

基于深度学习的多维特征微博情感分析

金志刚¹, 胡博宏^{1,2}, 张瑞¹

(1. 天津大学 电气自动化与信息工程学院, 天津, 300072;
2. 天津大学 国际工程师学院, 天津, 300072)

摘要: 提出 1 种基于卷积神经网络的多维特征微博情感分析新机制; 利用词向量计算文本的语义特征, 结合基于表情字符的情感特征, 利用卷积神经网络挖掘特征集合与情感标签间的深层次关联, 训练情感分类器; 结合微博文本的语义和情感特征, 同时利用卷积神经网络的抽象特征提取能力, 进而改善情感分析性能。研究表明: 引入表情字符的情感特征模型可使情感分析准确率提高 2.62%; 相比基于词典的机器学习模型, 新机制将情感分析准确率与 F 度量分别提升 21.29% 和 19.20%。

关键词: 情感分析; 卷积神经网络; 微博短文本; 表情字符

中图分类号: TP391

文献标志码: A

文章编号: 1672-7207(2018)05-1135-06

Analysis of Weibo sentiment with multi-dimensional features based on deep learning

JIN Zhigang¹, HU Bohong^{1,2}, ZHANG Rui¹

(1. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China;
2. Tianjin International Engineering Institute, Tianjin University, Tianjin 300072, China)

Abstract: A new mechanism of Weibo sentiment analysis based on convolutional neural networks with multi-dimensional features was proposed. The proposed mechanism combines semantic features from word vectors with sentiment features from emoticons, in which convolutional neural networks was used to mine deep correlation between features and labels. The performance of Weibo sentiment analysis was improved through mining multi-dimensional features and utilizing abstract features extraction ability of convolutional neural networks. The results show that the accuracy of sentiment analysis model based on emoticons increases by 2.62%. The accuracy and F measure increase by 21.29% and 19.20% respectively compared with that of machine learning model based on lexicon.

Key words: sentiment analysis; convolutional neural networks; Weibo short text; emoticons

随着互联网的飞速发展, 越来越多的用户乐于在网络上发布自己的观点、分享个人的生活状态, 同时获取实时新闻、资讯, 社交平台也因此发展迅猛。新

浪微博作为一个基于用户关系信息分享、传播以及获取的平台, 改变了传统社交网络交流方式, 给用户提供更便捷获取丰富信息的方式, 迅速成为了极具

收稿日期: 2017-05-21; 修回日期: 2017-06-30

基金项目 (Foundation item): 国家自然科学基金资助项目(61571318); 青海省科技项目(2015-ZJ-904); 海南省科技项目(ZDYF2016153) (Project(61571318) supported by the National Natural Science Foundation of China; Project(2015-ZJ-904) supported by the Science Foundation of Qinghai Province; Project(ZDYF2016153) supported by the Science Foundation of Hainan Province)

通信作者: 张瑞, 博士, 讲师, 从事社交网络数据挖掘研究; E-mail: hitchcockzhr@163.com

人气的社交网络平台与新媒体平台。由于微博平台的开放性,多数突发事件都是第一时间由身处现场的目击者发布到微博平台,并因此引发网友的高度关注,例如“巴黎恐怖袭击事件”、“上海浦东机场疑似爆炸事件”、“英国退欧”等。互联网用户通过微博第一时间了解热点事件,关注事态发展;相关部门通过微博通报事件最新进展,澄清事实,安抚民众情绪。因此对微博平台内容进行准确分析具有重要意义。Twitter 作为国外著名的社交平台,已成为学术界的热门研究对象。用于 Twitter 情感分析的传统方法是基于词典的情感分析,使用包含情感极性、否定以及程度副词的词典计算句子的情感倾向^[1-2]。PANG 等^[3]基于学习的情感分析方法,将文本情感分析看作文本分类问题,利用从情感词选取的大量嘈杂的标签文本作为训练集直接构建分类器。大量基于学习的情感分析方法过于依赖特征工程,例如 MOHAMMAD 等^[4]利用大量手工挑选的特征取得了当时的最佳分类性能,但手工特征需要大量的语法构词等专业知识,同时成本极高。因此,国内外研究者开始将深度学习模型应用于 Twitter 情感分析问题。SEVERYN 等^[5]使用词向量表示文本的语义特征,并使用卷积神经网络提取深层次特征训练分类器;TANG 等^[6]构建神经网络模型学习包含情感特征与语义特征的词向量表示,并与手工特征结合作为复合特征,训练分类器。此外,国内外研究者也将深度学习模型应用于英语短文本处理。KIM 等^[7]使用卷积神经网络对短文本进行建模,并在多个数据集上进行了多组实验,完成了句子级别的文本分类任务。ZHANG 等^[8]提出了字符级别的卷积神经网络模型并用于文本分类,该模型适用于多种语言,且其性能不亚于词袋模型、N-grams、TFIDF 等传统模型,以及基于词组的卷积神经网络模型和循环网络模型。WANG 等^[9]提出将卷积神经网络与 *k*-means 算法融合的半监督学习模型,使用少量带标签数据对短文本进行聚类。XU 等^[10]提出了利用卷积神经网络与 *k*-means 聚类算法结合的短文本聚类模型,该模型是自学习框架,无需任何标签数据。英语语料相比中文语料具有已分词的天然优势,此外,中英文在文化以及使用习惯上也具有很大差别。针对中文短文本的情

感分析并不多,大部分通过提取文本的情感和统计等特征,进而使用机器学习中的分类算法完成情感分析任务。黄发良等^[11]提出基于隐含狄利克雷分布(LDA)主题模型的结合表情符号与用户性格特征的多特征融合模型,实现了微博文本的主题与情感倾向的同步检测。张冬雯等^[12]通过 word2vec 工具扩充情感词典,根据情感词特征与词性特征,使用支持向量机训练文本情感分类模型。黄仁等^[13]也通过 word2vec 工具扩展情感词典,使用基于情感特征词的统计模型对文本进行情感评分。通过情感词典提取情感特征对情感词典要求严格。一方面,模型往往需要多种词典,如情感词词典、程度副词词典和否定词词典等;另一方面,互联网新词、网络用语日新月异,更新和维护词典也是一个难题。而统计特征的提取依赖自然语言的研究分析,需要语法构词等专业知识。当前,将深度学习模型应用于中文微博情感分析的研究不多。本文作者一方面使用 word2vec 工具计算词向量,提取短文本的语义特征,另一方面保留微博文本中的表情字符,作为情感特征,共同构成特征集合;再通过卷积神经网络模型提取深度抽象特征,训练分类器完成情感分类任务。

1 基于卷积神经网络的多维特征微博情感分析新机制

本文作者提出 1 种多维特征微博情感分析新机制。图 1 所示为多维特征微博情感分析新机制流程图。

微博文本的多维度特征包括语义特征与情感特征。语义特征通过对大规模语料进行无监督学习训练,计算词组的词向量表示。词向量不仅将词组映射至高维空间中实现向量化表示,同时词向量在高维空间中的余弦距离也表示词组间的相似度,蕴含了词组间的深层语义联系。情感特征由微博文本中的表情字符表示。对微博文本中的表情字符提取与转换,并进行随机向量化以匹配语义特征的词向量表示,共同构成微博文本特征集合。

将特征集合作为卷积神经网络的输入,通过卷积与池化运算,捕获局部特征并对局部特征进行筛选,

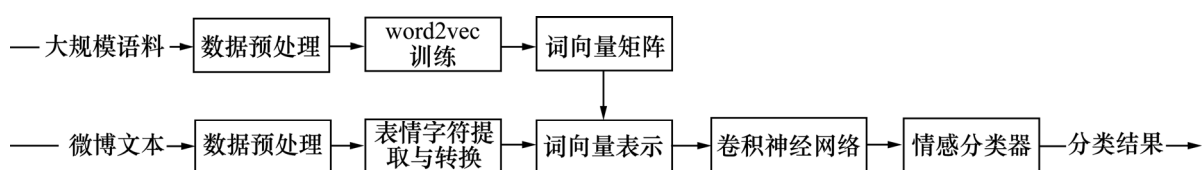


图 1 多维特征微博情感分析新机制流程图

Fig. 1 Flow chart of mechanism with multi-dimensional features for Weibo sentiment analysis

所得特征作为情感分类器的输入, 训练微博文本情感分类器。

2 卷积神经网络模型

卷积神经网络是深度学习中的常用模型之一, 已经在图像和语音分析领域取得了可观的效果。该模型受生物视觉系统启发, 通过池化和权值共享减少了网络中权值的数量, 降低了网络模型的复杂度。池化操作利用了自然文本的统计特性即局部的统计特性同样适用于其他部分的原理。同时, 卷积神经网络对平移、比例缩放、倾斜或者其他形式的变形具有高度不变性。

2.1 模型概况

本文作者借鉴文献[7]中使用的卷积神经网络模型, 所用的模型可分为 3 个部分, 如图 2 所示。

第 1 部分为输入层。输入层由词组对应的词向量构成的矩阵表示。经过数据预处理后, 一条微博文本 s 由 n 个词组 $\{x_1, x_2, \dots, x_n\}$ 构成, 其中主要由文本的语义特征词组构成, 当微博中使用了表情字符时, 存在与该表情对应的情感特征词组。每个词组 x_i 都是一个 d 维的向量, 其值由 word2vec 工具训练求得, 若该词组不在 word2vec 训练语料中, 则被随机初始化。因此, 1 条微博文本由 1 个 $n \times d$ 的矩阵表示。

第 2 部分为卷积层, 主要由卷积和池化运算构成。使用卷积层是为了提取微博文本的深层次情感模式特征, 同时配合池化运算减少计算量, 归一化特征长度。模型使用不同长度的卷积核提取不同的特征, 每个卷积核 w 是一个 $h \times d$ 的矩阵, 其中 h 是卷积核的长度。使用卷积核对微博文本进行卷积, 该过程可表示为

$$t_i = f(w * s_{i:i+h-1} + b) \tag{1}$$

其中: $s_{i:i+h-1}$ 表示由第 i 个词组至第 $i+h-1$ 个词组组成的连续文本片段; $*$ 表示卷积运算; f 为非线性激活函数 ReLU; b 为偏置项; t_i 为第 i 个卷积特征。当使用卷积核 w 对整个微博文本完成卷积后, 即可生成特征集合 $T = \{t_1, t_2, \dots, t_{n-h+1}\}$ 。由于不同微博文本的词组数量不同, 且不同卷积核的长度不同, 导致特征集合的长度也不同。因此, 采用池化操作选出特征集合的代表值, 既可保证对不同微博文本使用不同长度的卷积核卷积后的特征数量一致, 也可减少神经网络中的权值数量, 降低计算复杂度。模型采用最大值池化操作, 即选取特征集合中的最大值作为代表特征值。

$$\hat{t} = \max(\{t_1, t_2, \dots, t_{n-h+1}\}) \tag{2}$$

经过池化操作, 即可保证使用多个不同长度卷积核对微博文本卷积生成长度统一的特征向量。

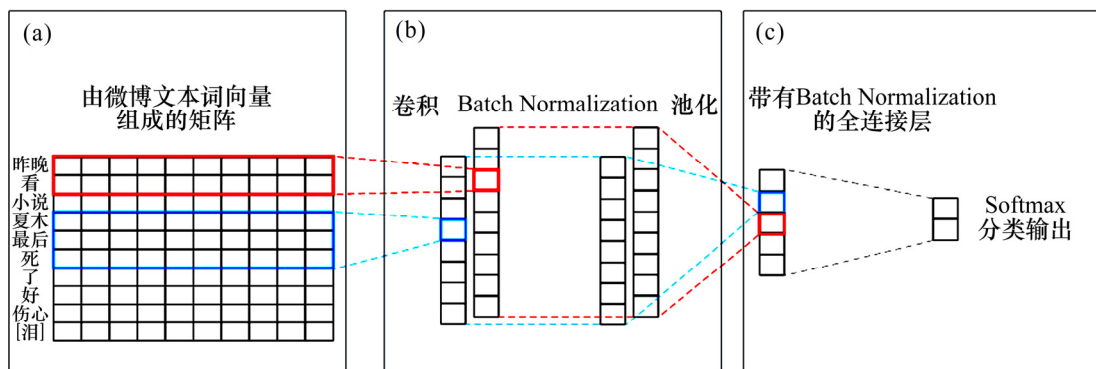
第 3 部分是分类层。将上一层生成的特征向量输入至全连接的 softmax 层, 计算属于各类别的概率分布:

$$P_j = P(y = j | X, b) = \frac{e^{W_j X + b_j}}{\sum_{i=1}^L e^{W_i X + b_i}} \tag{3}$$

式中: P_j 为属于第 j 类的概率; X 和 W 分别为分类层的输入和权值矩阵; b_i 和 b_j 分别为偏置项 b 的第 i 和 j 元素; L 为类别数量。

2.2 正则化

模型使用反向传播算法进行训练, 采用随机梯度下降算法(SGD)计算梯度, 使用 Adadelta 算法^[14]自动更新学习率。此外, 模型引入 Batch Normalization 算法对各层神经网络的输入进行正则化处理。通过在每层计算激活值前加入一个 Batch Normalization 归一化



(a) 输入层; (b) 卷积层; (c) 分类层

图 2 卷积神经网络结构

Fig. 2 Structure of convolutional neural networks

网络层,能够减少内部协变量偏移的影响,进而加速参数更新的收敛与网络的训练^[15]。

本模型中加入了 2 处 Batch Normalization 归一化网络层。一是在卷积运算后,对单个卷积核对应的卷积特征进行归一化,再进行池化运算;二是在全连接分类层中计算激活值之前加入归一化层,激活值 z 计算公式为

$$z = f(WX + b) \quad (4)$$

引入归一化层后,激活值计算公式为

$$z = f(g(WX)) \quad (5)$$

其中: g 为归一化函数。由于归一化运算对于偏置无效,因此省略偏置项。

3 微博文本多维度特征提取

3.1 基于词向量的语义特征

已有的研究表明,在无法对大规模带标签数据进行监督学习的情况下,使用无监督学习的方式初始化神经网络的输入值,是一个有效改善训练效果、加速训练收敛的方法^[16-17]。因此,本文作者使用 word2vec 工具计算微博文本的词向量,作为卷积神经网络输入的初始化值。

word2vec 工具^[18]是 Google 上一款开源的高效计算分布式词向量的工具。此工具以语料库为输入,计算出词向量表示。它以由文本数据构建的词汇表为训练数据,然后学习词组的高维向量表示,即将词组映射至有限维的高维空间中。与传统的 One-hot 表示模型相比,词向量是稠密的向量表示,并且更易计算向量间的距离度量,因此更适用于自然语言处理和机器学习,同时对微博文本的简短性和随意性具有更好的鲁棒性。

word2vec 工具包含 2 个计算词向量的模型即连续词袋模型(CBOW)和 Skip-Gram 模型。该工具基于神经网络语言模型,通过删除隐层以及利用分层 softmax 和负采样等技巧减少计算复杂度,同时优化训练结果^[19-20]。通过 word2vec 工具对分词后的中文语料进行训练,可计算每个词组的指定维度的向量表示。利用词组的词向量可以方便地计算词向量的余弦距离作为词组间的相似度度量,因此词向量表示了语料中词组间的深层语义联系。

3.2 基于微博表情字符的情感特征

微博平台为用户提供了大量的默认表情字符,帮

助用户更生动地表达当下的感受和体会。图 3 所示为微博默认表情。直观上,这些表情字符也是微博文本情感倾向的重要组成部分。由于传统的自然语言处理大多针对新闻、博客等正式语料,因此只关注对于文本的研究与处理,在文本预处理阶段只过滤出文本信息,而删除网页链接、特殊字符等数据,导致微博文本情感特征的缺失。



图 3 微博默认表情

Fig. 3 Default emoticons in Weibo

通过研究发现,微博平台的默认表情字符是由 HTML 图像标签构成。因此,在数据预处理阶段,对图像标签进行提取与转换,并将转换后的表情字符插入所属微博文本的原位置,然后用方括号对表情字符进行标注,以区分微博中的文本与表情,例如“[哈哈]”。将处理后的表情字符采用随机初始化的方式转换成词向量,与语义特征保持一致,进而实现语义特征与情感特征的结合。

4 实验与分析

4.1 数据集

实验所用数据集包含 word2vec 训练语料与微博数据集。

word2vec 训练语料使用搜狗实验室整理提供的新闻数据^[21],包含全网新闻数据与搜狐新闻数据共 2 706 229 条新闻语料。选取新闻预料中新闻正文进行分词处理,作为 word2vec 工具的训练数据,并设置词向量长度 $d=300$ 。训练完成后,共包含 565 345 个词组。对于未出现在词向量集合中的词组,将对其进行随机初始化。

对于微博数据集,由于模型需要考虑微博中的表

情字符, 并未发现合适的公开数据集, 因此通过自行采集的方式获取了约 10 000 条微博文本, 通过人工标注的方式将其分为积极和消极 2 类, 其中积极类微博 3 100 条, 消极类微博 3 800 条。最终, 随机选择 3 000 条积极类和 3 000 条消极类微博作为数据集。表 1 所示为实验所用微博样本示例。

表 1 微博文本示例

Table 1 Examples of Weibo text

积极文本	消极文本
看着都难过 大家不要乱骂人家。吸取教训 逝者安息 [蜡烛]	是啊 再吵架也要分场合和地点...完全得不到同情。哎 可惜剩下的家人会多伤心...
这个小说满满的都是回忆 很快就能在电影院里看到啦 [心]	昨晚看小说夏木最后死了, 好伤心[泪]
啊! 啊! 啊! 帅出新高度啊!	看个预告都能看哭, 在电影院感觉能哭傻

4.2 对比实验

对比实验包含 2 个部分: 第 1 部分是与传统基于词典的机器学习模型的对比; 第 2 部分是与无表情字符的卷积神经网络模型对比。

第 1 部分的对比实验采用基于情感词典与词性特征的机器学习情感分类算法。参考文献[12], 使用 word2vec 工具计算词组相似度, 进而对由 HowNet 词典和 IAR 词典组合构成的情感词典进行扩充, 同时使用 {形容词, 副词, 动词} 的词性组合作为文本的词性特征, 将提取的情感特征与词性特征进行主成分分析 (PCA) 降维处理, 最终通过支持向量机对特征进行训练, 完成中文情感分类任务。

为了验证微博文本中表情字符对情感分析的促进效果, 设计关于表情字符的对比实验, 即在文本预处理阶段将表情字符去除作为对比实验, 其余设置与主模型相同。

4.3 结果与讨论

实施 3 组情感分析实验。第 1 组为本文提出的微博情感分析新机制, 记为有情感特征的深度学习模型 其参数通过在小批量数据集上进行交叉验证后确定, 如表 2 所示。第 2 组为无感情特征的基于卷积神经网络的模型, 即在本文作者提出模型的基础上, 去除表情字符, 记为无情感特征的深度学习模型。第 3 组为基于情感词典的机器学习模型, 记为机器学习模型。

采用准确率与 F 值作为分类效果的度量。对于本文所属的二分类问题, 设类别分别为积极和消极, 令

表 2 实验参数

Table 2 Experiment parameters setting

参数	参数设置
卷积核维度/维	2,3,4,5
卷积单元数量/个	100
分类层输入节点数/个	50
迭代次数	25
迷你批维度/维	50
Adadelta 衰减参数	0.9

T_P 为原本属于积极且分类正确的数量, F_P 为原本属于积极但分类错误的数量, F_N 为原本属于消极但分类错误的数量, T_N 为原本属于消极且分类正确的数量。则准确率 P_{acc} 、精确率 P_{prec} 、召回率 P_{recall} 和 F 值的计算如下:

$$P_{acc} = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \quad (6)$$

$$P_{prec} = \frac{T_P}{T_P + F_P} \quad (7)$$

$$P_{recall} = \frac{T_P}{T_P + F_N} \quad (8)$$

$$F = \frac{2P_{prec} \cdot P_{recall}}{P_{prec} + P_{recall}} \quad (9)$$

每组实验均采用 10 折交叉验证的方式进行。表 3 所示为各组实验的实验结果。

表 3 对比实验结果

Table 3 Results of experiments

学习模型	准确率	精确率	召回率	F 值
有情感特征的深度学习模型	0.848 9	0.850 3	0.847 1	0.848 1
无情感特征的深度学习模型	0.827 2	0.835 8	0.830 9	0.833 3
机器学习模型	0.699 9	0.683 8	0.741 5	0.711 5

由表 3 可知: 与机器学习模型相比, 本文作者提出的微博情感分析新机制情感分析准确率与 F 值分别提升 0.149 0 和 0.136 6, 相对提升 21.29% 和 19.20%。可以得出结论, 卷积神经网络与多维度文本特征的引入对微博情感分析的改善效果显著。与无情感特征的深度学习模型相比, 发现准确率与 F 值分别提升 0.021 7 与 0.014 8, 相对提升 2.62% 与 1.78%, 考虑到并非所有微博文本都包含表情字符, 该改善效果比较可观, 可见, 引入微博表情字符有助于促进情感分析。

5 结论

1) 提出应用深度学习提取微博文本情感模式,且包含微博表情字符特征的情感分析新机制。通过词向量提取文本的语义特征,对微博表情字符进行提取并作为微博文本的情感特征。

2) 与无情感特征的模型相比,基于表情字符的情感特征模型对情感分析的准确率提高 2.62%。

3) 通过引入情感特征,利用卷积神经网络的抽象特征挖掘能力,训练情感分类器,使得本文提出模型的准确率与 F 度量相对于传统基于词典的机器学习模型分别提升 21.29% 和 19.20%。

参考文献:

- [1] TABOADA M, BROOKE J, TOFILOSKI M, et al. Lexicon-based methods for sentiment analysis[J]. *Computational Linguistics*, 2011, 37(2): 267–307.
- [2] THELWALL M, BUCKLEY K, PALTOGLOU G. Sentiment strength detection for the social web[J]. *Journal of the American Society for Information Science & Technology*, 2012, 63(1): 163–173.
- [3] PANG Bo, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques[C]// *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. PA, USA: ACL, 2002: 79–86.
- [4] MOHAMMAD S M, KIRITCHENKO S, ZHU Xiaodan. NRC-Canada: building the state-of-the-art in sentiment analysis of tweets[EB/OL]. 2013–08–28. <https://arxiv.org/abs/1308.6242>.
- [5] SEVERYN A, MOSCHITTI A. Twitter sentiment analysis with deep convolutional neural networks[C]// *The International ACM SIGIR Conference*. Santiago, Chile: ACM, 2015: 959–962.
- [6] TANG Duyu, WEI Furu, QIN Bing, et al. Coooolll: A deep learning system for twitter sentiment classification[C]// *Proceedings of the 8th International Workshop on Semantic Evaluation*. Dublin, Ireland: ACL, 2014: 208–212.
- [7] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. 2014–08–25. <https://arxiv.org/abs/1408.5882>.
- [8] ZHANG Xiang, ZHAO Junbo, LECUN Y. Character-level convolutional networks for text classification[C]// *Advances in Neural Information Processing Systems*. New York, USA: Curran Associates, Inc., 2015: 649–657.
- [9] WANG Zhiguo, MI Haitao, ITTYCHERIAH A. Semi-supervised clustering for short text via deep representation learning[EB/OL]. 2016–02–22. <http://arxiv.org/abs/1602.06797>.
- [10] XU Jiaming, WANG Peng, TIAN Guanhua, et al. Short text clustering via convolutional neural networks[C]// *Proceedings of NAACL-HLT*. Denver, Colorado, USA: ACL Anthology, 2015: 62–69.
- [11] 黄发良, 冯时, 王大玲, 等. 基于多特征融合的微博主题情感挖掘[J]. *计算机学报*, 2017, 40(4): 872–888.
HUANG Faliang, FENG Shi, WANG Daling, et al. Mining topic sentiment in microblogging based on multi-feature fusion[J]. *Chinese Journal of Computers*, 2017, 40(4): 872–888.
- [12] 张冬雯, 杨鹏飞, 许云峰. 基于 word2vec 和 SVMperf 的中文评论情感分类研究[J]. *计算机科学*, 2016, 43(S1): 418–421, 447.
ZHANG Dongwen, YANG Pengfei, XU Yunfeng. Research of Chinese comments sentiment classification based on Word2vec and SVMperf[J]. *Computer Science*, 2016, 43(S1): 418–421, 447.
- [13] 黄仁, 张卫. 基于 word2vec 的互联网商品评论情感倾向研究[J]. *计算机科学*, 2016, 43(S1): 387–389.
HUANG Ren, ZHANG Wei. Study on sentiment analysis of internet commodities review based on word2vec[J]. *Computer Science*, 2016, 43(S1): 387–389.
- [14] ZEILER M D. ADADELTA: an adaptive learning rate method[EB/OL]. 2012–12–22. <https://arxiv.org/abs/1212.5701>.
- [15] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[EB/OL]. 2015–02–11. <https://arxiv.org/abs/1502.03167>.
- [16] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. *Journal of Machine Learning Research*, 2011, 12(1): 2493–2537.
- [17] SOCHER R, PENNINGTON J, HUANG E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]// *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK: ACL, 2011: 151–161.
- [18] Google. Word2vec[EB/OL]. 2013–07–30. <https://code.google.com/p/word2vec/>.
- [19] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// *Advances in Neural Information Processing Systems*. Lake Tahoe, USA: Curran Associates, Inc, 2013: 3111–3119.
- [20] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. 2013–01–16. <https://arxiv.org/abs/1301.3781>.
- [21] 搜狗实验室. 新闻数据[EB/OL]. 2012–08–16. http://www.sogou.com/labs/resource/list_news.php.

(编辑 伍锦花)