doi:10.3969/j.issn.1673-9833.2013.02.019

基于SVM的中文微博观点句识别算法

杜 锐，朱艳辉，鲁 琳，王文华，邓 程，喻魁兰

（湖南工业大学 计算机与通信学院，湖南 株洲 412007）

摘 要：针对中文微博中的海量文本，提出了利用领域观点词词典和支持向量机的方法对中文微博中的观点句进行识别。构建领域观点词词典，统计了表示中文微博观点句的5个特征，选取特征1, 2, 3, 4进行观点句识别，并将基于支持向量机的3种不同特征组合识别算法与基于领域观点词词典的识别算法进行对比。算法对比结果表明，基于支持向量机的算法对微博观点句的识别效果较好，准确率68.75%，召回率48.71%，*F*值57.02%。

关键词：中文微博；支持向量机；观点句

中图分类号：TP391 **文献标志码**：A **文章编号**：1673-9833(2013)02-0089-05

The SVM-Based Algorithm for Chinese Micro-Blog Opinion Sentence Identification

Du Rui，Zhu Yanhui，Lu Lin，Wang Wenhua，Deng Cheng，Yu Kuilan

（School of Computer and Communication，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：For the mass texts in the micro-blog, uses the dictionary of opinion words and the method of support vector machine (SVM) to recognize the opinion sentence in Chinese micro-blog. Constructs the dictionary of opinion words, counts five features of Chinese micro-blog opinion sentences, selects four features to recognize opinion sentences, as well as compares the SVM-based algorithm and the algorithm of opinion words dictionary. The contrast results show that the SVM-based method is best in identifying the micro-blog opinion sentence, and the accuracy is 68.75 %, the recall rate is 48.71 % and the *F*-measure is 57.02 %.

Keywords：chinese micro-blog；support vector machine；opinion sentence

1 相关研究

随着互联网的快速发展，微博已经成为人们获取信息的重要渠道。第29次中国互联网络发展状况统计报告[1]显示，截止2011年12月底，我国微博的用户数量达到2.5亿，较2010年底增长了296.0%，网民使用率为48.7%。微博用户数量的不断扩大以及政府对微博管理力度的加强，使得微博在互联网中的地位显得越来越重要。微博中网民对热点事件的观点能给政府制定新政策提供参考依据；消费者通过微博对自己购买的产品进行评论，能为商家对商品质量和服务进行改进提供参考。

收稿日期：2013-01-03

基金项目：国家自然科学基金资助项目（61170102），湖南省自然科学基金资助项目（10JJ3002），国家社会科学基金资助项目（12BYY045），教育部人文社会科学研究青年基金资助项目（09YJCZH019），中国包装总公司科研基金资助项目（2008-XK13）

作者简介：杜 锐（1987-），男，湖北仙桃人，湖南工业大学硕士生，主要研究方向为文本分类和信息检索，E-mail：578781015@qq.com

中文微博不同于其它的专业论坛。首先中文微博对字符上的限制（最多能输入140个汉字）使得微博文本较短，因而中文微博中可提取的观点句特征比较少，如："官员怎么教出这样的孩子！"为一条含有观点的中文微博，但其中能够表示观点的特征较少；其次，由于微博在评论信息上的口语化以及汉语在表达方式上的多样化等特点使得观点句的特征表现不明显，如："#疯狂的大葱#大葱它肿么了？"，该微博中没有明显的观点词，但该微博是观点句，且"肿么了"一词在汉语的正式表达中并不常见，因此，微博中可表示的观点句特征并不明显；最后，由于微博中网络用语较多、用语不规范等，使得微博中的观点句特征表现不稳定，同一网络观点词在不同领域所表现的观点特征并不相同。因此，中文微博观点句的识别比其它专业论坛的观点句识别难度大。

目前，在微博观点句识别方面，Alexander Pak等人[2]选取n-gram和微博中的词性标注作为特征，利用朴素贝叶斯分类器对微博中的观点句进行识别研究，并与支持向量机（support vector machine，SVM）、条件随机场2种分类器进行比较，实验结果表明，基于朴素贝叶斯的微博观点句识别效果较好。Luciano Barbosa等人[3]采用微博中的词性信息、词本身的主观性、词的情感极性以及否定词作为特征，训练分类器，对微博主客观性进行分类。D. Davidiv等人[4]提取Twitter中的标签和表情符号作为训练集，训练了一个类似KNN（K-nearest neighbor algorithm）的分类器，对微博情感极性进行分类，但实验的准确率不是很理想。

综上所述，本文提出了一种基于SVM的中文微博观点句的识别方法。首先，采用多种分词技术构建自定义词库[5]，将自定义词库与知网中情感词词典和观点词词典[6]分别进行比较，将其相同的词或包含情感词词典或观点词词典的词作为候选观点词，利用连词词典对候选观点词进行扩充，然后在扩充后的候选观点词中加入网络观点词构造观点词集。将观点词集、"？"、"！"、"？"与观点词是否同现、"！"与观点词是否同现作为候选特征，在候选特征中分别选取3组不同的特征组合训练SVM分类器，分析其训练结果，选取F值最高的一组特征作为分类特征，利用该分类特征的SVM分类器对测试集中的微博进行观点句识别。

2 领域观点词词典构建

不同领域之间所使用的观点词存在明显的差异，同一观点词在不同领域所表达的观点也不尽相同。如"该机采用了高端笔记本流行的金属机身，'城市流光'纹理，外形边角处理十分圆滑"与"沙溢扮演的郭洋港在大学时就充满梦想，一心想着毕业后能大展鸿图，性格却决定他无法像袁浩东那样变得圆滑，只能选择一再躲避现实"，这2个句子都出现了"圆滑"，但出现在不同的领域，因此，其表达的观点不同。领域观点词能较好地区分同一观点词在不同领域所表达的不同观点，因而能较好地区分观点句与非观点句。基于此，本文构建了领域观点词词典，判断分词后的微博文本中是否含领域观点词，如果有则为观点句，否则为非观点句。构建领域观点词词典方法如下：

首先，利用多重分词系统构建自定义词库；然后，将各个自定义词库与知网中的情感词词典比较，选取相同的词或包含情感词词典中的词作为候选观点词，将候选观点词与知网中的观点词词典合并，并去掉重复词，得到领域候选观点词词典；最后，利用连词词典，对领域候选观点词词典进行扩充。

根据领域观点词词典判断中文微博中是否含领域观点词的方法，具体实现步骤如下：

Step 1 利用ICTCLAS分词系统[7]对预处理后的微博文本进行分词；

Step 2 判断分词后的微博文本中是否有连词和领域候选观点词；

Step 3 如果微博中有连词和领域候选观点词，则记录连词和领域候选观点词的位置，否则，返回Step 2；

Step 4 如果连词和领域候选观点词位置的距离之差小于4，则以连词为中心，设置窗口大小为[−3, 3]，判断窗口内是否存在与领域候选观点词词性相同的词，如果有，则抽取该词，并将该词加入领域候选观点词词典中，否则，返回Step 2；

Step 5 在上述扩充后的领域候选观点词词典中加入网络观点词，并去掉重复的词，最终构建领域观点词词典。

3 特征提取

针对微博中观点句的特征稀少、表现不明显等问题，本文对训练集中1 000条中文微博的标点符号在观点句与训练集中所占的比例进行了统计分析，最终选取了表示中文微博观点句的5个特征。

将原始语料进行预处理后，利用极易分词器对处理后的语料进行分词，然后提取训练数据集中观

点句的特征。由于微博中的语句较短，其特征表现稀疏，观点词表现不是很明显，但实际上它是观点句，如：“怎么这么多李刚？”“为人官，又为人父母，难道就不懂教好孩子吗？”，所以仅通过中文微博中的观点词难以判断该微博是否是观点句。经统计分析，在训练语料集中，含“！”的语句占训练数据集的 28.10%，含“！”的观点句占微博观点句的 28.23%，因此，“！”可作为微博观点句的特征。在训练语料集中，含“？”的语句占 23.97%，含“？”的观点句占总观点句的 24.71%，“？”的特征表现较明显，因此，“？”也可作为微博观点句的特征。但是，仅通过语句中含“！”或“？”也难以准确地判断该句子是否是观点句，如：“希望长期关注！！！！！！！！！！”，虽然含有“！”，但这条微博不是观点句。为了解决仅通过句子中含“！”或“？”来判断观点句的不准确问题，本文通过判断情感词或观点词是否与“？”或“！”同现来识别观点句。

定义如下特征来判断一条微博是否是观点句。

特征 1　判断微博中是否含有观点词或情感词，即

$$f_i(x)=\begin{cases}1, \text{存在观点词};\\ 0, \text{反之}。\end{cases} \quad (1)$$

特征 2　判断微博中是否含有“？”，即

$$q_i(x)=\begin{cases}1, \text{存在“?”};\\ 0, \text{反之}。\end{cases} \quad (2)$$

特征 3　判断微博中“！”的个数，即

$$h_i(x)=n,\ n\geqslant 0。 \quad (3)$$

特征 4　判断微博中观点词与“！”是否同现，即

$$p_i(x)=\begin{cases}1, \text{观点词与“!”同现};\\ 0, \text{反之}。\end{cases} \quad (4)$$

特征 5　判断微博中观点词与“？”是否同现，即

$$g_i(x)=\begin{cases}1, \text{观点词与“!”同现};\\ 0, \text{反之}。\end{cases} \quad (5)$$

例如　训练集中有如下 3 条微博：

1. #官二代求爱不成将少女毁容#太嚣张了，就跟“我的爸爸是李刚”差不多了。

2. 希望长期关注！！！！！！！！！！

3. #官二代求爱不成将少女毁容#是挺气愤的！

这 3 条微博的文本特征向量见表 1。

表 1　示例微博文本的向量化表示

Table 1　Examples of micro-blog text represented by vectors

序号	特征 1	特征 2	特征 3	特征 4	特征 5
1	1	0	0	0	0
2	0	0	10	0	0
3	1	0	1	1	0

由于有些特征在微博文本的向量化表示中比较稀疏，如“？”和观点词是否同现，而特征稀疏会影响分类的准确度，因此，对上述特征进行约简。本文选取了 3 组不同的特征组合进行观点句识别实验，通过实验发现，使用特征 1, 2, 3, 4, 5 进行观点句识别时，准确率和召回率相对较低；而使用特征 2, 3, 4, 5 进行观点句识别时，准确率较高，但召回率较低，识别效果不好；使用特征 1, 2, 3, 4 识别观点句时，准确率和召回率都较高，识别效果较好。由此可以看出，特征 5 对中文微博观点句的识别作用不大，因此，本文最终选择特征 1, 2, 3, 4 作为识别中文微博观点句的特征。

4　SVM 分类器介绍

SVM 将线性可分问题分为两类。中文微博观点句识别的结果只有两种：观点句和非观点句，因此，观点句的识别问题也是二分类问题。基于此，本文利用第2章中所提取的特征及Libsvm[8]中的C-SVC（C-support verctor classification）模型构造 SVM 分类器，该分类器可对微博文本进行观点句识别。

4.1　线性可分问题的分类原理

定义[9]　训练集 $T=\{(\vec{x}_1,y_1),(\vec{x}_2,y_2),\cdots,(\vec{x}_l,y_l)\}$，其中 $\vec{x}_i\in\mathbf{R}^n$，i=1, 2,⋯, l，表示第 i 个样本的特征向量，$y_i\in y=\{-1,1\}$，i=1, 2,⋯, l，表示第 i 个样本对应的类别。若存在 $\vec{w}\in\mathbf{R}^n$，$b\in\mathbf{R}$ 和正数 ε，当 y_i=1 时，有 $(\vec{w}\cdot\vec{x}_i)+b\geqslant\varepsilon$；当 y_i=−1 时，有 $(\vec{w}\cdot\vec{x}_i)+b\leqslant-\varepsilon$，称训练集 T 线性可分。

线性可分问题的分类采用最大间隔法，根据法方向 $\vec{w}$ 寻找两条支持直线 l_1, l_2，即这两条支持直线为分类问题的临界直线，且它们之间的距离最大。设 l_1 为 $(\vec{w}\cdot\vec{x}_i)+b=1$，$l_2$ 为 $(\vec{w}\cdot\vec{x}_i)+b=-1$，此时两条支持直线的距离为 $\dfrac{2}{\|\vec{w}\|}$，相应的分类直线为 $(\vec{w}\cdot\vec{x}_i)+b=0$。

在本文中，$\vec{x}_i$ 表示第 i 条微博的特征向量；y_i=1 表示第 i 条微博是观点句，y_i=−1 表示第 i 条微博不是观点句。

4.2　SVM分类器的构造

SVM 分类器的构造步骤如下：

步骤 1　将训练集中预处理后的微博文本进行分词处理，提取特征 1, 2, 3, 4。通过领域观点词词典，采用匹配的方式提取特征 1，即如果分词后的微博文本中存在与领域观点词词典中相同的词，且具有相同的词性，则该微博文本的特征值为 1，否则为 0；利用统计方法提取特征 2 和特征 3；通过特征 1 和特

征3的共现与否确定特征4。

步骤2 根据4个特征值，利用向量空间模型将文本向量化表示。

步骤3 利用Libsvm软件中C-SVC模型构造SVM分类器，训练所有训练样本，即指标向量$\boldsymbol{x}_i \in \mathbf{R}^n$，$i$=1, 2,…, l，最后输出其相应的y_i。C-SVC模型用于解决如下最优化问题，

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}}\left(\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{l}\xi_i\right),$$
$$\text{subject to}\quad y_i(\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,$$
$$\xi_i\geqslant 0,\ i=1,2,\cdots,l。 \tag{6}$$

式中：$\boldsymbol{w}$是$\mathbf{R}^n$空间中的权向量；

ξ_i是松弛变量$\boldsymbol{\xi}$的第i个分量；

$\Phi(x_i)$将$\boldsymbol{x}_i$向量映射到高维空间；

C是惩罚参数，C>0；

b是待求参数。

通常，将上述最优化问题转化为对偶问题进行求解，即

$$\min_{\alpha}\left(\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha}-\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha}\right),$$
$$\text{subject to}\quad y^{\mathrm{T}}\boldsymbol{\alpha}=0,\ 0\leqslant\alpha_i\leqslant C,\ i=1,\cdots,l。 \tag{7}$$

式中：α_i是向量$\boldsymbol{\alpha}$的第i个分量；

$\boldsymbol{Q}$是半正定矩阵，$Q_{ij}\equiv y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)$，其中$K(\boldsymbol{x}_i,\boldsymbol{x}_j)\equiv\Phi(\boldsymbol{x}_i)^{\mathrm{T}}\Phi(\boldsymbol{x}_j)$是核函数。

步骤4 选取径向基函数为核函数，

$$K(u,v)=\exp\left(-r|u-v|^2\right)。 \tag{8}$$

式中：u为训练集样本向量；v为测试集样本向量；利用训练集对参数r进行训练，通过实验发现，当参数r=0.25时，所构造的SVM分类器对中文微博观点句的识别效果最好。

最后，利用上述构造的SVM分类器，对测试集中的中文微博文本进行观点句识别。在测试集中，采取与上述步骤1、步骤2相同的方式对微博文本进行预处理、特征提取及文本向量化表示。对测试集中向量化表示的微博文本，通过上述SVM分类器分类后，如果该向量被标注为+1，则表示该条微博为观点句，相反，如果被标注为−1，则该条微博为非观点句。

5 实验结果和分析

将本文算法与基于领域观点词词典的中文微博观点句识别算法进行了对比，还对比了本算法采用不同的特征组合对观点句的识别效果。实验语料库为中国计算机学会中文信息技术委员会发布的《中文微博情感分析测评－样例数据集》[10]，实验平台为MyEclipse8.5，编程语言为Java，评价算法性能的3个指标为正确率（P）、召回率（R）和F值，计算公式如下：

$$P=\frac{C(y)}{R'(y)}; \tag{9}$$

$$R=\frac{C(y)}{T(y)}; \tag{10}$$

$$F=\frac{2PR}{P+R}。 \tag{11}$$

式中：$C(y)$为算法识别的观点句与人工标注观点句的匹配数目；

$R'(y)$为算法识别出的观点句总数；

$T(y)$为人工标注的观点句总数。

基于领域观点词词典的识别算法和基于不同特征组合的SVM算法对比结果见表2。

表2 不同识别算法及不同特征组合下观点句识别结果

Table 2 Results of different recognition algorithms with different features

识别算法	正确率/%	召回率/%	F值/%
基于领域观点词词典的识别算法	66.88	39.48	49.65
基于特征1, 2, 3, 4的SVM算法	68.75	48.71	57.02
基于特征2, 3, 4, 5的SVM算法	73.17	11.07	19.23
基于特征1, 2, 3, 4, 5的SVM算法	66.46	39.48	49.53

实验结果表明，当基于SVM的中文微博观点句识别算法选取合适的特征时，其识别效果比基于领域观点词词典的识别算法效果好。基于领域观点词词典的识别算法只是将含观点词的微博判断为观点句，但由于微博口语化和表达方式的多样性等特点，不含观点词的微博也可能是观点句，因此，该方法的准确率不高。而在基于SVM的识别算法中，当选取特征1, 2, 3, 4时，其充分利用了微博中的观点词和标点符号的特征，能有效地识别含观点词的观点句和不含观点词的观点句，因此，该方法的识别效果较好；当选取特征2, 3, 4, 5时，只考虑到了标点符号在微博观点句中的特征，没有考虑微博中观点词的特征，因此，召回率较低；当选取特征1, 2, 3, 4, 5时，虽然包含了微博中观点词和标点符号的特征，但由于“？”和观点词同现的特征稀疏，因此，该方法的准确率和召回率没有选取特征1, 2, 3, 4时高。

6 总结与展望

由于中文微博观点句的特征比较稀疏，构造

SVM 分类器时不再局限于选取某一个特征，可以是几个特征的组合，而本文通过实验发现，当选取特征 1, 2, 3, 4 时，SVM 分类器对中文微博的观点句识别效果较好。将本文算法与基于领域观点词词典的方法进行了比较，实验结果表明，在基于 SVM 的方法中选取特征 1, 2, 3, 4 的识别效果较好。

因此，本文接下来的工作是，一方面对领域观点词词典的构造进行完善，以提高识别的准确率；另一方面，针对是观点句但没有观点词的中文微博，采用基于句法分析的方式提取观点句特征，使用基于条件随机场的方法和朴素贝叶斯的方法进行观点句识别，同时考虑采用多分类器融合的方式对观点句与非观点句进行分类，以提高中文微博观点句识别的召回率和准确率。同时，利用基础情感词词典，结合机器学习的方法对观点句进行情感倾向性分析，以提高微博观点句识别效果。

参考文献：

[1] 中国互联网网络信息中心. 第29次中国互联网发展状况统计报告[EB/OL]. [2012-06-20]. http://www.cnnic.net.cn/research/bgxz/tjbg/201201/t20120116_23668.html.
China Internet Network Information Center. The 29th China Internet Development Statistics Report[EB/OL]. [2012-06-20]. http://www.cnnic.net.cn/research/bgxz/tjbg/201201/t20120116_23668.html.

[2] Pak Alexander，Paroubek Patrick. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]//Proceedings of International Conference on Language Resource and Evaluation. Lisbon：[s. n.]，2010：1320-1326.

[3] Barbosa Luciano，Feng Junlan. Robust Sentiment Detection on Twitter from Biased and Noisy Data[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing：[s. n.]，2010：36-44.

[4] 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：[s. n.], 2010：241-249.

[5] 朱艳辉，徐叶强，王文华，等. 中文评论文本观点抽取方法研究[C]// 第三届中文倾向性分析评测论文集. 山东：[出版者不详]，2011：126-135.
Zhu Yanhui，Xu Yeqiang，Wang Wenhua，et al. Research on Opinion Extraction of Chinese Review[C]//The Third Chinese Opinion Analysis Evaluation Proleedings. Shandong：[s. n.]，2011：126-135.

[6] 知网. 《知网》情感分析用词语集：Beta 版[EB/OL]. [2012-08-20]. http://www.keenage.com/html/c_index.html.
HowNet.“HowNet”Word Set for Sentiment Analysis：Beta Version[EB/OL]. [2012-08-20]. http://www.keenage.com/html/c_index.html.

[7] ICTCLAS汉语分词系统. ICTCLAS下载[EB/OL]. [2012-07-02]. http://ictclas.org/ictclas_download.aspx.
ICTCLAS Chinese Word Segmentation. Download ICTCLAS[EB/OL]. [2012-07-02]. http://ictclas.org/ictclas_download.aspx.

[8] Ku Lunwei，Wu Tungho，Lee Liying，et al. Construction of an Evaluation Corpus for Opinion Extraction[C]//Proceedings of the 5th NTCIR Workshop Meeting. Tokyo：[s. n.]，2005：513-520.

[9] 邓乃扬，田英杰. 支持向量机：理论、算法与拓展[M]. 北京：科学出版社，2009：45-50.
Deng Naiyang，Tian Yingjie. Support Vector Machines: Theory, Algorithms and Development[M]. Beijing：Science Press，2009：45-50.

[10] 中国计算机学会. 中文微博情感分析测评-样例数据集[EB/OL]. [2012-07-01]. http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html.
China Computer Federation. The Chinese Micro-Blog Emotional Analysis and Evaluation: Sample Data Sets[EB/OL]. [2012-07-01]. http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html.

（责任编辑：邓　彬）