基于 CNN-BIGRU 的中文文本情感分类模型

宋祖康,阎瑞霞
(上海工程技术大学 管理学院,上海 201620)

摘 要:在当今商业领域,对网络评论的情感分类一直是一个比较热门的研究方向,而为了克服传统机器学习方法所构建分类器会产生较大计算开销,精度表现较差的缺点,提出一种基于深度学习模型中卷积神经网络(CNN)与循环神经网络(RNN)模型的情感分类方法。在以往的研究中,卷积神经网络往往被用来提取文本的局部特征信息,但却容易忽视文本的长距离特征,而 RNN 则往往被用来提取句子的长距离依赖信息,但容易陷入梯度爆炸问题。因此,结合卷积神经网络对于局部特征信息的良好提取能力与循环神经网络对于长距离依赖信息的记忆能力,构建了一个 CNN-BIGRU 混合模型,用以提取文本的局部特征以及文本的长距离特征。其中循环神经网络模型使用了双向 GRU 模型,以避免 RNN 模型的梯度爆炸与梯度消失问题。在谭松波的酒店评论数据集上的实验结果表明,利用该模型,实验分类的准确率比单独使用卷积神经网络模型最高提升了 26.3%,比单独使用循环神经网络模型最高提升了 7.9%,从而提高了对中文文本情感分类的精度,并减少了计算开销。

关键词:卷积神经网络;循环神经网络;文本分析;情感分类
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2020)02-0166-05
doi:10.3969/j.issn.1673-629X.2020.02.032

Chinese Comment Sentiment Classification Model Based on CNN-BIGRU

SONG Zu-kang, YAN Rui-xia
(School of Management, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: In today's business field, the sentiment classification of online comments has always been a hot research direction. In order to overcome the shortcomings of the classifier constructed by the traditional machine learning method, such as large computational overhead and poor accuracy, a sentiment classification method based on the convolutional neural network (CNN) and recurrent neural network (RNN) in the deep learning model is proposed. In previous studies, CNN is often used to extract the local feature information of the text, but it is easy to ignore the long-distance feature of the text, while RNN is often used to extract the long-distance dependent information of the sentence, but it is easy to fall into the gradient explosion. Therefore, combining the great local feature information extraction of CNN and the memory of RNN to long-distance dependent information, we construct a CNN-BIGRU hybrid model to extract local feature and long-distance feature of text. A two-way GRU model is used in RNN model to avoid the gradient explosion and gradient disappearance of the RNN model. The experiment on Tan Songbo' hotel reviews data set shows that the classification accuracy of the proposed model is the highest by 26.3% compared with the CNN alone, and the highest by 7.9% compared with RNN alone, so as to improve the accuracy of the affection of Chinese text classification and reduce the computational overhead.

Key words: convolutional neural network; recurrent neural network; text analysis; sentiment classification

0 引 言

在自然语言处理领域,情感分析一直是一个比较热门的研究方向,伴随着互联网的发展,大量的商业评论涌现在各个平台上,商业评论大多夹杂着用户对商品的个人意见,因此对于这些评论文本情感极性的判别研究,可以帮助企业更好地了解自己产品或服务的客户满意度[1]。传统对文本的情感极性判别是从 20 世纪 90 年代开始的基于机器学习的方法。传统机器学习的方法主要分为两个步骤,首先人工构造特征来获取所需的文本信息。在这一步中,使用的传统方法是词袋模型(BOW)[2],将每个词依据事先建立的词典转换为 one-hot 向量,这种方法存在的一个缺点就

收稿日期:2019-02-28 **修回日期**:2019-06-28 **网络出版时间**:2019-11-07
基金项目:国家自然科学基金(71301100);上海市教委科研创新(14YZ140)
作者简介:宋祖康(1995-),男,硕士研究生,研究方向为自然语言处理、机器学习;阎瑞霞,博士,副教授,硕导,研究方向为粗糙集、模糊集、数据挖掘和机器学习。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191107.0912.042.html

是得到的文本向量具有高维度、高稀疏的特点,因此出现了一些降维的方法,如 TF-IDF、SVD 模型等等。另外为了使向量能够表现上下文的信息,还出现了 LDA、词嵌入(word embedding)[3]等模型,而词嵌入模型则更是将深度学习算法映入到自然语言领域的一个重要研究成果。在获取了所需的文本信息后,需要构建分类器对文本的情感极性进行分类。经典的机器学习方法基本都能够使用在文本分类之中,如支持向量机、随机森林、朴素贝叶斯等算法。

近年来,由于词向量等特征学习方法持续获得关注,深度学习在自然语言领域的发展也尤为迅速。深度学习模型可以自动从数据中提取特征,如 Bengio 等[4]利用深度学习思想构建的神经概率模型,将各种深层神经网络使用在大规模英文语料库上学习,完成了命名实体识别以及句法分析等多个自然语言处理的任务,CNN(convolutional netural network,卷积神经网络)与 RNN(recurrent neural network,循环神经网络)也被证明是情感分类任务上的有效模型。在文本的情感分类方面,Yuan S[5]、Vieira J P A[6]、Zhao Y[7]、Zhang Y[8]、Vo Q H[9]等利用循环神经网络与卷积神经网络等模型对短文本进行情感分类,获得了很好的效果,但是由于 RNN 模型的梯度爆炸问题,因此基于 RNN 模型的 LSTM 与 GRU 模型是目前比较常用的模型[10-14]。因此文中提出了基于 CNN-BIGRU 的中文文本情感分类模型,构建 CNN 模型提取句子的局部特征,使用 RNN 模型中的双向 GRU 模型提取句子的上下文长距离依赖特征,利用 Keras 开源库中的 Merge 层将两个模型融合。在谭松波的酒店评论语料上采用十折验证法进行实验验证,实验结果证明,该模型比传统的 RNN 与 CNN 模型在准确率与 F 值上都有显著提高。

1　基础知识

1.1　词向量模型

最早的词向量模型为 one-hot 词袋模型,将所有模型构成一个词典 D,用一个长向量来表示一个词,这种词向量容易造成维度灾难,并且不能很好地表示词的上下文关系,例如“他打了我”和“我打了他”的词向量是相同的,但是却有完全不同的意思[9]。而在近几年出现了一种分布表示词向量的模型,基本思想是把所有的词用固定长度的向量来表示,最近谷歌开源的 word2vec 词向量模型就属于这种分布式模型。在文中所构建的模型中,借鉴了分布式表示词向量的方法。

首先提取文章中所有的单词,按其出现的次数降序排序,接着将每个编号赋予一个 one-hot 词向量,然后使用 skip-gram 模型生成一个矩阵 $\boldsymbol{M}$。skip-gram 模型是一个一到多的词向量生成模型,在模型中,首先会随机初始化矩阵 $\boldsymbol{M}$,然后使用神经网络来训练矩阵 $\boldsymbol{M}$,标签为附近词对应的 one-hot 编码,通过这种方法可以生成固定长度的词向量,模型结构见图 1。

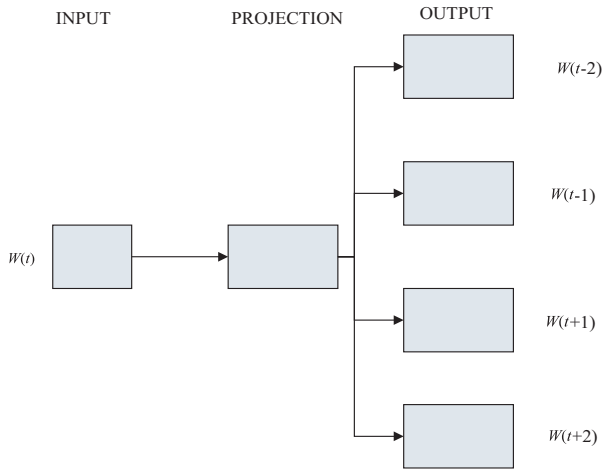

图 1　skip-gram 模型

1.2　CNN 模型

CNN 模型最初常应用在图像处理领域,卷积神经网络模型与一般神经网络模型相似,它是由输入层、隐藏层和输出层组成,可以捕捉数据的局部特征,主要是通过反向传播算法进行参数的优化。

CNN 网络的输入层是前文所述的词向量矩阵,即分词后的句子矩阵。假设一个句子在中文分词后有 M 个词,则这个词向量矩阵的行数为 M 行,每个词的词向量维度是 N 维,矩阵为一个 $M * N$ 的矩阵。CNN 网络的隐藏层包括有卷积层和池化层,卷积层是 CNN 网络的核心层,通过不同大小的卷积核,可以对文本依次进行卷积运算,得到确定个数的卷积映射。池化层往往与卷积层交替出现,主要是负责对特征的大小进行压缩,以简化网络的计算开销。池化层往往包括 average pooling 和 max pooling,池化层通常是输入卷积后的特征映射中的最大值,从这个角度上来讲,可以解决句子长短不一的问题。CNN 的输出层一般连接着一个 softmax 层,输出分类的概率以及最终的分类结果。

1.3　双向 GRU 模型

RNN 模型与 CNN 模型相同,其主要架构为输入层、隐藏层和输出层[15]。RNN 模型中,从单一方向输入一个词向量矩阵,而输出单元也将会是一个单一方向矩阵,大部分的工作将在隐藏单元中完成。由于 RNN 模型常常存在着梯度爆炸的问题,由此而衍生了来源于 RNN 模型的双向 LSTM 模型。GRU 模型是 LSTM 模型的一个变体,它将 LSTM 模型中的忘记门和输入门合成了一个单一的更新门,并且还混合了细

胞状态和隐藏状态以及一些其他的改动,最终输出的模型要比标准的 LSTM 模型简单,并且其效果基本一样。但是 GRU 模型的状态输出要少一个,在编码时使用 GRU 模型可以让代码变得简单一些。因此文中使用 GRU 模型,GRU 模型可以学习长期依赖信息[16-17]。它主要有两个门,一个是更新门,一个是输出门,在 GRU 模型中,当前单元的状态是通过计算求和上一个单元状态得到,也就是说,模型可以得到历史信息和当前信息,这在语言处理中对于提取上文信息有很大的帮助。但是标准的无论 LSTM 还是 GRU 模型都是以时间顺序处理时间序列,这样的话就会忽略掉下文的信息,因此文中采用了双向 GRU 模型。双向 GRU 模型是在单层 GRU 的基础上扩充了一层 GRU 模型,通过让两层以相反的方向流处理数据来获得上文信息以及下文信息,这样可以充分提取所有文本的信息,双向 GRU 模型结构见图 2。

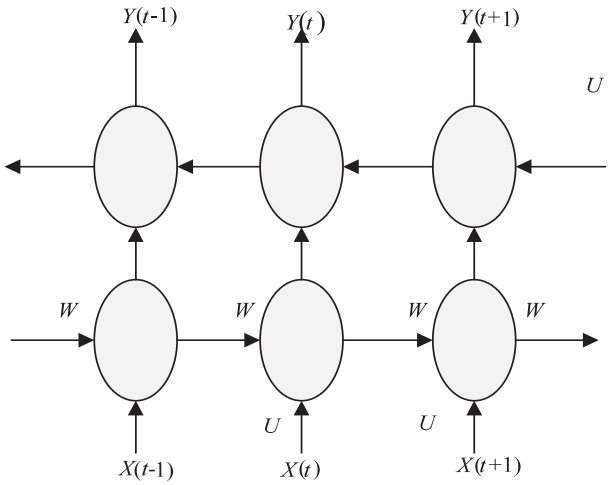

图 2 双向 GRU 模型

2 文中模型

通过 CNN 模型来解决文本分类问题,虽然从一定程度上可以解决维度灾难问题以及中文文本长短不一的问题,但是 CNN 模型只能提取局部特征。例如,在“这家酒店的周围环境虽然比较嘈杂,但总体还是不错的”这样一句话中,固定大小的卷积核,很可能只能提取到“环境-嘈杂”,而无法将“酒店-不错”这样带有反转色彩的评价提取出来,也就是说,CNN 模型无法解决句子的长距离依赖问题,并且,CNN 模型需要固定卷积核窗口的大小,对于卷积核的参数调节也比较麻烦。而 RNN 模型则可以很好地解决对序列的长距离依赖问题,因此文中提出一种新的文本分类模型 CNN-BIGRU(CNN-bidirectional-RNN),使用 CNN 模型作为辅助分类模型,以充分提取句子的局部特征,结合双向 GRU 模型以充分提取句子的长距离依赖特征,具体模型见图 3。

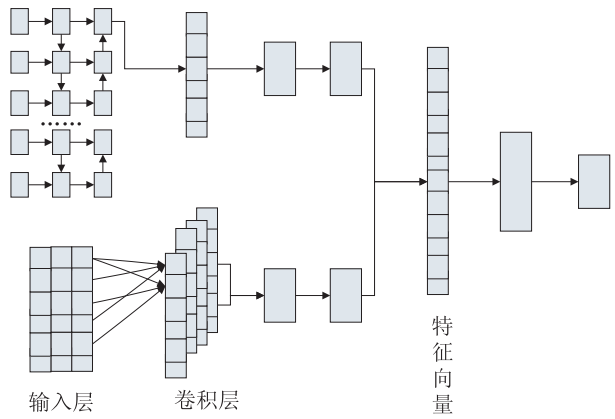

图 3 CNN-BIGRU 模型

(1)对评论语料进行预处理,去除一些无用词,并且使用 python 库中的 jieba 分词库对中文文本进行分词,输入 skip-gram 中训练得出词向量。

(2)分别搭建了两个卷积核步长为 2 和 3 的 CNN 卷积模型,以提取 2 个与 3 个词间距内文本信息,记为 2-gram 与 3-gram,将预处理后的句子矩阵输入 CNN 网络的输入层。文中所搭建 CNN 网络的隐藏层包括卷积层和池化层,卷积层为双层,对文本依次进行卷积运算,激活函数为 relu 函数,得到确定个数的卷积映射后将数据输入池化层,池化层为全局平均池化,以解决句子长短不一的问题。

(3)同时也搭建了一个双向 GRU 模型以提取文本的长距离依赖信息,与 CNN 模型相同,将预处理后得到的词向量矩阵按照顺序输入双向 GRU 模型输入层,假设每个词的词向量为 x_t,依据图 2 可以得到,更新门的计算(σ 为 sigmoid 函数)如下。

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{1}$$

输出门的计算公式如下:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{2}$$

GRU 单元状态更新公式如下:

$$\widetilde{h_t} = \tanh(W_{h_t} \cdot [r_t * h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \tag{4}$$

单元输出层公式如下:

$$y_t = \sigma(W_o \cdot h_t) \tag{5}$$

GRU 网络的计算使用了 BPTT 算法,将输入层记为 $\{x_1, x_2, \cdots, x_n\}$,隐藏层的输出记为 $\{s_1, s_2, \cdots, s_m\}$,BPTT 算法主要是将原来的网络折叠开,利用前面时间序列的影响对最后一个分类作判断,对应的每一层的计算方法如下:

$$t_i = W_{hx} x_i + W_{hh} h_{i-1} + b_h \tag{6}$$

$$h_i = e(t_i) \tag{7}$$

$$\mathrm{s_i} = \mathrm{W_{yh}} \mathrm{h_i} + \mathrm{b_y} \tag{8}$$

$$\mathrm{y_i} = \mathrm{g}(\mathrm{s_i}) \tag{9}$$

其中:

$$\tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{10}$$

$$g(x) = \mathrm{softmax}(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \tag{11}$$

与传统神经网络不同的是,GRU 的损失函数为交叉熵函数。

(4)通过 Keras 的 Merge 层将不同卷积核的 CNN 模型与双向 GRU 模型融合为一层模型,并且接入全连接层以及一个 sigmoid 层,最后实现语料的情感分析。

3　实例验证

3.1　实验数据

实验数据为谭松波博士的酒店评论语料,语料规模为 10 000 篇,全部是从携程网上自动采集,语料被分成四个子集,分别是:(1) ChnSentiCorp-Htl-ba-2000:平衡语料,正负类各 1 000 篇;(2) ChnSentiCorp-Htl-ba-4000:平衡语料,正负类各 2 000 篇;(3) ChnSentiCorp-Htl-ba-6000 平衡语料,正负类各 3 000 篇;(4) ChnSentiCorp-Htl-u nba-10000:非平衡语料,正类为 7 000 篇。选用其中的一个语料集 ChnSentiCorp- Htl-ba-6000,该语料包含 3 000 篇积极倾向语料与 3 000 篇消极倾向语料,文中采用十折验证法,将语料分成 10 份,其中 9 份为训练集,1 份为测试集,因此训练集为 5 400 篇,测试集为 600 篇。数据样例见表 1。

表 1　数据样例

编号	积极评论	编号	消极评论
1	住的豪华大床房,房间基本无可挑剔,就是阳台脏了些。早餐要早点去,晚了连座位都没有。还是比较满意的	1	房间比较差,尤其是洗手间,房间隔音和餐饮服务都不好
2	觉得是许昌最好的酒店,性价比相当高房间大,整洁,硬件也不错,服务相当的好,自助早餐也不错,离春秋广场很近,出去玩也方便	2	酒店老化,房间装修差,服务过于生硬,中午退房时间过不到两小时收取半日房费,而其他酒店基本从人性化考虑,不收此项费用
3	总体上来说还是不错的,一家四星级酒点,毕竟是许昌市最好的去处了,唯一的缺点是房间的隔音不太好,隔壁打麻将的声音吵的我好半天没有睡着	3	房间设施太过简陋,顶多是个普通的招待所的标准,卫生间太简陋,那个马桶简直脏的吓人,不是打扫的不干净,不知道是用了太多年还是质量太差,整个马桶都是乌其麻黑的

3.2　实验过程

文中主要使用 Keras 深度学习接口,以 Tensorflow 作为后台框架,实验器材为联想一体机,处理器为 Inter Core(TM) i5-4590S CPU @ 3.00 GHz,安装内存为 4.00 GB。

3.2.1　文本预处理

首先对文本进行数据清洗,构建数据词典,将一些停用词以及无关词去除,然后对数据进行分词,并将数据集顺序调整为一条积极评论一条消极评论,以便于使用十折验证法时训练集与验证集的数据分布均衡。然后对所有语料编制词典,赋予索引值,将索引矩阵输入 skip-gram 词向量训练模型,文中所取句子长度为 200 个词,语料超过 200 个词部分会被删去,不足两百个词部分对其进行补 0 操作。

3.2.2　搭建模型

CNN 为两层模型,卷积核分别选用 2 个、3 个以及 2 个与 3 个结合三种步长,RNN 为双向 GRU 模型,之后使用 Merge 层将两个模型做融合,添加 Dense 层,再添加 Sigmoid 层对文本进行分类。

卷积神经网络参数设置如表 2 所示。

3.3　实验结果

经过多次试验,发现 2-gram CNN-RNN、3-gram CNN-RNN 与 2-3-gramCNN-RNN 的迭代论数设为 20 轮、10 轮与 6 轮时,模型即可达到最优。

表 2　卷积神经网络参数设置

参数	参数值
卷积核尺寸	(3,200),(4,200),(5,200)
向量维度	200
Dropout 比例	0.5
迭代方法	Mini-batch

模型试验结果对比见表 3。

表 3　各模型实验结果对比

模型	迭代	精度/%	F 值
GRU	50	75.3	0.776
LSTM			
2-gram CNN	50	57.1	0.592
3-gram CNN	50	55.4	0.548
2-gram CNN-BIGRU	20	81.7	0.829
2-gram CNN-BILSTM			
3-gram CNN-BIGRU	10	80.6	0.834
3-gram CNN-BILSTM			
2-3-gram CNN-BIGRU	10	83.2	0.834
2-3-gram CNN-BILSTM			

注:2-gram 表示两个词间距,其余相同。

通过表 3 可以得出:

(1)相比较 RNN 模型,采用 CNN-BIGRU 模型的情感分类准确率在 2-3-gram 上提升了 7.9%,F 值提升了 0.058,在 2-gram 与3-gram 上准确率分别提高了 6.4% 与 5.3%, *F* 值分别提高了 0.053 与 0.058。分析其原因是双向 GRU 模型对上下文特征的提取使得模型对文本情感倾向的提取更加充分,而 CNN 模型作为辅助模型,使得模型对局部特征的关注也足够充分,因此模型的信息提取也更加完善。

(2)相比较 CNN 模型,文中模型相比较 2-gramCNN,准确率分别提高了 24.6%、23.5% 与 26.1%, *F* 值分别提高了 0.237、0.242 和 0.242,相比较 3-gramCNN,准确率分别提高了 26.3%、25.2% 和 27.8%, *F* 值分别提高了 0.281/0.286 和 0.286。分析其原因是一方面文中语料虽然是短文本,但文本长度都不是很短,而 CNN 模型更多地会关注局部特征,这样就使得 CNN 模型的表现很差,而文中模型在局部特征提取能力与长距离文本依赖信息抽取能力上的优势,使得它可以远远胜过传统的 CNN 模型。

4 结束语

文本情感倾向的判别是自然语言领域如今比较热门的一个方向,文中提出了一种结合传统 CNN、RNN 模型的 CNN-BIGRU 模型。该模型不仅兼具了 CNN 模型的局部特征提取能力,也将双向 RNN 模型的信息记忆能力融合进来。首先通过对语料的处理得到可以充分表示文本信息的矩阵向量;其次提出了改进后的 RNN-BI-CNN 模型,在词向量矩阵输入的基础上针对传统 RNN、CNN 与文中模型做了对比实验。实验结果表明,文中模型在情感倾向判别任务中有着更高的准确率与 *F* 值。

文中的创新点在于:采用词级别的句子情感分析方法,使用双向 GRU 模型兼具考虑了上下文关系,改善了机器学习等传统统计方法以及基于规则的情感分类方法对于人工的依赖,并且采用 CNN 模型作为辅助模型,避免了双向 GRU 模型过度关注长距离特征而忽视局部特征的缺点,利用深度学习的技术避免了人工构建规则进行分类的繁冗方法,并且在成熟数据集上取得了良好的表现。该研究目前仅对小数据集进行了二分类,可以为同一文本的多情感问题以及长文本的情感分类问题提供一定的研究思路。

参考文献:

[1] 林燕霞,谢湘生. 基于社会认同理论的微博群体用户画像[J]. 情报理论与实践,2018,41(3):142-148.

[2] MAAS A L,DALY R E,PHAM P T,et al. Learning word vectors for sentiment analysis[C]//Meeting of the association for computational linguistics:human language technologies. Portland,Oregon:Association for Computational Linguistics,2011:142-150.

[3] JIANG F,LIU Y Q,LUAN H B,et al. Microblog sentiment analysis with emoticon space model[J]. Journal of Computer Science and Technology,2015,30(5):1120-1129.

[4] BENGIO Y,DUCHARME R,VINCENT P,et al. A neural probabilistic language model[M]//Innovations in machine learning. Berlin:Springer,2006:137-186.

[5] YUAN S,WU X,XIANG Y. Incorporating pre-training in long short-term memory networks for tweets classification[C]//2016 IEEE 16th international conference on data mining. Barcelona:IEEE,2017:1329-1334.

[6] VIEIRA J P A,MOURA R S. An analysis of convolutional neural networks for sentence classification[C]//2017 XLIII Latin American computer conference (CLEI). Cordoba:IEEE,2017:1-5.

[7] ZHAO Y,QIN B,LIU T. Encoding syntactic representations with a neural network for sentiment collocation extraction[J]. Science China:Information Sciences,2017,60(11):110101.

[8] ZHANG Y,ER M J,VENKATESAN R,et al. Sentiment classification using comprehensive attention recurrent models[C]//International joint conference on neural networks. Vancouver,BC:IEEE,2016:1562-1569.

[9] VO Q H,NGUYEN H T,LE B,et al. Multi-channel LSTM-CNN model for Vietnamese sentiment analysis[C]//2017 9th international conference on knowledge and systems engineering (KSE). Hue:IEEE,2017:24-29.

[10] 张小川,余林峰,桑瑞婷,等. 融合 CNN 和 LDA 的短文本分类研究[J]. 软件工程,2018,21(6):17-20.

[11] 罗 帆,王厚峰. 结合 RNN 和 CNN 层次化网络的中文文本情感分类[J]. 北京大学学报:自然科学版,2018,54(3):459-465.

[12] 冯兴杰,张志伟,史金钏. 基于卷积神经网络和注意力模型的文本情感分析[J]. 计算机应用研究,2018,35(5):1434-1436.

[13] 任 勉,甘 刚. 基于双向 LSTM 模型的文本情感分类[J]. 计算机工程与设计,2018,39(7):2064-2068.

[14] 王汝娇,姬东鸿. 基于卷积神经网络与多特征融合的 Twitter 情感分类方法[J]. 计算机工程,2018,44(2):210-219.

[15] 刘红光,马双刚,刘桂锋. 基于降噪自动编码器的中文新闻文本分类方法研究[J]. 现代图书情报技术,2016(6):12-19.

[16] 李 杰,李 欢. 基于深度学习的短文本评论产品特征提取及情感分类研究[J]. 情报理论与实践,2018,41(2):143-148.

[17] 徐 凯,陈平华,刘双印. 基于 AdaBoost-Bayes 算法的中文文本分类系统[J]. 微电子学与计算机,2016,33(6):63-67.