文章编号: 1006-2475(2017)04-0073-05

基于卷积神经网络模型的互联网短文本情感分类

刘小明[1,2],张 英[1,2],郑秋生[1,2]

(1. 中原工学院计算机学院,河南 郑州 450007; 2. 计算机信息系统安全评估河南省工程实验室,河南 郑州 450007)

摘要: 情感分类旨在发现用户对热点事件的观点态度,但由于现今互联网短文本格式随意,语言规范性不够,所以目前传统方法的情感分类效果并不理想。面向大数据互联网短文本信息,本文提出一种基于深度卷积神经网络(Convolutional Neural Networks,CNNs)模型的互联网短文本分类。首先选择词向量作为原始特征,然后通过卷积神经网络进一步提取特征,最后训练出基于深度卷积神经网络的互联网短文本情感分类模型。实验结果表明,该模型不仅可以有效处理互联网短文本中的情感分类这一任务,而且明显提高了情感分类的准确率,平均提高约 5%。

关键词: 互联网短文本; 情感分类; 卷积神经网络; 自然语言处理; 深度学习

中图分类号: TP391 **文献标识码:** A **doi:** 10.3969/j.issn.1006-2475.2017.04.015

Sentiment Classification of Short Texts on Internet Based on Convolutional Neural Networks Model

LIU Xiao-ming[1,2], ZHANG Ying[1,2], ZHENG Qiu-sheng[1,2]

(1. School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China;
2. Henan Province Engineering Laboratory of Computer Information System Security Assessment, Zhengzhou 450007, China)

Abstract: Sentiment classification aims to find the users' views on hot issues, but now the format of the short texts on the Internet is not normative, the effect of traditional sentiment classification method is not ideal. Facing the information of the short texts on the Internet of big data, this paper puts forward a deep convolution neural network (CNNs) model of the short text on the Internet. First it uses the Skip-gram in the Word2vec training model as the feature vector, then further extracts feature vector into CNNs, finally trains the classification model of the depth convolution neural network. The experimental results show that, compared with classification methods of traditional machine learning, this method not only can effectively handle emotion classification of the short texts on the Internet, but also improves the accuracy of emotion classification significantly, the average increased by about 5%。

Key words: short texts on the Internet; sentiment classification; convolutional neural networks (CNNs); natural language processing; deep learning

0 引 言

近年来,随着社交网络的逐渐成熟和移动终端技术的迅猛发展,互联网文本作为一种网络传播的主要媒体形式,越来越受到人们的青睐。用户在互联网上表达观点传播思想,抒发个人情感的同时,产生了大量带有个人主观情感特征的信息,这些信息中包含着不同趋向的情感特征,进而对网络舆情的传播能产生巨大的影响。深度挖掘这些情感信息对于网络舆情监控和新闻专题追踪都有着重要意义[1]。因此,研究互联网短文本情感分类具有重要意义。

文本的情感分类是指对带有情感色彩的主观性文本进行分析、处理、归纳和判断的过程,本文分为以下 2 个子任务: 1) 主观性识别,指发现文本中具有观点的句子,进行主观性识别能有效减少干扰,帮助文本进行情感极性分类; 2) 情感极性判定,文本情感极性判定可以看成是一类特殊的文本分类问题,与文本分类的区别主要在于特征的选择上,通过将不同情感

收稿日期: 2016-08-17
基金项目: 国家自然科学基金资助项目(U1304611); 河南省科技攻关计划项目(132102310284); 河南省教育厅科学技术研究重点项目(14A520015); 郑州市科技攻关项目(131PPTGG416-4)
作者简介: 刘小明(1979-),男,河南许昌人,中原工学院计算机学院讲师,博士,研究方向: 机器学习,自然语言处理; 张英(1992-),女,河南洛阳人,硕士研究生,研究方向: 自然语言处理,情感分析; 郑秋生(1965-),男,河南郑州人,教授,硕士,研究方向: 信息安全,数据资源管理。

特征词视为不同的类别,利用类别间的差异建立分类器以区分不同的文本情感极性[2]。

目前,情感分析是自然语言处理领域的研究热点,针对文本情感分析的方法有很多,其中大多数是基于传统机器学习的最大熵(Maximum Entropy,MaxEnt)和支持向量机(Support Vector Machine,SVM)算法。这些传统方法对于普通规范文本情感分类取得了较好的效果,但由于互联网短文本存在主题不明确、情感发散、语言不规范、未登录词等较多等问题,使得这些现有的基于传统机器学习的方法无法学习到互联网文本深层语义信息,导致其情感分类准确率并不高。为了解决这一现状,本文采用基于深度学习的卷积神经网络(CNNs)模型的方法来研究互联网短文本中的情感分类问题,通过 CNNs 模型的逐层分析优化,提取更多文本特征,获得较好的短文本情感分类效果[3]。

1 相关研究

1.1 深度学习

深度学习是近年来兴起的机器学习研究的一个新领域,其利用多层神经网络结构模拟人脑对大量数据进行分析,提取有效特征。2006 年,Hinton[4] 等人最早提出了深度学习的概念,随后又用深度学习网络构造出高质量的语言模型用于处理自然语言问题;Lecun[5] 等人首次利用 CNNs 解决手写识别问题,CNNs 的概念由此被提出;2010 年,Mikolov[6-7] 等人使用 Log-Bilinear 模型将深度学习模型降低到可接受范围内,并在谷歌推出 Word2vec 将词语转换成词向量的工具;梁军[8] 等利用深度学习来做中文文本微博情感分析,采用 LSTM 递归网络发现特征并根据前后词语的关联性引入情感极性转移模型,进行情感分析并取得了不错的效果;而 Yoon Kim[9] 使用基于深度学习的 CNNs 模型处理文本情感分类任务,通过改进模型结构并使用特定的静态字向量来提高准确率,虽已取得了良好的效果,但该方法在处理中文文本中常出现的多重语义现象中使用特定的静态字向量,具有局限性,不能充分表达其深层语境特征。因此,本文考虑上下文语境特征并采用 Word2vec 深度学习模型的工具训练词向量,使其更适用于互联网文本情感分类问题。

1.2 情感分类

互联网短文本的情感分类问题具有很大的研究价值和应用价值,目前已在国内外广泛展开。其中,国内情感分类的研究方法主要有 2 类:1)使用传统机器学习的方法。这些方法主要有基于监督学习利用传统的最大熵、SVM、条件随机场(Conditional Random Field,CRF)、朴素贝叶斯(Naive Bayesian,NB)等算法,训练语料构造分类器进行情感分类。2)基于情感词典提取情感特征词计算文本的情感倾向性。其中,谢丽星[10] 等人提出了基于层次结构的多策略中文微博情感分析和特征抽取,分别使用表情符号的规则方法、情感词典的规则方法和基于 SVM 的层次结构的多策略方法进行对比实验,最终结果表明基于 SVM 的层次结构多策略方法效果最好;孙建旺[11] 等人针对微博文本的特点,选用极性副词和表情符号特征构建词典计算情感特征值,提出了基于词典和机器学习相结合的方法,并通过 SVM 模型将文本分为正、中和负 3 类情感;随后,郑妍[12] 等人提出了一种基于情感主题模型的特征选择方法,通过极性词和情感主题模型的特征选择来生成特征子集,该方法可有效提高跨领域文本情感分类的准确率。

虽然传统方法在情感分析问题中取得了不错的成绩,但这些方法的情感特征的选择过于依赖现有的情感词典或人工标注的语料库,需要大量人力资源对情感语料库进行不断的完善。因此,本文使用卷积神经网络模型处理文本信息,避免过于依赖人工标注的问题。

2 基于 CNNs 的情感分类模型分析

近年来,随着自然语言研究问题的深入发展,已有的基于机器学习的传统特征抽取方法已不能满足当前需求,有学者尝试利用深度学习的方法解决自然语言处理中的一些难题并取得不错的效果。为此,本文也使用深度学习模型处理文本情感分类问题。

2.1 Word2vec 简介

Word2vec 是 Google 推出的一种训练词向量模型的工具实现,由简单的三层神经网络构成。将词语转换为向量的表示方式的实现可谓是将 Deep Learning 算法引入 NLP 领域的一个核心技术。

将自然语言理解问题转化为机器学习问题的第一步就是要找一种方法把这些符号数学化,而 Word2vec 就是可以把一个词条用数字特征表示的工具,Word2vec 是一个神经网络,它可以在使用深度学习算法之前预处理文本。虽然它本身并没有实现深度学习,但是 Word2vec 把文本变成深度学习能够理解的向量形式。

Word2vec 在不需要人工干预的情况下创建特征,包括词的上下文特征。这些上下文来自于多个词的窗口。如果有足够多的数据、用法和上下文,Word2vec 能够基于这个词的出现情况高度精确地预测一个词的词义。

2.2 CNNs

作为深度学习的模型之一,CNNs 是首个成功训练多层网络结构的监督学习算法,它利用空间相对关

系减少参数数目以提高训练性能,其本质就是多层卷积运算。

传统前馈神经网络中,每个输入神经元对应连接下一层的输出神经元,这种方式叫做全连接层或称仿射层。而卷积神经网络利用一系列的卷积层、池化层以及一个全连接输出层构建模型。CNNs 首先通过输入层进行卷积来计算输出,相当于局部连接,每块局部的输入区域连接一个输出神经元,然后对每一层应用不同的滤波器,基于需要完成的任务自动学习滤波器的权重值,并通过池化层降采样汇总其结果,构建一个多层的网络来模仿人脑感知视觉信号的逐层处理机制,以实现视觉特征信号的自动提取与识别。

CNNs 早期用来对图像进行识别,由于其在局部特征抽取方面具有极大的优势,成为该领域的研究热点。近年来 CNNs 已被用于文本分类、要素识别等自然语言处理任务中并取得了不错的效果,因此,本文参考 Kim[9] 等人的工作,引入卷积神经网络模型来处理针对中文语料的情感分类问题。

2.3　CNNs 应用于情感分类

本文采用的单层 CNNs 模型结构如图 1 所示,由卷积层和池化层以及全连接层构成。首先,输入数据通过滤波器和可加偏置进行卷积,卷积后由卷积核提取出局部特征映射图,然后,由池化层再进行加权值,加偏置,降采样提取出每张特征图中的典型特征,最后通过全连接层映射得到输出向量。

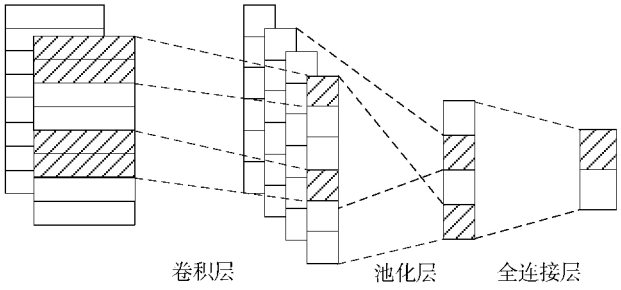

图 1　CNNs 模型结构图

若 $X_i \in R^k$ 为第 i 个词对应的 k 维向量,则一个长度为 n 的句子表示为:

$$X_{1:n} = X_1 \oplus X_2 \oplus \cdots \oplus X_n \tag{1}$$

其中⊕为连接运算符,$X_{i:i+j}$ 表示由词向量 X_i,X_{i+1},…,X_{i+j} 组成的特征矩阵。滤波器 $w \in R^{hk}$ 是指对窗口为 h 的 k 维词向量进行卷积操作。例如,对特征 c_i 卷积,生成词向量 $X_{i:i+h-1}$ 的窗口:

$$c_i = f(w \cdot X_{i:i+h-1} + b) \tag{2}$$

其中 $b \in R$ 为偏置项,f 为非线性激活函数[9]。神经网络中常用的激活函数有多种,例如 sigmod 函数、tanh 函数等。为加快训练收敛速度,本实验中采用 ReLu 函数作为激活函数:

$$f(x) = \max(0, x) \tag{3}$$

对输入文本的矩阵$\{X_{1:h}, X_{2:h+1}, \cdots, X_{n-h+1:n}\}$进行过滤,就可以得到一张特征图:

$$C = [c_1, c_2, \cdots, C_{n-h+1}], c \in R^{n-h+1} \tag{4}$$

由于文本特征关联到词的前后文信息,因此要提取出更多的特征,需要通过改变窗口 h 设计不同大小的卷积核,再经过卷积操作得到不同尺寸的特征图,提取出每种卷积核的二维特征。

在自然语言处理中,文本情感分类大多数依赖于句子的情感特征。因此,本文使用不同大小的卷积核提取出更多不同特征图,然后对此特征图进行 max-pooling 操作,从而从输入数据中提取出最优局部特征。由于卷积核的大小不同,卷积所得到的特征图尺寸也不同,要用多个不同大小的窗口进行 max-pooling 操作,并按照 $\hat{c} = \max\{c\}$ 取最大值,过滤出特征图中的最优特征。最后,通过全连接层将所有得到的局部最优特征连接到最后一层的输出结点,使得更加充分地考虑提取出来的所有特征,完成情感分类任务。

2.4　输入处理

卷积神经网络通常是处理由二维矩阵组成的数据。因此,需要将文本数据处理成二维矩阵的形式作为模型的输入数据。首先对文本数据进行预处理,把每个句子进行分词处理,然后将每个词用 Word2vec 转换成对应的词向量形式表示,用每个句子中的若干个词组成对应矩阵的每行,即构成了一个二维矩阵。若用 k 维的词向量表示 n 个词的句子,则输入为 k × n 的二维数据矩阵。假设文本长度最长不超过 n,则为满足输入二维矩阵,长度小于 n 的需要进行补零处理,如图 2 所示,若文本数据的长度不足 n,不足位补零,维度为 k。

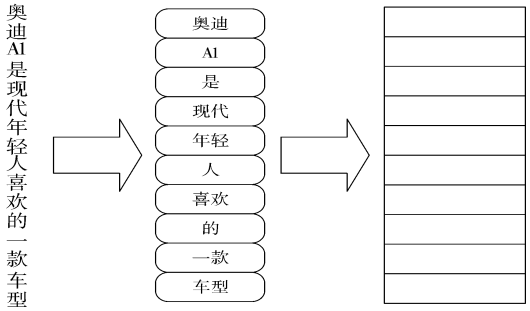

图 2　二维特征矩阵图

这样输入文本通过 Word2vec 训练成词向量,再根据句子进行不同的纵向排列组合,便成为适合 CNNs 处理的二维特征矩阵,即模型的输入数据。

2.5　训练模型

对于卷积神经网络模型的训练过程,不仅要寻找到最优解,也要考虑模型的训练效率,因此,本文采用 mini-batch 梯度下降法训练模型。该方法每次只选择一小部分样本参与运算更新权值,这样在寻找到最优

解的同时收敛速度也较快。

为了防止过拟合,利用 L2 正则化约束网络参数。在全连接层训练过程中引入 Dropout 策略,即每次迭代后随机放弃一部分训练好的参数,经过反复实验,将 Dropout 的值设置为 0.5,即随机放弃一半参数。最后,为了得到稳定可靠的模型,采用多次迭代交叉验证实验结果,最终结果取多次实验的平均值[13]。

3 文本情感分类实验

本文为验证该模型的可行性设计了实验,首先进行主观性识别,使用 CNNs 模型判断文本中的句子是否为主观句,即带有个人感情色彩的评论或断言,然后对过滤掉数据集中的客观句进行情感极性判定实验,用 CNNs 模型对剩余数据进行训练,构造情感分类器,判断主观句的情感正负极性。

3.1 实验设计

实验总体过程如图 3 所示。

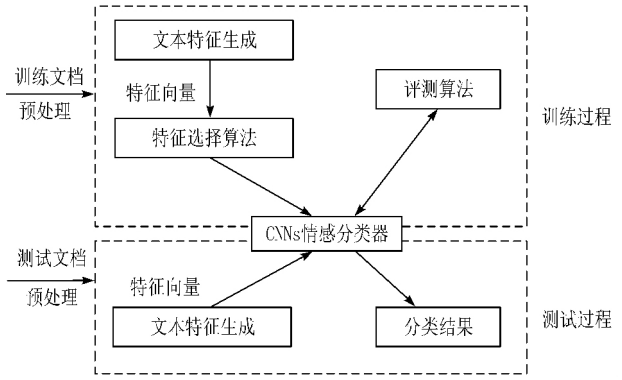

图 3 实验方案

图 3 为实验总体设计,其中输入数据为训练文档和测试文档,首先对输入数据进行预处理,然后生成特征向量,使用训练文本的特征向量训练 CNNs 情感分类器,最后用 CNNs 情感分类器对测试文档进行分类并检验分类效果。

3.2 数据选择与处理

为了验证模型的可行性,本文选取 COAE2014 中 Task4 提供的 5000 条测试数据集作为实验数据,选择其中 4000 条作为训练数据,其余 1000 条作为测试数据,采用人工标注的方法对训练数据集进行情感极性标注,其中情感极性为正的有 1971 条,情感极性为负的有 2029 条。

由于语料中存在大量的虚词,而汉语中的这些虚词包括副词、介词、连词、助词、叹词以及拟声词等词性,其中除副词、叹词对情感倾向性有较大的影响需作为情感特征考虑外,其余词性的词语可列入停用词表不予考虑[14]。因此,情感特征模型训练方法如下:

1) 使用 NLPIR 中文分词工具对实验数据集进行分词和词性标注预处理并过滤掉其中的标点符号和特殊字符等。

2) 根据标注的词性过滤掉其中介词、连词以及助词等不影响情感极性的虚词。

3) 对过滤后的数据集用 Word2vec 工具训练成不同维度的情感特征向量。

3.3 模型参数

卷积窗口的大小不同,获取的局部信息也不同,因此,考虑到文本的上下文信息,设置窗口大小 h 为 5。当短文本最长长度不超过 n 时,h = 5,则特征图的长度为 n + 1 - 5。其余参数值设置如表 1 所示。

表 1 CNNs 模型参数设置

参数	参数值
卷积窗口大小	5
过滤器数	100
batch	50
Dropout	0.5
迭代次数	10

由于向量维度的增加会成倍地增长整个模型的复杂度,因此,特征向量维度的选择对实验结果有着重要的影响,本文对比 50 维、100 维、150 维和 200 维 4 种不同维度的二维词向量矩阵,并使用其余 1000 条未标注数据作为测试数据进行预测。同时采用十折交叉法,将样本均分为 10 份并进行 10 次实验,最终结果为 10 次实验的平均值,实验结果如图 4 所示。

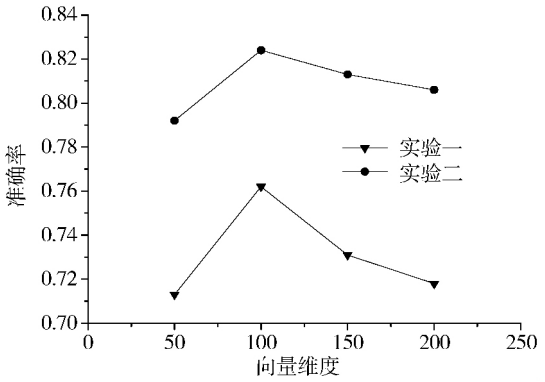

图 4 不同维度向量的结果

由图 4 可以看出,词向量维度选择 100 维时可达到最佳实验结果。

3.4 对比实验

在词向量 100 维的基础上,本文选择与基于传统机器学习的 SVM 模型和基于深度学习的 RNN(Recursive Neural Network)[8] 模型以及 LSTM(Long Short Term Memory)[15] 模型进行对比。模型具体方法设计如下:

SVM 模型: 本文选取 unigram 作为情感特征,使用 Tfidf 计算特征值,并用 libSVM 工具进行 SVM 分类实验。

RNN 模型: 使用 Word2vec 训练词向量并用 RNN

模型训练分类器进行对比实验。

LSTM 模型: 同样使用 Word2vec 训练词向量并用 LSTM 模型训练分类器进行对比实验。

CNNs-Rand 模型: 随机初始化构造特征词向量，其余 CNNs 模型不变进行对比实验。

CNNs-Word2vec 模型: 使用 Word2vec 工具训练特征向量进行实验。

3.5　实验结果分析

根据 COAE 评测提供的标准作为实验结果的评价指标，依照评价标准计算准确率、召回率和 F 值。其中，实验结果如表 2 所示，图 5 为 6 种方法的准确率对比图。

表 2　情感分类实验结果

方法	准确率	召回率	F 值
SVM	0.825	0.819	0.813
RNN	0.839	0.836	0.835
LSTM	0.846	0.842	0.843
CNNs + Rand	0.867	0.853	0.862
CNNs + Word2vec	0.893	0.875	0.887

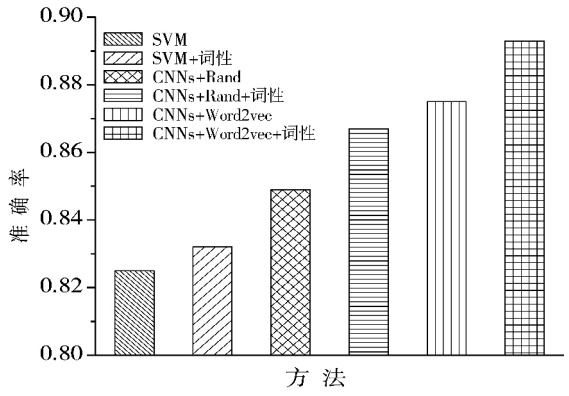

图 5　准确率对比图

对比表 2 实验结果发现，在情感分类任务中，使用 CNNs + Word2vec 模型有较好的实验结果，证明该模型处理互联网短文本情感分类问题具有可行性。

通过对比图 5 中传统 SVM 模型和 CNNs 模型实验结果发现，CNNs 模型的准确率相较于传统 SVM 分类方法有较多提升，其主要原因是本文使用的 CNNs 模型不同于传统 SVM 分类方法而是通过自身模型提取特征，这在特征提取上比传统方法有很大的优势。

通过对比基于深度学习的 RNN 和 LSTM 模型与 CNN 模型的实验结果可以发现，使用 CNNs 模型结果的准确率明显优于其余 2 种模型，这说明 CNNs 模型比其余深度学习模型更适用于处理文本情感分类问题。

比较 CNNs + Rand 和 CNNs + Word2vec 的实验结果可以发现，同样使用 CNNs 模型，基于 Word2vec 训练的特征词向量其准确率明显优于随机构造的词向量。其主要原因在于，针对中文互联网短文本，使用 Word2vec 工具训练的词向量要比随机词向量能够抽取更多原始数据的特征。

4　结束语

情感分析一直是自然语言处理中的重要任务之一。本文针对互联网短文本，采用基于深度学习的 CNNs 模型处理情感分类问题，该模型不仅能够挖掘更多文本特征，弥补传统方法的不足，同时，也比其余方法在准确率以及算法的性能上具有更多优势。

由于神经网络模型的结构较为复杂，其在情感分析上的应用研究还存在很多问题需要进一步探讨。因此，下一步笔者将深入研究模型结构，调整优化参数，降低模型结构风险，从而使神经网络模型能够更加适用于处理情感分析等自然语言问题。

参考文献:

[1] 杜振雷. 面向微博短文本的情感分析研究 [D]. 北京: 北京信息科技大学, 2013.
[2] 薛璐影. 文本情感分类相关问题研究 [D]. 哈尔滨: 哈尔滨工业大学, 2010.
[3] 赵妍妍, 秦兵, 刘挺. 文本情感分析 [J]. 软件学报, 2010, 21(8): 1834-1848.
[4] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.
[5] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[6] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model [C] // INTERSPEECH 2010, Conference of the International Speech Communication Association. 2010: 1045-1048.
[7] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26(1): 3111-3119.
[8] 梁军, 柴玉梅, 原慧斌, 等. 基于极性转移和 LSTM 递归网络的情感分析 [J]. 中文信息学报, 2015, 29(5): 152-159.
[9] Kim Y. Convolutional Neural Networks for Sentence Classification [EB/OL]. https://arxiv.org/abs/1408.5882, 2014-08-25.
[10] 谢丽星, 周明, 孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取 [J]. 中文信息学报, 2012, 26(1): 73-83.
[11] 孙建旺, 吕学强, 张雷瀚. 基于词典与机器学习的中文微博情感分析研究 [J]. 计算机应用与软件, 2014, 31(7): 177-181.
[12] 郑妍, 庞琳, 毕慧, 等. 基于情感主题模型的特征选择方法 [J]. 山东大学学报(理学版), 2014, 49(11): 74-81.
[13] 蔡慧苹, 王丽丹, 段书凯. 基于 Word Embedding 和 CNN 的情感分类模型 [J]. 计算机应用研究, 2016, 33(10): 2902-2905.
[14] 詹红英, 朱学锋. 面向自然语言处理的汉语虚词研究与广义虚词知识库构建 [J]. 当代语言学, 2009(2): 124-135.
[15] Wang Peilu, Qian Yao, Soong F K, et al. A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding [EB/OL]. https://arxiv.org/abs/1511.00215, 2015-11-01.