doi: 10.12052/gdutxb.180016

基于THUCTC的金融语料情感分析模型优化

饶东宁，黄思宏
（广东工业大学　计算机学院，广东　广州 510006）

摘要: 近几年, 情感分析技术引起人们的兴趣, 在金融应用上, 可以作为投资者投资前的参考. 但是现有方法存在应用过于专一、数据偏差、结果过于笼统和不够精确的问题. 因此本文优化一个通用的中文文本分类器, 用于对在线评论数据和股票新闻数据进行情感分析. 收集整理了2万条数据作为语料库, 每条数据分别由3个人进行独立标注. 之后对THUCTC进行优化, 具体从3个方面对中文文本分类器进行优化, 首先是词语切分, 使用词干词典方法结合不同的分词法, 实验比较后得到二分法为最好的结果; 其次, 为分类器选择最好的内核, 发现Liblinear内核对即时性要求较高的投资人更好, 另一方面Libsvm在提高准确率方面更有优势; 最后在金融导向的情绪字典方面, 它由Chi-square和TF-IDF方法构建, 可用在普通文本分类器上. 通过这种方式, 本文的结果可以被推广且不会失去准确性.

关键词: 情感分析；文本分类；股价趋势预测；中文分词
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1007–7162(2018)03–0037–06

Model Optimization of Financial Corpus Sentiment Analysis Based on THUCTC

Rao Dong-ning, Huang Si-hong
(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Sentiment analysis has attracted interest recently. In financial applications, it can be a reference for investors. However, existing approaches are either so specific as to cause data drift or too general to be precise. Therefore, a general Chinese text classifier for online reviews and news on stocks is optimized. A corpus with 20000 items is first collected. Then, each item is labeled by three persons as ground truth. After that, the THUCTC is optimized, thus optimizing a general Chinese text classifier in three aspects. First, by tokenization, the THUCTC is modified to a 2-gram with a stemming dictionary method and got better results. Second, the best kernel is selected for classifier. The Liblinear kernel is found to be better for people pressed for time. On the other hand, the Libsvm kernel is good at promoting accuracy. Third, a finance-oriented sentiment dictionary is set based on Chi-square and TF-IDF approach. It can be used by on-the-shelf general text classifiers. In this way, the result can be generalized without the loss of preciseness.

Key words: sentiment analysis; text categorization; stock price trend prediction; Chinese word segmentation

近几年，金融领域情感分析研究在现代金融体系中越来越重要. 2015年，Hassan Saif等[1]提出一种情感圈的方式来对Twitter上的文本进行情感分析. 2016年，Qin等[2]提出一种CWTM (character-word topic model)模型来处理中文字符与中文单词之间的关系. 2016年，Qian Q等[3]将LSTM算法改进，对电影评论数据集和斯坦福情感分类树进行情感分析. 可见当前情感分析是研究热点之一，越来越受到研究者重视.

本文选用THUCTC (THU Chinese Text Classification)为文本分类工具. THUCTC[4-5]是清华大学自然语言处理实验室开发的一款高效的分类工具. 它选用二字串bigram作为特征单元和Chi-square[6]降维方法，采用的权重计算方法为tf-idf，对于

收稿日期： 2018-01-01
基金项目： 广东省自然科学基金资助项目(2016A030313084，2016A030313700)
作者简介： 饶东宁(1977–)，男，副教授，博士，主要研究方向为金融智能、智能规划.
通信作者： 黄思宏(1993–)，男，硕士研究生，主要研究方向为金融智能. E-mail: 584654848@qq.com

长文本THUCTC有良好的普适性. 但是网络论坛平台大部分评论数据都为短语料（字符少于30），且金融语料又有着自身的特性，所以对于金融评论短语料分类THUCTC显得不够好.

本文主要工作如下. 首先对网络股市论坛平台数据进行抽取、分析和人工进行标注，共得到有效金融评论数据2万条. 每条数据都由3个人独立进行标注并通过评分程序[7]给出每条数据最终标注结果. 其次本文使用清华大学开发的THUCTC中文文本分类器，针对金融语料特性通过正则表达式方法在程序中增加Stemming词干词典对股票专有歧义词进行替换. 然后调整*n*-gram分词算法和特征数，使用Libsvm与Liblinear两种分类算法在调整基础上反复训练分类模型并测试比较.

本文内容安排如下. 第一节将介绍本篇论文相关工作，包括背景知识及金融语料分类的当今现状. 第二节将陈述金融数据语料的分析、获取、标注原则和词干词典方法. 介绍Libsvm与Liblinear分类原理，描述THUCTC对金融语料的优化调整，以及调整后工作流程. 第三节是实验设计与分析. 最后是结束语与未来工作展望.

1 相关工作

行为金融理论[8]是新兴经济学理论，它结合行为科学和心理学，通过投资人个体行为动机等角度分析和预测金融市场波动. 该理论试图通过研究市场活动主体在市场中的投资行为来挖掘出不同个体在不同环境中投资规律与决策行为，以此构造一个能反映个体投资心理行为与市场波动的模型. 随着金融异常现象的累积，许多研究人员开始尝试从心理学角度分析金融问题. 2014年，G Bekaert等[9]通过对2007~2009年金融危机时期的研究，得出投资者存在羊群效应. 2016年，M Giannetti等[10]通过研究得出负面信息对投资者投资的影响范围不仅仅是单一个体，还具有扩散性. 2016年，E Avdis[11]提出在信息充足时代人们会综合更多信息进行投资决策. 2016年，RM Edelen等[12]通过研究发现投资者信息行为偏好造成了股票的异常收益. 2016年，TY Chang等[13]通过研究，提出公司在宣布较好业绩收益后会得到更多投资者对其进行投资的现象.

情感分析（又称倾向性分析）目的是将文本中带有情感色彩的数据进行分析、处理、归纳和推理，最终得到文本情感指向. 其中数据预处理包括标记、过滤停止词、标注词性、分词、抽取和特征表示. 情感分析技术近几年在股票趋势预测上发展迅速，特别是对于短篇博文上. 研究案例如Ruan X等[14]在2016年人工对Twitter数据进行标注，利用不同神经网络对标注后数据进行情感分析. 张对等[15]对股吧上30支股票数据进行预处理，然后用SVM进行分类从而预测股市走向. 反观国内，相关研究还比较少. 2017年，江腾蛟等[16]设计出规则对金融数据情感词对进行抽取，实验结果表明在金融数据中存在大量情感词，这些情感词为金融数据的情感分析提供了基础. 饶东宁等[17]使用改进的加权中心度算法对社交网络进行量化，计算得出节点的重要性. 林穗等[18]将优化后的线性模型应用在广告投放系统中，提高了广告投放的精准度.

2 基于THUCTC的金融语料情感分析

2.1 整体流程

通过之前章节描述得到THUCTC在金融语料情感分析上的总体流程. 本文引入Stemming词干词典且调整多个参数，然后分别与两种分类算法相结合，逐一验证Stemming词干词典、参数调整与分类器选择对情感分析结果的影响. 在数据预处理上增加基于正则式技术的Stemming词干词典并调整*n*-gram、特征保留数，之后分别与Libsvm、Liblinear两个分类算法结合进行比较实验. 实验方案如下：方案1，增加使用Stemming词干词典方法；方案2，调整*n*-gram和特征保留参数；方案3，使用Stemming方法并同时调整*n*-gram和特征保留参数.

针对不同方案有以下的推论. 如果方案1提升了分类效果，则说明增加Stemming方法对分类器分类效果有积极影响. 如果方案2提升了分类效果，说明调整分类器参数对分类器分类效果有积极影响. 如果方案3提升了分类效果，说明Stemming方法、分词法与特征数相组合的优化方法对情感分类结果产生积极影响.

整体流程如图1所示.

1) 数据获取. 通过爬虫程序，从股吧获取2015~2017年间10万条评论数据.

2) 数据预处理. 先过滤与金融无关数据以及标点符号、空白符等，然后从剩余数据中随机抽取2万条数据进行人工标注. 每条数据由3个金融专业人员独立进行标注，标注完成后用投票系统给出每条数据最终标注结果.

3) 词干替换. 总结一批股市专有且经常出现的人造词语，在程序读取数据时利用正则表达式对其

进行匹配并替换为具有该词特殊含义的常用词.

4) 分词. 使用2-gram、3-gram、4-gram实验方法做出比较.

5) 特征与情感词提取. 使用Chi-square特征降维与tf-idf权重计算相结合的方法来提取特征词并赋予相应权重.

6) 用不同方法训练模型做出比较. 选择台湾大学林智仁教授所开发的Libsvm分类器和Liblinear分类器，使用两个分类器进行相同实验.

7) 模型测试. 通过*n*-fold交叉验证方法进行模型测试，做出比较并计算出3类（积极、消极、中性）情感倾向概率值等指标.

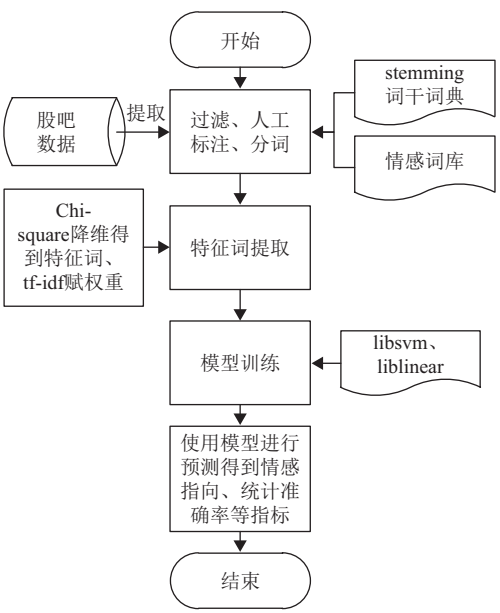

图 1　整体流程

Fig.1　Overall process

2.2　数据采集处理与分词

2.2.1　金融语料采集过滤与人工标注

金融语料采集过滤与人工标注是本实验基础. 据调查[19]可知，全国至少有35%的投资人会从股吧获取相关信息作为自己投资决策参考. 从这可以看出丰富的情感信息存储在股吧等社交平台上. 本文使用爬虫程序收集2015~2017间10万条股吧评论数据作为源数据，然后进行无关数据过滤. 语料标注是实验数据收集的关键点. 从处理后的源数据随机抽取2万条金融数据进行标注，每条语料标注都由3个金融专业人员独立进行，标注后用投票系统得出该条语料标注最终结果.

2.2.2　词干替换与中文分词

本文构建词干词典的目的是减少数据歧义性，对多种*n*-gram分词法进行比较是为了找到金融短语料最佳的分词方法. 通过在程序读取数据时加入正则表达式的方法，对影响准确率的人造词进行替换. 由于股票评论数据存在着许多专有人造词，这些词在股票市场中出现往往有特殊含义，不能以词表面意思去解释. 因此总结一批人造词词干，当程序读取数据时会先将这些人造词词干转换成具有该词特殊含义的普通词进行读取保存. 同理将评论数据中股票名称替换为对应股票代码，减少数据歧义，以此提升下一步分词和分类工作的准确率，然后通过2-gram、3-gram和4-gram这3种分词方法做对比实验.

2.3　特征词典构建

特征词典采用Chi-square降维和tf-idf方法组合进行构建. 首先通过Chi-square降维得到特征词. Chi-square形式化函数为

$$x^2 = \sum \frac{(A-T)^2}{T}. \tag{1}$$

其中，A为标注数据实际值，T为程序对标注数据的推测值. 通过计算得到的x^2表示实际值与推测值之间的差异程度，x^2值大小与关系程度大小呈正相关. 本文提取相关程度大的词作为特征词，然后通过tf-idf对特征词赋予权重并保存到特征词典中，为下一步模型训练提供条件. 选择Chi-square降维和tf-idf方法组合构建特征词典的原因在于，虽然Chi-square降维在特征选取上具有高效性，但却存在低频词缺陷. 因此，选择Chi-square降维与tf-idf方法组合进行使用来扬长避短.

2.4　模型训练

使用不同方法训练模型可以更好地分析金融语料特性和在不同环境下的适用性. 本文选用台湾大学林智仁教授所开发的Libsvm和Liblinear两种方法[20]进行比较，它们提供多种参数进行调整可适用于不同领域.

Libsvm默认基于RBF核函数. 高斯径向基函数(RBF)是一种局部性强的核函数，可以将一个样本映射到一个更高维空间内，形式化函数为

$$K\left(x_i, x_j\right) = \exp\left[\frac{-\left\|x_i - x_j\right\|^2}{2\sigma^2}\right]. \tag{2}$$

其中参数σ^2为高斯核函数方差，σ控制了函数径向作用范围. σ过小易出现“过拟合”，σ过大则可能出现“欠拟合”.

Liblinear基于线性核函数，特征空间到输入空间

维度是一致的，主要用于线性可分情况. 其优点是参数少、速度快，对于线性可分数据效果理想. 形式化函数为

$$K\left(x_i, x_j\right) = x_i \times x_j. \tag{3}$$

2.5 模型测试

根据2.2和2.4节的结论，本文使用不同n-gram算法分别与Liblinear和Libsvm进行组合实验. 实验结果发现Libsvm虽然得到最高准确率，但其计算时间较长且在不同n-gram方法下准确率波动大，所以相对于Liblinear方法较不好.

3 实验及结果分析

3.1 数据收集与标注

从股吧爬取10万条有效数据进行处理，收集数据时间区间为2015年7月~2017年2月. 首先过滤金融无关数据，再从中随机抽取2万条数据作为最终实验数据，然后对数据进行人工标注，每条数据由3个金融专业人员进行标注. 标注的内容有情感极性（积极、消极、中性）、来源（官方公告、论坛、新闻、其他）、情感词主体（主语、宾语）、主客观（主观句、客观句）和是否存在否定词. 最后通过投票程序进行打分得出每条数据最终标注结果. 在数据收集时发现金融语料的一些自身特性，例如语料长度一般不长（不会超过30个字符），语料中存在许多股票市场特有词汇，例如：飘红、跳水等.

本文采用词干词典对股票市场特有词汇做相对应替换处理. 为防止股市专有词语（例如：跳水、红三兵和抢搭车等）的歧义性对之后分词和分类准确率造成影响，在程序中设计增加一套正则表达式，用来匹配数据中歧义性较强的股市词语，并将这些词语替换成具有相同含义的普通词语给程序进行读取处理和保存，提高程序对数据的理解. 人造词替换举例见表1.

表1 部分人造词替换

Tab.1 Part of artificial words replacement

替换前	替换后
熊市	下跌
牛市	上涨
跳水	大跌
狂拉	大涨

3.2 实验设置

根据标注项目将情感分析的结果分为3类，这也是当前较为通用详细的分类方法. 首先，三分类定义如下

$$C = \begin{cases} 0, & P > N且P > M; \\ 1, & N > P且N > M; \\ 2, & M > P且M > N. \end{cases} \tag{4}$$

式(4)中，P、N、M分别表示积极、消极、中性. C表示三分类类别，分别以0, 1, 2表示. 当积极概率最大，则C为0；当消极概率最大，则C为1；当中性概率最大，则C为2. 由于需要对训练出的模型进行验证分析，本文使用了n-fold交叉验证法，这里n=5. 依次使用5等份中的4份数据作为训练集，1份数据作为测试集来测试模型的准确率、召回率. 为验证第二节中的推论，将设置如下实验：实验1，分类算法参数不变，在数据预处理中增加Stemming方法；实验2，在实验1基础上继续调整n-gram分词法参数分别与Libsvm和Liblinear结合；实验3，在实验2基础上调整特征保留数.

3.3 实验结果与分析

通过数据预处理与分类算法参数组合上的调整进行实验，从而得到不同的模型. 主要为Libsvm和Liblinear两种预测模型，参数在这两种模型上进行调整，以此做出比较并得到结论. 各评价指标公式分别为

$$\text{准确率(precision)}: P_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j}. \tag{5}$$

$$\text{召回率(recall)}: R_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FN}_j}. \tag{6}$$

$$F_1 : F_\beta = \frac{\left(\beta^2 + 1\right) \times \mathrm{pricision} \times \mathrm{recall}}{\beta^2 \times \mathrm{pricision} + \mathrm{recall}}. \tag{7}$$

结合式(4)的定义，式(5)~(7)中，TP表示人工标注和分类程序得到相同的C值；FP表示分类程序得到一个C值，而对应人工标注值与其不同；FN表示人工对数据标注一个C值，而程序得到值与之不同. 在F_1测试值公式中，β是一个可以调整准确率和召回率的重要参数，在本论文中β取值为1.

实验分为两步进行. 采用了n-fold交叉验证法，这里n=5. 具体的做法是，把2万条语料分成5等份，依次取1份作为测试子集，其余4份作为训练子集来构造多个分类器并计算分类准确率等指标，最后计算出5个分类器准确率、召回率等指标的平均值作为结果. 第一步，依次将5等份中的4份语料数据作为训练数据结合Stemming法进行数据预处理，并多次以不

同方法训练得到分类模型. 第二步，依次将5等份中的1份数据作为测试数据，直接使用第一步得到的多个模型进行分类，比较程序标注结果与人工标注结果，从而得出不同模型多个准确率、召回率指标并计算出对应的平均值，然后在不同n-gram方法和特征数间再次进行比较，得到数据比较结果见表2.

表 2　不同组合得到的三分类结果

Tab.2　Results in different combinations

组合	liblinear					libsvm				
	准确率	召回率	F_1	用时/ms	标准差	准确率	召回率	F_1	用时/ms	标准差
InitData+InitProgram	0.531	0.542	0.536	1 237	0.048	0.554	0.567	0.561	4 765	0.045
InitData+2-gram+3000	0.558	0.562	0.560	1 114	0.064	0.610	0.614	0.612	9 467	0.071
InitData+3-gram+3000	0.496	0.518	0.507	1 460	0.044	0.474	0.479	0.476	4 816	0.052
InitData+4-gram+3000	0.454	0.504	0.478	1 355	0.057	0.456	0.481	0.468	4 731	0.063
Stemming+InitProgram	0.648	0.654	0.651	1 265	0.057	0.590	0.604	0.597	2 119	0.045
Stemming+2-gram+3000	0.684	0.715	0.699	1 140	0.025	0.705	0.716	0.710	5 697	0.041
Stemming+3-gram+3000	0.499	0.648	0.564	1 167	0.033	0.447	0.523	0.482	4 084	0.064
Stemming+4-gram+3000	0.531	0.658	0.588	1 068	0.051	0.387	0.474	0.425	4 217	0.063

表2中InitData和InitProgram分别表示原始数据和原始程序. Stemming表示经过词干词典处理后的数据. OptProgram表示优化调整参数后的程序. 图2反映在特征数变化过程中分类准确率变化趋势，可以看出在特征数超过3 000后准确率没有明显提升.

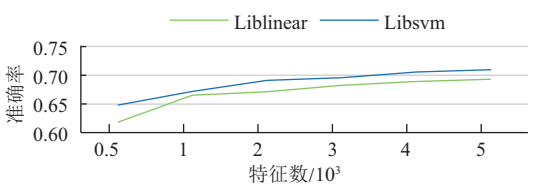

图 2　特征数对分类准确率的影响

Fig.2　The influence of feature number on classification accuracy

从表2可以看出：(1) 加入Stemming词干词典对原始数据进行预处理后在2-gram下的分类准确率有明显的提升. (2) 对程序参数进行调整后准确率得到进一步提升. (3) 2-gram相对于3-gram和4-gram得到较好准确率. (4) Liblinear分类算法在运行时间上远小于Libsvm分类. 针对以上4点可以得出：首先，基于Stemming词干词典的数据预处理与分类程序调参对于分类准确率有着积极影响. 其次，2-gram分词法针对金融短语料有着最好分词效果. 最后，实验结果体现了Libsvm在分类准确率上的优势和Liblinear在计算时间上的优势.

4　结语

本文通过自己收集并标注的2万条金融短语料对THUCTC分类器进行优化调整. 为提升THUCTC对金融短语料的分类效果，在数据处理中增加一套基于正则表达式的Stemming词干词典方法，并调整n-gram和特征保留参数，以此结合Libsvm与Liblinear进行分类模型构建. 实验结果表明增加Stemming词干词典方法并调整分类器参数的组合优化可以得到更好的结果.

未来工作还需要进一步改进与完善，主要体现在两个方面. 一方面，在语料标注信息中，目前主要使用情感极性. 未来尝试将来源、主体、主客观、是否存在否定词等标注信息也组合到情感分析处理中. 另一方面，考虑如何将语料中的表情、语音等非结构化数据与结构化数据联合使用.

参考文献：

[1] SAIF H, HE Y, FERNANDEZ M, *et al*. Contextual semantics for sentiment analysis of Twitter [J]. Information Processing & Management, 2015, 52(1): 5-19.

[2] QIN Z, CONG Y, WAN T. Topic modeling of Chinese language beyond a bag-of-words [J]. Computer Speech & Language, 2016, 40: 60-78.

[3] QIAN Q, HUANG M, LEI J, *et al*. Linguistically regularized LSTMs for Sentiment Classification[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Canada: ACL, 2017: 1679-1689.

[4] LI J, SUN M S. Scalable term selection for text categorization[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.Prague: EMNLP-CoNLL, 2007: 774-782.

[5] LI J Y, SUN M S, *et al*. A comparison and semi-quantitative analysis of words and character-bigrams as features in Chinese text categorization[C]//Proceedings of the 21st In-

ternational Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney: ACL, 2006: 17-21.

[6] 时永宾, 余青松. 基于共现词卡方值的关键词提取算法[J]. 计算机工程, 2016, 42(6): 191-195.
SHI Y B, YU Q S. Key words extraction algorithm based on Chi-square value of co-concurrence words [J]. Computer Engineering, 2016, 42(6): 191-195.

[7] ZHU J, WANG H, ZHU M, *et al*. Aspect-based opinion polling from customer reviews [J]. IEEE Transactions on Affective Computing, 2011, 2(1): 37-49.

[8] 李心丹. 行为金融理论: 研究体系及展望[J]. 金融研究, 2005(1): 175-190.
LI X D. Behavioral finance theory: Research system and prospects [J]. Journal of Financial Research, 2005(1): 175-190.

[9] BEKAERT G, EHRMANN M, FRATZSCHER M, *et al*. The global crisis and equity market contagion [J]. Journal of Finance, 2014, 69(6): 2597-2649.

[10] GIANNETTI M, WANG T Y. Corporate scandals and household stock market participation [J]. Social Science Electronic Publishing, 2016, 71(6): 2591-2636.

[11] AVDIS E. Information tradeoffs in dynamic financial markets [J]. Journal of Financial Economics, 2016, 122(3): 568-584.

[12] EDELEN R M, INCE O S, KADLEC G B. Institutional investors and stock return anomalies [J]. Journal of Financial Economics, 2016, 119(3): 472-488.

[13] CHANG T Y, HARTZMARK S M, SOLOMON D H, *et al*. Being surprised by the unsurprising: earnings seasonality and stock returns [J]. Social Science Electronic Publishing, 2016, 30(8): 281-323.

[14] RUAN X, WILSON S, MIHALCEA R. Finding optimists and pessimists on Twitter[C]// Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016: 320-325.

[15] 张对. 网络股评影响股市走势吗——基于股票情感分析的视角[J]. 现代经济信息, 2015(1): 355-357.
ZHANG D. Internet stock analysts do affect the stock market trend_stock-based sentiment analysis perspective [J]. Modern Economic Information, 2015(1): 355-357.

[16] 江腾蛟, 万常选, 刘德喜, 等. 基于语义分析的评价对象-情感词对抽取[J]. 计算机学报, 2017, 40(3): 617-633.
JIANG T J, WANG C X, LIU D X, *et al*. Extracting target-opinion pairs based on semantic analysis [J]. Chinese Journal of Computers, 2017, 40(3): 617-633.

[17] 饶东宁, 温远丽, 魏来, 等. 基于Spark平台的社交网络在不同文化环境中的中心度加权算法[J]. 广东工业大学学报, 2017, 34(3): 15-20.
RAO D N, WEN Y L, WEI L, *et al*. A weighted centrality algorithm for social networks based on Spark platform in different cultural environments [J]. Journal of Guangdong University of Technology, 2017, 34(3): 15-20.

[18] 林穗, 赵菲. 基于Spark的线性模型在广告投放系统中的应用研究[J]. 广东工业大学学报, 2016, 33(5): 28-33.
LIN S, ZHAO F. An application research of linear model in the advertising system based on Spark [J]. Journal of Guangdong University of Technology, 2016, 33(5): 28-33.

[19] 王洪伟, 郑丽娟, 刘仲英, 等. 中文网络评论的情感特征项选择研究[J]. 信息系统学报, 2012(1): 76-86.
WANG H W, ZHENG L J, LIU Z Y, *et al*. Emotional feature selection of Chinese web comments [J]. China Journal of Information Systems, 2012(1): 76-86.

[20] CATAL C, GULDAN S. Product review management software based on multiple classifiers [J]. Iet Software, 2017, 11(3): 89-92.