

基于表情符号的情感词典的构建研究

林江豪^{1,2}, 顾也力¹, 周咏梅^{2,3}, 阳爱民^{2,3}, 陈 锦^{1,2}

(1. 广东外语外贸大学, 广东 广州 510006;

2. 广东外语外贸大学 语言工程与计算实验室, 广东 广州 510006;

3. 广东外语外贸大学 信息科学与技术学院, 广东 广州 510006)

摘 要:情感词典是文本情感分析的基础资源。利用表情符号明显的情感表达作用,提出一种基于种子表情符和 SO-PMI 算法结合的情感词典构建方法。选择 44 个情感明显、内容丰富的表情符号词作为种子情感集合。构建过程融合了 TF-IDF 值在词汇重要程度的度量作用,有效选择候选情感词集。基于 SO-PMI 算法,在大量语料中计算候选情感词汇与种子表情符号之间的情感共现信息,进而确定词汇的情感权值和极性。在 500 万条微博语料中,计算并构建情感词典 SentiNet,共有情感词汇 13 814 个,其中正向词汇 6 885 个,负向词汇 6 929 个。将 SentiNet 应用于微博文本情感分析任务中,实验结果表明,SentiNet 能实现情感词的情感表示,并可应用于大规模的微博语料情感分析任务。该方法融合了情感词的重要度衡量优势和种子表情符号集的情感表达优势,证明了获得的情感权值有效。

关键词:情感词典;情感词;情感权值;种子表情符号;SO-PMI;TF-IDF

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2019)06-0181-05

doi:10.3969/j.issn.1673-629X.2019.06.037

Research on Building Sentiment Lexicon Based on Emoticons

LIN Jiang-hao^{1,2}, GU Ye-li¹, ZHOU Yong-mei^{2,3}, YANG Ai-min^{2,3}, CHEN Jin^{1,2}

(1. Guangdong University of Foreign Studies, Guangzhou 510006, China;

2. Laboratory for Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006, China;

3. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China)

Abstract: Sentiment lexicon is the basic resource of text sentiment analysis. By using the advantages of the obvious emotion expression of emoticons, we propose a construction method of sentiment lexicon via seed emoticons and SO-PMI method. First of all, forty-four sentimental emoticons, which possess obvious sentiment and rich content, are choose as a set of seed words. Then, candidate sentimental words among the microblog texts are acquired via the measuring value TF-IDF. Based on the SO-PMI method, the sentimental concurrence information between the candidate sentimental words and the seed emoticons can be calculated in a large set of texts, and then the sentimental weight and polarity of the candidate sentimental words is determined. Subsequently, the sentimental weight of the candidate sentimental words is calculated based on five million microblog texts. And the sentiment lexicon (SentiNet) is built, with a size of 13 814 sentiment words, including 6 885 positive words and 6 929 negative words. Finally, SentiNet is applied into the polarity classification of sentimental text analysis. The experiment shows that SentiNet can represent sentiment of sentimental words and is more adaptable into massive microblog text sentiment analysis. The proposed method combines the importance measure advantage of affective words with the sentimental expression advantage of seed emoticons, and the sentimental weight is effective.

Key words: sentiment lexicon; sentiment word; sentimental weight; seed emoticons; SO-PMI; TF-IDF

收稿日期:2018-07-27

修回日期:2018-11-28

网络出版时间:2018-11-22

基金项目:教育部人文社会科学项目(14YJA740011);广东省哲学社会科学“十二五”规划项目(GD15YTS01);广东省科技计划项目(2017A040406025);广州市哲学社会科学“十三五”规划2018年度课题(2018GZQN27);广东外语外贸大学教改项目(GWJY2017046)

作者简介:林江豪(1985-),男,助理研究员,CCF会员(15663M),研究方向为文本情感分析;顾也力,教授,通讯作者,研究方向为日本文化、中日关系;周咏梅,教授,研究方向为机器学习、机器应用。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20181122.1506.004.html>

0 引言

文本情感分析有利于观点挖掘、产品口碑分析、舆情分析等实际应用。词语作为用户表达观点的最小单元,富含情感信息。因此,构建高品质的情感词典,能有效应用于文本情感分析^[1]。由于微博文本具有口语化的特点,并且来自多个领域,导致用户在微博文本中使用到的情感词差异性非常大,加大了微博文本情感分析的难度。因此,构建能应用于微博文本情感分析的情感词典具有重要的价值。情感词是组成情感词典的单元,其存在形式一般为情感词,情感倾向和情感权值。如国外知名的 SentiwordNet^[2] 分别给出情感词的正向、中性和负向三种极性的情感权重。国内的 HowNet^[3] 则用+1 来表示词汇的正向情感,-1 来表示负向情感。

现有的情感词典构建方法主要有基于情感词典的方法^[4-9]、基于种子情感词集方法^[10-13]、基于机器学习的方法^[11-16]、基于词向量的方法^[17-18]等。文献[4]提出了基于情感词典的情感特征提取及其在文本情感分析中的应用方法。Dragut 等^[5]以多情感词典中词汇极性不同的现象,自动构建了领域情感词典。在短文本情感特征提取中,Vo 等提出了利用神经网络和情感词典结合的方法^[6]。基于 HowNet 中情感词汇的情感信息,文献[7]提出了语义相似度和语义相关场两种计算方法,通过计算情感候选词与 HowNet 中情感词汇的语义相似度,得到词汇的情感倾向。同样利用 HowNet,柳位平等利用义原计算的优势,根据词与正、负向种子词的语义相似度差,计算获得情感倾向^[8]。文献[9]结合了 HowNet 和 SentiWordNet,对词语进行义元分解并计算其情感值。以情感种子词集为基础,利用 SO-PMI 算法,在特定语料环境中,可计算获得词汇的情感倾向和权重^[10-13]。基于机器学习算法,在特定语料中对词汇信息进行统计和计算,也可获得词汇的情感信息。如文献[14]提出利用页面、页面社区和页面社区的所属类别,将单词语义特征映射到这些类别上,获得词汇的类别属性。文献[15]在新闻和评论中进行对比分析,再将情感向通用领域扩展,得到通用的情感特征。文献[16]通过利用评论中的普通特征训练情感分类器,再基于 spectral 聚类将词汇的情感映射到扩展特征。现有的研究为情感词典构建提供了新思路,特别是在微博语料中进行情感词典构建,微博中的表情符号带有明显的情感特征,如用户喜欢👍([赞])表示赞同,用😭([泪])表示伤心等情感,可作为有效的基础情感信息,进而拓展计算词语的情感权值。

因此,文中利用微博表情符号的情感表达作用,选择情感表情符号作为基准情感信息,利用 TF-IDF 和

SO-PMI 的计算优势,实现情感词的识别与情感权值的计算,并通过微博文本情感分析任务,验证该方法的有效性。

1 基于表情符号的情感词典构建方法

1.1 SO-PMI 算法

点间互信息算法(pointwise mutual information, PMI)可用于计算语料库中两个词语之间的语义相似度。基本思想是统计词语在文本中的共现率,共现率越高其语义关联度越高,反之则语义关联度越低。给定语料库中,通过 PMI 算法,词语 w_1 与 w_2 间的 PMI 值可用两个词在语料库中共现的概率 $P(w_1 \& w_2)$ 和两个词在语料库中单独出现的概率 $P(w_1)$ 与 $P(w_2)$ 进行表示,具体计算如式 1 所示。

词的语料库中出现的概率可以使用词的文档频次来计算。

$$\text{PMI}(w_1, w_2) = \log_2 \frac{P(w_1 \& w_2)}{P(w_1) \cdot P(w_2)} \quad (1)$$

情感倾向点互信息算法(semantic orientation pointwise mutual information, SO-PMI)是由 PMI 算法扩展而来,通过引入计算词语的情感信息,达到词语情感倾向计算的目的。给出正面种子词集 W_p 和负面种子词集 W_n ,则候选情感词 w_i 的情感倾向值(SO)可采用式 2 计算。

$$\text{SO}(w_i) = \sum_{w_p \in W_p} \text{PMI}(w_i, w_p) - \sum_{w_n \in W_n} \text{PMI}(w_i, w_n) \quad (2)$$

SO 值大于 0 的为正面词汇,小于 0 的为负面词汇。通常将情感倾向值进行线性变化,使情感词的情感权值为介于 $[-1, 1]$ 之间的实数,如式 3:

$$\text{SO}'(w_i) = \begin{cases} \frac{\text{SO}(w_i)}{\max_{i=1}(\text{SO}(w_i))}, & \text{SO}(w_i) > 0 \\ -1 \cdot \frac{\text{SO}(w_i)}{\min_{i=1}(\text{SO}(w_i))}, & \text{SO}(w_i) < 0 \end{cases} \quad (3)$$

为了过滤掉情感表达较弱的词汇,在式 3 中加入约束条件。设定情感阈值 θ ($0 < \theta < 1$),认为情感强度在 θ 以外的词汇为非情感词汇,具体计算如式 4:

$$\text{SO}_{\text{new}}(w_i) = \begin{cases} \frac{\text{SO}(w_i)}{\max_{i=1}(\text{SO}(w_i))}, & \text{SO}(w_i) \in (+\theta, +1] \\ -1 \cdot \frac{\text{SO}(w_i)}{\min_{i=1}(\text{SO}(w_i))}, & \text{SO}(w_i) \in [-1, -\theta] \\ \text{非情感词}, & \text{SO}(w_i) \in [-\theta, +\theta] \end{cases} \quad (4)$$

情感阈值 θ 的取值直接关系到情感词典的规模和范围。 θ 太小容易产生太多的噪音情感词,影响情感词典的质量;取值过大容易过滤到太多词汇,约束情感词典的规模。

文中通过大量实验,最终设定阈值 $\theta = 0.35$,可取得较好的情感词典构建效果。

1.2 情感词典构建过程

情感词典构建过程中,首先选定正、负种子表情符号集合 W_p 和 W_n 。接着对微博语料 Weibo_texts 进行分词和 TF-IDF 值的计算,计算结果可用 $W = \{(w_1, \text{tf-idf}_1), (w_2, \text{tf-idf}_2), \dots, (w_m, \text{tf-idf}_m)\}$ 表示;采用

阈值过滤方法,选择 W 中 TF-IDF 值高于阈值的词汇作为候选词集 $WL = \{(w_1, \text{tf-idf}_1), (w_2, \text{tf-idf}_2), \dots, (w_n, \text{tf-idf}_n)\} (n \leq m)$, WL 是表示在语料中具有重要区分度的词集,但词集中词汇的情感权重未确定。该算法主要原理是通过计算词集 WL 中每一个词 w_i 与 W_p 、 W_n 中各个表情符号在语料中的情感倾向点互信息,再与词 w_i 的 TF-IDF 值 tf-idf_i 相乘,得到词 w_i 的情感特征权重;最终获得情感词典 $\text{SentiNet} = \{(w_1, \text{wt}_1), \dots, (w_m, \text{wt}_m)\}$,实现了对词汇情感表达的抽象表示,方便计算机实现情感计算。具体过程如图 1 所示。

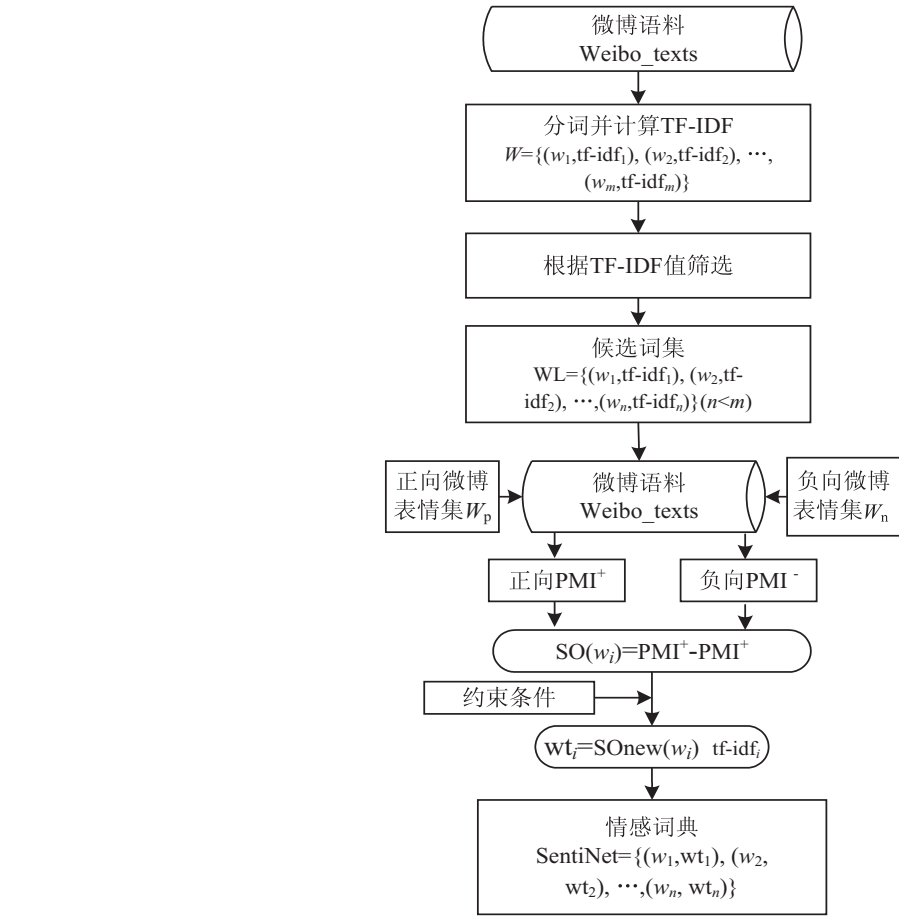


图1 基于表情符号的情感词典构建

根据图 1,利用 TF-IDF 值的重要性度量和情感表情符号的情感强度,实现了情感词的权值计算。算法描述如下:

算法:基于种子表情符的情感词典自动构建算法。
输入:微博语料集 Weibo_texts;正向表情符号集 W_p ;负向表情符号集 W_n ;

输出:SentiNet。

步骤 1:初始化 $\text{Senti2vec} = \emptyset$;

步骤 2:将 Weibo_texts 进行分词、去标点符号等预处理,计算词汇的 TF-IDF 值,得到词集 $W = \{(w_1, \text{tf-idf}_1), (w_2, \text{tf-idf}_2), \dots, (w_m, \text{tf-idf}_m)\}$;

步骤 3:对每一个 $(w_i, \text{tf-idf}_i) ((w_i, \text{tf-idf}_i) \in W)$,如果 $\text{tf-idf}_i \geq a (a \in [0,1])$,则 $(w_i, \text{tf-idf}_i) \rightarrow WL$;得到 $WL = \{(w_1, \text{tf-idf}_1), (w_2, \text{tf-idf}_2), \dots, (w_n, \text{tf-idf}_n)\} (n \leq m)$;

步骤 4:对每一个 $(w_i, \text{tf-idf}_i) ((w_i, \text{tf-idf}_i) \in WL)$,在 Weibo_texts 中计算获得 $\text{SO}(w_i)$;如果 $\text{SO}(w_i)$ 满足式 4 中的情感词范围,则计算 $\text{SOnew}(w_i)$,进而采用式 5 计算获得 wt_i ;

$$\text{wt}_i \leftarrow \text{SOnew}(w_i) \times \text{tf-idf}_i \tag{5}$$

步骤 5:输出 $\text{SentiNet} = \{(w_1, \text{wt}_1), \dots, (w_m, \text{wt}_m)\}$ 。

模型的输出 SentiNet,在情感权值计算过程中,一方面考虑了 TF-IDF 值的重要性度量,另一方面以种子表情符号的情感信息作为基础,实现更好的融合。种子表情符号不受语料的领域约束,使得提出的方法能在情感权值计算方面更具有适应性。

2 实验结果与分析

2.1 语料采集与预处理

文中使用的微博语料来自北京理工大学搜索挖掘实验室张华平博士的微博开放语料(Weibo_texts),包含了 500 万条微博语料,用于情感词典的构建。同时,从新浪微博上采集的 4 130 个用户的 298 295 条个人微博。过滤不含有表情符号的微博和不含情感词的微博,最后人工筛选 4 000 条并对语料进行情感极性标注,作为微博文本情感分析实验语料。语料为平衡语料,其中正、负向情感的微博语料各 2 000 条,用于微博情感分析实验,验证构建的情感词典在情感分析应用中的有效性。

2.2 种子表情符号的选择

种子表情符号的有效选择是情感词典构建的基础。文中主要采用以下两种选择规则:一是微博语料中的高频表情符号,有利于提高表情符号的使用覆盖率;二是情感极性比较明显的表情符号,有利于提升情感词极性计算结果的准确性。

通过调用新浪微博 API 获取到 1 999 个微博表情,对采集到的微博语料中的表情符号进行频率统计,选择出现频率较高并且情感明显的表情符号作为种子表情符号集,共 44 个表情符号,其中正、负向种子表情符号各 22 个(见表 1)。

表 1 种子表情符号

正向情感表情符号(22 个)	负向情感表情符号(22 个)
[心]、[哈哈]、[爱你]、[嘻嘻]、[偷笑]、[鼓掌]、[蛋糕]、[开心]、[花心]、[good]、[酷]、[害羞]、[给力]、[奥特曼]、[威武]、[笑哈哈]、[爱]、[可爱]、[赞]、[呵呵]、[亲亲]、[礼物]	[泪]、[可怜]、[花心]、[衰]、[抓狂]、[汗]、[生病]、[蜡烛]、[吃惊]、[哼]、[怒]、[晕]、[阴险]、[悲伤]、[委屈]、[黑线]、[囧]、[伤心]、[泪流满面]、[失望]、[鄙视]、[鬼脸]

2.3 情感词典构建结果及其验证

利用文中提出的算法构建获得情感词汇 13 814 个,其中正向词汇 6 885 个,负向词汇 6 929 个。将构建的情感词典应用于微博语料情感分析实验。实验数据为人工标注的平衡微博语料,共 4 000 条,正负向微博文本各 2 000 条。随机取正向语料 100 条和负向语料 1 000 条构建平衡训练语料库。其余的语料用于微博文本情感分类器的测试。

基于支持向量机(SVM)这种监督式学习的方法,构建了微博文本情感分类器。分类过程中,首先对微博文本进行分词等文本预处理操作;接着基于传统的向量空间模型(vector space model, VSM)对文本进行向量表示,对出现在 SentiNet 中的词汇用情感权重表示,其他的用 0 表示,向量的维度是 SentiNet 的长度。

同时直接利用 SentiNet 中词汇的情感权值对微博语料进行情感分析,主要采用情感加权(SO-SUM)和情感乘积(SO-MUL)的方法,也就是将每一条微博进行分词等预处理后,直接扫描出现在 SentiNet 中的情感词,将每个情感词的权值分别进行求和与乘积运算,最终每条微博的情感值大于 0,则分类为正向,否则为负向。

为了进一步验证情感词典,与国内知名的 HowNet 情感词典进行对比。评价指标采用微平均 F_1 值。采用折叠交叉实验的方式,迭代 10 次,最终取平均值作为实验结果,如表 2 所示。

表 2 微博文本情感分类结果

情感分类方法	F_1 (100%)	方差
SentiNet+SVM	63.25	2.54
SentiNet+SO-SUM	62.36	1.89
SentiNet+SO-MUL	45.87	2.05
HowNet+SVM	55.74	3.05
HowNet+SO-SUM	45.21	2.36
HowNet+SO-MUL	40.56	2.78

观察语料和实验结果发现,由于微博口语化、转折词、程度副词等对分类效果也有一定的影响,为了验证所构建情感词典的有效性,对这些影响暂不考虑。因此,仅获取了微博文本中出现的情感词作为情感特征,导致总体的 F_1 值偏低。采用 SentiNet+SVM 的方法可取得较好的分类效果($F_1 = 63.25\%$)。对比分析了 SentiNet 和 HowNet 两种情感词典在分类中的效果,如图 2 所示。

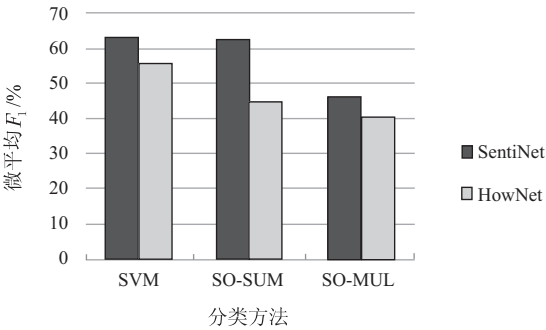


图 2 微博情感分类结果对比

从图 2 中可以看出,同样的分类方法 SentiNet 可取得比 HowNet 更好的分类性能。主要原因是,

SentiNet 是从语料中计算获得,情感词汇的覆盖面更广泛一些,能提取到更多有效的情感特征。SO-SUM 方法与分类器 SVM 方法效果相当,说明文中的情感权重计算结果是有效的。同时,SO-SUM 方法具有不需要训练,可直接应用于大规模的语料分类的优势。实验结果表明,文中方法能对词汇中的情感词汇进行有效的表示。

3 结束语

基于词汇的 TF-IDF 值,选择语料中具有重要度区分的词汇作为候选情感词集。提出基于种子表情符号和 SO-PMI 算法的权重计算方法实现情感词汇的情感权值计算,最终构建情感词典 SentiNet。该方法融合了情感词的重要度衡量优势和种子表情符号集的情感表达优势,在大量微博语料中实现了情感词的权值计算。基于微博文本情感分析的实验证明了该方法的可行性,构建的 SentiNet 有效。下一步将研究表情符号和情感词汇相结合的种子词集,分析种子情感集合对情感词典构建的影响,进一步提升 SentiNet 的规模和质量。

参考文献:

- [1] XU Ge, MENG Xinfan, WANG Houfeng. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources[C]//Proceedings of the 23rd international conference on computational linguistics. Beijing: ACM, 2010: 1209-1217.
- [2] BACCIANELLA S, ESULI A, SEBASTIANI F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining [C]//International conference on language resources and evaluation. [s. l.]: [s. n.], 2010: 83-90.
- [3] DAI Liuling, LIU Bin, XIA Yuning, et al. Measuring semantic similarity between words using HowNet [C]//Proceedings of the 2008 international conference on computer science and information technology. Singapore: IEEE, 2008: 601-605.
- [4] TABOADA M, BROOKE J, TOFILOSKI M, et al. Lexicon-

based methods for sentiment analysis[J]. Computational Linguistics, 2011, 37(2): 267-307.

- [5] DRAGUT E C, WANG Hong, SISTLA P, et al. Polarity consistency checking for domain independent sentiment dictionaries[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3): 838-851.
- [6] VO D T, ZHANG Y. Don't count, predict! an automatic approach to learning sentiment lexicons for short text[C]//The 54th annual meeting of the association for computational linguistics. Berlin, Germany: ACL, 2016: 219.
- [7] 朱嫣岚, 闵 锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.
- [8] 柳位平, 朱艳辉, 栗春亮, 等. 中文基础情感词典构建方法研究[J]. 计算机应用, 2009, 29(11): 2875-2877.
- [9] 周咏梅, 阳爱民, 杨佳能. 一种新闻评论情感词典的构建方法[J]. 计算机科学, 2014, 41(8): 67-69.
- [10] YANG Aimin, LIN Jianghao, ZHOU Yongmei, et al. Research on building a Chinese sentiment lexicon based on SO-PMI[J]. Applied Mechanics and Materials, 2013, 263-266: 1688-1693.
- [11] 周咏梅, 阳爱民, 林江豪. 中文微博情感词典构建方法[J]. 山东大学学报: 工学版, 2014, 44(3): 36-40.
- [12] WANG Guangwei, ARAKI K. Modifying SO-PMI for Japanese weblog opinion mining by using a balancing factor and detecting neutral expressions [C]//Proceedings of NAACL HLT. Rochester, New York: ACL, 2007: 189-192.
- [13] 王义真, 郑 啸, 后 盾, 等. 基于 SVM 的高维混合特征短文本情感分类[J]. 计算机技术与发展, 2018, 28(2): 88-93.
- [14] 彭丽针, 吴扬扬. 基于维基百科社区挖掘的词语语义相似度计算[J]. 计算机科学, 2016, 43(4): 45-49.
- [15] 陶富民, 高 军, 王腾蛟, 等. 面向话题的新闻评论的情感特征选取[J]. 中文信息学报, 2010, 24(3): 37-43.
- [16] 李素科, 蒋严冰. 基于情感特征聚类的半监督情感分类[J]. 计算机研究与发展, 2013, 50(12): 2570-2577.
- [17] 林江豪, 周咏梅, 阳爱民, 等. 基于词向量的领域情感词典构建[J]. 山东大学学报: 工学版, 2018, 48(3): 40-47.
- [18] 彭昀磊, 牛 耘. 基于词向量的特征词选择[J]. 计算机技术与发展, 2018, 28(06): 7-11.