基于情感词典与 LDA 模型的股市文本情感分析

延 丰 杜腾飞 毛建华 刘学锋
(上海大学通信与信息工程学院 上海 200444)

摘 要：建立了一种基于股票情感词典与 LDA 分析股票文本情感倾向的模型。针对股票文本情感分析中情感词典不全面与句子分析片面的问题，构建较为全面的股票情感词典，同时以句子的倾向性、程度性与相关性三方面分析股票文本情感。引入针对股票的词语、程度性词语与转折性词语构建较为全面的情感词典；抽取预处理之后的股票文本句子的情感词；利用句子算法计算句子倾向、程度向量，并对句子向量利用支持向量机(SVM)和 K 均值算法分类；利用 LDA(latent dirichlet allocation)对情感词计算文档--主题、文档--词语概率分布，以此概率分布获取句子的相关性；综合句子的倾向性、程度性、相关性计算句子情感；最后，通过句子情感获取股票文本的情感倾向比例。通过对百度新闻经济板块收集的股票文本进行实验并与其他算法比较，该模型对句子与文本分类准确率提高到 82.78% 与 84.14%。

关键词：股票；文本情感分析；情感词典；LDA；支持向量机；K 均值
中图分类号：TN9 **文献标识码**：A **国家标准学科分类代码**：510.4030

Stock text sentiment analysis based on emotion dictionary and LDA model

Yan Feng Du Tengfei Mao Jianhua Liu Xuefeng
(School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China)

Abstract: This paper improvesan analysis model ofstock text sentiment orientation based on stock emotion dictionary and LDA. For the problem of the incomplete stock dictionary and the unilateral analysing of sentence, this paper constructs a relatively complete stock emotional dictionary, and analyses the emotional tendency of stock text from the three aspects of the tendency, the degree and the correlation of sentence. This paper builds a more complete emotional dictionary by introducing the stock words, the degree words and the turning words. Then it extracts the sentiment words from stock text sentences after the processing. It educes the sentence tendency and degree vector from the emotional words in the sentence by the sentence algorithm, and uses SVM and K mean algorithm to classify sentence vector. The paper gets the words distribution of the topic and the topic distribution of document from the sentiment words by LDA model, and obtains the correlation of the sentence by this probability distribution. Finally it synthesizes sentence tendency, degree, correlation to obtain the sentence emotion, and acquires the emotional tendency of stock text through the sentence emotion. At last, This paper collects the text of Baidu news in the sector of the economy as the experimental material, and does experiment and compares with other algorithms. Experimental results show that the accuracy of sentence and article classification are 82.78% and 84.14%.

Keywords: stock; text sentiment analysis; emotion dictionary; LDA; SVM; K-means

0 引 言

散户是目前中国股票市场的重要参与者，其股票操作往往表现出追涨杀跌特点，反映出明显的情绪特征。投资者情绪研究成为证券市场研究的重要内容之一。当前研究主要集中于投资者情绪与股票特征、证券市场指数、管理层业绩预告策略、开放式基金业绩与激励机制激励效率等相关性方面的研究。尽管近几年在文本情感分析研究领域已取得丰硕研究成果[1-3]，但大部分集中于对社会、产品、商业分析等文本的情感研究[3-4]，而对投资者评论的情感分析研究相对缺乏。目前，文本情感分析方法主要有基于词典、基于语句、基于 LDA[5]模型的 3 种方法。

基于词典方法通常对抽取的文本情感词进行分类。肖江等人[6]通过构建领域情感词典和副词情感词典能，实现了对中文微博情感的更好更准确的分析。王新宇[7]一种基于旅游情感词典和机器学习相结合的旅游网络点评的情感

收稿日期：2017-04

倾向分析方法。张建华等人[8]基于抽取情感词与 LDA 特征表示的情感分析方法,对产品评论进行褒贬二元分类。但是考虑的情感词语不全面,转折词、主张词、程度词等没有考虑到;同时当前缺乏针对股票的情感词典,需要构建专门的股票词典。

基于语句方法通常对文本中语句进行分类。但是研究方面比较单一,只考虑句子倾向性或只考虑句子相关性。传统文本句子级情感研究一般只考虑句子的倾向性分类,赵妍妍等人[1]的句子褒贬二元分类。贾电如等人[9]采用浅层次句法结构分析和深层次语义分析相结合的算法计算相似度,以提高主观题自动评分的效率和准确率。但仅仅考虑语句语义整体相似度,而没有考虑到句子在整个段落、篇章中的情感倾向分析[10]。高雪霞等人[11]从语句权重计算和冗余处理等方面进行改进,实现了一个基于语句类似度优化计算的自动摘要算法,但没有考虑关键词语与整个主题的权重关系,致使自动摘要易漏掉关键信息。而句子对于情感的倾向性的程度大小(命名为句子程度性)计算目前还很缺乏。但是句子程度性是句子在文本中表现出的态度、情感、立场的痕迹,传达了句子的评价、感受的大小,因此计算句子程度性成为可能。

基于LDA模型通常利用LDA模型挖掘文本内的主题与词之间的分布关系,挖掘出文本中潜在的语义知识。孙艳等人[12]通过 LDA 模型构建 UTSU 模型。王鹏等人[13]提出基于 LDA 主题模型的文本聚类和聚簇方法。Ding 等人[14]基于 LDA 构建的 HDP-LDA 模型,自动确定真实的语句,提取情感词。

基于以上的问题构建股票词典,情感词语不全面,研究句子单方面情感,没有考虑句子倾向性的中性分类,没有考虑句子的程度性,没有考虑词语和句子与主题的联系。本文提出一种基于股票情感词典与句子的股票文本情感倾向性分析模型。该模型依据建立的股票情感词典对股票文本进行词汇抽取,且对抽取的词汇通过句子算法建立句子倾向与程度向量,之后通过支持向量机(SVM)与 K 均值划分句子分类,获取句子倾向性与程度性分类;通过 LDA 模型建立主题--词语与文本--主题分布,计算句子相关性分布;最后,综合句子的倾向性、程度性与相关性获取句子情感,通过句子情感得出投资者倾向比例。

1　研究方法

模型包括 4 个部分:股票词典与词汇级别划分、句子倾向性与程度性、句子相关性、文本情感。如图 1 研究流程所示。

图 1 中,在对股票文本进行词语抽取前,为了更好地获取文本情感词语,须对股票文本进行预处理,主要步骤包括:文本切分为句子、句子以逗号与分号切分为分句、分词、词语标记、删除停顿词。

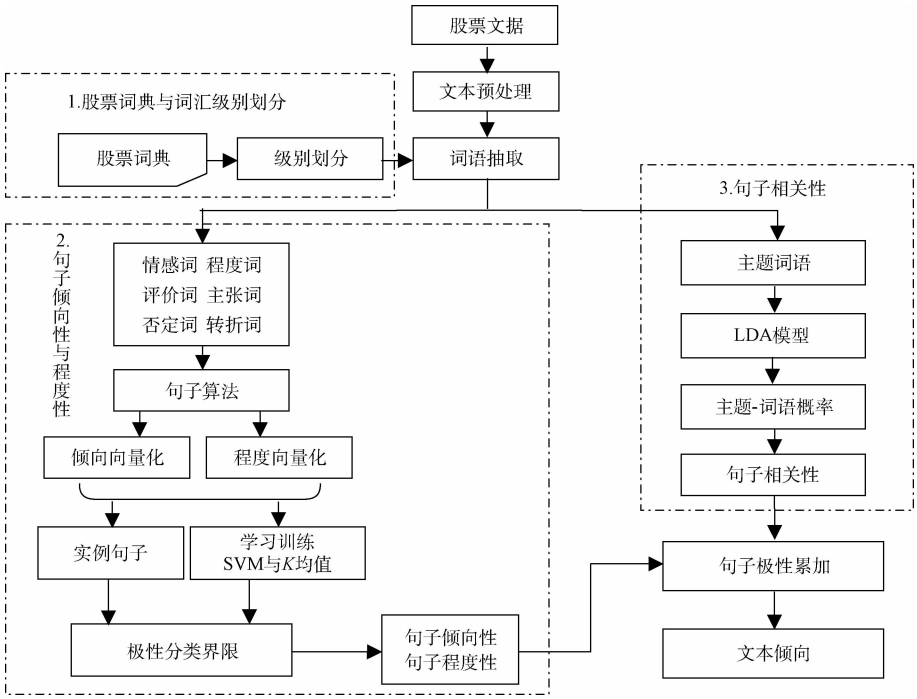

图 1　研究流程

1.1　股票词典与极性级别划分

目前,国内外还没有完全覆盖所有领域的情感词典,针对股票的情感词典更加稀少,而情感词典又是文本分析中最基础的工作。为解决当前缺乏针对建立股票的情感词典的问题。本文收集 1 500 篇含有股票极性的文本,提取针对股票的极性词语,比如“上涨、涨幅、下跌”。使得情

感词典更加适应股票文本分析。

针对文本情感分析中情感词语考虑不全面的问题。本文引入包含正面与负面极性词、主张词、程度词的台湾大学 NTUSD-简体中文情感极性词典[15]。该词典主要拥有 2 810 个正极性词语、8 276 个负极性词语、还包含 219 个程度词、38 个主张词、3730 个评价词等词语。

情感词的极性判断是也文本分析的重要方面。目前，判决界限模糊，本文结合文献[16]中的情感词极性程度分级对建立的股票词典进行划分为 5 个级别。

1.2　句子倾向性与程度性

针对没有考虑句子的程度性的问题，本文引入句子程度性算法，从而解决这个问题。针对没有考虑句子倾向性的中性分类的问题，本文结合一步三分法的思想[17]，在句子极性分类中分为把句子分为褒、贬、中性三大分类。

$$Y=\begin{cases}(\sum_{i=0}^{i=n}S\text{ 情感词}+\sum_{i=0}^{i=n}S\text{ 评价词})\times(\prod_{i=0}^{i=n}S\text{ 否定词}), & \text{单分句}\\ \begin{cases}Y\text{ 前句分数}+Y\text{ 后句分数}, & \text{两分句极性相同}(Y\text{ 参考单分句})\\ \alpha\times Y\text{ 前分句分数}+(2-\alpha)\times Y\text{ 后分句分数}, & \text{两分句极性相反}(Y\text{ 参考单分句},\alpha\text{ 转折因子}\end{cases}\\ \sum_{i=0}^{i=n}Y[i]\text{ 分句分数}, & \text{分句数量}\geqslant 3(Y\text{ 参考为单分句})\end{cases}\tag{1}$$

$$X=\prod_{i=0}^{i=n}(S[i]\text{ 转折词}\times S[i]\text{ 程度词}\times S[i]\text{ 主张词})\tag{2}$$

式中：α 为转折因子，由句子中前后两分句的分数差距决定，即 $\beta=\|Y\text{ 前分句}|-|Y\text{ 后分句}\|$，$\alpha$ 主要调整前后分句极性不相同的句子，使得前后分句极性更加明显；而对于前后分句极性相同的句子，进行前后分句增强叠加。α 公式如式(3)所示。

$$\alpha_1=\begin{cases}0.8 & 0\leqslant\beta\leqslant 0.2\\ 0.7 & 0.2<\beta\leqslant 0.4\\ 0.6 & 0.4<\beta\leqslant 0.6\\ 0.5 & 0.6<\beta\leqslant 1\end{cases}\tag{3}$$

$$\alpha_2=-\beta+1$$

$$\alpha_3=\sqrt{1-\beta^2}$$

$$\alpha_4=\sqrt{1-(\beta-1)^2}+1$$

1.2.1　句子向量分类

基于机器学习的情感倾向分类方法是通过机器学习对标注的语料进行学习训练并生成倾向分类器，从而对文本进行分类。目前主流的分类方法有支持向量机 SVM、朴素贝叶斯和最大熵等。邹明[18]采用 SVM 算法对恶意发帖进行分类，进而对帖子进行文本情感分析，显示了 SVM 分类方法效果最好。在机器学习领域，SVM 是一个有监督的学习模型，通常用来进行模式识别、分类、以及回归分析。选择不同的核函数，可以生成不同的 SVM，常用的核函数有以下 4 种：线性核函数、多项式核函数、径向基核函数、二层神经网络核函数。经实验线性核函数对文本有最好的分类效果。通过 SVM，对 X 与 Y 句子作为特征表示向量进行机器学习训练，使得句子分成正面、中性与负面三大类。

1.2.2　句子向量

定义 1 S 代表词语分数；Y 为句子的倾向性分数；X 为句子程度性分数；Z 为句子的相关性分数；对经过文本处理后的股票文本提取情感词、评价词、否定词，以此三类词语计算得到句子倾向性；提取程度词、主张词、转折词，以此三类词语计算得到句子程度性；从而尽量全面地考虑到情感词语。对提取到的词语，通过句子算法获取句子倾向性分数与程度性分数。句子算法如式(1)、(2)所示。

K-means[19]算法是硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，它是将数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。

K-means 算法过程如下：1)初始化。输入基因表达矩阵作为对象集，输入指定聚类类数，并在对象集中随机选取 N 个对象作为初始聚类中心。设定迭代中止条件，比如最大循环次数或者聚类中心收敛误差容限。2)进行迭代。根据相似度准则将数据对象分配到最接近的聚类中心，从而形成一类。初始化隶属度矩阵。3)更新聚类中心。然后以每一类的平均向量作为新的聚类中心，重新分配数据对象。4)反复执行 2)和 3)直至满足中止条件。其评价标准如下：

$$J(c,u)=\sum_{i=1}^{m}\|x^{(i)}-u_{c(i)}\|^2$$

C 个聚类中心，μ_c 为聚类中心。

本文设定同时给 X 与 Y 形成的坐标系，给予 3 个中心点，进行 K-means 算法，直到分类结果不变，分成三大类。

最后，综合 SVM 与 K-means 相同部分作为分类结果。从而形成 6 个分布：Y 的正面、中性、负面极性分布，以 Y_j $(j=1,2,3)$代指；X 的包含 Y 的正面、中性、负面属性的程度分布，以 $X_j(j=1,2,3)$代指。

1.3　句子相关性

针对没有考虑词语——主题与文档——主题的联系的问题，本文引入 LDA 模型，计算主题——词语与文档——主题的概率分布，通过这两种概率计算出包含这些词语的句子的相关性。与文本主题相关性有关的词语，包括情感词、否定词、评价词、程度词、主张词与转折词等等。通过 LDA 模型获得主题——词语的概率分布，进而通过计算出包含这些词语的句子的相关性。

1.3.1　LDA 模型

LDA 模型是一种主题模型,由文档、主题、词 3 层结构组成。对于语料库中的每篇文档,LDA 定义了如下生成过程(generative process)[13-14]:1)对每一篇文档,从主题分布中抽取一个主题;2)从上述被抽到的主题所对应的单词分布中抽取一个单词;3)重复上述过程直至遍历文档中的每一个单词。

语料库中的每一篇文档与 T 个主题的一个多项分布相对应,将该多项分布记为 θ。每个主题又与词汇表中的 V 个单词的一个多项分布相对应,将这个多项分布记为 ϕ。LDA 模型如图 2 所示。

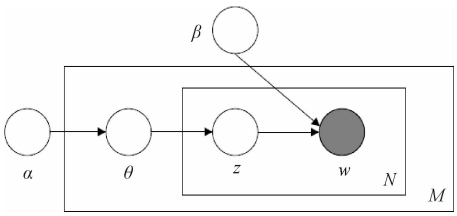

图 2　LDA 图模型

符号	表示
α 与 β	语料级别的参数
Z 与 w	单词级别变量
θ	文档级别的变量
N 与 M	评论中的单词总数与评论总数

LDA 联合概率为:

$$P(\theta,z,w \mid \alpha,\beta) = P(\theta \mid \alpha)\prod_{n=1}^{n=N} P(z_n \mid \theta)P(w_n \mid z_n,\beta)$$

1.3.2　相关性

LDA 模型通常有 EM 和 Gibbs[20]抽样两种方法,但 Gibbs 抽样计算量小,且能得到满意的输出结果,本文采用 Gibbs 方法。引用文献[8]中的实验最好效果,设置 topic 的个数 $K=15$[8],同时设置 $\alpha=2,\beta=0.5$,迭代次数为 2 000 次。其中每个短文本包含各自的文本主题。依次可以获取:主题——词语的概率分布,计为 E;文档——主题分布,计为 H。通过式(4)计算主题-句子的概率分布。

$$Z_{[i]} = \sum_{i=0}^{i=n} P(C_i \mid E,H) \tag{4}$$

1.4　文本比例

为对股票文本中的倾向比例进行统计,本文对 1.2 与 1.3 节的结果——句子的倾向性、程度性与相关性通过算法(5)获得三类句子情感;再对句子情感进行算法(6)计算,得出整篇股票文本的 3 种倾向所占的比例,其中 $X_{[i]}$ 与 $Y_{[i]}$ 决定 j 的取值。

$$S_{正,中,负面} = \sum_{i=0}^{i=n} P(X_{[i]} \mid X_j) \cdot P(Y_{[i]} \mid Y_j) \cdot P(Z_{[i]} \mid Z) \tag{5}$$

最后,统计文本倾向比例:

$$\begin{aligned} Y_{全文正面比例} &= S_{正面}/(S_{正面}+S_{中性}+S_{负面}) \\ Y_{全文中性比例} &= S_{中性}/(S_{正面}+S_{中性}+S_{负面}) \\ Y_{全文负面比例} &= S_{负面}/(S_{正面}+S_{中性}+S_{负面}) \end{aligned} \tag{6}$$

2　实验及结果分析

2.1　数据集收集

本文研究对象选择百度新闻经济板块的股票文本。通过收集投资者评价文本,对文本词语进行归纳总结,同时对比台湾大学 NTUSD-简体中文情感极性词典,补充股票相关的词语:正极性词语补充 104 个、负极性词语补充 248 个;同时,添加 29 个转折词、30 个否定词、681 个停顿词。从而建立起较全面的股票词典。同时基于文献[16]中的情感词极性程度分级,划分出的 5 个级别,部分例子如表 1 所示。

表 1　部分词语分类等级

词语	进场	斩获	多头	讥刺
分数	1.2	1.2	1.3	−1.3
词语	开高	涨停板	企稳	阴跌
分数	1.3	1.5	1.1	−1.3

本文选取 2014 年 9 月到 12 月百度新闻中的经济板块为实验数据,对文本句子分类进行实验。为获得标准数据集,选择 5 个人对抽样句子做情感标记,表 2 给出了两两之间的标注相同的百分比。

表 2　人工句子标记比较

对比人	A^B	A^C	A^D	A^E	B^C
百分比/%	87.8	89.6	90.5	93.5	90.3
对比人	B^D	B^E	C^D	C^E	D^E
百分比/%	91.7	92.1	91.5	94.2	94.7

从表 2 中可以看出 D 和 E 的情感标注相同百分比是最高的,但是为获取标准句子数据集,本文综合 5 个人的情感标记,选取 5 个人相同的情感标记句子,共计 2 725 条,其中正面情感句子 1 264 条、中性情感句子 424 条、负面情感句子 1 037 条。

基于以上的获取标准句子数据集的方法,选取 1 279 篇文本作为标准文本数据集,其中正面文本 735 篇、负面文本 544 篇。

2.2　评测方法

对实验数据集中的句子与文本,进行情感极性分析,并与标准数据集的标注进行对比。而评价方法[21]目前通常采用准确率(Precision)、召回率(Recall),利用 F 测度(F-measure)。

$$Precision = \frac{\text{检测正确的句子或文章}}{\text{总的句子或文章}}$$

$$Recall = \frac{\text{检测正确的句子或文章}}{\text{总的正确的句子或文章}}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

2.3　实验结果

本文通过对转折因子 α 的 4 个公式进行实验，图 3 所示为所获取的句子极性准确率对比。从图表中看出 α_3 的准确率最高，因此本文选取 α_3 作为转折因子。

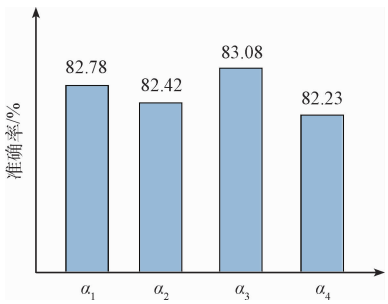

图 3　α 准确率

分别用本文方法、一步三分类结合 SVM 方法[17]与 OpinionFinder[22]对上述数据集进行实验。表 3 给出了 3 种算法对文本句子分类的结果。表 4 给出了通过 3 种算法获取到句子分类结果后，结合 1.4 节的文本比例计算文本倾向比例的结果。

表 3　句子分类比较　　(%)

针对句子方法	*Precision*	*Recall*	*F*
一步三分类+SVM 分类	66.87	80.21	72.94
OpinionFinder	73.35	83.44	78.07
本文方法	82.78	87.72	85.18

表 4　文本分类比较　　(%)

针对文本方法	*Precision*	*Recall*	*F*
一步三分类+SVM	71.41	82.27	76.46
OpinionFinder	78.35	84.46	81.27
本文方法	84.14	87.65	85.86

从表 3 中可以看出本文方法准确率为 82.78%，相对于一步三分类结合 SVM 算法的 66.87%提高了 15.91%，相比于 OpinionFinder 提高了 9.43%；同时本文的召回率也得到了一定程度的提高。这表明本文算法得到的股票文本句子分类正确率得到了提高。

从表 4 中可以看出通过本文方法判断文本极性的准确率相较于前两种算法得到了一定的提高，分别提高了 12.73%与 5.79%，同时召回率也有提高。主要是因为本文方法利用三维方式表示句子形态：倾向性、程度性、相关性。

3　结　　论

本文提出了一种基于股票情感词典与句子的股票文本情感倾向的模型。首先解决了构建股票词典，考虑到情感词语句子的全面性，对句子倾向性的中性进行了分类，另外，还解决了词语和句子与主题的联系等问题。通过建立针对股票的全面多级性情感词典，利用三维方式表示句子情感，凸显多分句的前后极性。以百度新闻经济板块收集的股票文本为实验数据，并与其他算法比较，结果表明，本文的方法分析股票文本的句子与文本情感更加准确。下一步将加大股票文本测试数据集，并对算法进一步优化，并与其它情感分析系统进行对比，重点是引入同义词，引入混合模型对词汇、句子进行上下文联系，分析句子的位置等。

参考文献

[1]　赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010,21(8):1834-1848.

[2]　WU J, GAO W, ZHANG B, et al. Online web sentiment analysis on campus network[C]. Fourth International Symposium on Computational Intelligence and Design, IEEE, 2011:379-382.

[3]　KARAMIBEKR M, GHORBANI A A. Sentiment analysis of social issues[C]. International Conference on Social Informatics. IEEE, 2012:215-221.

[4]　DEHKHARGHANI R, YILMAZ C. Automatically identifying a software product′s quality attributes through sentiment analysis of tweets[C]. International Workshop on Natural Language Analysis in Software Engineering. IEEE, 2013:25-30.

[5]　张培晶，宋蕾. 基于 LDA 的微博文本主题建模方法研究评述[J]. 图书情报工作，2012,56(24):120-126.

[6]　肖江，丁星，何荣杰. 基于领域情感词典的中文微博情感分析[J]. 电子设计工程，2015,23(12):18-21.

[7]　王新宇. 基于情感词典与机器学习的旅游网络评价情感分析研究[J]. 计算机与数字工程，2015,44(4):578-582.

[8]　张建华，梁正友. 基于情感词抽取与 LDA 特征表示的情感分析方法[J]. 计算机与现代化，2014(5):79-83.

[9]　贾电如，李阳明. 基于语句结构及语义相似度计算主观题评分算法的研究[J]. 信息化纵横，2015(5):5-7.

[10]　黄萱菁，张奇，吴苑斌. 文本情感倾向分析[J]. 中文信息学报，2011,25(6):118-126.

[11]　高雪霞，贾海龙. 基于语句类似度优化计算改进自动摘要算法研究[J]. 计算机应用与软件，2013,30(9):160-162,182.

[12]　孙艳，周学广，付伟. 基于主题情感混合模型的无监督

文本情感分析[J]. 北京大学学报(自然科学版),2013,49(1):102-108.

[13] 王鹏,高铖,陈晓. 基于 LDA 模型的文本聚类研究[J]. 情报科学,2015,33(1):63-68.

[14] DING W, SONG X, GUO L, et al. A novel hybrid HDP-LDA model for sentiment analysis[C]. IEEE/WIC/ACM International Joint Conferences on Web Intelligence, IEEE, 2013:329-336.

[15] 阳爱民,林江豪,周咏梅. 中文文本情感词典构建方法[J]. 计算机科学与探索,2013,7(11): 1033-1039.

[16] 任远,巢文涵,周庆,等. 基于话题自适应的中文微博情感分析[J]. 计算机科学,2013, 40(11): 231-235.

[17] 周胜臣,瞿文婷,石英子,等. 中文微博情感分析究研综述[J]. 计算机应用与软件,2013, 30(3):161-164.

[18] 邹明. 基于情感分析的恶意发帖检测方法研究[J]. 电脑知识与技术,2014,10(7):1403-1406.

[19] YAO M Y,PI D C,CONG X X. Chiness text clusteringalgorithm based k-means [C]. International Conference on Services Science, Management and Engineering, 2010.

[20] XIAO H, STIBOR T. Efficient collapsed Gibbs sampling for latent dirichlet allocation[J]. Journal of Machine Learning Research, 2010, 13:63-78.

[21] 陈涛,徐睿峰,吴明芬,等. 一种基于情感句模的文本情感分类方法[J]. 中文信息学报,2013,27(5): 67-74.

[22] BOLLEN J, MAO H, ZENG X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2010, 2(1):1-8.

作者简介

延丰,硕士研究生,主要研究方向为 Web 文本数据挖掘与可视化。

E-mail:314627745@qq.com