[8-10]。

关于微博的情感分类研究,由于 Twitter 的用户量大并且知名度高,目前国外学者主要对 Twitter 进行微博情感分类研究^[11-16]。而在国内,微博情感分析研究还处于起步阶段,主要研究的对象是新浪、腾讯和网易微博等^[5,7,9]。

文献[11]基于四个基础词典开发了一个 Twitter 情感分类系统,使用四个词典分别对微博进行情感值计算,最后使用四个结果的加权值作为该条微博的情感分析结果。文献[13]利用 3 个 Twitter 的情感分析网站 Twendz、Twitter Sentiment 和 TweetFeel 中已分好类的数据作为数据源,利用分类偏见信息和噪声标记进行建模,并且结合微博的抽象特征进行情感分类,结果相对于传统的结果有一定程度的提高。文献[5]提出了一种基于 SVM 的层次结构的多策略方法,即“一步三分类”方法,对微博不进行分句处理,直接选取微博中的极性特征进行 SVM 模型训练,然后基于模型训练的结果进行情感分类,由于微博的内容简短性和网络用语的不规范性,使得文中的方法对于微博的情感分析识别率较低。文献[9]提出了一种基于情感词典和规则的方法,构建了情感词权重词典(WD)、否定词词典(NWD)、程度副词词典(DWD)和感叹词词典(IWD),分别给词典中的词分配一个权重,然后基于每条微博的情感权重值进行情感分类,文中虽然考虑了否定词的作用,但是并没有考虑否定词出现的特征,并且文中并没有考虑双重否定在中文情感表达中的作用。

针对微博的内容简短性和语句表达的不规范性,本文提出了一种基于极性词典的中文微博客情感分类方法,首先构建基础情感词典、表情符号词典、否定词典和双重否定词典,然后基于微博的特性进行基础情感词典扩展,最后使用基于极性词典的情感分类算法对微博进行情感分类。

2 情感极性词典构建

极性词典是文本情感分析和倾向性分析的基础。传统的文本主要通过情感词来表达情感倾向性,和传统的文本不同,微博中还包含着大量的口语化用词和表情符号。因此,本文构建了一组比较全面、高效的情感极性词典,包括基础情感字典和表情符号字典,每个词典中的词按正负极性强度分为四类。分别为弱正向情感 EmotionA、强正向情感 EmotionB、弱负向情感 EmotionC 和强负向情感 EmotionD。同时,考虑到否定词和双重否定词在中文情感表达中的作用,本文还建立了否定词典和双重否定词典。

2.1 基础情感词典构建

网络情感表达涉及到各个领域的内容,每个领域的一些特征有所不同,因此有些词在不同领域表达的情感也不同,如“这个喇叭的声音很大”和“这个冰箱的声音很大”,前者所表述的意思是喇叭性能的优点,即发出的声音大,而后者想要表述的是冰箱的缺点,即噪音大。总的来说,大部分的极性词没有领域差

异,构建一个跨领域的基础情感词典是非常重要的。构建的基础词典主要利用了《知网》^[17]提供的情感词集。由于微博内容比较短,因此微博中情感词的出现频次相对较少,同时缺乏上下文的相关信息,导致情感分析时对于情感词的极性比较敏感。本文只选用了其中的部分中文正向情感词和中文负向情感词,建立了基础情感词典(BL),并把正向情感词分为 EmotionA 和 EmotionB;把负向情感词分为 EmotionC 和 EmotionD。

2.2 基础情感词典扩展

《知网》中的情感词典对于传统的情感词分析比较实用,但是由于微博中的用语不规范以及大量网络用词的出现,基于传统的情感词的微博情感分析往往不能取得较好的效果。由此,本文对 2.1 节构造的基础情感词典进行扩展。

为了能够比较完善地扩展基础情感词典并考虑网络用语和相关领域词,本文从当前比较流行的四大中文微博服务网站新浪、腾讯、网易和搜狐微博中随机抽取约 100 000 条微博,然后人工从中抽取当前常用的并且含有情感极性的网络词,如“腹黑”、“尼玛”、“伤不起”、“稀饭”等等;同时,抽取出常用的含有情感极性的领域用词,如篮球领域中的“抱大腿”、“抱团”、“伪球迷”,影视领域中的“狗血”、“烂片”、“泡沫剧”等。

通过人工标注并多次校对,将情感词极性分别分为强性情感和弱性情感,加入到基础情感词典中。经统计,词典一共包含 2 199 个情感词,其中弱正向情感词 342 个,强正向情感词 345 个,弱负向情感词 848 个,强负向情感词 664 个。基础情感词典各情感词集的总数分布如表 1 所示。

表 1 基础情感词典分布

{ EmotionA }	{ EmotionB }	{ EmotionC }	{ EmotionD }	总计
342	345	848	664	2 199

其中,{ EmotionA }表示弱正向情感词集,{ EmotionB }表示强正向情感词集,{ EmotionC }表示弱负向情感词集,{ EmotionD }表示强负向情感词集。

2.3 表情符号词典


表情是一种比语言更直观的个人情感表达方式,微博服务提供的表情符号具有一目了然的特点,微博用户可以直接使用表情符号来表达自己的喜悦、悲伤、愤怒和失望等情感。包含表情符号是微博内容的一大特征,因此建立表情符号词典是很有意义的(虽然在一些情况下,例如讽刺的句子中,基于表情符号的情感分析会失效,但是在大多数情况下,还是有效的)。本文分析了新浪微博中表情符号的极性,筛选明确表达个人情感的表情符号,构建了微博表情符号词典(EL)。新浪微博表情用文本表示为“[…]”的形式,如的文本表示为“[偷笑]”。新浪微博表情符号词典的组成如表 2 所示。

表 2 微博表情符号词典

表情符号词典		
正向情感词典	{ EmotionA }	[din 脸红]、[hold 住]、[不好意思]、[爱心传递]…
	{ EmotionB }	[din 兴奋]、[din 鬼脸]、[cai 晃头]、[cai 开心]…
负向情感词典	{ EmotionC }	[伤心]、[感冒]、[生病]、[失望]、[cai 插眼]、[bed 凌乱]…
	{ EmotionD }	[抓狂]、[鄙视]、[怒骂]、[怒]、[吐]、[杀死你]…

其中, $\{EmotionA\}$ 的个数为 10, $\{EmotionB\}$ 的个数为 43, $\{EmotionC\}$ 的个数为 33, $\{EmotionD\}$ 的个数为 6。

2.4 否定词典和双重否定词典

使用否定词和双重否定词是汉语语言的特色, 微博内容中的用语也不例外。否定词使得词的情感极性发生改变, 双重否定词不改变情感极性, 但情感语气有加强的作用。否定词典 (NL) 和双重否定词典 (DNL) 如表 3 所示。

表 3 否定词典和双重否定词典

否定词典 和 双重否定词典	
否定词典	不可以、怎么不、几乎不、从来不、从不、不用、不曾、不必、不会、很少、极少、没有、不是、难以、放下、终止、停止、放弃、反对、不、甬、勿、别、未、反、没、否
双重否定词典	绝非不、并非不、不是不、不能不、不会不、不可不、不要不、不得不、没有不、无不、不无

3 基于极性词典的微博情感分类

定义 1

$W_{EmotionA} = 1$ $W_{EmotionB} = 2$ $W_{EmotionC} = -1$ $W_{EmotionD} = -2$
其中 $W_{EmotionA}$ 、 $W_{EmotionB}$ 、 $W_{EmotionC}$ 和 $W_{EmotionD}$ 分别表示 EmotionA、EmotionB、EmotionC 和 EmotionD 中单个情感词的权重为 1、2、-1 和 -2。

在第 2 节中构建的词典的基础上, 本文融合汉语语言学特征和微博情感表达特征, 将否定词和双重否定词引入到微博情感词权重计算中; 同时, 考虑到对于含有多个分句的微博, 提出了一种新的基于极性词典的中文微博情感分类方法。

3.1 否定和双重权重计算

定义 2

情感词块: 对于微博中的情感词 $Emotion_i$, 将 $Emotion_i$ 及其左邻近三窗口组成的单元称为一个情感词块, 记为 $Block_Emotion_i$ 。

使用否定和双重否定是常见的汉语语言学现象, 其中在一个情感词之前使用否定表明了对当前词义的否定, 而双重否定表示肯定, 并且有加强语气的作用。否定词和双重否定词对微博的情感极性有很大的作用, 因此, 本文提出了一种基于否定词和双重否定词的情感词块 ($Block_Emotion$) 权重计算方法为:

$$W(Block_Emotion_i) = \begin{cases} f(Emotion_i) & \text{condition1} \\ f'(Emotion_i) & \text{condition2} \\ W_{Emotion_i} & \text{otherwise} \end{cases} \quad (1)$$

$$f(Emotion_i) = \begin{cases} W_{Emotion_i \times \lambda_1} & Emotion_i \in \{EmotionA\} \cup \{EmotionC\} \\ W_{Emotion_i \times \lambda_2} & Emotion_i \in \{EmotionB\} \cup \{EmotionD\} \end{cases} \quad (2)$$

$$f'(Emotion_i) = W_{Emotion_i} \times (-1) \quad (3)$$

其中, $W(Block_Emotion_i)$ 表示情感词块 $Block_Emotion_i$ 在微博中的情感权重值; condition1 表示微博中的情感词块 $Block_Emotion_i$ 中包含双重否定词典中的任意一个词; condition2 表示微博中的情感词块 $Block_Emotion_i$ 中包含否定词典中的任意一个词; $W_{Emotion_i}$ 的值如定义 1 所示; λ_1 和 λ_2 为双重否定词对不同极性强度情感词的影响因子。

3.2 微博分句权重计算

当前微博情感分析主要有分句的方法和不分句的方法^[5], 本文采用分句的方法对微博进行情感权重计算。对于微博中的

每一个分句 S_i , S_i 的情感权重等于分句中所有的情感词块权重和表情符号权重的总和。微博分句权重计算方法为:

$$W_{S_i} = W_{expression} + W_{basic} \quad (4)$$

$$W_{expression} = \sum_{i=1}^n W_{Emotion_i} \quad (5)$$

$$W_{basic} = \sum_{i=1}^m W(Block_Emotion_i) \quad (6)$$

其中, $W_{expression}$ 表示当前句子的所有表情符号的权重之和, n 表示句子中包含表情符号的个数, 因为表情符号很少涉及否定和双重否定的语法特征, 所以只将表情符号权重相加。 W_{basic} 表示当前句子的基础极性词典的权重之和, m 表示句子中包含的基础情感词的个数, 基础词情感的权重在微博分句中涉及到否定词和双重否定词的作用, 使用式 (1) 计算其权值。

感叹句有加强该语句的情感语气的作用, 本文认为, 微博中的一个句子, 如果以感叹号结尾, 则该句子的情感值 W_{S_i} 加倍, 即 $W_{S_i} = W_{S_i} \times 2$ 。因此, 分句的最终情感权值计算为:

$$W_{S_i} = \begin{cases} 2 \times W_{S_i} & S_i \text{ contains "!"} \\ W_{S_i} & \text{otherwise} \end{cases} \quad (7)$$

3.3 微博权重计算

在一段传统的文本段落中, 一般首先在首句点名主旨, 然后中间部分陈述和主题相关的部分内容, 最后结尾作总结, 即所谓的“首尾呼应”。微博和传统文本有很大的不同, 在相对较长的微博中 (含两个以上的分句), “主题漂移”现象比较严重, 多个主题的评论存在于一条微博中, 一般最后的一句和作者所要表达的情感相关。

例如微博“雷霆不争气, 终究还是太嫩了, 每每到关键时刻处理球的时候就显得很幼稚, 三个少爷都是。杜兰特缺少主宰比赛的霸气, 不然的话他不会让韦少这么独, 哈登到了总决赛怎么‘格登’一声说不见就不见了呢? 难道天意让詹姆斯夺冠。好吧, 祝贺他, 虽然他不是那么让人喜欢!”。

这条微博一共包含四句话, 第一句和第二句都属于批评的评论, 属于负向情感句, 第三句是中性的评价, 第四句是属于正向情感句, 并且第四句和主题“詹姆斯夺冠”相关。

文献[9]经过统计, 结果表明在 50% 以上的微博中, 最后一个分句 (尾句) 的情感极性能代表整条微博的情感极性。从“最后一句情感表达最接近作者的思想”这一个角度出发, 本文认为一条微博的最后一句最能体现该条微博的情感。因此, 提出了一种基于尾句优先的微博情感权重计算算法:

$$W_{tweet} = \begin{cases} W_{S_n} & |W_{S_n}| > 0 \\ \sum_{i=1}^{n-1} W_{S_i} & \text{otherwise} \end{cases} \quad (8)$$

其中 n 表示微博中的分句数, 当微博最后分句 S_n 的情感权值不为 0 时, 将最后一句的情感权值作为微博 tweet 的情感权值; 否则, 微博的情感权值等于前 $n-1$ 句分句权值之和。

$|W_{tweet}|$ 越大, 表明情感越强烈, 反之越弱。微博的情感极性计算为:

$$E_{tweet} = \begin{cases} +1 & W_{tweet} > 0 \\ 0 & W_{tweet} = 0 \\ -1 & W_{tweet} < 0 \end{cases} \quad (9)$$

其中, “+1”表示当前微博是正向情感, “0”表示当前微博是中

性情感,“-1”表示当前微博是负向情感。

4 实验结果及分析

4.1 实验设置

本文使用爬虫工具抓取了新浪微博中三大类共六个话题相关的数据,分别是节日类:“端午节”、“中元节(鬼节)”;名人事件类:“詹姆斯夺冠”、“俞灏明复出”;影视剧类:“天涯明月刀”、“轩辕剑”。各话题中消息的情感分布情况如表 4 所示。

表 4 话题消息的情感分布

话题	正向情感条数	负向情感条数	中性情感条数	总条数
端午节	418	333	249	1 000
中元节	246	377	377	1 000
詹姆斯夺冠	718	169	113	1 000
俞灏明复出	597	280	123	1 000
天涯明月刀	396	462	142	1 000
轩辕剑	382	310	308	1 000
总计	2 757	1 931	1 312	6 000

同时,3.1 节中的双重否定的两个影响因子 λ_1 和 λ_2 分别设置为 2 和 1.5。

4.2 实验对比及分析

4.2.1 与传统的基于字典的方法比较

文献[9]使用基于极性词典的方法对中文微博进行情感分析,仅考虑了否定词在情感分类中的作用,并没有考虑双重否定词的作用;同时,文中发现了中文微博中不同位置的分句对微博情感表达的作用,使用简单的线性加权计算微博的情感值。本文分别使用未考虑双重否定词(EABL-DN)和未考虑分句位置的基于情感词典分析方法(EABL-SP)作对比实验与文中提出的基于极性词典的中文微博客情感分析方法(EABL)作对比,评测指标使用准确率。实验结果如表 5 所示。

表 5 三种方法实验结果对比

方法	清明节	中元节	詹姆斯 夺冠	俞灏明 复出	天涯 明月刀	轩辕剑	总计
EABL-DN	79.1%	76.3%	77.6%	76.7%	81.8%	77.2%	78.16%
EABL-SP	80.8%	75.5%	79.7%	75.2%	81.1%	76.4%	78.16%
EABL	81.1%	76.5%	79.1%	78.4%	82.2%	77%	79.05%

由表 5 中可知,相比于未考虑双重否定词和未考虑分句位置的基于情感词典分析方法,本文提出的方法在六个话题下,情感分类的准确率都有所提升。结果证明考虑微博情感表达中双重否定词和分句的不同位置的权重,能够提高微博情感分类的准确率。

4.2.2 与支持向量机和集中特征选取的方法比较

文献[4,5,7]通过实验证明支持向量机(SVM)方法相对于其他机器学习模型,能够获得较好的情感识别准确率。因此,本文使用 SVM 方法和文中提出的基于极性词典的中文微博客情感分析方法(EABL)作对比,评测方法选用十折交叉验证法,评测指标使用的是准确率。

选用四个特征作为 SVM 的分类特征,特征描述如表 6 所示。

表 6 基于 SVM 的特征描述

特征号	特征类型	特征内容
1	不分句的情感极性	正向情感词数、负向情感词数
2	不分句的情感极性	正、负向情感词数,否定、双重否定词数
3	尾句的情感极性	尾句正向情感词数、负向情感词数
4	尾句的情感极性	尾句正、负向情感词数,否定、双重否定词数

注:表 6 中的正向情感词数和负向情感词数分别包含正向表情符号数和负向表情符号数。

基于特征 1、2、3 和 4 的 SVM 分类方法分别称为 SVM_1、SVM_2、SVM_3 和 SVM_4。本文的方法和四种基于 SVM 的方法在六个话题下的情感分类实验结果如表 7 所示。

由表 7 可知,本文中提出的方法在准确率上都优于其他的方法,在六个话题下微博情感分类的最高准确率为 82.2%,平均准确率为 79.05%。其中,尤其在节日类话题和影视剧话题类下,本文的方法有明显的优势;在名人事件(“詹姆斯夺冠”和“天涯明月刀”)中,EABL 方法准确率也略高于其他四种方法。

表 7 五种方法的实验结果对比

话题	SVM_1	SVM_2	SVM_3	SVM_4	EABL
清明节	59.4%	61.4%	55.1%	56.1%	81.1%
中元节	53.8%	57.3%	50.0%	52.9%	76.5%
詹姆斯夺冠	76.0%	75.1%	74.4%	73.6%	79.1%
俞灏明复出	75.4%	74.8%	73.5%	71.0%	78.4%
天涯明月刀	69.7%	71.2%	66.8%	68.0%	82.2%
轩辕剑	53.2%	59.2%	50.7%	52.5%	77%
总计	64.6%	66.5%	61.75%	62.35%	79.05%

4.3 错误分析

由于汉语用语比较灵活,因此基于词典的情感分析方法不可能解决所有的问题,下面是本文方法的常见错误及分析。

(1) 反讽语气的语句

如微博“韵达快递真是太极品了,端午节前一天我的快递就到南宁了,结果端午节放假不送,我忍了。第二天莫名其妙又没送,今天我打电话去问,他说今天一定送到,结果下午还没到,我又电话他,他说,下雨不送,最后要我自己去他们站点领!!太极品了”。显然,本条微博是采用一种讽刺的方法来表达对快递公司的不满,这是一个纯粹的负向情感评论。然而,根据微博中出现的情感词“极品”,本文将其分为正向情感类,这是由讽刺语气导致的错误。

(2) 领域情感词

如微博“为了庆祝热火夺冠,中午做了一大碗螃蟹,看来我果然是一个不折不扣的‘詹黑’”。在体育领域,尤其是篮球领域中,“**黑”与“**密”相反,前者是指专门攻击某个球星的球迷,而后者表示是某个球星的支持者(“粉丝”)。由于情感词典中只有“庆祝”这个情感性,并没有“詹黑”,所以这条微博被分为正向情感类。事实上,它应该被分为负向情感类。

(3) 网络新词

本文中的词典虽然考虑了部分网络新词,但是也不可能包含所有的网络新词。如“#天涯明月刀#昨天的剧情甚是虐身虐心啊,期待今晚,小红啊,你肿么办啊,块好吧”,根据情感词“期待”,本文将其分为正向情感类;然而考虑到网络新词“虐身”和“虐心”,微博应该被分为负向情感类。

ZigBee 无线定位网络,利用收集到的 RSSI 信号,分别进行传统质心算法、加权质心算法和修正加权质心算法的 Matlab 仿真。仿真结果显示加权质心算法较大幅度地提高了定位精度,与传统的质心算法相比精度提高了 0.68 米,与加权质心算法相比提高了 0.52 米,定位精度为 3.5 米左右。且不需要改进硬件设备,满足无线传感器网络网络定位的基本要求,可以实现区域定位,有一定的实用价值。

参 考 文 献

- [1] 陈维克,李文锋,首珩,等.基于 RSSI 的无线传感器网络加权质心定位算法[J].武汉理工大学学报,2006,20(12):2695-2700.
- [2] 李文忠,段朝玉,等. Zigbee 2006 无线网络与无线定位实战[M].北京:北京航空航天大学出版社,2008.
- [3] 彭渤.基于 RSSI 测距误差补偿的无线传感器网络定位算法研究[D].大连:大连理工大学,2008:28-30.
- [4] 邵丽鹏,朱冬梅,杨丹.基于 Zigbee 的加权质心定位算法的仿真与实现[J].传感技术学报,2010,23(1):149-152.
- [5] 李瑶怡,赫晓星,刘守印.基于路径损耗模型参数实时估计的无线定位方法[J].传感技术学报,2010,23(9):1328-1333.
- [6] Daiya V. Experimental Analysis of RSSI for Distance and Position Estimation[C]//IEEE International Conference on Recent Trends in Information Technology, ICRTIT 2011 MIT, Anna University, Chennai, 2011.
- [7] Priwgharm R. A Comparative Study on Indoor Localization based on RSSI Measurement in Wireless Sensor Network[C]//2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE), 2011.
- [8] 任维政,徐连明.基于 RSSI 的测距差分修正定位算法[J].传感技术学报,2008(7):1247-1250.
- [9] Heurtefeux K. Is RSSI a good choice for localization in Wireless Sensor Network? [C]//2012 26th IEEE International Conference on Advanced Information Networking and Applications, 2012.
- [10] 周艳,李海成.基于 RSSI 无线传感器网络空间定位算法[J].通信学报,2009,30(6):75-79.

(上接第 37 页)

(4) 陈述事实

有些微博本身是对事实的陈述,即属于中性情感,如“詹姆斯夺冠主要靠这套阵容,热火其实是惧怕有强力内线的球队,或者是内线防守强的球队,而雷霆内线防守的整体性太差,帕金斯太慢,18 卡防守太注重盖帽。”因为微博中出现“惧怕”这个负向情感词,所以被错误地分为负向情感类。

(5) 无情感词

有些微博中虽然没有出现情感词,但是从一些事件的描述中可以区分它的情感极性。如“詹姆斯夺冠我再也不相信爱情了,我要绝食三天了…”。由于该条微博中并没有出现情感词典中的情感词,因此被归为中性情感类。然而,从“不相信爱情”和“绝食三天”可以分析出,这应该分为负向情感类。

5 结 语

本文通过分析中文微博客中情感的表达特点,构建了四个词典,分别是基础情感词典、表情符号词典、否定词典和双重否定词典;同时,考虑到否定词和双重否定词在情感表达中的作用,将否定词和双重否定词引入到情感分析权重计算中。实验证明,本文的方法在中文微博客情感分析上取得了比较好的效

果。和基于机器学习的方法 SVM 方法相对比,本文提出的方法能获得更高的准确率。然而根据分类错误分析,本文还有许多可以改进的地方。

第一,研究情感词典的自动扩建方法,能够自动识别网络新词和领域相关情感词汇,并加入到情感词典中,更好地弥补手工情感词表中领域情感词和网络新词覆盖的不足。

第二,融合社交网络中微博用户“关注”和“被关注”关系或者微博消息的转发和回复关系特征,构建图模型,对本文规则进行扩展。

总之,中文微博的情感分析是一个比较热的研究领域,还有很多方法需要进行深入的研究。

参 考 文 献

- [1] 陆文星,王燕飞.中文文本情感分析研究综述[J].计算机应用研究,2012,29(6):2014-2017.
- [2] 赵妍妍,秦兵,刘挺.文本情感分析综述[J].软件学报,2010,21(8):1834-1848.
- [3] 徐军,丁宇新,王晓龙.使用机器学习方法进行新闻的情感自动分类[J].中文信息学报,2007,21(6):95-100.
- [4] 李寿山,黄居仁.基于 Stacking 组合分类方法的中文情感分类研究[J].中文信息学报,2010,24(5):56-61.
- [5] 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,2012,24(1):73-83.
- [6] Annett M, Kondrak G. A comparison of sentiment analysis techniques: polarizing movie blogs [C]//Canadian AI 2008. LNCS (LNAI), Springer, Heidelberg, 2008, 5032:25-35.
- [7] 刘志明,刘鲁.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.
- [8] 张成功,刘培玉,朱振方,等.一种基于极性词典的情感分析方法[J].山东大学学报:理学版,2012,47(3):47-50.
- [9] Shen Y, Li S C, Zheng L, et al. Emotion mining research on micro-blog [C]//2009 1st IEEE Symposium on Web Society, 2009:71-75.
- [10] Velikovich L, Goldensohn S B, Hannan K, et al. The viability of web-derived polarity lexicons [C]//Processing HLT' 10 Human Language Technologies; The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistic. Stroudsburg, PA, USA, 2010:777-785.
- [11] Aditya Joshi, Balamurali A R, Pushpak Bhattacharyya, et al. C-feel-i: a sentiment analyzer for micro-blog [C]//Proceeding of the ACL-HLT 2011 System Demonstration. Portland, Oregon, USA, 2011:127-132.
- [12] Jiang L, Yu M, Zhou M, et al. Target-dependent twitter sentiment classification [C]//Proceeding of the 49th Annual Meeting of the Association for Computational Linguistic, Stroudsburg, PA, USA, 2011:151-160.
- [13] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data [C]//Proceedings of the 23th International Conference on Computational Linguistic; Posters, Stroudsburg, PA, USA, 2010:36-44.
- [14] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data [C]//Proceedings of the Workshop on Language in Social Media, Stroudsburg, PA, USA, 2011:30-38.
- [15] Wang X L, Wei F R, Liu X H, et al. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach [C]//Proceedings of the 20th ACM international conference on Information and knowledge management, New York, USA, ACM Press, 2011:1031-1040.
- [16] Khuc V N, Shivade C, Rammath R, et al. Towards building large-scale distributed system for twitter sentiment analysis [C]//Proceedings of the 27th Annual ACM Symposium on Applied Computing, New York, USA, ACM Press, 2012:459-464.
- [17] 董振东,董强.知网[OL].http://www.keenage.com.

作者:	王勇 , 吕学强 , 姬连春 , 肖诗斌 , Wang Yong , L Xueqiang , Ji Lianchun , Xiao Shibin
作者单位:	王勇, 吕学强, 肖诗斌, Wang Yong, L Xueqiang, Xiao Shibin(北京信息科技大学网络文化与数字传播北京市重点实验室 北京 100101) , 姬连春, Ji Lianchun(新华网络股份有限公司 北京 100101)
刊名:	计算机应用与软件
英文刊名:	<div></div> Computer Applications and Software
年, 卷(期):	2014(1)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_jsjyyrj201401011.aspx