DOI:10.3969/j.issn.1673-4785.201208020
网络出版地址:http://www.cnki.net/kcms/doi/10.3969/j.issn.1673-4785.201208020.html

语义规则在微博热点话题情感分析中的应用

赵文清[1],侯小可[1],沙海虹[2]
(1. 华北电力大学(保定) 控制与计算机工程学院,河北 保定 071003; 2. 英业达集团(北京)电子技术有限公司 开发部,北京 100086)

摘 要:近来,针对微博热点话题的情感分析研究得到了广泛关注,而基于监督的学习方法在分析文本时会忽视词语的上下文联系。根据中文微博的特点,提出了一种基于语义规则的方法对微博热点话题进行情感分析。该方法首先需要人工整理出程度副词表、否定词表和微博中默认表情符号的褒贬分类。然后在情感词语计算的基础上,考虑上下文中否定词和程度词对修饰情感词语的情感倾向和情感强度的影响,同时也设定规则计算表情符号对一条微博的情感倾向判断的作用。最后与基于情感词典的方法做实验对比,实验结果表明该方法在文本情感倾向性识别的准确率上有了一定提高。

关键词:微博;热点话题;情感分析;语义规则;情感词典
中图分类号:TP391.1 **文献标志码**:A **文章编号**:1673-4785(2014)01-0121-05

中文引用格式:赵文清,侯小可,沙海虹.语义规则在微博热点话题情感分析中的应用[J]. 智能系统学报, 2014, 9(1): 121-125.
英文引用格式:ZHAO Wenqing, HOU Xiaoke, SHA Haihong. Application of semantic rules to sentiment analysis of microblog hot topics[J]. CAAI Transactions on Intelligent Systems, 2014, 9(1): 121-125.

Application of semantic rules to sentiment analysis of microblog hot topics

ZHAO Wenqing[1], HOU Xiaoke[1], SHA Haihong[2]
(1. School of Control and Computer Engineering, North China Electric Power University (Baoding), Baoding 071003, China; 2. Inventec (Beijing) Electronics Technology Co., Ltd., Beijing 100086, China)

Abstract: The research on the sentiment analysis for microblog hot topics has attracted much attention recently, while the studying method on the basis of supervision neglects the context of a word in the analysis of text. According to the characteristics of Chinese microblogs, a method based on semantic rules is proposed for sentiment analysis of microblog hot topics. As for the method, firstly, we need to manually sort out a degree adverb list, a negative word list and the appraisal category of the expression symbols defaulted in a microblog. Secondly, on the basis of the calculation of sentiment words, we consider the impact of negative words and degree words in the context of the emotional tendency and strength decorating sentiment words; in addition, we also set rules for calculating the influence of the expression symbol on the sentiment tendency judgment of a piece of microblog. Finally, our proposed method is compared with the method based on the emotional dictionary. The experimental results show that the proposed method improves the identification accuracy of the text sentiment tendency.

Keywords: microblog; hot topics; sentiment analysis; semantic rules; emotional dictionary

随着微博的飞速发展,微博作为一种通过关注机制分享简短实时信息的广播式社交网络平台,吸引了越来越多的网民参与。从 2011 年 12 月底中国互联网中心显示的报告可知,我国拥有微博的人数已达到 2.5 亿人次,占我国网民总数的 48.7%,比 2010 年增加了 296%。微博改变了公众信息获取的方式,是一种能够观察和了解正在发生什么的实时民意调查系统。中国的微博已由一种单纯的社交工具,变成舆论监督的利器,便于决策者做出决策。

目前关于微博情感分析的研究主要集中在英文微博方面,而面向中文微博的情感分析研究还在起步阶段。情感分析的主流方法依然是基于监督学习

收稿日期:2012-08-14. 网络出版日期:2014-02-20.
基金项目:国家自然科学基金资助项目(70671039);中央高校基本科研业务费专项资金资助项目(12MS121).
通信作者:侯小可. E-mail:houxiaoke2008@163.com.

的方法[1],2009 年 Alec 等[2]首次尝试通过训练表情符号特征,使用比较常用的 3 种机器学习的方法(naive Bayes、maximum entropy 和 SVM))对 Twitter 消息进行情感分类,并取得了不错的效果。而 Davidiv 等[3]把 Tweets 中的标签和表情符号作为特征,通过训练一个基于监督学习的、与 K 近邻(KNN)相似的分类器,来判断 Tweets 消息的情感倾向性。Jiang 等[4]采用内容、情感词典和主题相关 3 类特征对 Twitters 进行主客观和情感极性的分类,并提出了基于图模型的算法来优化情感分类结果。在中文微博的情感分析研究方面,谢丽星[5]通过从新浪微博提供的 API 接口抓取实验数据,对微博的链接、表情、情感词和上下文等主题无关的特征的有效性及多种分类方法进行了研究,最终选定 4 种特征共用及基于 SVM 的方法对微博消息进行了情感分类。刘志明等[6]通过对比 3 种机器学习算法、3 种特征选取算法和 3 种特征项权重计算方法,最终选用 SVM、IG 以及 TF-IDF 作为特征项权重对中文微博进行情感分析。

以上是基于监督的机器学习方法对微博进行情感分析,这种方法的第 1 步是将文本向量化,文本向量化的前提是特征之间的相互独立,这样势必会造成在分析文本时忽视词语的上下文联系。与基于监督学习的情感分析相比,基于规则和无监督学习[7-8]的研究不是很多。沈阳等[9]从饭否网抓取实验数据,通过定义态度词典、权重词典、程度词典、否定词典和连接词典来计算每条微博的情感指数,该方法存在的缺点是每个句子中的修饰程度词和否定词只是简单的统计,并没有针对到具体所修饰的情感词。机器学习的方法需要大规模标注的训练集,同时对训练集的质量要求也很高。

基于以上分析,本文提出一种基于语义规则的方法对微博热点话题进行情感分析。事实上,文本情感分析和文本分类最大的区别就在于语义相关性和上下文相关性,使用情感词语和词语的上下文关系来进行文本情感分析才更为合理。

1 微博情感分析

传统文本(如新闻网页、博客等)虽然只是简单的文字描述,但文本一般较长。与传统文本不同,微博文本简短,字数一般在 140 字以内,并且形式多样,一条微博中除了文字信息,还可以包含网页链接、图片信息、标签、表情符号等。微博的这些新特征对微博文本的情感分析会产生一定影响,例如人们判断一条微博的情感的第一反应就是通过表情符号,通常一条微博的情感与它所含表情符号的情感是相符的。

本文从新浪微博抓取某个热点话题的相关评论,对其采用基于情感词典和基于语义规则 2 种方法进行情感分析。

1.1 基于情感词典的方法

词典资源是基于情感词典方法的前提,本文使用中文最具权威的知网[10]词典资源,知网于 2007 年发布了最新版本“情感分析用词语集(beta 版)”,其中中文情感分析用词语集包含中文正面情感词语 836 个、负面情感词语 1 254 个、中文正面评价词语 3 730 个、负面评价词语 3 116 个。知网虽然对情感词语进行了褒贬分类,但是没有标注情感极性强度。本文将知网中褒义词语的情感极性值设为 0.8,贬义词语的情感极性值设为-0.8。

基于情感词典的方法首先对每条微博进行分词、词性标注等预处理,然后依据情感词典判断每条微博中出现的所有情感词以及其强度,并采用极性累加的方法计算每条微博的情感极性,如式(1):

$$P(T)=\sum_{i=1}^{n}P(w_i) \tag{1}$$

式中:w_i 为一条微博中所含的情感词;$P(w_i)$ 为一个情感词的情感极性;$P(T)$ 为一条微博的情感极性,若结果大于零,表明微博为褒义倾向,若结果小于零,表明结果为贬义倾向,否则为中性。

1.2 基于语义规则的方法

基于情感词典的方法是对独立的词语进行分析,也就是把词语从句子中孤立出来,忽略词语的上下文关系,因此,称之为词语的原极性。如果孤立地看待这些词语,并不能正确地反映微博消息的情感倾向,必须将上下文的联系考虑进来,才能够提高分析的准确度。因此,在词语情感计算的基础上,应该考虑上下文中能够改变词语情感倾向或者情感强度的修饰副词等。本文将会改变词语极性强度的修饰副词分为 2 类:第 1 类是否定词,它会改变极性倾向,如“不”;第 2 类是程度词,它会改变极性强度,如“很”、“非常”等[11]。

另外,微博消息文本有其自身的特征,如包含网页链接(http://t.cn/zWpsuJx)、表情符号()、标签(#孙杨#)等,本文只考虑与微博消息文本的情感极性相关的特征,如表情符号特征,而像网页链接、标签这些特征对微博文本的情感极性影响不大的,则不予以考虑。

1.2.1 情感词

情感词是判断微博文本是否具有情感倾向的一

个重要特征。根据人们写作习惯和大量语料分析得知,人们在微博中发表的观点和情感大多是通过情感词的形式实现,情感词的褒贬也通常代表这句子的褒贬。

一般情况下微博文本中都是比较简单的句子,情感词的倾向就直接决定了这条微博的情感倾向,情感词的数量和情感强度对每条微博文本的情感倾向有较大的影响,因此仍然采用极性累加的方法,即通过情感词极性累加公式(1)来计算每条微博的情感极性。

1.2.2　程度副词特征

知网的中文情感分析用词表中提供了程度级别词语,以知网程度级别词语为基础,参考蔺璜对程度副词的分类[12],人工整理所使用的程度副词,并把程度词语分为 3 个级别。第 1 级的程度词对所修饰的情感词的情感强度大大加强,例如“极其”、“最”。第 2 级的程度词对所修饰的情感词的情感强度是加强作用,如“很”、“非常”。第 3 级的程度词对所修饰的情感词的情感强度是削弱作用,如“有点”、“稍微”。3 个级别程度词对所修饰情感词的情感强度扩大倍数分别设置为第 1 级 2 倍,第 2 级 1.5 倍,第 3 级 0.5 倍。

如果句子中情感词语前面有程度词修饰,那么被修饰的情感词语的情感强度必然发生改变,进而会影响到这个句子的情感强度。一个程度副词后面可以有多个情感词,同样一个情感词也可以被多个程度副词所修饰。本文处理程度副词的方法是把情感强度加到其后修饰的第 1 个情感词上,情感强度对情感词 w_i 的影响因子 γ 定义为

$$\gamma(w_i) = \prod_{k=1}^{m} D(d_k)$$

式中:$D(d_k)$ 为程度副词的情感强度扩大倍数。

1.2.3　否定词特征

本文参考郝雷红对否定副词范围的界定[13],选取“不是”、“不会”、“不要”、“没有”等 30 个常见否定词作为否定副词表,并将其极性强度设置为-1。

否定词在句子的情感倾向性判断上有着重要作用。如果褒义词前面出现否定词,整个词汇的语义就会发生逆转,进而影响整个句子的情感倾向性。例如“我喜欢你”,在情感词前面加上否定词“不”,整个句子的情感极性就会发生改变。本文处理否定词的方法是将否定加到其后的第 1 个情感词上,当一个情感词前面出现不只一个否定词时,则根据否定词出现的次数来判断情感词的极性,出现奇数次则情感词的极性逆转,否则情感词的极性不发生改变。因此,否定词对情感词 w_i 的影响因子 η 定义为

$$\eta(w_i) = (-1)^n$$

1.2.4　表情符号特征

很多微博用户习惯在发布消息时加上一些表情符号,这些表情符号通常是由微博平台提供,方便用户的使用。微博消息中的表情符号被抓取后的表现形式变为中括号加文本,如表情符号“  ”相应文本为“嘻嘻”。用户选择不同的表情符号表达了不同的情感色彩,因此,本文把新浪微博平台提供的常用表情符号分为正向和负向 2 类。

一般情况下,如果一条微博消息包含表情符号,那么首先选择通过表情符号来判断一条微博的情感倾向。一条微博中可以包含多个表情符号,因此,首先需要对一条微博消息中的正向表情符号个数 e_P 和负向表情符号个数 e_N 进行统计,并把表情符号对整条微博的情感倾向影响因素 δ 定义为

$$\delta = e_P - e_N$$

式中:当 $\delta>0$ 时,就认为该微博为褒义倾向;当 $\delta<0$ 时,表明该消息为贬义倾向;否则通过其他特征来判定该条微博消息的情感倾向。

1.2.5　情感计算

因为微博文本内容较短,一般都在 140 字以内,通常也只包含一两句话,并且句法分析技术直接用于微博文本存在错误率较高的问题,所以本文省略了对微博文本进行句法分析,不再分句,从而直接对整条微博进行处理。综合考虑以上几个特征使用式(2)对一条微博的情感倾向值 $P(T)$ 进行计算:

$$P(T) = \begin{cases} \left|\sum_{i=1}^{n} P(w_i) \times \gamma(w_i) \times \eta(w_i)\right|, & \delta > 0 \\ -\left|\sum_{i=1}^{n} P(w_i) \times \gamma(w_i) \times \eta(w_i)\right|, & \delta < 0 \\ \sum_{i=1}^{n} P(w_i) \times \gamma(w_i) \times \eta(w_i), & \delta = 0 \end{cases} \tag{2}$$

若 $P(T)$ 计算结果大于零,表明微博为褒义倾向;若结果小于零,表明微博为贬义倾向;否则为中性。

2　实验结果与分析

利用 ROST 虚拟学习团队开发的新浪微博搜索数据抓取工具,抓取了在伦敦奥运会期间新浪微博上的热点话题“国羽女双输球”的 50 页共 924 条相关评论,经过人工识别,其中 13 条评论与此话题无关,443 条是负面评论,412 条是正面评论,还有 56 条是中性评论。一条微博中包含褒贬 2 种态度的取

其中一种较为明显的态度,否则归为中性。

首先对每条微博评论进行分词、词性标注等预处理,然后分别采用上节介绍的基于情感词典的方法和基于语义规则的方法分析处理微博评论,最后分别得到正面、负面和中性的评论数目,并与人工识别的结果相比较和计算2种方法的准确率(P)。实验结果如表1所示。

表1 2种方法得到的评论数目和准确率

Table 1 Reviewer number and correct ratio of two methods

类别	情感词典			语义规则		
	自动识别数	正确数	准确率 P/%	自动识别数	正确数	准确率 P/%
正面	385	254	65.9	406	281	69.2
负面	487	306	62.8	470	318	67.6
中性	39	7	17.9	35	8	22.8
合计	911	567	62.2	911	607	66.6

通过表1,可以看到基于语义规则的方法比基于情感词典的方法的准确率有了明显的提升,说明在微博消息处理中上下文语义关系和表情符号对一条微博的情感倾向的判断有着重要的影响,同时也说明了基于语义规则的方法是有效可行的。

本文从召回率(R)和F值(F-measure)2个方面对所提方法和情感词典方法做了进一步比较,结果如表2所示。

召回率(R)表示为

$$R = \frac{A}{A + C}$$

式中:A表示分类正确的文本数目,C表示分类错误和没有被分类的文本数目。F值的表达式为

$$\text{F-measure} = \frac{2 \times P \times R}{P + R}$$

表2 2种方法得到的召回率和F值

Table 2 Recall and F-measure of two methods %

类别	情感词典		语义规则	
	召回率	F值	召回率	F值
正面	61.7	63.7	68.2	68.7
负面	69.1	65.8	71.8	69.6
中性	12.5	14.7	14.3	17.6

由表2可以看出,基于语义规则的方法比基于情感词典的方法在召回率和F值方面都有了明显的提升,基于语义规则的方法能得到最好的分类效果,其F-measure分别为正面68.7%、负面69.6%和中性17.6%。F-measure是对准确率与召回率的综合评估。表2进一步说明了在微博情感分析中,基于语义规则的方法是有效可行的。

对造成情感判断错误的原因进行分析,由于网络中出现的一些网络词语导致情感词、修饰词未能识别,对分类结果造成影响,另外由于微博中的反讽、隐喻等表达方式也对分类结果造成一定影响。

3 结束语

本文根据中文微博的特点,在情感词的基础上综合考虑了微博消息文本中的修饰程度副词、否定词和表情符号特征,提出一种基于语义规则的方法对微博热点话题进行情感分析,并与基于情感词典的方法做了实验对比,实验结果证明了所提方法的有效性和可行性。

另外,虽然该方法在微博消息文本情感倾向性识别的准确率上有了一定提高,但整体水平不是很高,主要是由于微博文本情感分析中缺少情感所属对象的识别,以及微博上反讽和网络新语的出现,这些都是情感分析中的难点,未来的工作将考虑上属情况对微博情感分析做进一步研究。

参考文献:

[1]PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA, 2002: 79-86.

[2]郑斐然,苗夺谦,张志飞,等.一种中文微博新闻话题检测的方法[J].计算机科学, 2012, 39(1): 138-141.
ZHENG Feiran, MIAO Duoqian, ZHANG Zhifei, et al. News topic detection approach on Chinese microblog[J]. Computer Science, 2012, 39(1): 138-141.

[3]DAVIDIV D, TSUR O, RAPPOPORT A. Enhanced sentiment learning using Twitter hashtags and smileys[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China, 2010: 241-249.

[4]JIANG Long, YU Mo, ZHOU Ming, et al. Target-dependent Twitter sentiment classification[C]//The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA: 151-160.

[5]谢丽星.基于SVM的中文微博情感分析研究[D].北京:清华大学, 2011.
XIE Lixing. Sentiment analysis of Chinese microblog using SVM[D]. Beijing: Tsinghua University, 2011.

[6]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.
LIU Zhiming, LIU Lu. Empirical study of sentiment classification for Chinese microblog based on machine learning[J]. Computer Engineering and Applications, 2012, 48(1): 1-4.

[7]TURNEY P D. Thumbs up or thumbs down? Semantic orien-

tation applied to unsupervised classification of reviews [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 417-424.

[8]朱嫣岚,闵锦,周雅倩,等.基于 HowNet 的词汇语义倾向计算[J].中文信息学报, 2006, 20(1): 14-20.
ZHU Yanlan, MIN Jin, ZHOU Yaqian, et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.

[9] SHEN Yang, LI Shuchen, ZHENG Ling, et al. Emotion mining research on micro-blog[C]//2009 1st IEEE Symposium on Web Society. Lanzhou, China, 2009: 71-75.

[10]张华平. NLPIR 微博内容语料库-23 万条[EB/OL]. [2012-08-05]. http://www.nlpir.org/? action-viewnews-itemid-231.2012,02,14/2012,02,18.

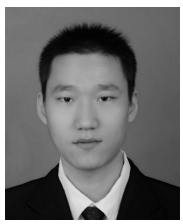
[11]吴泽衡.基于话题检测和情感分析的互联网热点分析与监控技术研究[D].广州:华南理工大学, 2011.
WU Zeheng. Research on internet hotspot analysis and monitoring technologies based on topic detection and sentiment analysis[D]. Guangzhou: South China University of Technology, 2011.

[12]蔺璜,郭姝慧.程度副词的特点范围与分类[J].山西大学学报:哲学社会科学版, 2003, 26(2): 71-74.
LIN Huang, GUO Shuhui. On the characteristics, range and classification of adverbs of degree[J]. Journal of Shanxi University: Philosophy & Social Science, 2003, 26(2): 71-74.

[13]郝雷红.现代汉语否定副词研究[D].北京:首都师范大学, 2003.
HAO Leihong. Research on negative aderbs of modern Chinese[D]. Beijing: Capital Normal University, 2003.

作者简介:

赵文清,女,1973 年生,副教授,中国人工智能学会粗糙集与软计算专业委员会委员,主要研究方向为机器学习、数据挖掘、贝叶斯网络学习等。获河北省科技进步三等奖 1 项、国家发明专利 1 项,发表学术论文 30 余篇,出版教材 3 部。



侯小可,男,1985 年生,硕士研究生,主要研究方向为人工智能、数据挖掘等。

沙海虹,男,1971 年生,高级工程师,主要研究方向为人工智能、数据挖掘。

2014 年全国模式识别学术会议
Chinese Conference on Pattern Recognition 2014

全国模式识别学术会议(Chinese Conference on Pattern Recognition, CCPR)旨在为国内学者提供一个学术交流和成果展示的平台,促进国内模式识别研究和应用的发展。该会议 2007 年和 2008 年在北京、2009 年在南京、2010 年在重庆、2012 年在北京举行,得到了国内同行的积极响应。2012 年起,全国模式识别学术会议改为 2 年举行 1 次。第 6 届全国模式识别会议(6th Chinese Conference on Pattern Recognition, CCPR 2014)将于 2014 年 11 月 3—5 日在长沙湖南大学召开。会议面向国内外同行征集模式识别领域的高质量学术论文,任何促进模式识别技术发展、进步和应用的相关成果都属于我们征集的范畴。同时,鼓励探讨本领域研究热点、长远规划和学科前沿的论文投稿。

重要日期:

投稿截止:2014-05-30

录用通知:2014-07-30

最终稿提交:2014-08-25

会议日期:2014-11-17—19

会议网站:http://eeit.hnu.cn/ccpr2014/index.html