基于词典与机器学习的中文微博情感分析研究

孙建旺 吕学强 张雷瀚
(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)

摘 要 随着 Web2.0 时代的兴起,与微博相关的研究得到学术界和工业界的广泛关注。选取微博文本中的动词和形容词作为特征;提出基于层次结构的特征降维方法;采用设计的基于表情符号的方法计算特征极性值;在此基础上,提出基于特征极性值的位置权重计算方法,借助 SVM 作为机器学习模型将微博文本分为正面、负面和中性三类。实验结果表明,提出的方法能够比较有效地对中文微博文本进行情感分类。

关键词 微博 表情符号 极性值 位置权重 情感分类

中图分类号 TP391.1 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2014.07.045

ON SENTIMENT ANALYSIS OF CHINESE MICROBLOGGING BASED ON LEXICON AND MACHINE LEARNING

Sun Jianwang Lü Xueqiang Zhang Leihan
(*Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China*)

Abstract Along with the rising of Web2.0 age, the studies related to microblogging have drawn wide attentions from both the academia and industry communities. By selecting verbs and adjectives in microblogging texts as the features, we put forward a hierarchical structure-based feature dimensionality reduction approach. The designed emoticon-based method is adopted to calculate the feature polarity value. On this basis, the position weight calculation method based on feature polarity value is proposed. And with the help of SVM as the machine learning model, the approach classifies the microblogging texts into positive, negative and neutral categories separately. Experimental results show that the proposed approach can effectively make sentiment classification on Chinese microblogging texts.

Keywords Microblogging Emoticon Polarity value Position weight Sentiment classification

0 引 言

微博是一种新的信息发布及社交网络平台。自问世以来,迅速吸引了大众的眼光,得以蓬勃发展。据 CNNIC 统计显示,截止 2011 年 12 月底,我国微博用户数达到 2.5 亿,较上一年底增长了 296.0%,网民使用率为 48.7%[1]。微博消息量大、更新速度快,吸引了大批学者对其进行研究,针对微博的自然语言处理研究已成为一个新的研究热点及前沿性课题,微博情感分析就是其中的一个热点课题。微博情感分析是将微博文本按其主观倾向性将其分为正向、负向和中性三类。

目前,在情感分析方面,主要使用的技术分为两大类:一类是采用情感词典的方法[2],借助情感词典统计待分析文本中正向情感词和负向情感词的数目,根据他们的差值分析文本的情感极性;另一类是采用机器学习的方法[3],标注训练语料和测试语料,使用支持向量机、最大熵、KNN 等分类器进行情感分类。Wang[4]等人构建一个 Twitter 情感分析系统,能够实时地对有关总统选举的评论信息进行情感倾向性分析。Agarwal[5]等人以词语的极性及其词性为特征,借助树内核模型对微博文本进行情感分类研究,并取得了一定的成果。Jiang[6]等人采用主题相关和无关的方式对微博文本进行情感极性分类,分为正向情感和负向情感。中文微博与英文微博有较大的区别,中文微博文本的主题较为发散,内容更加丰富繁杂,行文习惯与英文微博也有明显的不同,导致上述方法不能很好地应用于中文微博文本的情感分类研究。刘志明等人[7]使用三种机器学习算法、三种特征选择方法以及三种特征权重计算方法对微博进行了情感分类方面的研究,但该方法没有考虑中文微博文本的行文特点,忽略了微博中的表情符号对整条微博文本情感极性的影响。谢丽星[8]等人提出了基于层次结构的多策略中文微博情感分析方法,与基于表情符号的规则方法相比,分类效果有了一定的提高,但该方法忽略了微博文本中特征的极性信息。基于情感词典的方法无法解决未登录词问题。中文微博文本内容较短、主题发散、未登录词多、用语不规范等特点,导致目前针对中文微博文本的情感分类效果不理想。

鉴于基于词典和基于机器学习的方法各自存在的不足,针对中文微博内容较短、口语化严重、主题分散等特点,提出了基于词典和机器学习相结合的方法,用于中文微博方面的情感分

收稿日期:2013 - 01 - 19。国家自然科学基金项目(61171159;61271304);北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目(KZ201311232037)。**孙建旺**,硕士生,主研领域:中文信息处理。**吕学强**,教授。**张雷瀚**,硕士生。

析研究。

1 极性词典的构建

基于词典的情感分析方法,需要一个标有极性的情感词典。在中文情感分析方面,目前国内还没有一部像英文《General Inquirer》[9]比较完善的情感词典;另外,情感词的极性往往受到其前面的极性副词的影响,同时表情符号对于整条微博文本的极性具有重要影响,因而还需要一部极性副词词典和表情符号词典。

1.1 情感词典的构建

借助已有的资源尝试建立一个相对比较完善的中文情感词典。已有资源主要包括:《学生褒贬义词典》中的正负情感词,《知网》提供的正负情感词以及搜狗实验室提供的互联网词库SogouW。

对《知网》提供的情感词进行人工筛选,并进行词性标注,得到一部词典,简称"知网词典",用 *HNetD* 表示。同样对《学生褒贬义词典》中的词语进行人工筛选,得到学生词典,用 *StuD* 表示。对 *HNetD* 和 *StuD* 进行合并去重操作,得到候选情感词典,用 *CanD* 表示,则可以用如下公式表示:

$$CanD = HNetD \cup StuD$$

借助 SogouW 对 *CanD* 做进一步的筛选和完善操作,最终得到本文需要的情感词典(用 *MD* 表示)。为了直观,将极性值限定在[-1,1]之间,本文规定,正向情感词与负向情感词的极性值分别为 0.8 和 -0.8。*MD* 中的词条是由情感词(用 *mpw* 表示)、词性(用 *mpos* 表示)和极性值(用 *mpol* 表示)构成的三元组,$MD = \{ < mpw_1, mpos_1, mpol_1 >, < mpw_2, mpos_2, mpol_2 >, \cdots, < mpw_{n1}, mpos_{n1}, mpol_{n1} > \}$。构建的部分情感词典如表 1 所示。

表 1 情感词典

mpw	mpos	mpol
爱戴	v	0.8
爱国	v	0.8
暧昧	a	-0.8
安详	a	0.8
…	…	…

1.2 极性副词词典的构建

定义 1 修饰情感词的副词称为极性副词。在进行极性分析时,情感词的极性强度通常受到其前面的极性副词的影响较大,极性副词包括极性否定副词和极性程度副词。

在文献[10]中对否定副词范围界定的基础上,加入了若干否定副词,最终选取"不、甭、别、不必"等 36 个否定副词构建极性否定副词词典,并规定极性值为 -0.8。采用蔺璜[11]对程度副词范围的界定构建极性程度副词词典,用 *PEnD* 表示,根据程度的强弱将极性值依次标为 0.9、0.7、0.5 和 -0.5。*PEnD* 中的词条是由程度副词(用 *ea* 表示)和极性值(用 *pol* 表示)构成的二元组,则 $PEnD = \{ < ea_1, pol_1 >, < ea_2, pol_2 >, \cdots, < ea_{n2}, pol_{n2} > \}$。构建的部分副词词典如表 2 所示。

表 2 程度副词词典

ea	pol
最为	0.9
更加	0.7
比较	0.5
稍为	-0.5
…	…

1.3 表情符号词典的构建

微博文本,尤其是评论性微博与普通文本相比,除了具有内容短、未登录词多、口语化严重等特点外,还通常含有较为丰富的表情符号。微博文本中的表情符号能够较为简洁、直观地表达人们情感,对于反映作者的情感或立场具有十分重要的作用。因而本文借助新浪微博中的表情符号构建表情符号词典,并根据表情符号表达的情感倾向性将其分为正向和负向两类。

2 词典与机器学习相结合的微博文本情感分析

针对微博文本内容较短、未登录词较多、口语化严重等特点,提出了基于词典与机器学习相结合的方法,用于中文微博文本的情感倾向性分析研究。采用向量空间模型表示微博文本,以动词、形容作为特征,根据提出的基于层次结构的特征降维方法对特征空间进行降维,借助构建的词典计算特征的极性值,根据提出的基于表情符号的方法计算特征极性值,采用设计的基于特征极性值方法计算位置权重,借助 SVM 机器学习模型将中文微博文本分为正向、负向和中性三类。

2.1 基于层次结构的特征降维方法

通过对微博文本的分析发现,文本中的动词和形容词往往是能够反映文本情感倾向性的情感词,因而选取动词和形容词作为特征。特征空间包含了微博文本集合中所有的动词和形容词,当训练文本集较大时,将导致特征空间的维数非常高。同时,中文微博文本通常仅含有十几个词甚至几个词或表情符号,导致特征向量中绝大多数维上的值为零,造成特征空间严重的数据稀疏性。因而必须对特征空间进行降维。

目前,特征降维有特征选择和特征抽取两种方法。由于特征抽取,存储和计算量大,不适合对文本的处理。χ^2 统计法在特征选择方面具有良好的性能,经过 χ^2 统计法的特征降维后,特征空间的高维性与数据稀疏性依然比较严重,需要对特征空间做进一步的降维处理。层次聚类算法[12]在词语聚类方面具有良好的效果。据此,提出融合 χ^2 统计法与层次聚类算法的层次结构的特征降维方法。采用 χ^2 统计法进行特征选择,初步降低特征空间的维数,借助层次聚类算法对特征空间进行降维,进一步降低特征空间的维数,最终达到特征降维的目的。

设类簇 i 用 c_i 表示,类簇间的平均相似度用 $sim(c_i, c_j)$ 表示,$sim(c_i, c_j)$ 计算方法如下:

$$sim(c_i, c_j) = \frac{\sum_{w_1 \in c_i} \sum_{w_2 \in c_j} sim(w_1, w_2)}{|c_i| \times |c_j|} \tag{1}$$

其中,w_1 与 w_2 分别是类簇 c_i 与 c_j 中的特征项,$sim(w_1, w_2)$ 为 w_1 与 w_2 的语义相似度[13],$|c_i|$ 为类簇 i 中特征项的个数。

设向量空间用 *FS* 表示,为了提高对 *FS* 的聚类效果,提出 *FS* 中特征项的褒贬义性将 *FS* 分为褒义、贬义与中性三个子空间,分别用 *PosFs*、*NegFS* 和 *NeuFS* 表示。然后分别对 *PosFs*、*NegFS* 和 *NeuFS* 三个子空间进行聚类操作。最终,将三个子空间的聚类结果合并在一起,可以用如下公式表示:

$$\begin{aligned} Cluster(FS) = Merge(&Cluster(PosFS) \\ &Cluster(NegFS), Cluster(NeuFS)) \end{aligned} \tag{2}$$

其中,$Cluster(X)$ 是对向量空间 X 进行聚类操作,$Merge(x, y, z)$ 是对 x、y 和 z 进行合并操作。

2.2 基于表情符号的特征极性值计算

定义 2 出现在情感词典中的特征称为极性特征。

定义 3 不在情感词典中的特征称为中性特征。

定义 4 极性特征的极性值为由情感词及其修饰副词构成的极性短语的极性值。极性特征的极性值的绝对值越大,其情感极性越强,反之越弱。参考文献[14]给出了关于极性短语的极性值计算方法,计算方法如表 3 所示。

表 3　极性短语的极性值计算

极性短语	强度计算公式	例句	强度
S = PW	E(PW)	她长得好看	0.80
S = NA + PW	E(PW) * E(NA)	她长得不好看	-0.64
S = NA1 + NA2 + PW	E(PW) * E(NA2) * E(NA1)	她长得不是不好看	0.512
S = DA + PW	若 PW 是正面: E(PW) + (1 - E(PW)) * L(DA) 若 PW 是负面: E(PW) + (-1 - E(PW)) * L(DA)	她长得很好看她长得很丑	0.94 -0.94
S = DA2 + DA1 + PW	E(PW) + (1 - E(PW)) * L(DA1) + [1 - E(PW) - (1 - E(PW)) * L(DA1)] * L(DA2)	她长得十分很好看	0.982
S = NA + DA + PW	E(PW) + (1 - E(PW)) * (L(DA) - 0.2)	她长得不很好看	0.90
S = DA + NA + PW	E(PW) * E(NA) + (-1 - E(PW) * E(NA)) * L(DA)	她长得很不好看	-0.892

一个极性特征在一个文本中可能出现若干次,每次出现时,其前面的极性副词及极性副词之间的顺序都不尽相同,导致每次计算得到的极性值都有所不同。由此,本文取每次计算得到的极性值的算术平均值作为该极性特征的最终极性值。则极性特征 t_i 的极性值计算公式如下:

$$E(t_i) = \frac{\sum_{k=1}^{tf_{ij}} E(t_{ik})}{tf_{ij}} \tag{3}$$

其中,$E(t_{ik})$表示第 k 次出现在文本中特征项 t_i 的极性值。

评论性微博文本中的表情符号对于反映作者的情感或立场具有十分重要的作用,能够增强整条文本的情感极性。本文假设命题"如果微博文本中正向表情符号的数目大于负向表情符号的数目,则微博文本中的正向极性特征的极性值得到增强,反之,负向极性特征的极性值得到增强"成立。设特征 t_i 得到增强后的极性值用 $IE(t_i)$表示,则本文设计的基于表情符号的极性特征的极性值计算方法如下:

$$IE(t_i) = \begin{cases} [1 + (pN - nN)] \times E(t_i) & \text{if } pN - nN \geqslant 0 \quad E(t_i) \geqslant 0 \\ [1 - (pN - nN)] \times E(t_i) & \text{if } pN - nN < 0 \quad E(t_i) < 0 \end{cases} \tag{4}$$

其中,pN 为微博文本中正向表情符号的数目,nN 为负向表情符号的数目。

通常来讲,中性特征的极性值应该为零,但为了和出现次数为零的特征项区分开来,本文规定中性特征的极性值是个非常小的常数,公式如下:

$$IE(t_i) = \gamma \tag{5}$$

其中,γ 为 0 到 0.1 之间的一个常数。

2.3　基于特征极性值的位置权重计算

特征权重用于衡量某个特征项在文档表示中的重要程度或者区分能力的强弱。目前最常用的是 TF-IDF 方法,TF-IDF 的计算公式如下:

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_i} \tag{6}$$

其中,tf_{ij}表示特征项 t_i 在微博文本 wb_j 中出现的次数,n_i 表示含有特征项 t_i 的文档数,N 为全部微博文本的数量,上面的公式进行归一化后变为:

$$w_{ij} = \frac{tf_{ij} \times \log \frac{N}{n_i}}{\sqrt{\sum_{t_i \in wb_j} \left(tf_{ij} \times \log \frac{N}{n_i} \right)^2}} \tag{7}$$

通过对中文评论性微博文本的分析发现,文本中的首句、中间部分以及尾句的极性对整条微博文本极性的贡献不尽相同,微博文本中的首句、尾句对整条微博文本极性的贡献相对较大。而式(7)忽略了这些信息,这也是造成微博文本情感分类效果不佳的重要原因之一。

据此,引入位置系数,对式(7)进行改进。将微博文本(用 wb_j 表示)分为首句(用 p_s 表示)、中间部分(用 p_m 表示)以及尾句(用 p_e 表示)三部分,则 $wb_j = (p_s, p_m, p_e)$。在计算特征权重时,每部分赋予一定的权值 pw_r,用于体现该部分的特征对 wb_j 的贡献程度,称 pw_r 为 wb_j 的 p_r 部分的位置系数。位置系数满足如下约束:

$$\sum_{r \in \{s,m,e\}} pw_r = 1 \tag{8}$$

其中,$0 < pw_r < 1$。则特征项 t_i 在 wb_j 中出现的频次 tf_{ij}可以用如下方法计算:

$$tf_{ij} = \sum_{r \in \{s,m,e\}} pw_r \times tf_{ir} \tag{9}$$

其中,tf_{ir}为特征项 t_i 在 wb_j 的 p_r 部分中出现的次数。则引入位置系数的 TF-IDF 的计算公式变为:

$$w_{ij} = \frac{\sum_{r \in \{s,m,e\}} pw_r \times tf_{ir} \times \log \frac{N}{n_i}}{\sqrt{\sum_{t_i \in D_j} \left(\sum_{r \in \{s,m,e\}} pw_r \times tf_{ir} \times \log \frac{N}{n_i} \right)^2}} \tag{10}$$

式(10)采用简单的词频统计来计算特征项的权重,而忽略了特征项本身的极性,然而特征项本身的极性通常反映了作者的观点,对判定文本的性质有十分重要的作用。因而,本文将特征项的极性值作为权重计算的一部分,在式(10)的基础上做进一步的改进。

由此,本文设计的基于特征极性值的位置权重计算方法如下:

$$w_{ij} = \frac{\sum_{r \in \{s,m,e\}} pw_r \times tf_{ir} \times IE(t_i) \times \log \frac{N}{n_i}}{\sqrt{\sum_{t_i \in D_j} \left(\sum_{r \in \{s,m,e\}} pw_r \times tf_{ir} \times \log \frac{N}{n_i} \right)^2}} \tag{11}$$

2.4　情感分类算法

本文将微博文本分为正向、负向和中性三类,因而微博情感分析可以转化为文本分类问题。支持向量机 SVM 是近几年发展起来的新型分类方法,主要用于解决文本分类问题。SVM 模型的基本思想:将输入空间转换到一个高维空间,使非线性样本变得线性可分,最终求取最优分类面,防止对训练数据的过拟合。SVM 方法具有结构风险最小化的归纳原则,同时可以控制整个样本集的期望风险。因此它解决了其他机器学习方法对训练数据产生过适应的缺点,而且引入核函数的概念,根据前文所

提出的特征选择算法,获取合适的特征,可以很好地解决训练文本特征空间的高维性和数据稀疏性问题,在短文本分类方面相比其它机器学习模型具有一定优势。因而,本文借助 SVM 作为微博情感分类算法。

3 实验结果及分析

3.1 实验数据

利用新浪微博 API 抓取了两个领域的微博评论:名人新闻领域—关于"刘翔退赛"微博评论;产品领域—关于"iphone4s"的微博评论。经过去重后每个领域各选取了 2 000 条微博评论作为语料。人工对这些语料进行了标注,其中名人新闻领域中正向情感微博数为 686 条、负向情感微博数为 1 036 条、中性情感微博数为 278 条;产品领域中正向情感微博数为 954 条、负向情感微博数为 679 条、中性情感微博数为 367 条。语料的统计结果如表 4 所示:

表 4 语料的统计结果

主题	训练语料				测试语料				总数
	正	中	负	总数	正	中	负	总数	
刘翔退赛	486	178	636	1 300	400	100	200	700	2 000
iphone4s	554	267	479	1 300	350	100	250	700	2 000

分词之前,先对微博语料做了一下预处理。预处理包括:剔除重复的微博,删掉微博中"【"与"】"、"#"与"#"符号之间的内容(包括符号本身);除去"回复@ * * * *:"和"@ * * * *"部分内容;除去"("和")"之间的内容,除去 URL 链接和英文单词。分词以后,进行停用词过滤,排除干扰特征。

3.2 实验结果

本文以正确率、召回率、F 值和宏平均值作为评价指标。设 m_{right_i} 是正确的分到类别 c_i 中的微博文本的数量,m_{wrong_i} 是其他类的微博文本被错误的分到类别 c_i 中的微博文本的数量,m_{all_i} 是类别 c_i 中实际含有的微博文本的数量。

则类别 c_i 的正确率为:

$$precision_i = \frac{m_{right_i}}{m_{right_i} + m_{wrong_i}} \times 100\% \tag{12}$$

类别 c_i 的召回率为:

$$recall_i = \frac{m_{right_i}}{m_{all_i}} \times 100\% \tag{13}$$

类别 c_i 的 F 值为:

$$F_i = \frac{2 \times precision_i \times recall_i}{precision_i + recall_i} \tag{14}$$

分类的平均准确率可以由如下方法计算:

$$precision = \frac{\sum_{i=1}^{m} precison_i}{m} \tag{15}$$

分类的平均召回率可以由如下方法计算:

$$recall = \frac{\sum_{i=1}^{m} recall_i}{m} \tag{16}$$

宏平均值公式为:

$$MacroF = \frac{\sum_{i=1}^{m} F_i}{m} \tag{17}$$

为了验证本文方法的有效性,设计了三组实验,分别利用文献[15]中基于 SVM 的方法、文献[16]中基于情感词典的方法和本文提出的方法对中文微博文本进行情感分类研究。

其中,本文方法中位置权重不同部分的系数分别为 0.3、0.2、0.5,实验结果如表 5 – 表 7 所示。

表 5 对比实验结果 – 正确率

	刘翔退赛			iphone4s		
	SVM	情感词典	本文方法	SVM	情感词典	本文方法
正向情感微博	71.28%	67.65%	73.76%	70.09%	64.10%	73.45%
中性情感微博	54.74%	52.69%	59.38%	56.84%	47.52%	59.14%
负向情感微博	78.78%	74.69%	83.33%	65.35%	57.66%	69.96%
平均正确率	68.27%	65.01%	72.16%	64.09%	56.43%	67.52%

表 6 对比实验结果-召回率

	刘翔退赛			iphone4s		
	SVM	情感词典	本文方法	SVM	情感词典	本文方法
正向情感微博	69.5%	69%	74.5%	70.29%	64.29%	74.29%
中性情感微博	52%	49%	57%	54%	48%	55%
负向情感微博	80.75%	75.25%	83.75%	66.4%	57.2%	70.8%
平均召回率	67.41%	64.42%	71.75%	63.56%	56.5%	66.7%

表 7 对比实验结果-F 值

	刘翔退赛			iphone4s		
	SVM	情感词典	本文方法	SVM	情感词典	本文方法
正向情感微博	70.38%	68.32%	74.13%	70.19%	64.19%	73.87%
中性情感微博	53.33%	50.78%	58.17%	55.38%	47.76%	56.99%
负向情感微博	79.75%	74.97%	83.54%	65.87%	57.43%	70.38%
MacroF	67.82%	64.69%	71.95%	63.81%	56.46%	67.08%

3.3 结果分析

从表 5、表 6、表 7 可以看出,本文方法的正确率、召回率、F 值、$MacroF$ 值都明显高于另外两种方法;此外,基于情感词典分类效果明显低于另外两种方法。原因有以下几点:

(1) 与文献[15]中的基于 SVM 的方法相比,本文采用基于层次结构的特征降维方法,有效地降低了特征空间的维数,减少了特征空间的高维性与稀疏性对分类精度的影响。

(2) 针对中文微博文本的特点,提出了基于表情符号的特征极性计算方法,使得权重计算在中文微博情感分类方面与文献[15]中的方法相比,更加合理。

(3) 与文献[15]中的基于 SVM 的方法相比,引入了位置权重,将中文微博文本分为首句、中间部分和尾句三部分,且尾句对微博文本极性的贡献相对较大,一定程度上迎合了汉语的行文习惯(与英文微博不同)。

(4) 与文献[15]中的基于 SVM 的方法相比,在计算特征的

权重时,本文将特征的极性值作为权重计算的一部分,使得计算的权重在情感分类方面更加合理。

(5) 与基于情感词典的方法相比,本文方法解决了未登录词的问题。

(6) 中文微博文本内容较短、口语化严重、用语不规范、网络新词较多以及主题分散等特点严重影响基于情感词典方法的分类效果。

另外,从表中还可以看出,中性微博文本分类的正确率明显低于另外两类的正确率,表明分类的正确率与训练语料的规模成正相关。

从上表可以发现,"刘翔退赛"的分类效果优于"iphone4s"的分类效果,且在"iphone4s"的评论当中,基于情感词典的分类效果下降的最为明显,正确率下降了 8.58 个百分点。经过分析,原因有以下几点:

(1) 在对"刘翔退赛"新闻时间的评论中,使用的表达方式较为单一,结构相对简单,未登录词相对少一些。

(2) 在对"iphone4s"的评论中,涉及到的专业术语较多,而未登录词较多,句子结构相对复杂。

(3) 在对"iphone4s"的评论中,有些情感词的极性与情感词典中的极性不一致。例如"虽然非常智能,但玩时间长了手机就会变得很热…"中的"热"字,在情感词典中为正向情感词,而在该条微博中的极性表现为"负向"。

此外,从表 7 可以发现,对"iphone4s"评论的分类效果与"刘翔退赛"相比,文献[15]中基于 SVM 的宏平均值下降了 4.01个百分点,基于词典的方法下降了 8.23 个百分点,本文方法下降了 4.87 个百分点。表明本文方法对不同领域的中文微博评论情感分类效果的稳定性弱于基于 SVM 方法,但相差不大,明显优于基于情感词典的方法。

4　结　语

本文提出了基于词典与机器学习相结合的方法解决中文微博情感分类问题。提出了基于层次结构的降维方法,针对微博文本的特点,设计了基于表情符号的特征极性值计算方法,在此基础上,设计基于特征极性值的位置权重计算方法,借助 SVM 作为机器学习模型将微博文本分为正面、负面和中性三类。当然本文还存在较大的提升空间。例如网络中的一些新词,如"给力"、"坑爹"、"鸭梨"等,现有的分词系统是无法识别的,而它们对中文微博情感分类有比较大的影响,因而需要采用新的方法来匹配识别他们。

参　考　文　献

[1] CNNIC(中国互联网信息中心). 第 29 次中国互联网络发展状况统计报告[R]. 北京:中国互联网络信息中心(CNNIC),2012.

[2] Lunwei Ku, Tungho Wu, Liying Lee, et al. Construction of an Evaluation Corpus for Opinion Extraction[C]//NTCIR-5 Japan. 2005:513 - 520.

[3] Dasgupta S, Ng V. Mine the Easy. Classify the Hard:S Semi-Supervised Approach to Automatic Sentiment:Classification[C]//ACL'09:701 - 709.

[4] Hao Wang, Dogan Can, Abe Kazemzadeh. A system for real-time Twitter sentiment analysis of 2012 U. S. presidential election cycle[C]// Proceedings of the ACL 2012 System Demonstrations, 2012:115 - 120.

[5] Apoorv Agarwal, Boyi Xie, Ilia Vovsha. Sentiment analysis of Twitter data[C]// Proceedings of the Workshop on Languages in Social Media, 2011:30 - 38.

[6] Long Jiang, Mo Yu, Ming Zhou. Target-dependent Twitter sentiment classification[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011:151 - 160.

[7] 刘鲁,刘志明. 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] 谢丽星,周明,孙茂松,等. 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] http://www. wjh. harvard. edu/ ~ inquirer/.

[10] 郝雷红. 现代汉语否定副词研究[D]. 首都师范大学,2003.

[11] 蔺璜,郭姝慧. 程度副词的特点范围与分类[J]. 山西大学学报,2003,26(2):71 - 74.

[12] 段明秀,杨路明. 对层次聚类算法的改进[J]. 湖南理工学院学报:自然科学版,2008,21(2):28 - 29,36.

[13] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. 计算机语言学及中文信息处理,2007,31(7):59 - 76.

[14] 张成功,刘培玉,朱振方,等. 基于词典的中文倾向性分析报告[C]//第三届中文倾向性分析评测,2011:149 - 156.

[15] Na Jin Cheon, Khoo Christopher, Wu Paul Horng Jyh. Use of Negation Phrases in Automatic Sentiment Classification of Product Reviews[J]. Library Collections, Acquisitions and Technical Services, 2005, 29(2): 180 - 191.

[16] 张成功,刘培玉等,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,2012,47(3):47 - 50.

(上接第 144 页)

[11] Leuven Sebastiaan Van, Schevensteen Kris Van, Dams Tim, et al. An implementation of multiple region-of-interest models in H. 264/AVC [J]. Signal Processing for Image Enhancement and Multimedia Processing, 2008, 31(Ⅳ):215 - 225.

[12] Zhang Tianruo, Liu Chen, Wang Minghui, et al. Region-of-interest based H. 264 encoder for videophone with a hardware macroblock level face detector[C]//IEEE International Workshop on Multimedia Signal Processing, Rio De Janeiro, 5 - 7 Oct. 2009:1 - 6.

[13] Huang Chungming, Lin Chungwei. Multiple priority region-of- interest H. 264 video compression using constraint variable bitrate control[J]. Optical Engineering, 2009, 48(4):047004.

[14] JVT Reference Software, Version JM8. 6 [CP/OL]. http://iphome. hhi. de/suehring/tml/download/.

[15] Kannur Avin Kumar, Li Baoxin. Power-aware content-adaptive H. 264 video encoding[C]//IEEE International Conference on Acoustics, Speech and Processing (ICASSP), Taipei, 19 - 24 April 2009:925 - 928.

[16] Shatique Muhammad, Bauer Lars, Henkel Jörg. 3-Tier dynamically adaptive power-aware motion estimator for H. 264/AVC video encoding [C]//IEEE International Symposium on Low Power Electronics and Design, Bangalore, 11 - 13 Aug. 2008:147 - 152.