基于多线索混合词典的微博情绪识别

潘明慧,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:微博等社交媒体为人们情绪表达提供了重要平台,分析微博的情绪倾向具有重要的商业价值和社会意义。文中提出了基于词典的规则方法识别微博所表达的喜、哀、怒、惧、恶、惊六种情绪。针对情绪表达的重要线索表情符利用互信息法生成了表情符词典,与传统情绪词典相结合,制定了针对否定用法的规则对微博进行分析。建立了第一个包含六种情绪的人工标注微博数据集。实验表明,传统的情绪词典虽然收录了大量词汇,但对于社交媒体文本分析的准确率和覆盖率都不高。表情符词典的应用显著地提高了微博情绪分析的精度和覆盖率。

关键词:微博;情绪分析;情绪词典;表情符
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2014)09-0028-05
doi:10.3969/j.issn.1673-629X.2014.09.006

Emotion Recognition of Micro-blogs Based on a Hybrid Lexicon

PAN Ming-hui, NIU Yun
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: The proliferation of micro-blogs has created a popular digital platform where people are able to express emotions and share feelings. Analysis of emotions in micro-blogs would be potentially beneficial to companies and the society. In this paper, a lexicon-based approach is proposed to identify six emotions in micro-blog text, including joy, sadness, anger, fear, disgust and surprise. A lexicon of emoticons is built based on the mutual information method between emoticons and emotions. Combined with a traditional emotion lexicon in this approach, negation rules are made to process negations in emotion expression to analyze mirco-blog. The first corpus of Chinese micro-blogs manually annotated with the six emotions is built as the test set. The experimental results show that the traditional lexicon has a moderate accuracy and coverage in analysis of micro-blog text. The combination of the two lexicons greatly improves the accuracy and coverage.

Key words: micro-blog; emotion analysis; emotion lexicon; emoticons

0 引 言

随着互联网的迅速发展,微博、博客、论坛等社交媒体成为大众抒发情感和获取信息的重要途径。其中微博以140字的特色使人们能迅速分享生活中的喜怒哀乐而受到越来越多用户的欢迎,已成为抒发个人情绪的重要平台。对微博所表达情绪的自动分析能够帮助用户分析自身情绪以及波动变化情况,检测抑郁症的发生,也便于迅速掌握大众的情绪走向,对预测民众需求有着重要的现实意义。自动情绪分析也是自然语言处理领域很多研究课题的重要组成部分,例如,自然语言界面[1]、电子学习环境[2]、安全信息学[3]等。而微博作为一种新兴的媒介在形式上和传统文本有着明显的区别:首先,文本简短,表达口语化;其次,包含有表情符、转发、评论、链接等多种信息的表达方式。其中表情符能够生动醒目地表达丰富的情绪。综上所述,针对微博特殊的表达方式进行情绪分析显得尤为重要。

文中对微博文本中所表达的情绪进行自动分析判别。情绪(Emotion)是人各种的感觉、思想和行为的一种综合的心理和生理状态,是对外界刺激所产生的心理反应,以及附带的生理反应,如:高兴、生气、伤心等。情绪分析(Emotion Analysis)即自动判别文本所

收稿日期:2013-11-02 **修回日期**:2014-02-15 **网络出版时间**:2014-07-17
基金项目:国家自然科学基金青年科学基金项目(61202132);教育部高等学校博士学科点专项基金资助项目(20103218120024);校青年科创基金(NS2012073)
作者简介:潘明慧(1988-),女,江苏仪征人,硕士生,研究方向为自然语言处理;牛 耘,博士,副教授,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.20140717.1233.043.html

表达的情绪状态。Ekman[4]通过研究人的面部表情,提出了六种基本情绪状态:喜(joy)、哀(sadness)、怒(anger)、惧(fear)、恶(disgust)、惊(surprise)。这六种基本情绪分类被自然语言处理领域的自动情绪识别研究所广泛采纳[5-7]。文中也将以这六类为目标分析微博文本中表达的情绪。

文中提出一种基于情绪词典的方法识别一条微博表达的情绪。重点分析了微博情绪表达中的两个重要线索,即微博中大量出现的表情符以及情绪表达中的否定用法。对表情符进行了细致的分类,并制定了情绪分析中否定用法的分析规则。实验表明,该方法有效地利用了表情符及否定用法,提高了情绪识别的精度。

1　相关工作

1.1　英文社交媒体中的情绪分析

针对英文的情绪分析已取得了一些初步的研究成果。所采用的方法主要可分为机器学习的方法和基于规则的方法。机器学习的方法主要利用 n-元(n-gram)为特征建立分类器[8-10]。Mohammad[8]以 Ekman 六种情绪为目标情绪集合,判断博客表达了其中的哪一种情绪。该文对以情绪词典中的情绪词为特征以及以 n-元为特征的分类器进行了对比试验,发现情绪词典对于不同领域的适应性更强。基于规则的方法是情绪分析研究采用的另一种主要策略。Golder 和 Macy[9]主要采用基于词典的方法对上百万的不同地域、不同文化背景的博主发表的 Twitter 微博进行了自动情绪分析,清晰地识别出人们情绪随时间呈周期性变化的模式。Paltoglou 等[10]利用基于情绪词典的规则方法对 Twitter 等社交媒体上的微博进行积极和消极两类情绪分类以及情绪强度的分析。实验结果表明多数情况下这种方法优于监督的机器学习方法。

1.2　中文社交媒体中的情绪分析

在对社交媒体的情感分析中,谢丽星等[11]仅考察微博中的评论是正面或负面的。该文重点研究了基于支持向量机的层次结构多策略方法,提取的特征包括主题无关和主题相关的特征。杨亮等[12]通过监测微博文本中情感词数量以及所表达情感的变化来建立情感分布语言模型,由此发现微博中的热点事件。

目前针对中文社交媒体中的情绪的研究还在初级阶段,而中、英文微博在表达方式上存在很大区别,使得英文情绪分析技术难以直接用于中文。因此,针对中文微博中表达的情绪开展研究有重要的现实意义。

2　情绪词典

情绪词典最基本的形式由情绪词条及该词表达的情绪组成。如"开心 喜"。文中方法采用的词典包括两个部分,情绪词词典部分以及表情符词典部分。

2.1　中文情绪词典

目前只有为数很少的中文情绪词典可供下载使用。包含 Ekman 六类情绪且词汇量较大的中文情绪词典有大连理工的中文情感词汇本体库(DUTIR)[13]。因而文中用它来进行微博的情绪分析。DUTIR 情感本体库在 Ekman 的六类情绪的基础上加入了情感类别"好",共含有情绪词 27 466 个,每个情绪词拥有至少一个情绪标签。新增的"好"类包含有"尊敬""赞扬""相信""喜爱""祝愿"五个子类。文中认为它们也表达了高兴的情绪,因而将其和乐类合并作为喜类。表 1 显示了 DUTIR 中情绪种类及相应的词条数(拥有两个以上情绪标签的词未统计在表中。如"娇弱"既属于"喜爱",又属于"憎恶")。

表 1　DUTIR 六类情绪词及示例

情绪类别	词数	示例
喜(乐、好)	13 031	喜悦、笑眯眯
哀	1 905	忧伤、心如刀割
怒	333	气愤、大发雷霆
惧	1 072	惶恐、畏惧、害怕
恶	9 811	反感、可耻、恶心
惊	201	吃惊、愕然、惊奇

2.2　表情符词典

微博平台提供了大量的表情符,微博中出现的表情符往往与博主的情绪存在直接的对应关系,因而对情绪分析起着至关重要的作用。利用新浪 API 获得的表情符共 1 764 个,而这个集合随着流行的变化仍在不断地更新扩展。其中有些符号并不表达情绪,如数字"3"。对于表达情绪的符号,微博平台也没有给出其情绪标签。很多符号由于个人理解不同因而采用人工标注很难对其情绪进行划分,如"[抱]"。还有些符号看似表达的情绪很明确,然而对其进行细致的情绪分类则并不容易。比如"[泪]",既可以表达哀,也可在惧的时候使用。因此,如何从大量表情符中挑选能够表达情绪的符号,并合理划分其情绪种类是有效利用表情符的难点。

通过观察发现一条微博中常常包含一个以上的表情符,而这些符号经常传达相同的情绪。比如"我的熊好可爱[哈哈]哈哈。有木有很乖? 谢谢我家姐姐! 么么!!! [泪][泪][心]"中的四个表情符都表达了开心。基于这些观察,提出利用语料库来获取表情符表达的情绪。即采用互信息法判断一个符号是否带有情绪色彩,并判别其表达的情绪为 Ekman 六类中的哪一种,

进而建立表情符词典，然后不断迭代扩充表情符词典。具体而言，先对每一种情绪选取少量表情符作为种子；然后计算其他表情符与种子之间的互信息，从而得到该表情符与种子所代表的情绪之间的关系。互信息值越高表明该表情符与相应情绪之间的关系越密切；最后比较该表情符与各情绪之间的互信息值，将其情绪类别确定为与之具有最高互信息的情绪的种类。

设情绪集合为 $EF = \{ef_1, \cdots, ef_i, \cdots, ef_m\}$，其中 m 为情绪类别的总数，$ef_i = i$，$i \in \{1,2,\cdots,6\}$，代表 Ekman 六类情绪。通过以下步骤建立表情符的情绪词典：

(1)对每种情绪 ef_i 选出少量种子表情符。

(2)对一个未知类别的表情符 ω，按照以下公式计算它与某种情绪的互信息值：

$$MI(\omega, ef_i) = \max_{\tau \in ef_i} \left\{ \log \frac{P(\omega,\tau)}{P(\omega)P(\tau)} \right\} \quad (1)$$

其中，τ 为 ef_i 中的已知表情符；$P(\omega,\tau)$ 是 ω 和 τ 同时出现的概率；$P(\omega)$ 是 ω 出现的概率；$P(\tau)$ 是 τ 出现的概率。

(3)比较 ω 与每种情绪的互信息值，判定 ω 的情绪类别：

$$ef(\omega) = \begin{cases} ef_i, \max\limits_{1 \leqslant j \leqslant m \cap j \neq i} \{MI(\omega, ef_j)\} < MI(\omega, ef_i) \\ \quad \text{and } MI(\omega, ef_i) > \alpha(\alpha = 0.5) \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

因此，对于表情符 ω，如果 $MI(\omega, ef_i)$ 中存在唯一的最大值，那么可以判断 ω 的情绪类别为 ef_i。否则，无法判别。

(4)当 $ef(\omega)$ 不为 0 的时候，将表情符 ω 添加到 $ef(\omega)$ 类中。

(5)如果未知表情符集为空或者未知表情符集中所有表情符的情绪类别判定都为 0，则程序停止，否则回到(2)。

为了应用互信息法，从新浪微博首页表情符中默认的常用表情中提取少量人工可明确判断为表达六种情绪的典型符号作为种子。表 2 列出了种子的数量以及例子。并用新浪 API 获得表情符共 1 764 个，构成未知情绪的表情符集合，称为目标表情符集。进一步下载了 20 000 条含有表情符的微博语料库，其中使用了 463 个表情符，共计 53 652 次。利用这个语料库，应用互信息法得到表情符词典，称为 EmoDic。

结果表明大量人工难以判断的表情符被划分到了具体的情绪种类，如表 2 中的“[奥特曼]”、“[感冒]”、“[汗]”。值得注意的是所选的惧的两个种子表情符在上述表情符语料库中都未出现，因而互信息法未能选出更多的表达惧的表情符。

表 2 互信息法产生 EmoDic

情绪类别	种子个数	示例	EmoDic 中表情符个数	示例
喜	4	[哈哈]	139	[爱你]、[奥特曼]
哀	3	[悲伤]	78	[失望]、[委屈]
怒	3	[怒]	32	[闭嘴]、[抓狂]
惧	2	--	2	--
恶	1	[吐]	32	[感冒]、[骷髅]
惊	1	[吃惊]	30	[汗]、[xkl 石化]

3 基于词典的微博情绪分析方法

文中提出一种基于规则的方法，利用词典进行微博情绪自动分析，其中词典包括 DUTIR 和 EmoDic 两个部分。

3.1 情绪分析规则

给定一条微博文本 t，假设待判断的情绪种类集合为 $E = \{e_1, \cdots, e_i, \cdots, e_m\}$，其中 m 为情绪类别的总数，那么对 t 的情绪判断过程如下：

(1)使用中文分词系统对 t 进行分词处理得到单词序列 q。

(2)对 q 中的单词/表情符与情绪词典中的情绪词/表情符进行匹配；对于每种情绪类别 e_i，统计匹配到的该类的情绪词/表情符个数 n_i。

(3)对于每种情绪类别 e_i，计算 t 所对应的情绪值，公式如下：

$$e(t) = \begin{cases} e_i, n_i > \max\limits_{j \neq i \cap 1 \leqslant j \leqslant m} \{n_j\} \\ 0, \text{otherwise} \end{cases} \quad (3)$$

对于微博文本 t，如果情绪词/表情符个数集合 $\{n_1, \cdots, n_i, \cdots, n_m\}$ 中存在唯一的最大值，那么可以判断 t 的情绪。否则，无法判别 t 的情绪类别。相应的，如果 t 未匹配到词典中任何的情绪词/表情符，t 的情绪也无法判别。对于后一种情况来说 t 未被词典覆盖。定义词典的覆盖率为被词典所覆盖的微博的条数占微博总条数的百分比。

3.2 否定用法

在情绪的表述中否定词的出现常常会对所表达的情绪产生影响。在情绪分析中制定恰当的否定规则，将有效减少因为否定造成的情绪倾向的误判，从而提高情绪识别的准确率。对于细致的情绪识别，否定的情况比较复杂。比如，“害怕”表示惧，但是“不会/让/人/害怕”表示不害怕，但很难将这个否定用法归入六种情绪中的一种。“开心”表达了高兴的情绪，但是有否定词“不”的修饰“不/开心”，则既可能表达伤心难过，又可能表达生气。针对这些情况，文中采取中性化策略来处理否定用法。当一个情绪词被否定词修饰

时,忽略该词的情绪色彩,即按中性词对待。算法中将否定词的作用范围定义为该词前后的标点符号之间。另外,否定用法还有两种特殊情况:

(1)当否定词出现在句尾时往往并不起否定作用。如"很多/人/喜欢/吃/好/不/。"

(2)双重否定多表示肯定。如"我/真/没/办法/不/爱/他/!"这句中的"没"和"不"不但没有起到否定的作用,反而加强了"爱"的情绪。

这两种情况下忽略否定词。否定用法处理的具体算法如下:

输入:一条微博 t 分词后得到分词序列 $q(q_1,\cdots,w_1,\cdots,q_i,\cdots,w_j)$ //其中 w_i 指标点

输入:NegationList//否定词表

DUTIR$_j$ //情绪词表,下标表示其情绪类别

输出:t 的情绪类别 e_i

初始化 for all $n_i \leftarrow 0$// n_i 是情绪词计数器

for all $z_i(q_i,\cdots,q_j) \leftarrow$ 相邻的两个标点 w_i 和 w_{i+1} 之间的单词序列

for all z_i do

否定词计数器 $c \leftarrow 0$

for all $q_i \in z_i$

if $q_i \in$ NegationList then

$c \leftarrow c + 1$

else if $q_i \in$ DUTIR$_j$ then

if $c\%2 == 0$ then

$n_j \leftarrow n_j + 1$

end if

$c \leftarrow 0$

end if

end for

end for

if $n_i > n_j \in \{n_1,\cdots,n_{i-1},n_{i+1},\cdots\}$ then return e_i

else return 0

end if

4　微博情绪分析数据集

文中使用新浪 API 抓取微博文本,按话题选取微博,并由两名标注人员各自独立进行情绪标注。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3显示了十种话题微博在六种情绪类别中的分布。

如表3所示,一个事件往往存在占主流的情绪。比如,主题为"武广高铁"的微博文本中喜的情绪占大多数,达到75.2%。"燃油涨价"中怒的情绪占大多数,为71.2%。"KFC 异物"中恶的情绪占多数,为59%。可以看出,情绪分析清晰地显示出大众对社会热点问题的反应。

表3　十种话题微博在各类情绪中的分布

情绪类别	话题										
	科比	燃油涨价	留学生澳洲遇袭	武广高铁	普京离婚	理想与现实	复旦投毒	樱桃蛆虫	KFC异物	少年父亲	合计
喜	210	19	13	152	22	18	7	32	--	--	473
哀	12	14	20	23	18	31	41	9	5	--	173
怒	33	99	76	24	1	--	11	19	13	--	276
惧	1	7	101	3	--	2	12	17	4	--	147
恶	--	--	--	--	7	1	3	59	75	--	145
惊	--	--	--	--	42	--	7	9	30	33	121
合计	256	139	210	202	90	52	81	145	127	33	1 335

5　实验结果与分析

文中使用了 NLPIR[14] 汉语分词系统(ICTCLAS2013)。利用其中的用户词典功能将表情符以"[嘻嘻] joy"的格式存储进用户词典。

5.1　表情符词典 EmoDic 测试

本节对互信息法提取的表情符进行测试。这里所用的词典中仅包括表情符部分。为了测试互信息法的有效性,从新浪微博首页表情符中默认的常用表情中人工选取所有能判断为表达六种情绪的符号形成人工词典。实验比较了人工词典以及用互信息法得到的表情符构成的词典 EmoDic,结果如表4所示。

表4　EmoDic 测试结果

情绪类别	精确度(precision)		召回率(recall)		F-值	
	人工词典	EmoDic	人工词典	EmoDic	人工词典	EmoDic
喜	94.9	94	35.5	39.7	51.7	55.9
哀	82.5	82.1	30.1	37	44.1	51
怒	95.3	73.9	14.9	18.5	25.7	29.6
惧	0	0	0	0	0	0
恶	88.1	83.3	35.9	34.5	51	48.8
惊	80	64.8	23.1	28.9	35.9	40

由于人工方法及互信息法均未选出表达惧所特有的表情符,因而仅利用表情符来识别情绪时惧类的精确度和召回率都为零。如表4所示,两个词典对喜和哀的判断都取得了较高的精确度。而人工词典在其他三种情绪中取得了更高的精确度。召回率方面,EmoDic 在除恶外的其他情绪中的表现都优于人工词典。说明互信息法选出的表情符能够有效识别更多带有情绪色彩的微博。同样的结果也出现在最终的 F-值中。互信息词典取得了平均高于人工词典5个百分点的 F-值。

5.2　否定用法测试

文中对于微博情绪表达中的否定用法采取了中性

化策略进行处理。这里测试了否定规则在识别微博情绪中的作用。采用的词典为DUTIR(不包括表情符词典部分)。实验结果表明,不应用否定规则时取得了47.1%的准确率(accuracy),而添加否定规则后准确率达到48.5%,提高了约1.5个百分点。否定规则对情绪判断的准确率有所提高,但提高不多。

5.3 基于词典的微博情绪分析

本节对基于词典的微博情绪分析方法进行了综合评估。表5显示了在微博数据集上的准确率和覆盖率。其中,DUTIR指仅使用词典DUTIR,+EmoDIc指添加表情符部分EmoDIc。而*2为将修饰内容权重增加一倍。

表5 基于词典的微博情绪分析结果

方法	准确率/%	覆盖率/%
DUTIR	47.1	63.6
DUTIR+否定	48.5	63.6
DUTIR+EmoDIc	60.9	79.0
DUTIR+EmoDIc*2	62.8	79.0
DUTIR+否定+EmoDIc*2	63.9	79.0

由表5可以看出,在传统的情绪词词典中加入互信息法得到的表情符词典不仅大幅提升了系统的准确率(约15个百分点)且覆盖率也有很大增加(约16个百分点)。在微博数据集中含有表情符的微博占全部微博的36%。可以看出,表情符是微博情绪表达的重要组成部分,充分利用表情符对识别微博情绪起到了重要的作用。另外,当添加了表情符词典后,应用否定规则对情绪判断的准确率进一步提高。

表6给出了六种情绪的F-值。DUTIR对喜的判断结果最高,而其他情绪中都比较低。尤其是怒和惊,分别只有3.6%和4.7%。这主要有两个原因:一方面DUTIR中这两类词的数量比较少,只占到了情绪词总数的约1%。在和实验数据集中的微博的匹配中,怒和惊两类情绪词与微博文本匹配分别只匹配到10和9个,进一步说明DUTIR中这两类情绪词对微博文本覆盖很有限。另一方面表达惊的微博中经常出现"不会吧"或"不是吧"。这些表达方式中不包含情绪词,因而无法与词典匹配。在词典中增加表情符部分EmoDic后六种情绪的平均F-值提高了约32个百分点。在此基础上应用否定规则进一步提高了F-值,达到平均56.3%。六种情绪中喜的F-值最高,表明该类无论是情绪词还是表情符的收集都更全面。

表6 四种方法的F值结果

情绪类别	DUTIR	DUTIR+否定	DUTIR+EmoDIc*2	DUTIR+否定+EmoDIc*2
喜	49.2	50.1	76.3	77.4
哀	18.9	21.6	64.8	67.9
怒	3.6	2.9	37.8	37.4
惧	31.6	33.3	40.3	45.2
恶	33.6	32.4	57.8	57.2
惊	4.7	4.7	53.5	52.8
平均值	23.6	24.2	55.1	56.3

6 结束语

微博已成为个人表达观点、分享心情的重要平台。自动分析微博的情绪倾向有助于预测事件走向、对大众需求做出迅速及时的反应,也能够帮助微博用户分析自身的情绪变化以调整个人状态。文中建立了一个包含喜、哀、怒、惧、恶、惊六种情绪的人工标注微博数据集。并提出了基于词典的规则方法来识别微博中的这六种情绪。方法中针对情绪分析的两个重要线索表情符和否定用法分别制定了相应的策略。利用互信息建立了表情符词典,建立了否定用法的处理规则以消除被否定的情绪词对整个微博情绪判断的影响。实验表明,传统的情绪词典虽然收录了大量词汇,但是对于社交媒体如微博文本情绪分析的覆盖率和准确率都不高。而表情符词典的应用显著地提高了微博情绪分析的精度。

参考文献:

[1] Cosatto E, Ostermann J, Graf H P, et al. Lifelike talking faces for interactive services[J]. Proceedings of the IEEE, 2003, 91(9): 1406-1429.

[2] Ryan S, Scott B, Freeman H, et al. The virtual university: the Internet and resource - based learning [M]. London, UK: Kogan Page, 2000.

[3] Abbasi A. Affect intensity analysis of dark web forums[C]// Proc of IEEE conference on intelligence and security informatics. New Brunswick, NJ: IEEE, 2007: 282-288.

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

[5] SemEval2007[EB/OL]. 2007. http://nlp.cs.swarthmore.ed-u/semeval/.

[6] 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.

[7] Bellegarda J R. 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.

[8] Mohammad S. Portable features for classifying emotional text [C]//Proceedings of the NAACL HLT 2012. [s.l.]: [s.n.],

(下转第36页)

90%以上。实验结果表明,添加语气词和标点特征可以从不同程度上提高情感分类精度。

表2 选取不同文本特征的分类精度

特征编号	提取的特征	正确分类文本数/测试文本总数	分类精度(Accuracy)/%
0	情感词	174/200	87
1	形容词	166/200	83
2	副词	146/200	73
3	动词	100/200	50
4	语气词	100/200	50
5	标点	104/200	57
0+4	情感词+语气词	197/200	98.5
0+5	情感词+标点	185/200	92.5
0+4+5	情感词+语气词+标点	199/200	99.5
1+4	形容词+语气词	172/200	86
1+5	形容词+标点	172/200	86
1+4+5	形容词+语气词+标点	169/200	84.5
2+4	副词+语气词	148/200	74
2+5	副词+标点	143/200	71.5
2+4+5	副词+语气词+标点	148/200	74
4+5	语气词+标点	128/200	64

传统的文本情感分类大多只提取出了情感词作为基本的特征项,且在使用停用词表时把类似于“呢”“啊”的语气词和类似于“!”“~”的标点符号过滤掉,忽略了语气词和标点对情感表达的重要作用。文中的实验研究表明了虽然语气词和标点单独作为特征项时效果并不理想,但与其他文本特征组合使用时,则可以提高情感分类的精度。

4 结束语

由于人类情感的复杂性,再加上存在着文本的语义歧义等难题,文本情感倾向性识别具有较高的难度。文中主要研究了选取不同的文本特征对文本情感分类精度的影响。研究的文本特征包括了传统实验中忽略掉的语气词和标点特征。实验结果显示,选取情感词、形容词、副词作为特征项对情感分类具有较好的效果,在此基础上添加语气词和标点特征可以有效地提高情感分类的精度。文中提取的文本特征没有考虑句型、修辞等语义信息,这些以后还需要进一步深入研究。

参考文献:

[1] 黄萱菁,赵 军.中文文本情感倾向性分析[J].中国计算机学会通讯,2008,4(2):39-46.

[2] Kim S M, Hovy E. Automatic detection of opinion bearing words and sentences[C]//Proceedings of IJCNLP-05. [s. l.]:[s. n.],2005:61-66.

[3] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques[C]//Proceedings of the 2002 conference on empirical methods in natural language processing. New Jersey: ACL, 2002:79-86.

[4] 吴 琼,谭松波,张 刚,等.跨领域倾向性分析相关技术研究[J].中文信息学报,2010,24(1):77-83.

[5] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11):613-620.

[6] Lewis D D. An evaluation of phrasal and clustered representations on a text categorization task[C]//Proceedings of the fifteenth annual international ACM SIGIR conference on research and development in information retrieval. [s. l.]:[s. n.],1992:37-50.

[7] Sharma A, Dey S. A comparative study of feature selection and machine learning techniques for sentiment analysis[C]//Proceedings of the 2012 ACM research in applied computation symposium. San Antonio, Texas: ACM, 2012:1-7.

[8] 王素格,魏英杰.停用词表对中文文本情感分类的影响[J].情报学报,2008,27(2):175-179.

[9] 于津凯,王映雪,陈怀楚.一种基于N-gram改进的文本特征提取算法[J].图书情报工作,2004,48(8):48-50.

[10] 刘丽珍,宋瀚涛.文本分类中的特征选取[J].计算机工程,2004,30(4):14-15.

[11] 姜 鹤,陈丽亚.SVM文本分类中一种新的特征提取方法[J].计算机技术与发展,2010,20(3):17-19.

[12] 张希娟,王会珍,朱靖波.面向文本分类的基于最小冗余原则的特征选取[J].中文信息学报,2007,21(5):56-60.

[13] 李媛媛,马永强.基于潜在语义索引的文本特征词权重计算方法[J].计算机应用,2008,28(6):1460-1462.

[14] 刘金岭.基于主题的中文短信文本分类研究[J].计算机工程,2010,36(4):30-32.

(上接第32页)

2012:587-591.

[9] Golder S A, Macy M W. Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures[J]. Science, 2011, 333(6051):1878-1881.

[10] Paltoglou G, Twitter T M. MySpace, Digg: unsupervised sentiment analysis in social media[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(4):66-66.

[11] 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.

[12] 杨 亮,林 原,林鸿飞.基于情感分布的微博热点事件发现[J].中文信息学报,2012,26(1):84-90.

[13] 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.

[14] NLPIR[EB/OL]. 2013. http://ictclas.nlpir.org/.