基于情绪强度的中文微博情绪分析

王世泓,牛 耘
(南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

摘 要:由于中文情绪表达的多样性,以及微博情绪的丰富性和敏感性,情绪词在表达情绪时存在强弱差别,相同的情绪词在不同的语料语境中也可能表达不同的情绪强度。因此文中提出了基于情绪强度的中文微博情绪分析,并根据语料上下文计算出情绪词的情绪相似度,基于情绪相似度自动标注了情绪强度,利用情绪强度进行微博文本的情绪分析。实验结果表明,对情绪词进行情绪强度的标注可以更细致地识别出微博中的主要情绪,进一步提高微博情绪分析的准确率。

关键词:情绪词;情绪强度;情绪相似度;微博情绪
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2015)06-0137-04
doi:10.3969/j.issn.1673-629X.2015.06.030

Analysis of Chinese Micro-blog Emotion Based on Emotional Strength

WANG Shi-hong, NIU Yun
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: Due to the diversity of expressions of emotion in Chinese, and the richness and sensitivity of micro-blogs emotion, emotional words may express emotions at different levels of strength, an emotional word may express different levels of emotional strength in different corpora. The Chinese micro-blogs emotion analysis is proposed based on emotional strength, calculate the emotional similarity of emotional words based on corpus context, and use it to annotate the emotional strength of words. Finally, analyze the emotion of Chinese micro-blogs based on emotional strength. The experimental results show that emotional strength is effective in identifying the major emotion of micro-blogs, and improves the accuracy of micro-blog emotional analysis further.

Key words: emotional words; emotional strength; emotional similarity; micro-blog emotion

1 概 述

微博已成现代人交流和传播信息的重要平台,微博中涌现出大量琐碎的文本数据蕴含着丰富信息资源,引起国内外学者对其展开了一系列研究,微博情绪分析就是其中一个热点话题。它主要是判断出微博文本中所表达的情绪。对情绪的划分,不同领域划分方法各不相同,情绪识别领域普遍按照 Ekman[1-3] 通过研究人的面部表情提出的六类情绪进行划分,六类情绪包括:joy, sad, anger, fear, digust, surprise。

微博文本不同于传统文本[4-5],显著的特点有:

(1)文本简短,特征稀疏;

(2)语言表达口语化,语法极不规则;

(3)文本中使用大量网络用语和一些表情图片。

这些特点不仅使传统文本的分析方法无法适用于微博文本分析,而且给微博文本分析带来了新挑战

现有的微博情绪分析方法主要是基于机器学习的有监督方法[6]和基于词典的无监督方法[7]。其中,基于机器学习的有监督方法的工作有:曹海涛[8]结合 PAD 情感模型,利用支持向量机的分类方法对微博文本进行情感分析,该方法与其他机器学习方法相比提高比较显著。Danisman 和 Alpkocak[9] 使用向量空间模型进行了六类情绪分类,获得的 F 值大约是 32%。Mohammad[10] 以情绪词为特征和以 n -元为特征的分类器进行六类情绪的分类,最好的 F 值结果在 52% 左右。Teng Zhi[11] 等提出利用粗集合理论和支持向量机对文本进行情绪分析,实验取得了较好的分析结果。虽然有监督方法已经取得了一定的成果,但其依赖于人工标注语料,代价昂贵。基于词典的无监督方法克服了有监督方法的问题,利用情绪词典来进行微博的

收稿日期:2014-08-08 **修回日期**:2014-11-13 **网络出版时间**:2015-05-06
基金项目:国家自然科学基金资助项目(61202132,61170043);国家"973"重点基础研究发展计划项目(2013CD744904)
作者简介:王世泓(1990-),女,硕士研究生,CCF 会员,研究方向为自然语言处理;牛 耘,博士,副教授,CCF 会员,研究方向为自然语言处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150506.1645.037.html

情绪分析。如,Fen Fuji 等从语料中标注出词典,通过计算句子与情绪词典的相似度,利用先验知识来判断句子的情绪[12]。实验取得了 62.7% 的 F 值。张晶等[13]以常用情绪词和情绪短语为基础构建情绪词典,结合情绪规则库来分析微博情绪。但这些方法对情绪词典的应用都仅限于情绪词的情绪类别。

情绪词在语料中不仅有情绪类别之分,在表达情绪时也有强弱之分。例如:“买了喜欢的鞋和衣服,超满足,不开心统统散去”和“底线不可突破,诉求必须满足,这才是正确的行事方式。”两条微博中都有 joy 类情绪词“满足”,前一条中其表达了强烈的 joy 情绪,而后一条中却没有。可见情绪强度的标注对微博的分析是很有帮助的。因此,文中提出了基于语料上下文自动标注情绪强度,并利用标注的情绪强度对微博进行情绪分析。实验结果表明,情绪强度的标注确实提高了微博情绪分析的准确率。

2 词 典

CLIWC[14]和大连理工大学的情感词汇本体库[15]是中文情绪词典中广泛使用的两个情绪词典。文中按照一定的规则将这两个词典进行合并,并基于合并的情绪词典进行了情绪分析。

CLIWC 词典中有五个词语类别表达了情绪,分别是:正向情绪词、负向情绪词、生气词、伤心词和焦虑词。其中,与 Ekman 的情绪分类吻合的只有“生气词”和“伤心词”两个词语类别,对另外三个词语类别按照 Ekman 情绪分类做了新的划分:正向情绪词划分为 joy 类的情绪词;负向情绪词和焦虑词两个词语类别,对其进行了人工标注,在这两个词语类别中标注出 fear、disgust 和 surprise 三个情绪类的情绪词。形成了情绪词集 CliwcSix。中文情感词汇本体库的情绪共分为七类:好、乐、哀、怒、惧、恶、惊。共含词语 27 466 个。对照 Ekman 的情绪分类,文中把“好”和“乐”两类情绪词合并为 joy 类的情绪词,并形成了情绪词集 DutirSix。文中将 CliwcSix 和 DutirSix 两个词集进行了合并,形成新的情绪词典 EmoCD。合并时去掉两个词集中同时出现但情绪分类不同的情绪词。表 1 列出了 EmoCD 中每类情绪的情绪词分布情况。

表 1 情绪词典 EmoCD 的分布情况

	喜	哀	怒	惧	恶	惊	合计
数量	13 166	2 319	472	1 171	10 041	224	27 393

3 情绪词的情绪强弱计算

文中通过计算情绪词与情绪类之间的相似度来衡量情绪词的情绪强弱。并基于六维情绪向量来建立各个情绪类的共现模式。其中,每一维对应了 Ekman 情绪类别的一个情绪状态。

3.1 词情绪向量的生成

词情绪向量反映了一个情绪词在语料中与各个情绪类的情绪词的共现情况,由于高频词的干扰,情绪词与各个情绪类的情绪词之间的共现模式不能简单地使用频率来表示,还需要考虑共现的情绪词在语料中出现的情况。文中基于六维情绪向量的模式表示词情绪向量,情绪词 ew_j 的词情绪向量记为 $\mathbf{WV}_j$,表示如下:

$$\mathbf{WV}_j = (W_{j,e_1}, W_{j,e_2}, \cdots, W_{j,e_6})$$

每一维的值 W_{j,e_i},表示情绪词 ew_j 与 e_i 情绪类的情绪词共现的权重,计算公式如下:

$$W_{j,e_i} = \frac{\mathrm{tf}_{j,e_i}}{\log\left(\sum_{ew_z \in \mathrm{WD}_{j,e_i}} (n(\mathrm{ew}_z))\right)}$$

其中,tf_{j,e_i} 表示情绪词 ew_j 与情绪类 e_i 的情绪词在语料中共现的频率和;WD_{j,e_i} 是所有与情绪词 ew_j 共现的 e_i 情绪类的情绪词集合;$n(\mathrm{ew}_z)$ 表示情绪词 ew_z 在语料中出现的总次数,ew_z 是与情绪词 ew_j 共现的 e_i 情绪类的情绪词。

考虑共现情绪词在语料中出现的频率,可以有效地防止高频词的干扰。

3.2 基情绪向量组的构造

文中基于六维情绪向量的模式给每个情绪类都创建了一组情绪向量,称为基情绪向量组。它包括两个部分:简单基情绪向量和扩展基情绪向量。简单基情绪向量代表某个情绪类的情绪趋势。扩展基情绪向量反映了某个情绪类在语料中与各个情绪类的共现情况。基情绪向量组的创建为情绪词的词情绪向量提供了比较的依据。定义方法如下:

(1)简单基情绪向量记为 $\mathbf{SSV}_{e_i}$,表示 e_i 情绪类的简单基情绪向量。定义方法为:情绪向量中对应情绪类的维的值置 1,其他情绪类的维置 0。如,情绪类 e_1 的简单标准情绪向量为 $\mathbf{SSV}_{e_1} = (1,0,0,0,0,0)$,其他情绪类的简单基情绪向量以此类推。

(2)扩展基情绪向量 $\mathbf{ESV}_{e_i}$,表示 e_i 情绪类的扩展基情绪向量。定义方法为:首先,统计出 e_i 情绪类在语料中出现次数>3 的情绪词,再计算出这些情绪词在语料中与各个情绪类的情绪词的共现频率,形成六维的频率向量 $\mathbf{TF}$,最后再求出这些频率向量的平均值作为扩展情绪向量。具体公式如下:

$$\mathbf{ESV}_{e_i} = \frac{1}{M}\left(\sum_{\substack{(\mathrm{ew}_h \in W(e_i)) \\ \wedge (n(\mathrm{ew}_h) > 3)}} (\mathbf{TF}_h)\right)$$

其中,$\mathbf{TF}_h$ 表示情绪词 ew_h 的频率向量;$W(e_i)$ 表示 e_i 情绪类的情绪词集;M 表示 $W(e_i)$ 中情绪词的个数;$n(\mathrm{ew}_h)$ 表示情绪词 ew_h 在语料中出现的次数。

3.3 情绪词的情绪强度的标注

情绪词的情绪强度是情绪词在语料中表达情绪的强弱程度。当情绪词的情绪向量与基情绪向量组的情绪相似度越大,文中就认为它在语料中表达情绪的程度就越强。反之情绪相似度越小,它在语料中表达情绪的程度就越弱。文中通过情绪阈值来具体区分情绪词的情绪强度。步骤如下:

(1)计算情绪词的词情绪向量与基情绪向量组的情绪相似度,公式如下:

$$\mathrm{sim}_{j,e_i}=\frac{1}{2}(\mathrm{sim}(\mathrm{WV}_j,\mathbf{SSV}_{e_i})+\mathrm{sim}(\mathrm{WV}_j,\mathbf{ESV}_{e_i}))$$

其中,情绪相似度 sim 的公式计算采用 consine 相似度; sim_{j,e_i} 表示情绪词 ew_j 与 e_i 情绪类的基情绪向量组的相似度。每个情绪词与六个情绪类的基情绪向量组计算得到六个情绪相似度值,将其中最大的相似度值记为 Msim_j ,对应的情绪类记为 Memo_j ,将该情绪词在词典中划分的情绪类记为 emo_j 。

(2)情绪阈值是情绪强度划分的分界线,下文简称阈值。因为各个情绪类的情绪词在语料中的分布情况不同,所以每个情绪类的情绪词划分情绪强度的阈值也不相同,因此文中对每个情绪类都计算一个阈值。阈值定义为 e_i 情绪类的情绪词在语料中出现的频率占所有情绪词在语料中出现的总频率的比例,并记为 γ_{e_i} 。具体公式如下:

$$\gamma_{e_i}=0.5+\frac{\mathrm{total}_{e_i}}{\mathrm{total}}$$

其中, γ_{e_i} 表示 e_i 情绪类的阈值; total_{e_i} 表示 e_i 情绪类的情绪词在语料中出现的总频率; total 表示所有情绪词在语料中出现的总频率。

(3)情绪词情绪强度的标注。文中将情绪词的情绪强度划分为强、中和弱三个等级,并根据阈值将情绪强度按照两种规则进行标注。其中,当情绪类的阈值大于等于 1 时,文中认为该情绪类的情绪词在语料中分布极不平衡,在各情绪类的微博中出现频率都较为分散,容易对其他情绪类的情绪词产生干扰,按规则一进行情绪标注。当情绪类的阈值小于 1 时,认为该情绪类的情绪词在语料中的分布相对平衡,按规则二进行情绪强度标注。具体规则定义如下:

规则一:

$$\mathrm{strength}(\mathrm{ew}_j)=\begin{cases}3,\mathrm{no}\\ 2,\mathrm{Msim}_j>0.95\ \wedge\ \mathrm{Memo}_j==\mathrm{emo}_j\ \wedge\\ \qquad n(\mathrm{ew}_j)>3\\ 1,\mathrm{otherwise}\end{cases}$$

规则二:

$$\mathrm{strength}(\mathrm{ew}_j)=\begin{cases}3,(\mathrm{Msim}_j>\gamma_{e_i})\ \wedge\ (n(\mathrm{ew}_j)>3)\ \wedge\\ \qquad (\mathrm{Memo}_j==\mathrm{emo}_j)\\ 2,\mathrm{Memo}_j==\mathrm{emo}_j\\ 1,\mathrm{otherwise}\end{cases}$$

其中, $n(\mathrm{ew}_j)$ 表示情绪词 ew_j 在语料中出现频率; $\mathrm{strength}(\mathrm{ew}_j)$ 表示情绪词 ew_j 在语料中情绪强度。

3.4 基于情绪强度的规则方法

(1)通过词典匹配出微博中所包含的各情绪类的情绪词,并将各情绪类的所有情绪词的情绪强度相加,得到该微博各情绪类的得分,形成微博的情绪分数集。

(2)若微博情绪分数集中存在唯一的最大值时,那么微博的情绪就判断为最大值对应的情绪。否则,微博中不存在最大值或微博未匹配到任何的情绪词,认为微博的情绪不能判断,该微博未被词典覆盖。

4 实验结果及分析

文中采用文本分类通用评测方法对文本工作进行评测,包括覆盖率、准确率、精确率(p)、召回率(r)、f -score(f)。

4.1 实验语料

文中从新浪 API 抓取微博文本。由两名标注人员各自独立对文本进行情绪标注。每条微博标注为喜、哀、怒、惧、恶、惊和其他共七类中的一类。将两名标注员标注结果一致的微博文本提取出来作为实验数据集以保证数据的可靠性。表 2 是得到的实验语料中六类情绪的微博分布情况。

表 2　六类情绪的微博分布情况

情绪类别	喜	哀	怒	惧	恶	惊	合计
数量	473	173	276	147	145	121	1 335

4.2 基于情绪强度的微博情绪分析

实验使用 EmoCD 情绪词典,利用规则方法基于情绪强度来对微博语料进行情绪分析,并将实验结果和基于词频的规则方法的结果进行对比。

4.2.1 情绪阈值的计算

根据公式计算出语料中各情绪类的情绪类阈值,具体情况如表 3 所示。在表 3 中,joy 类的情绪类阈值超过了 1,而 fear 和 surprise 类的情绪类阈值只有 0.5,这说明语料中各情绪类的情绪词分布极不均衡,joy 类情绪词在语料中出现的频率和占情绪词总频率的 50% 以上,而 fear 和 surprise 类的情绪词出现频率还不到10% 。情绪词在语料中的不均衡分布可能导致不

表 3　情绪类阈值

阈值参数	γ_{e_1}	γ_{e_2}	γ_{e_3}	γ_{e_4}	γ_{e_5}	γ_{e_6}
参数值	1	0.6	0.6	0.5	0.7	0.5

同情绪的情绪词之间的相互干扰,以及各情绪类识别的不均衡性,给多类的微博情绪识别带来了挑战。

4.2.2 文中方法与传统方法对比

本节将文中方法与传统的基于词典的规则方法的实验结果进行比较。其中,A 代表传统的基于词典的规则方法,B 代表文中方法,两组实验的覆盖率和准确率见表 4。表 4 中,B 组实验的覆盖率比 A 组提高了 7.3%,准确率也提高了 4.8%。这表明情绪词在表达情绪时确有强弱差别,划分情绪词的情绪强度可以更细致地区别出微博文本中的情绪,从而有效地提高了多类的微博情绪的识别。

表 4 两组实验的情绪分类结果

词典	方法	覆盖率/%	准确率/%
EmoCD	A	68.5	37.4
	B	75.8	42.2

下面通过精确率、召回率和 F -score 值更全面地看一下该方法带来的利弊,结果如表 5 所示。

表 5 两组实验的六类情绪分类结果比较

参数	方法	喜	哀	怒	惧	恶	惊	平均
召回率/%	A	74.6	18.5	11.6	17.0	22.8	19.8	27.4
	B	71.5	28.3	19.6	27.9	35.9	24.0	34.5
精确率/%	A	57.6	42.6	78.0	80.6	26.8	75.0	60.4
	B	62.0	46.2	79.4	74.5	27.1	63.0	58.7
F -score/%	A	65.0	25.8	20.2	28.1	24.6	31.4	32.5
	B	66.4	35.1	31.4	40.6	30.9	34.7	39.9

表 5 中,B 组实验的召回率和 F -score 的平均值相对 A 组实验分别提高了 7.1% 和 7.4%,说明文中方法比传统的基于词典的方法更准确地分析了微博表达的情绪。另外,B 组实验中各情绪类的召回率相对 A 组实验分布更均衡,这说明 B 组实验中各情绪类都得到了更充分地识别,这对多类的微博情绪分析具有重要的意义。此外,B 组实验中惧和惊两类情绪的精确率有所下降,通过分析发现,惧和惊两个情绪类在语料中出现频率较少,情绪强度标注比较高,这在一定程度上提高了两个情绪类对其他情绪类的干扰,从而导致精确率的下降。由此可见,对情绪词的情绪强度标注是一个复杂的问题,需要更深入的探讨。

5 结束语

微博情绪的分析有助于人们了解生活状态,预测事件的走向。文中提出了基于情绪强度的微博情绪分析。通过对语料上下文的分析计算了情绪词的情绪相似度,并利用情绪相似度进行了情绪强度的自动标注,再将标注出的情绪强度应用到基于规则方法的微博情绪分析中。实验结果表明,基于情绪强度的方法确实提高了微博情绪分析的准确率和覆盖率,同时更均衡了各情绪类的微博情绪识别,这对多情绪类的微博情绪分析具有重要意义。

参考文献:

[1] Ekman P. Facial expression and emotion[J]. American Psychologist,1993,48(4):384-392.

[2] Ghazi D, Inkpen D, Szpakowicz S. Hierarchical versus flat classification of emotions in text[C]//Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Los Angeles, California:[s. n.],2010:140-146.

[3] Bellegarda J. Emotion analysis using latent affective folding and embedding[C]//Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Los Angeles, California:[s. n.],2010:1-9.

[4] 张剑峰,夏云庆,姚建民. 微博文本处理研究综述[J]. 中文信息学报,2012,26(4):21-27.

[5] 阳 蓉. 从网络语言的特征谈其价值与前景[J]. 西昌学院学报:社会科学版,2006,18(2):35-38.

[6] 庞剑锋,卜东波,白 硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001,18(9):23-26.

[7] 牛 耘,潘明慧,魏 欧,等. 基于词典的中文微博情绪识别[J]. 计算机科学,2014,41(9):253-258.

[8] 曹海涛. 基于 PAD 模型的中文微博情感分析研究[D]. 大连:大连理工大学,2013.

[9] Danisman T, Alpkocak A. Feeler:emotion classification of text using vector space model[C]//Proc of AISB 2008. Aberdeen:[s. n.],2008:53-59.

[10] Mohammad S. From once upon a time to happily ever after:tracking emotions in novels and fairy tales[C]//Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences and humanities. [s. l.]:Association for Computational Linguistics,2011:105-114.

[11] Teng Zhi, Ren Fuji, Kuroiwa S. Emotion recognition from text based on the rough set theory and the support vector machines[C]//Proc of international conference on natural language processing and knowledge engineering. Beijing:IEEE,2007:36-41.

[12] Quan Changqin, Ren F. Recognizing sentence emotions based on polynomial kernel method using Ren-CECps[C]//Proc of international conference on natural language processing and knowledge engineering. Dalian:IEEE,2009:1-7.

[13] 张 晶,朱 波,梁琳琳,等. 基于情绪因子的中文微博情绪识别与分类[J]. 北京大学学报:自然科学版,2014,50(1):79-84.

[14] C-LIWC[EB/OL]. 2013. http://c-liwc.blogbus.com/.

[15] 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.

基于情绪强度的中文微博情绪分析

作者： 王世泓， 牛耘， WANG Shi-hong， NIU Yun
作者单位： 南京航空航天大学 计算机科学与技术学院,江苏 南京,210016
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(6)