CN43-1258/TP
ISSN 1007-130X

计算机工程与科学
COMPUTER ENGINEERING & SCIENCE

2012 年第 34 卷第 9 期
Vol. 34, No. 9, 2012

文章编号:1007-130X(2012)09-0160-06

一种基于朴素贝叶斯的微博情感分类*

Classification of Microblog Sentiment Based on Naïve Bayesian

林江豪[1],阳爱民[2],周咏梅[2],陈 锦[3],蔡泽键[2]

LIN Jiang-hao[1], YANG Ai-min[2], ZHOU Yong-mei[2], CHEN Jin[3], CAI Ze-jian[2]

(1. 广东外语外贸大学国际工商管理学院,广东 广州 510006;
2. 广东外语外贸大学思科信息学院,广东 广州 510006;
3. 广东外语外贸大学英语语言文化学院,广东 广州 510006)

(1. School of Management, Guangdong University of Foreign Studies, Guangzhou 510006;
2. Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006;
3. School of English Language and Culture,
Guangdong University of Foreign Studies, Guangzhou 510006, China)

摘 要:本文基于二次情感特征提取算法,利用句法依存关系进行一次文本情感特征提取,在此基础上,利用情感词典,进行二次情感特征提取。构建朴素贝叶斯分类器,对采集的热门话题微博和酒店评论进行文本情感倾向性分类。主要比较了表情符号、标点符号,基于情感词典的特征提取和基于二次情感特征提取方法,在不同的组合下的分类性能,寻找更佳的微博文本情感分类预处理方法。并与酒店评论情感分类结果对比、分析,发现影响微博情感分类性能的原因。实验结果表明,二次特征提取方法在分类上取得更高的 F_1。实验最佳的分类预处理方式是"表情符号+标点符号+二次情感特征提取+BOOL 值"。同时发现,朴素贝叶斯在酒店评论情感分类取得更高的分类性能,主要是微博评价对象多样化造成的。

Abstract: Based on the twice sentiment feature extraction approach, this paper uses syntactic dependency as the first extraction method and semantic lexicon as the second. A sentiment classifier based on naïve Bayesian is constructed in order to classify the inclination of emotions from the collected hot topic data in Chinese microblog and hotel remarks. The experiments mainly compare the classification performance of different combination groups including emoticons, punctuation, extraction methods based on semantic lexicon feature and those based on twice sentiment feature to find out better pretreatment methods for sentiment classification of microblog text. Besides, the experiments also compare and analyze the sentiment classification results between microblog text and hotel remarks to seek out the reasons for influencing the classification performance of microblog sentiment. The results indicate that the twice sentiment feature extraction gain the higher F1. And the performance of "emoticons + punctuation + twice sentiment feature extraction + BOOL" is the best pretreatment method. Meanwhile, it also shows the reason why the classifier based on naïve Bayesian obtains higher classification performance in hotel remarks is probably that the topic in microblog is various.

关键词:微博;文本情感分类;二次情感特征提取;朴素贝叶斯

Key words: microblog; text sentiment classification; twice sentiment feature extraction; naïve Bayes-

* **收稿日期**:2012-04-13;**修订日期**:2012-06-25
基金项目:国家社科基金资助项目(12BYY045);教育部人文社会科学研究青年资助项目(10YJCZH247);广东省科技计划资助项目(2010B031000014);广东外语外贸大学研究生科研创新资助项目;广东外语外贸大学大学生创新实验资助项目
通讯地址:510006 广东省广州市广州大学城广东外语外贸大学学生公寓 1 栋 513 房
Address: Room 513, 1st Building of Dormitory, Guangdong University of Foreign Studies, Guangzhou Higher Education Mega Center, Guangzhou, Guangdong 510006, P. R. China

ian

doi:10.3969/j.issn.1007-130X.2012.09.029
中图分类号:TP393 **文献标识码**:A

1 引言

微博(Micro Blog)是当代社会人们发布信息的一个重要网络工具。微博带着用户的情感信息,对微博的情感进行分类研究,有利于微博监控、舆情发现、舆论引导等工作的实现,研究意义重大。本文将微博所表达的情感倾向(Sentiment Orientation)分为正面、负面两类,对微博进行情感分类研究。

目前,微博情感分析研究主要是针对英文微博[1~5],针对中文微博的研究工作尚处于起步阶段[6,7]。英文微博研究主要针对 Twitter 上的微博消息(即 Tweets)所做的研究工作[1~5]。在英文微博上的情感分析研究可以为两类:主题无关的情感分析与主题相关的情感分析。与中文微博的情感分析方面相关的研究工作尚处于起步阶段,主要是基于机器学习方法,如支持向量机(Support Vector Machine,简称 SVM)进行情感分类研究[7];还有基于微博文本情感强度[6]的微博情感分类。文本在文本情感分类研究中,主要采用机器学习方法构建不同分类器,运用如基于情感词典[8]、基于句法路径[9]等不同的情感特征选择方法,来进行语料的情感分类实验。

本文考虑微博自身的特点,采集热门话题下的微博,将微博视为一个观点整体和细分微博观点两种方式,分析了表情符号在这两种方式下的分类效果。基于情感词典和基于二次情感特征提取的方法进行情感特征提取,在布尔型(BOOL)值和词频两种特征权值中,构建朴素贝叶斯,比较分析了不同条件下的分类性能。研究结果表明,基于二次的情感特征提取方法优于基于情感词典的情感特征提取方法,并可快速有效地提取情感特征,利于文本向量的降维,简洁有效地表示文本,提高计算性能,分类结果具有更高的微平均值 F_1。同时发现,考虑标点符号和表情符号,对提高分类性能有一定的作用。

2 微博预处理

微博的长度限制在 140 个汉字以内,属于短篇章。在进行微博预处理时,采用两种方式:(1)去掉标点符号,直接抽取微博情感特征向量;(2)考虑标点符号,以句子为单位向量,提取句子单位情感向量,组合成微博情感特征向量。由于网络语言有很多的表情符号,表情符号是用户情感表现的快捷模式,如"☺"、":D"表示正面情感,"☹"、":-("表示负面情感。当然也存在表情符号仅仅是符号,与微博的情感并没有关联的情况。为了能获得更准确的分类效果,本文比较、分析了考虑表情符号与不考虑表情符号在以上两种预处理方式下的分类性能。

本文的预处理主要包括了微博文本分词、文本情感特征选择、情感特征权值的计算、文本的向量表示等工作。采用两种文本情感特征提取方法:基于情感词典的方法和基于二次特征提取的方法。

2.1 微博文本分词

微博分词时,以《现代汉语常用词表(草案)》[10]为分词支撑,采用最大匹配算法对文本进行分词,设置最大匹配步长为 4 个汉字,只对中文内容进行分词处理。

2.2 基于情感词典的特征选择

微博中的情感词是微博情感的表征,情感词的极性对微博的情感极性起到决定性的作用。微博口语化的特点,使得情感词汇的提取非常复杂。本文利用情感词典,对分词后所得的稀疏矩阵进行降维,去掉无用的词汇,提取能显著表征文本情感类别的词汇。

选用的基础情感词典[11]共有 5 281 个情感词汇。其中褒义词 2 807 个,贬义词 2 474 个。考虑到微博上的一些流行词汇,如"杯具"、"洗具"等带有情感色彩的流行网络用语在情感分类中的重要性。本文收集了网络词汇,形成微博通用的网络词汇情感词典。

2.3 基于句法依存的特征选择

仅依据情感词典进行特征选择,忽略了文本之间的句法关系,可能会造成情感特征与评价主体的不对称。因为出现在情感句中的评价词语并不一定总能表现出一定的情感倾向性[9]。如"好吧,我对国足表示无语。"句子中具有褒义情感色彩的词语"好",并不是修饰评价主体"国足",如果选为情

感特征，将直接影响分类结果。因此，本文在基础情感词典的基础上，利用哈工大社会计算与信息检索研究中心的语言技术平台(Language Technology Platform，简称 LTP)进行句法分析，并借鉴文献[9]中基于句法路径精确匹配的情感特征选择方法。

句法路径是指在句法树上连接任意两个节点之间的句法结构。在这里，句法路径特指情感词语与评价对象之间的有向句法结构。如图 1 所示，分别是 S1 和 S2 的句法路径图。

S1：林书豪的投篮很酷。

S2：林书豪的投篮酷。

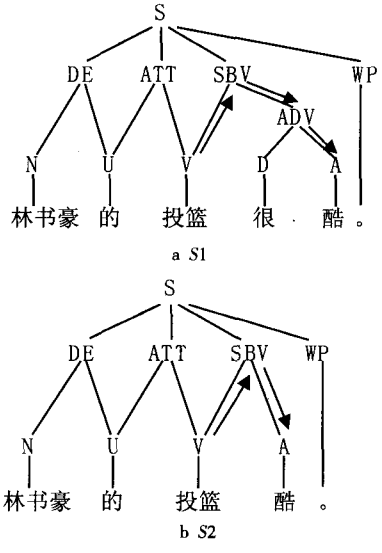

图 2 句法路径图

分析 S1 与 S2 的句法路径，S1 的句法路径是 V↑SBV↓ADV↓A，S2 的句法路径是 V↑SBV↓A。S1 与 S2 所表达的内容都是对“林书豪的投篮”表示“酷”，属于褒义情感。但是，两者的语法树和语法路径却是不同的，仔细分析 S1 与 S2 发现，S1 中用程度副词(D)“很”来修饰形容词(A)“酷”形成图 2a 中的结构 ADV，更增强了情感色彩。而本文只将情感类别设定为褒义和贬义两种，因此可以用 S2 的句法路径来泛化 S1 的句法路径，即将(V↑SBV↓ADV↓A)泛化为(V↑SBV↓A)。对句法进行泛化，一方面有利于统计高频句法，方便标注重点句法；另一方面，句法更简单清晰，降低分析的复杂度。为了更准确地分析，本文归纳出常用的程度副词系及常用词，如表 1 所示，利用该表更快速有效地泛化。

基于句法路径的情感特征选择时，构建句法树，提取评价对象，根据统计的句法路径表进行句法匹配，获取句子的句法结构，在此基础上，选择用于修饰评价对象的情感词语作为情感特征。

表 1 常用程度副词表

词系	常用词
“最”系	最、最为
“更”系	更、更加、更为、还、越、越发、越加、愈、愈发、愈加、格外
“比较”系	比较、较、较为
“稍微”系	稍、稍微、稍稍、略、略略、略微、多少
“太”系	太$_1$、好、多(么)、分外、万分
“过于”系	过于、太$_2$、过
“极”系	极、极为、极其、极度
“很”系	很、挺、怪、老、蛮、十分、非常、相当、异常、颇、颇为
“有点”系	有点儿、有些

2.4 特征权值计算

在微博情感特征选择后，进行情感特征权值的计算，采用 BOOL 型和词频两种特征权值计算方法。词频和 BOOL 型权值是比较简单的特征权重表示方法，适合用于朴素贝叶斯分类器，BOOL 型权值计算方法如式(1)所示。

$$bool(w_i)=\begin{cases}1, freq(w_i,d_j)>0\\0, freq(w_i,d_j)=0\end{cases} \tag{1}$$

其中 $freq(w_i,d_j)$ 是词 w_i 出现在文本 d_j 中的频次，即为词频。

2.5 情感文本的向量表示

文本 d 可以表示成若干词的集合 $d=\{w_1,w_2,\cdots,w_n\}$，词的特征权值为向量，则文本集 D 可视作文本为行、词汇为列的矩阵。因为每篇文本词汇量不多，是一个稀疏的矩阵，为了节省存储空间，在实际存储时采用“*word-index*:*weight*”的格式存储，每两个不同词的向量之间用空格隔开。其中 *word-index* 为词的索引值，测试语料矩阵中的词索引必须与训练语料中的相互对应，*weight* 为词在文本中的权值，之间用冒号隔开。一个文本一行，就形成了一个矩阵文本。

3 朴素贝叶斯分类器

情感文本集经过处理后得到的向量矩阵，就可以作为情感分类器训练和评测的数据，本文采用朴素贝叶斯方法[8]来构建文本情感分类器。它是一种基于概率的学习算法，基于假设的先验概率，给定假设下观察不同特征的概率。定义文本 $d=\{w_1,w_2,\cdots,w_n\}$ 的类别属于 $C=\{c_P,c_N\}$，在特征

相互独立的情况下,考虑特征词的权值,其分类方法如公式(2)所示。

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \left\{ P(c_j) \prod_{i=1}^{n} P(w_i, c_j)^{wt(w_i)} \right\} \quad (2)$$

其中,$P(c_j)$是类别 c_j 的先验概率,$P(w_i, c_j)$是特征词 w_i 在类别 c_j 中的后验概率,$wt(w_i)$是测试语料中词 w_i 的权值,当采用 BOOL 型权值时,$wt(w_i)=1$。

先验概率 $P(c_j)$是指预先给定类别一个出现的概率,如果不进行估计,可以视每类出现的概率是相同的。如果进行预先估计,可以人工直接估计或在训练语料的基础上通过概率计算估计。本文采用的是根据已正确标注的训练语料预先估计,估计方法如公式(3)所示,其中 $Doc(c_j)$是属于类别 c_j 的文档数。

$$P(c_j) = \frac{Doc(c_j)}{\sum_{c_j \in C} Doc(c_j)} \quad (3)$$

后验概率 $P(w_i, c_j)$是指特征词 w_i 出现在类别 c_j 中的概率,可以从训练语料中通过计算进行估计。普遍采用词 w_i 在属于类别 c_j 的文本中的权值之和除以类别 c_j 的文本中所有词的权值之和,估计方法如公式(4)所示,其中 $Weight(w_i, c_j)$是词 w_i 在属于类别 c_j 的文本中的权值之和。

$$P(w_i, c_j) = \frac{Weight(w_i, c_j)}{\sum_{i=1}^{n} Weight(w_i, c_j)} \quad (4)$$

如果训练语料中词 w_i 在类别 c_j 的所有文本中都不存在,则 $P(w_i, c_j)=0$。一般是由两种情况引起的,一是词 w_i 确实不能表示类别 c_j;二是训练语料不够全面。然而,如果这种情况出现,后验概率的乘积会是 0,则另一类就占据了统治地位,如果在所有类别中都为 0,就没有办法再进行分类。为避免 $P(w_i, c_j)$等于 0,本文采用 Laplace 转换,由此后验概率计算方法改进如公式(5)所示。

$$P(w_i, c_j) = \frac{Weight(w_i, c_j) + \delta}{\sum_{i=1}^{n} Weight(w_i, c_j) + \delta \mid V \mid},$$

$$V = \sum_{c_j \in C} \sum_{i=1}^{n} Weight(w_i, c_j), \delta = 1 / \mid V \mid \quad (5)$$

采用 Laplace 转换,一般 δ 取 1,常数 V 取所有词的权值总和。但是,当 $\delta=1$ 时存在一些问题,增大了训练语料中未出现的特征词的存在概率,并且缩小了出现词的概率。为了解决这个问题,本文取 $\delta=1/|V|$,等效于当特征词不存在时,后验概率为一个极小的存在概率。当特征词存在时,对原有概率影响也不大。

朴素贝叶斯分类器是解决相应机器学习问题的最有实际价值的方法之一。在多数情况下与其他学习算法性能相当,在某些情况下还优于其他算法。其分类算法实现比较简单,分类效率也比较高,在文本分类方面表现比较好。在利用朴素贝叶斯分类器进行文本分类时,需要先进行训练,估计类别的先验概率和特征的后验概率,再进行分类。

4 实验结果及分析

4.1 语料采集及预处理

本文利用腾讯微博公开的 API 接口,采集热门话题下的微博,研究基于热门话题的微博情感分类。获取体育类热门话题"林书豪"的 400 000 多条微博,采用人工交叉标注方法(3 个人对相同的语料标注),选择了其中标注结果一致,并且具有情感色彩的 2 000 条微博。分为测试语料和训练语料,测试语料 1 000 条微博,正面的微博数为 356 条,负面的微博数为 644 条;训练语料 1 000 条微博,正面的微博数为 325 条,负面的微博数为 675 条。

为了比较微博与驴评网的酒店评论情感分类结果,本文设计驴评网的评论自动采集系统,采集北京、上海、广州三大城市的酒店评论语料,共有 250 000多条。酒店评论的内容是用户对酒店的评论和对酒店的周围环境、酒店服务、房间卫生、设施设备四个方面进行评分,再形成综合分数,最高分为 5 分,最低分为 1 分。本文以用户的评分作为语料情感标注的标准,随机选其中 1 000 条综合分数为 1 分的作为负面评论,其中 1 000 条综合分数为 5 分的作为正面评论。构建平衡语料库,训练语料库和测试语料库的语料均为 1 000 条,正面语料与负面语料的比例是 1:1。

4.2 评价指标

在对分类器的性能进行评测时,本文采用了查准率(Precision)、召回率(Recall)和微平均(F_1)作为评价分类结果的指标,计算公式如式(6)、式(7)和式(8)所示。

$$Precision = \frac{\sum_{c_i \in C} True(c_i)}{\sum_{c_i \in C} Doc(c_i)} \quad (6)$$

$$Recall = \frac{\sum_{c_i \in C} True(c_i)}{\sum_{c_i \in C} Response(c_i)} \quad (7)$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} * 100\% \quad (8)$$

其中 $True(c_i)$是分类为 c_i 并且正确的文档数，$Response(c_i)$是分类为 c_i 的文档数。对于正向和负向概率相等或者没有特征的文本，将其作为没有情感倾向的客观描述，不做分类。如果测试语料中每一个文本都能进行分类，则分类器整体的查准率、查全率和微平均是相等的。如果出现无法分类的文本，分类器的查准率、查全率和微平均不相等。

4.3 实验结果及分析

4.3.1 不考虑表情符号

在选用情感特征时，认为微博中的表情符号与微博文本的情感符号无关，忽略微博中表情符号，仅仅提取微博中的中文文本进行情感分类。分别采用基于情感词典的微博情感特征提取和基于二次的情感特征提取方法，特征权值采用词频和BOOL型权值，使用训练语料，进行分类器训练。分类结果如表2所示。

表2 不考虑表情符号分类结果(F_1:100%)

	词频		BOOL	
	情感词典	二次提取	情感词典	二次提取
不考虑标点符号	68.26	68.53	69.13	70.79
考虑标点符号	67.62	69.73	69.21	71.84

不考虑表情符号的分类结果中，F_1 指标值显示，二次情感特征提取算法较基于情感词典的特征提取方法取得更好的分类性能，同时在实验过程中也发现，二次情感特征提取方法可有效避免噪声情感词的影响。在微博情感分类中，考虑标点符号的分类优于不考虑标点符号。主要原因是，微博虽为短篇章文本，但微博有转发、评论等功能特点，同一条微博，可能小部分内容是在评论原创微博，其他内容才是自己的观点。因此，在微博情感分类中，必须考虑博文的标点符号及分句。在特征权值方面，BOOL特征权值计算方法取得较好的分类性能。

4.3.2 考虑表情符号

参考文献[12]对Twitter微博情感分类中的表情符号的情感类别划分、特征权值计算和分类过程，本文将微博情感符号作为微博表情符号单位向量，与文本向量组成微博情感向量。因此，需要对训练语料和测试语料中的表情符号进行标注，在训练贝叶斯分类器的时候，将训练语料输入，训练获得分类器。基于情感词典和二次的情感特征提取方法，采用词频和BOOL型权值，将测试语料进行分类实验，实验结果如表3所示。

表3 考虑表情符号分类结果(F_1:100%)

	词频		BOOL	
	情感词典	二次提取	情感词典	二次提取
不考虑标点符号	70.09	72.22	70.04	72.36
考虑标点符号	70.61	73.17	70.32	74.54

实验在考虑标点符号和不考虑标点符号两种情况下，加入表情符号单位向量，对测试语料进行分类。分类结果表明，考虑标点符号及表情符号的情感取得较好的分类结果；在特征提取方法上，二次情感特征提取方法在基于主题的微博分类中，取得较高的 F_1 值；在特征权值计算方法选择上，BOOL型值优于词频。

4.3.3 与酒店评论情感分类实验对比

由于酒店评论中极少出现表情符号，忽略表情符号对情感分类的影响。实验将酒店的四个评价对象泛化为评价对象酒店，考虑将酒店评论视为一个情感观点整体或评论观点细分后再组合两种情况下，采用基于情感词典和二次情感特征提取方法分别进行情感特征提取，分别计算词频和BOOL情感特征权值，训练分类器后，测试语料的分类结果(F_1)如表4所示。

表4 酒店评论分类结果(F_1:100%)

	词频		BOOL	
	情感词典	二次提取	情感词典	二次提取
不考虑标点符号	75.17	77.31	75.23	76.87
考虑标点符号	75.01	76.84	75.19	76.33

实验结果表明，在酒店评论的情感分类中，是否将情感观点细分，采用词频或BOOL值作为情感特征值对分类的性能影响不大。二次情感特征提取在酒店评论情感分类中，取得较好的分类性能。

对比微博与酒店评论的情感分类结果，同为短文本情感分类，但朴素贝叶斯在酒店评论情感分类中，取得更好的分类效果。主要原因是酒店评论的评价对象相对专一化，评论中的情感词能有效表征评论的情感倾向；而微博内容口语化，指代复杂化，评价对象多样化，虽采用二次特征提取，仍存在情感词与评价对象不对称的情况，对分类结果造成影响。这一点也体现于微博在细分情感观点情况下，分类效果高于将微博视为一个整体观点，而在酒店评论情感分类时，观点细分与否，对分类影响不大。

4.3.4 实验总结

本文主要研究基于热门话题的微博情感分类，微博情感分类结果显示，在特征权值计算方面，

BOOL 型权值优于词频权值,主要是微博自身短篇章、口语化等特点造成的;在特征提取方面,二次情感特征提取方法优于基于情感词典的提取方法,微博虽为短篇章文本,但转发、评论等特点造成了篇章语义或者评价对象的不一致性;在标点符号方面,考虑标点符号优于不考虑标点符号,微博的评价对象多样性造成的,所以要将句与句分开,形成单位情感向量;表情符号是网络语言的简化表示,带着较强的情感特征,考虑表情符号获得更好的微博情感分类性能。通过与酒店评论情感分类结果的对比,发现朴素贝叶斯在酒店评论情感分类中,取得更好的分类效果。同时,在微博情感分类研究中,必须深入思考微博评价对象多样化的处理方法,才能有效、快速地进行情感特征提取,提高分类性能。综上所述,对基于主题的微博情感倾向性分类中,"考虑表情符号+考虑标点符号+二次情感特征提取+BOOL 值"的分类预处理方式,取得较好的分类结果。

5 结束语

本文采用朴素贝叶斯分类器对微博和酒店评论进行情感倾向性分类对比、研究。通过分类实验,结合微博自身的特点,在微博热门话题情感倾向性分类时,"考虑表情符号+考虑标点符号+二次情感特征提取+BOOL 值"预处理方式具有较好的分类效果。在今后的微博情感分类研究中,深度思考评价对象多样化的处理,是提高分类性能的有效方法之一。另外,深入特定领域的微博情感倾向性分类研究,可以实现舆论分析、舆情发现等工作。现有的微博主要采用人工标注方式,工作量巨大,以后可研究情感分类在微博情感极性自动标注方面的应用。

参考文献:

[1] Davidiv D, Tsur O, Rappoport A. Enhanced Sentiment Learning Using Twitter Hash-Tags and Smileys[C]// Proc of COLING'10, 2010: 241-249.

[2] Barbosa L, Feng J. Robust Sentiment Detection on Twitter from Biased and Noisy Data[C]// Proc of COLING'10, 2010: 36-44.

[3] Jiang L, Yu M, Zhou M, et al. Target-Dependent Twitter Sentiment Classification[C]// Proc of the 49th Annual Meeting of the Applicational Linguistics, 2011: 151-160.

[4] Tan L K-W, Na J-C, Chang K Y. Sentence-Level Sentiment Polarity Classification Using a Linguistic Approach[C]// Proc of ICADL'11, 2011: 77-87.

[5] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]// Proc of LREC'10, 2010: 1320-1327.

[6] Xin M J, Wu H X. A Public Opinion Classification Algorithm Based on Micro-Blog Text Sentiment Intensity: Design and Implementation[C]// Proc of MECS'11, 2011: 48-54.

[7] 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.

[8] 杨鼎,阳爱民. 一种基于情感词典和朴素贝叶斯的中文文本情感分类方法[J]. 计算机应用研究, 2010, 27(10): 3737-3739.

[9] 赵妍妍,秦兵,车万翔,等. 基于句法路径的情感评价单元识别[J]. 软件学报, 2011, 22(5): 887-898.

[10] 柳位平,朱艳辉,栗春亮,等. 中文基础情感词词典构建方法研究[J]. 计算机应用, 2009, 29(11): 2882-2884.

[11] 《现代汉语常用词表》课题组. 现代汉语常用词表(草案)[M]. 北京:商务印书馆, 2008.

[12] Agarwal A, Xie B, Vovsha I, et al. Sentiment Analysis of Twitter Data[C]// Proc of Association for Computational Linguistics, 2009: 30-38.

林江豪(1985-),男,广东揭阳人,硕士生,CCF 会员(E200015663G),研究方向为机器学习和人工智能。**E-mail**: lin_hao@foxmail.com

LIN Jiang-hao, born in 1985, MS candidate, CCF member(E200015663G), his research interests include machine learning, and artificial intelligence.