计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):207−211 [doi: 10.15888/j.cnki.csa.006687]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 PLSA 的新闻评论情绪类别自动标注方法①

林江豪[1,3], 顾也力[2], 周咏梅[1,3], 阳爱民[1,3]

[1](广东外语外贸大学 语言工程与计算实验室, 广州 510006)
[2](广东外语外贸大学 东方语言文化学院, 广州 510420)
[3](广东外语外贸大学 信息科学与技术学院, 广州 510006)
通讯作者: 顾也力, E-mail: guyeli2018@foxmail.com

摘　要: 针对大规模语料手动标注困难的问题, 提出利用概率潜在语义分析 (PLSA) 模型的新闻评论自动标注方法. 利用 PLSA 计算获得语料集的“文档-主题”和“词语-主题”概率矩阵; 基于情感本体库和“词语-主题”概率矩阵, 认为某一类情绪词汇出现的概率最高的主题与词汇的情绪类别相同, 对主题进行情绪类别标注; 最后, 基于“文档-主题”概率矩阵, 认为出现在某一主题概率最高的文档与主题的情绪类别相同, 通过“词汇-主题-文档”三者的关系, 达到自动标注的效果. 实验结果表明, 本文提出的方法准确率可达到 90% 以上.

关键词: 语料库; 情绪类别; PLSA 模型; 语料标注; 自动标注

引用格式: 林江豪,顾也力,周咏梅,阳爱民.基于 PLSA 的新闻评论情绪类别自动标注方法.计算机系统应用,2019,28(1):207–211. http://www.c-s-a.org.cn/1003-3254/6687.html

Automatic Annotation of News Comments Emotion Based on PLSA

LIN Jiang-Hao[1,3], GU Ye-Li[2], ZHOU Yong-Mei[1,3], YANG Ai-Min[1,3]

[1](Laboratory for Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006, China)
[2](Faculty of Asian Languages and Cultures, Guangdong University of Foreign Studies, Guangzhou 510420, China)
[3](School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China)

Abstract: In order to solve the problem of manually annotating large-scale corpus, this study, based on the model of Probabilistic Latent Semantic Analysis (PLSA), proposed a method of automatic emotional annotation for news comments. First of all, the “doc-topic” and “word-topic” probability matrixes were computed by PLSA model. Then, drawing upon the “word-topic” together with the ontology lexicon, the emotional categories of the topics were annotated, with the presupposition that the emotional category of words is similar to those of words within the topic which occurs most frequently. Finally, the automatic annotation was made via the “doc-topic”, with the assumption that the emotional category of topics is equivalent to those of topics within the document which occurs most frequently. The experimental results showed that the accurate rate of the method proposed by this study reached about 90%.

Key words: corpus; emotional category; PLSA model; corpus annotation; automatic annotation

网络新闻是社会事件传播的载体, 人们通过评论新闻参与事件, 形成了海量信息. 对新闻评论文本进行情绪分析可应用到舆情管理、民意调查、商业营销情报等领域, 有着广阔的应用空间和发展前景[1,2]. 新闻评

① 基金项目: 教育部人文社会科学研究项目 (14YJA740011); 广州市哲学社会科学“十三五”规划 2018 年度课题 (2018GZQN27); 广东省科技计划项目 (2017A040406025); 国家自然科学基金 (61877013)
Foundation item: Humanities and Social Science Research Program of Ministry of Education (14YJA740011); Fund for Year 2018, Philosophic Social Science of Thirteenth Five-Year Plan, Guangzhou Municipality (2018GZQN27); Science and Technology Program of Guangdong Province (2017A040406025); National Natural Science Foundation of China (61877013)
收稿时间: 2018-05-21; 修改时间: 2018-06-15; 采用时间: 2018-06-19; csa 在线出版时间: 2018-12-26

论语料库是进行新闻评论情绪分析研究的重要基础, 要提高语料的利用价值, 关键在于语料的标注, 所谓标注[3]就是对语料库中的原始语料进行加工, 把各种表示语言特征的附码标注在相应的语言成分上, 以便于计算机的识读. 然而, 规模庞大的评论文本如果通过人工标注, 是非常困难的. 当前在文本情绪分析研究领域还没有标准的语料库, 特别是细粒度的情绪标注语料库更是缺乏, 这在一定程度上影响了该领域的研究. 为了减轻标注人员的负担, 提高标注的效率和精确度, 减少标注的错误率, 非常有必要研究自动标注方法, 以便协助标注人员的工作. 因此, 探索研究新闻评论文本情绪类别自动标注方法是一项非常重要的工作.

目前, 在文本自动标注领域, 文献[4]提出了一种汉语意见型主观性文本标注语料库的构建方法, 重点讨论了语料的选取、标注、存储、检索和统计等工作. 阳爱民等提出了中文微博语料情感类别自动标注的方法, 包括基于关键词的、基于概率求和的和基于概率乘积的 3 种自动标注方法和一种集成标注方法, 实验验证了集成方法的准确率可达到 90% 以上[5]. 文献[6]对网络新闻评论数据的特点进行归纳总结, 选取不同的特征集、特征维度、权重计算方法和词性等因素进行文本情感自动标注. 文献[7]使用机器学习方法进行新闻的情感自动标注, 发现选择具有语义倾向的词汇作为特征项、对否定词正确处理和采用二值作为特征项权重能提高分类的准确率, 准确率能达到 90%. 文献[8]基于语义的方法, 实现了微博情感类别的自动标注. 文献[9]通过抽取主题句, 将抽取到的主题句与 JST 模型结合, 对新闻评论文本进行情感标注. 吴江等基于语义规则, 对金融领域的文本进行情感标注[10]. Khoo 等[11]案例分析了基于认知理论的在线新闻文本情感标注方法. Moreo 等[12]提出了一种基于词典的新闻评论情感自动标注系统 (LCN-SA), 多维度分析网民的情感倾向. Penalver-Martinez 等[13]运用本体论, 提出了基于特征的观点挖掘方法. 国内已公开的人工标注语料包括 NLPCC 评测的语料、谭松波等人标注的酒店评论语料等.

在现有语料情感类别自动标注研究中, 主要将标注结果设定为正向和负向来进行研究, 这种方法下自动标注的语料不能适用于细粒度的文本情绪分析. 鉴于此, 本文以新闻评论的乐、好、怒、哀、惧、恶、惊七类情绪作为标注的类别, 提出一种基于概率潜在语义分析 (Probabilistic Latent Semantic Analysis, PLSA) 的新闻评论情绪类别自动标注方法. 这种方法通过利用 PLSA 计算获得语料集的“文档-主题”和“词语-主题”概率矩阵, 基于“词语-主题-文档”之间的概率转换关系, 认为某一类情绪词汇出现的概率最高的主题与词汇的情绪类别相同, 对主题进行情绪类别标注; 认为出现在某一主题概率最高的文档与主题的情绪类别相同, 对文档进行情绪类别标注, 达到自动标注的效果. 文章的内容安排如下, 第 1 节重点介绍了基于 PLSA 的新闻评论情绪自动标注方法; 第 2 节给出了基于 PLSA 的概率矩阵抽取方法; 第 3 节对文本提出的方法进行实验验证和分析; 第 4 节给出了结论以及下一步的改进方向.

1 基于 PLSA 的新闻评论情绪自动标注方法概述

本文采用如图 1 所示的新闻评论情绪自动标注方法, 首先将采集的新闻评论集进行文本预处理, 分词后过滤掉停用词和无用词, 进行词频统计, 获得“文档-词汇”矩阵; 接着, 利用 PLSA 模型计算, 获得“词汇-主题”和“文档-主题”概率矩阵; 结合情感本体库[14]和“词汇-主题”, 认为某一类情绪词汇出现的概率最高的主题与词汇的情绪类别相同, 对主题进行情绪类别标注; 最后, 基于“文档-主题”概率矩阵, 认为出现在某一主题概率最高的文档与主题的情绪类别相同, 达到对文档情绪类别的自动标注.

根据图 1, 用 $[M_{\text{word-topic}}]_{m\times k}$ 和“文档-主题”矩阵 $[M_{\text{doc-topic}}]_{n\times k}$ 来表示分类为 k 个主题的 PLSA 计算结果, 其中 $[M_{\text{word-topic}}]_{m\times k}$ 表示词汇在对应主题中的概率, 也即词汇对主题的贡献度, 则对于词汇 $word_j$ 在 k 个主题中的概率 p 有 $p_j^1+p_j^2+\cdots+p_j^k=1$. 分别对每个主题下的词语概率按照由大到小排序, 利用情感本体库 OL, 抽取概率高的情绪词汇, 对情绪词汇的情绪强度直接加总计算, 得到主题在每一类情绪中的强度, 则主题在 m 类情绪中的权重分布 $Et=\{e_1,e_2,e_3,\cdots,e_m\}$, 通过判断 Et 中的最大值, 获得主题的情绪类别. 同理, 利用“文档-主题”矩阵, 认为对主题贡献度高的文档与主题的情绪类别相同, 对文档的情绪类别进行标注. 则基于 PLSA 的新闻评论情绪自动标注算法如算法 1.

算法 1. 基于 PLSA 的新闻评论情绪自动标注算法

输入: 情感本体库 OL, 语料集 $Data\_set$
输出: $[doc, e]_m$

步骤 1. 对 *Data_set* 进行预处理, 包括分词、词频统计等, 获得“文档-词频”矩阵 M_i;
步骤 2. 计算 $PLSA(M_i)$→“词汇-主题”矩阵$[M_{\text{word-topic}}]_{m\times k}$ 和“文档-主题”矩阵$[M_{\text{doc-topic}}]_{n\times k}$;
步骤 3. 逐列对$[M_{\text{word-topic}}]_{m\times k}$ 进行排序, 获取每个主题 z^j 中概率较高的情绪词汇, 得到 $Z^k=\{[w_1,w_2,w_3,\cdots,w_o]^1,[w_1,w_2,w_3,\cdots,w_p]^2,\cdots,[w_1,w_2,w_3,\cdots,w_q]^k\}$;
步骤 4. 在情感本体库查询情绪词的权重, 得到主题的情绪权重矩阵 $EZ^k=\{[wt_1,wt_2,wt_3,\cdots,wt_o]^1,[wt_1,wt_2,wt_3,\cdots,wt_p]^2,\cdots,[wt_1,wt_2,wt_3,\cdots,wt_q]^k\}$;
步骤 5. 对每个主题 z^j 的情绪权重进行加总, 得到 $Et^k=\{[e_1,e_2,e_3,\cdots,e_m]^1,[e_1,e_2,e_3,\cdots,e_m]^2,\cdots,[e_1,e_2,e_3,\cdots,e_m]^k\}$;
步骤 6. 逐列对 Et^k 进行排序, 获得情绪强度最强的类别为对应主题的情绪, 则主题情绪标注结果为 ZE^k;
步骤 7. 逐列对$[M_{\text{doc-topic}}]_{n\times k}$ 进行排序, 结合 ZE^k, 将对主题贡献度高文档的情绪类别标注为主题的情绪类别, 对每一个 *doc* 获得对应的情绪类别 *e*;
结束. 输出$[doc, e]_n$.

算法的最终输出为$[doc, e]_m$, 为验证该标注结果的准确性, 本文采集了凤凰网涉及中日关系的新闻评论语料, 自动筛选出含有两个情绪词汇以上的评论语料, 进行人工标注, 与自动标注结果对比验证$[doc, e]_m$ 的准确率.

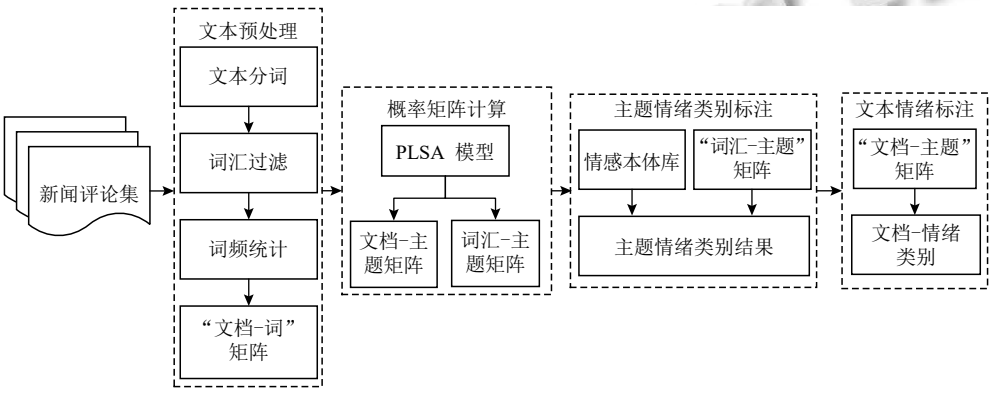

图 1 基于 PLSA 的新闻评论情绪自动标注过程

2 基于 PLSA 的概率矩阵抽取方法

PLSA 模型是由 Hofmann 在 1999 年提出的, 首先给定文档集$D=\{d_1,d_2,\cdots,d_n\}$和词集$W=\{w_1,w_2,\cdots,w_m\}$, 用 $freq(d_i, w_j)$ 表示词 w_j 在文档 d_i 中出现的概率, 则“文档-词语”共现矩阵 $M_{D\text{-}W}=[freq(d_i, w_j)]$. 假设主题类别$Z=\{z_1,z_2,\cdots,z_k\}$, k 为主题个数. PLSA 模型假设词与文档之间、话题与文档或者词之间的概率服从条件独立, 由此得到相应的联合分布概率为:

$$P(d_i,z_k,w_j)=P(d_i)P(z_k|d_i)P(w_j|z_k) \tag{1}$$

$P(d_i)$ 表示选择文档 d_i 的概率, $P(z_k|d_i)$ 表示某个主题 z_k 在给定文档 d_i 下出现的概率; $P(w_j|z_k)$ 表示词 w_j 在给定主题 z_k 下出现的概率, 本文基于该“词语-主题”的概率分布获取事件 Ev_t, 根据贝叶斯法则可得到:

$$P(d_i,w_j)=P(d_i)\sum_{l=1}^{k}P(z_l|d_i)P(w_j|z_l) \tag{2}$$

采用最大期望算法 (Expectation Maximization, EM) 算法对潜在语义模型进行拟合[13]. 用随机数初始化之后, 交替执行 E 步骤和 M 步骤进行迭代计算. E 步骤计算 (d_i, w_j) 所产生的潜在语义 z_k 的先验概率:

$$P(z_k|d_i,w_j)=\frac{P(z_k|d_i)P(w_j|z_k)}{\sum_{l=1}^{k}P(z_l|d_i)P(w_j|z_l)} \tag{3}$$

在 M 步骤中, 根据 $P(z|d, w)$ 对 $P(w|z)$ 和 $P(z|d)$ 矩阵重新估计:

$$P(w_j|d_i)=\frac{\sum_{i=1}^{N}\text{freq}(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N}\text{freq}(d_i,w_m)P(z_k|d_i,w_m)} \tag{4}$$

$$P(z_k|d_i)=\frac{\sum_{j=1}^{M}\text{freq}(d_i,w_j)P(z_k|d_i,w_j)}{\text{freq}(d_i)} \tag{5}$$

似然函数的对数如下:

当似然函数 L 期望值的增加量小于阈值时, 迭代终止. 此时得到一个最优解$P(w|z)=\left[P(w_j|z_k)\right]_{m\times k}$和$P(z|d)=[P(z_k|d_i)]_{k\times n}$.

$$L=\sum_{i=1}^{N}\sum_{j=1}^{M}\mathrm{freq}\left(d_i,w_j\right)\log\left(P\left(d_i,w_j\right)\right)$$
$$\propto\sum_{i=1}^{N}\sum_{j=1}^{M}\mathrm{freq}\left(d_i,w_j\right)\log\left(\sum_{l=1}^{k}P(z_l|d_i)P\left(w_j|z_l\right)\right) \tag{6}$$

3 实验结果及分析

3.1 实验数据采集

实验采集了凤凰网 (http://www.ifeng.com/) 涉及中日关系新闻“习近平应约会见日本首相安倍晋三”, 新闻为 2014 年 APEC 期间发布的, 共有 2346 条新闻评论. 由于本文利用情感本体库结合“词汇-主题”矩阵实现主题的情绪类别标注, 因此对不含情绪词汇和评论长度小于 10 的评论直接去掉, 取其中 1500 条来进行标注. 请 3 名研究人员对语料进行人工标注, 标注为 7 类情绪, 对于标注结果采用投票方式, 有两人标注结果为一致, 则认为语料标注有效, 最终取 1200 条新闻评论作为本文的实验语料.

3.2 PLSA 中主题数的确定

利用 PLSA 进行计算时, 需要先设定主题的数量, 而主题数的确定受到语料的规模和内容的影响. 对于主题明确的语料, 设定了正确的主题数量, 能有效提升主题的自动情绪标注准确性, 进而提高文档的情绪类别自动标注准确率. 鉴于此, 本文利用情感本体库 *OL* 作为情绪标注的基础, 而本体库中将情绪分为 7 类, 分别是乐、好、怒、哀、惧、恶、惊. 因此, 本文主题的数量的设定从 7 类开始, 一直增加到 28 类, 探索最适合于选定语料的主题数量.

对于主题 z^j 的情绪强度向量 $V_{zj}=[e_1,e_2,e_3,\cdots,e_m]^j$, 如果 V_{zj} 中除了情绪最强的类别之外, 其他情绪类别 e_i 的强度不能同时满足公式 (7), 则认为该主题的情绪类别不能准确分类.

$$\max(V_{zj})-e_i\geq\max(V_{zj})\times10\% \tag{7}$$

因此, 主题情绪类别自动标注准确率的计算公式如公式 (8) 所示.

$$\text{准确率}=\frac{\text{标注正确的主题数}}{\text{主题数}} \tag{8}$$

根据公式 (8), 在本文选定的语料集中, 计算主题情绪自动标注的准确率随着主题数增加的结果, 如图 2 所示, x 轴表示主题数量, y 轴表示自动标注准确率.

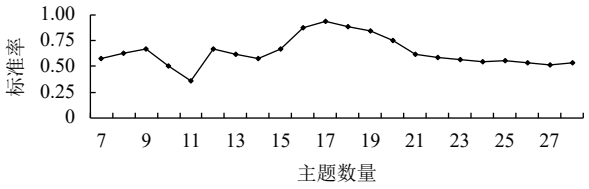

图 2 主题数与标注准确率

实验结果表明, 在主题数设置为 17, 准确率达到 0.94, 仅有 1 个主题不能被准确标注, 说明此时主题与主题之间的情绪区分度最高, 可利用主题情绪对文档情绪进行标注. 因此, 本文设定 PLSA 模型的主题数为 17, 获取 17 个主题中概率较高的情绪词汇 $Z^{17}=\left\{[w_1,w_2,\cdots,w_o]^1,\left[w_1,w_2,\cdots,w_p\right]^2,\cdots,\left[w_1,w_2,\cdots,w_q\right]^{17}\right\}$.

3.3 基于 PLSA 的文本情绪类别自动标结果

将主题数设定为 17, 利用 PLSA 的分析结果, 对主题进行情绪标注, 能正确标注的主题数为 16 个, 则 7 类情绪对应的主题数如表 1 所示, 主流情绪为乐、好. 进一步观察语料发现, 多数网民对中日两国友好关系和共同发展, 表示支持和高兴. 同时, 网民也对中日的钓鱼岛、靖国神社、南京大屠杀等中日历史事件表现出其他情绪.

利用表 1 中主题情绪类别标注结果, 根据算法 1 中的步骤 7, 对文档进行情绪标注, 各类情绪下语料标注的准确率如表 2 所示.

从表 2 的实验结果可以看出, 每一类情绪对应的文档自动标注准确率均高于 85%, 最高准确率达到 93.7%, 7 类情绪的平均准确率达到 88.98%, 总体的准确率达到 90.87%. 说明了采用 PLSA 可有效对文档进行细粒度的情绪类别自动标注, 特别是大规模语料, 可以快速地实现语料的自动标注.

4 结论与展望

文本自动标注能有效解决手工标注的困难, 本文提出一种基于 PLSA 的新闻评论文本情绪类别自动标注方法, 这种方法主要利用的 PLSA 计算输出“文档-主题”和“词语-主题”概率矩阵, 基于“词汇-主题-文档”三者的关系, 认为某一类情绪词汇出现的概率最高的主题与词汇的情绪类别相同, 对主题进行情绪类别标注; 认为出现在某一主题概率最高的文档与主题的情绪类别相同, 对进行情绪类别标注. 实验结果表明, 这种方法是可行的和有效的, 标注准确率达到 90% 以上.

表 1 主题情绪标注结果

情绪类别	主题数	文档数	部分关键词
乐	2	122	愿意, 呵呵, 精神, 安全, 增长, 解放, 高兴, 激发, 游戏, 胜仗, 放心, 幸福
好	4	327	一定, 相信, 牺牲, 支持, 希望, 实话, 真正, 确实, 和平, 坚决, 保卫, 英雄, 肯定, 光荣, 不可, 保家卫国, 血性, 值得, 全面, 意志, 根本, 发展, 强大, 对手, 不在话下, 一针见血, 教育
怒	1	65	抗议, 血债, 爆发, 报仇, 叫嚣, 复仇, 愤怒, 义愤填膺, 急眼
哀	2	145	打击, 不行, 炮灰, 失去, 不足, 悲哀, 为国捐躯, 可惜, 纪念, 伤害, 道歉, 摧毁
惧	2	124	可怕, 厉害, 害怕, 小心, 耻辱, 危亡, 流血, 担忧, 惧怕, 兵器, 致命, 震慑, 有毒, 心虚, 恐惧, 警戒, 交战
恶	5	417	鬼子, 抵制, 侵略, 偷袭, 汉奸, 敌人, 超生, 严重, 倭寇, 无耻, 担心, 扯淡, 对付, 仇恨, 屁话, 解决, 不好, 贪官, 欺负, 狗屁, 自私, 笑话, 退缩, 讨厌, 嚣张, 挑衅, 可笑, 残暴, 幼稚, 无知, 挑拨离间, 恨之入骨, 武断, 逃跑, 沙场, 论调, 借口, 胡说八道, 恶棍, 仇敌, 鄙视, 自欺欺人, 投降, 挑唆, 手段, 攻击, 帝国主义, 别有用心, 卑鄙, 报复
惊	0	0	低估, 悬念, 奇怪, 原来, 神奇, 惊诧

表 2 新闻评论情绪类别自动标注结果

情绪类别	乐	好	怒	哀	惧	恶	惊	总体
准确率 (%)	90.10	92.60	85.30	85.90	86.30	93.70	-	90.87

由于语料的规模和内容对 PLSA 的分析有一定的影响, 同时本体库的词汇涵盖面以及领域适应性等问题, 都会影响标注的效果. 因此, 本文的下一步工作是探索领域自适应性的语料标注方法, 拓展本体库, 利用句法依存关系等抽取领域关键情绪词汇, 提升自动标注的准确率.

参考文献

1 Yang AM, Lin JH, Zhou YM, *et al*. Research on building a Chinese sentiment lexicon based on SO-PMI. Applied Mechanics and Materials, 2013, 263–266: 1688–1693. [doi: 10.4028/www.scientific.net/AMM.263-266.1688]

2 Yang AM, Zhou YM, Lin JH. A method of Chinese texts sentiment classification based on Bayesian algorithm. Applied Mechanics and Materials, 2013, 263–266: 2185–2190. [doi: 10.4028/www.scientific.net/AMM.263-266.2185]

3 崔刚, 盛永梅. 语料库中语料的标注. 清华大学学报 (哲学社会科学版), 2000, 15(1): 89–94. [doi: 10.13613/j.cnki.qhdz.000730]

4 宋鸿彦, 刘军, 姚天昉, 等. 汉语意见型主观性文本标注语料库的构建. 中文信息学报, 2009, 23(2): 123–128. [doi: 10.3969/j.issn.1003-0077.2009.02.018]

5 阳爱民, 周咏梅, 周剑峰. 中文微博语料情感类别自动标注方法. 计算机应用, 2014, 34(8): 2188–2191.

6 周杰, 林琛, 李弼程. 基于机器学习的网络新闻评论情感分类研究. 计算机应用, 2010, 30(4): 1011–1014.

7 徐军, 丁宇新, 王晓龙. 使用机器学习方法进行新闻的情感自动分类. 中文信息学报, 2007, 21(6): 95–100. [doi: 10.3969/j.issn.1003-0077.2007.06.013]

8 杨佳能, 阳爱民, 周咏梅. 基于语义分析的中文微博情感分类方法. 山东大学学报 (理学版), 2014, 49(11): 14–21. [doi: 10.6040/j.issn.1671-9352.3.2014.069]

9 潘云仙, 袁方. 基于 JST 模型的新闻文本的情感分类研究. 郑州大学学报 (理学版), 2015, 47(1): 64–68. [doi: 10.3969/j.issn.1671-6841.2015.01.014]

10 吴江, 唐常杰, 李太勇, 等. 基于语义规则的 Web 金融文本情感分析. 计算机应用, 2014, 34(2): 481–485, 495.

11 Khoo CSG, Nourbakhsh A, Na JC. Sentiment analysis of online news text: A case study of appraisal theory. Online Information Review, 2012, 36(6): 858–878. [doi: 10.1108/14684521211287936]

12 Moreo A, Romero M, Castro JL, *et al*. Lexicon-based comments-oriented news sentiment analyzer system. Expert Systems with Applications, 2012, 39(10): 9166–9180. [doi: 10.1016/j.eswa.2012.02.057]

13 Peñalver-Martinez I, Garcia-Sanchez F, Valencia-Garcia R, *et al*. Feature-based opinion mining through ontologies. Expert Systems with Applications, 2014, 41(13): 5995–6008. [doi: 10.1016/j.eswa.2014.03.022]

14 徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造. 情报学报 , 2008, 27(2): 180–185. [doi: 10.3969/j.issn.1000-0135.2008.02.004]