doi:10.3969/j.issn.1001-4616.2023.02.015

基于深度学习和注意力机制的微博情感分析

周湘贞[1,2],李　帅[2],隋　栋[3]

(1.郑州升达经贸管理学院信息工程学院,河南 郑州 451191)
(2.北京航空航天大学计算机学院,北京 100191)
(3.北京建筑大学电气与信息工程学院,北京 102406)

[**摘要**] 为了提高微博情感分析的性能,采用深度学习算法中的循环神经网络用于情感分类,并采用注意力机制对词特征进行选择加权,以增强循环神经网络的分类的准确率. 首先,将微博语料进行去噪、分词、向量化等处理,形成微博初始样本. 然后,构建循环神经网络的微博分类模型,通过隐藏层节点循环,并结合历史时刻及当前时刻隐藏层输出获得词特征向量. 接着,注意力机制用于词特征相似计算及选择加权构建句子特征,并采用 Softmax 函数获得分类结果. 最后,通过微博情感分类仿真测试验证了所提方法的可靠性. 实验结果表明,相比常用微博情感分类算法,通过合理设置注意力机制窗口大小,所提方法在不同词向量规模样本下均表现出更高的分类性能.

[**关键词**] 微博情感,深度学习,循环神经网络,注意力机制
[**中图分类号**]TP312;G254　[**文献标志码**]A　[**文章编号**]1001-4616(2023)02-0115-07

Microblog Emotion Analysis Based on Deep Learning and Attention Mechanism

Zhou Xiangzhen[1,2], Li Shuai[2], Sui Dong[3]

(1.School of Information Engineering, Zhengzhou Shengda College of Economics and Trade Management, Zhengzhou 451191, China)
(2.School of Computer Science, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)
(3.School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102406, China)

Abstract: In order to improve the performance of Weibo's sentiment analysis, the recurrent neural network in the deep learning algorithm is used for sentiment classification, and the attention mechanism is used to select and weight the word features, so as to enhance the classification accuracy of the recurrent neural network. First, the Weibo corpus is denoised, segmented and vectorized to form an initial sample of Weibo. Then, the Weibo classification model of recurrent neural network is constructed, and the word feature vector is obtained through the node circulation of hidden layer and the output of hidden layer at historical moment and current moment. Then, the attention mechanism is used to calculate the similarity of word features and select weights to construct sentence features, and the classification results are obtained by Softmax function. Finally, the reliability of the proposed method is verified by the Weibo emotion classification simulation test. Experimental results show that, compared with the commonly used Weibo emotion classification algorithm, the proposed method shows higher classification performance under different word vector size samples by setting the attention mechanism window size reasonably.

Key words: microblog emotion, deep learning, recurrent neural network, attention mechanism

微博平台作为用户情感发布的重要平台之一,蕴含大量的可用数据资源,通过对微博数据资源的深度挖掘,既可以分析出热点微博上升的趋势,又可以分析出某个话题的用户的情感类别等[1].

微博情感分类作为微博数据管理的重要内容之一,是近年来的研究热点[2],通过对不规则程度高的海量微博数据进行清洗去噪、分词及向量化等处理,构建可用于深度挖掘的标准样本格式,然后通过分类方法和深度学习算法等对样本进行计算和训练,从而获得各样本的情感类别,最终实现微博的情感类别管

收稿日期:2023-01-03.
基金项目:国家自然科学青年基金项目(61702026)、河南省 2022 年度科技厅科技攻关项目(222102210290)、校级 2021 应用基础研究与应用专项项目(SD-ZDIAN2021-05).
通讯作者:隋栋,博士后,讲师,研究方向:人工智能与大数据. E-mail:619543699@qq.com

理. 当前,关于微博情感分类的研究较多. 朱亚军等[3]采用支持向量机(support vector machine,SVM)用于微博情感分类,其在语料规范性强和类别较少的条件下分类效果较好,但是应对不规则语料时分类准确率有待改进. 冯媛媛等[4]采用双向长短时记忆网络(bidirectional short-short memory network,Bi-LSTM)用于多情感类别的分类,并采用自注意力机制对样本特征进行筛选,取得了较高的分类准确率,但没有分析不同词特征规模和不同注意力窗口大小的影响.

与上述基于 SVM 或 Bi-LSTM 的情感分类方法不同,本文采用深度学习中的循环神经网络(recurrent neural network,RNN)算法作为分类器,能够实现复杂多维特征提取,并且能够适应大规模样本特征分析. 此外,本文还将注意力机制(attention mechanism,AM)引入到微博文本的词特征提取过程,从而提出了基于 RNN-AM 的微博情感分析方法. 这是因为注意力机制在多种类型的样本特征分析中优势明显,有助于提高对微博文本重点特征的分类训练准确度. 实验结果表明,本文所提方法有效提高了多类别微博文本的分类性能,且稳定性高.

1　注意力机制

注意力机制旨在对重点特征的深度挖掘,摒弃了非重点特征的无效训练[5]. 通过对重点特征的有效计算,满足对特征挖掘的需求,而滤出无效特征训练,降低运算复杂度. 注意力机制主要是在查询(Q)、关键字(K)和权重(V)三者之间进行运算,下面对注意力机制进行数学描述.

设系统 K 的总数为 L,先计算 Q 与所有 K 之间的相似程度 Similarity(Q,K_i) $(i=1,2,\cdots,L)$,将相似结果记为得分 Score(Q,K_i),其计算方法因为所选模型的差异有所不同[6],以下介绍 3 种模型的 Score(Q,K_i) 计算方法.

双线性:

$$\mathrm{Score}(Q,K_i)=K_iWQ, \tag{1}$$

点积:

$$\mathrm{Score}(Q,K_i)=K_iQ, \tag{2}$$

缩放点积:

$$\mathrm{Score}(Q,K_i)=\frac{K_iQ}{\sqrt{d}}, \tag{3}$$

式中,d 表示特征维度,W 表示线性变量.

相比于双线性和点积,缩放点积模型的计算复杂度有一定的增加,但是模型的分辨率更高,有利于提取对微博样本的词特征进行提取. 因此,本研究选择了缩放点积模型. 设 V_i 表示第 i 个 K 的权重值,其计算方式为:

$$V_i=\frac{\exp(\mathrm{Score}(Q,K_i))}{\sum_{j=1}^{L}\exp(\mathrm{Score}(Q,K_j))}. \tag{4}$$

根据 V_i 值,计算注意力机制结果:

$$\mathrm{Attention}(Q,K,V)=\sum_{i=1}^{L}V_iK_i. \tag{5}$$

2　本文所提方法

2.1　循环神经网络

RNN 相比于普通神经网络,主要差异体现在网络的输出 o 与历史输入和历史隐藏层输出有关,通过历史输入对当前输出的影响,较大限度地反映了历史数据特征对当前时间段的连续性影响. 设 x 和 o 分别为输入及隐藏层输出,核心结构如图 1.

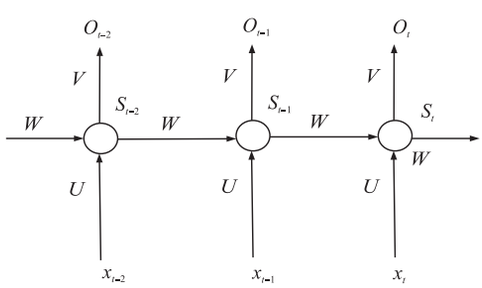

图 1　RNN 循环结构

Fig. 1　RNN cycle structure

U、V 和 W 均为权重,图 1 中 t 时刻的输出与 $t-1$ 和 $t-2$

均有关. 根据实际需要,还可以加大隐藏层循环规模,往更前的历史时刻进行扩展,这种时间叠加所带来的训练影响更能够保留训练样本的上下文信息,从而获得更精确的训练结果,这也正是 RNN 优于普通神经网络的原因[7].

设 n 个样本 $x_i(i=1,2,\cdots,n)$ 经过 RNN 的隐藏层后:

$$f(x)=w_1x_1+w_2x_2+\cdots+w_nx_n+b=\sum_{i=1}^{n}w_ix_i+b, \tag{6}$$

式中,w_i 和 b 分别表示所有样本点与隐藏层的连接权重系数及偏置.

将 $f(x)$ 输入至转换函数 $G(\cdot)$ 后得:

$$g(x)=G(\sum_{i=1}^{n}w_ix_i+b). \tag{7}$$

由图 1 知,RNN 的 t 时刻输出 s_t 与 s_{t-1} 和 x_t 有关,通过隐藏层激励 $f(\cdot)$ 得到[8]:

$$s_t=f(Ux_t+Ws_{t-1}+h_t), \tag{8}$$

式中,h_t 是 t 时刻激励的偏置.

$s(t)$ 通过 softmax 函数获得输出[9]:

$$o_t=\mathrm{softmax}(Vs_t+b_o), \tag{9}$$

式中,b_o 为输出偏置.

关于 RNN 的 U、V 和 W 求解,可以采用前向和反向迭代两种方式,而前者是 RNN 特有,其实现方法参照公式(10),后者是 NN 的普遍求解方式[10].

$$o_t=g(Vs_t)=Vf(Ux_t+Ws_{t-1})=Vf(Ux_t+Wf(Ux_{t-1}+Ws_{t-2}))=Vf(Ux_t+Wf(Ux_{t-1}+Wf(Ux_{t-2}+Ws_{t-3})))=$$
$$Vf(Ux_t+Wf(Ux_{t-1}+Wf(Ux_{t-2}+Wf(Ux_{t-3}+\cdots)))), \tag{10}$$

式(10)在迭代过程中滤除了常量 h_t 和 b_o,通过不断累积计算,根据样本的输入与输出值,则可以获得 U、V 和 W 值,从而确定稳定的 RNN 结构.

反向迭代主要通过不断缩小误差 δ_k 值来实现,δ_k 可以通过样本 k 的实际值 d_k 与 RNN 训练的结果 y_k 得到:

$$\delta_k=(d_k-y_k)y_k(1-y_k). \tag{11}$$

隐藏层节点 h_j 与 y_k 之间的权重更新为[11]:

$$\Delta w_{jk}(n)=\eta(\Delta w_{jk}(n-1)+1)\delta_kh_j, \tag{12}$$

式中,η 是学习率.

根据 $\Delta w_{jk}(n)$ 更新后,获得节点 h_j 与 y_k 之间的最新权重 $w_{jk}(n+1)$ 值:

$$w_{jk}(n+1)=w_{jk}(n)+\Delta w_{jk}(n). \tag{13}$$

隐藏层的偏置更新方法为:

$$\Delta b_k(n)=\alpha(\Delta b_k(n-1)+1)\delta_k, \tag{14}$$

式中,α 为偏置更新率.

根据 $\Delta b_k(n)$ 获得最新偏置 $b_k(n+1)$ 值:

$$b_k(n+1)=b_k(n)+\Delta b_k(n), \tag{15}$$

计算误差和 E 的值[12]:

$$E=\frac{1}{2}\sum_{k=1}^{M}(d_k-y_k)^2, \tag{16}$$

式中,M 表示输出节点总数.

由于公式(10)在计算时滤出了偏置的迭代,其主要计算都放在了 U、V 和 W 的求解上,相当于前向迭代仅进行了 RNN 部分参数的更新. 然而,反向迭代需将 RNN 的所有参数都进行更新,因此在求解模型参数的完整度方面,后者胜于前者,但反向求解的效率及复杂度明显高于前者,在实际使用时,按需求选择 RNN 参数求解方式. 由于本研究中微博情感分类的主要目的是提升分类准确率,因此采用了前向迭代方式.

2.2 基于 RNN-AM 的微博情感分析流程

在微博情感分析中,微博原始语料一般是由符号、图片、字母或汉字组成的句子,若要建立适合于 RNN 训练分析的数据样本,就需要对这些句子进行分词处理,而处理时,本文暂只对文字进行挖掘分析,

暂不考虑句子中的其他部分.

分词处理采用分词工具实现,接着采用 Word2vec 构建特征向量[13],从而为 RNN 训练提供可使用的数据样本.

在 RNN 的微博情感分析中,当训练步骤进行至式(8)后,并不直接进行 softmax 的分类,而是采用注意力机制对微博样本的词特征进行加权求和生成句子特征,其实现方式主要是通过式(4)和式(5),最后根据句子特征结果进行 softmax 运算获得情感分类结果.

3　实例仿真

为了验证 RNN-AM 的微博文本情感分类性能,选取了常用 4 类微博情感分类集进行不同维度的性能仿真,仿真集如表 1 所示. 首先,对不同词向量规模的样本进行 RNN-AM 微博情感分类仿真,验证本文算法应对不同规模微博文本量的分类性能;然后,差异化设置注意力机制窗口比例,验证不同注意力强度下的分类性能;接着分别采用 RNN 算法和 RNN-AM 进行情感分类仿真;最后,将 RNN-AM 算法与其他 3 种算法进行对比仿真.

3.1　不同词特征规模的微博情感分类性能

在 4 类样本集中,微博文本长度差异较大,而通过分词后其产生的词特征量数量差异明显,分别选择不同规模的词特征进行 RNN-AM 分类性能仿真,其中注意力机制采用全窗口模式,将所有词特征纳入加权求和.

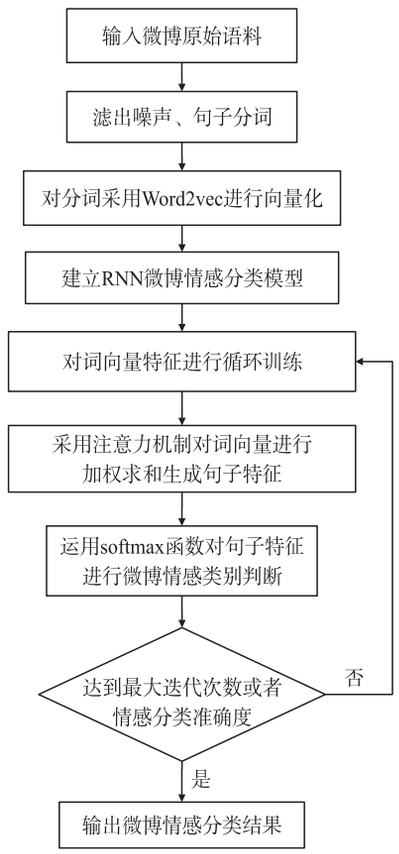

图 2　基于 RNN-AM 的微博情感分析流程

Fig. 2　Weibo's emotional analysis process based on RNN-AM

表 1　仿真集

Table 1　Simulation set

样本集	样本记录数	类别
MicroblogPCU	1308	3
NLPCC2013	1316	5
COAE2014	1401	5
SMP2014	1023	6

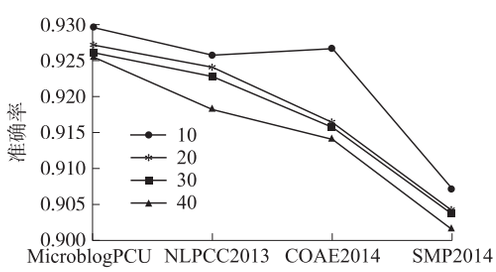

图 3　不同词向量规模下的 RRN-AM 分类准确率

Fig. 3　Accuracy of RNN-AM classification under different word vector scales

从图 3 得,对于相同的微博样本集,RNN-AM 算法在不同的词特征规模下的分类准确率差异较小,对比 4 类集在不同词特征规模下的分类准确率情况,词特征规模增大后,准确率略微有下降,这说明 RRN-AM 对不同规模微博数据的情感分类稳定性较高. 横向对比,RNN-AM 算法在 MicroblogPCU 的分类准确率最佳,范围约为[0.925, 0.93],而在分类准确率较低的 SMP2014 集也在 0.9 以上.

表 2　不同词特征规模的召回率和 *F*1 值

Table 2　Recall rate and *F*1 value of different word feature scales

数据集	词特征数量	召回率	*F*1 值	数据集	词特征数量	召回率	*F*1 值
MicroblogPCU	10	0.915 5	0.911 6	NLPCC2 013	10	0.911 2	0.901 2
	20	0.914 7	0.912 1		20	0.909 8	0.904 4
	30	0.914 4	0.915 3		30	0.907 5	0.092 9
	40	0.912 2	0.904 7		40	0.906 7	0.900 5
COAE2 014	10	0.905 9	0.899 6	SMP2 014	10	0.901 3	0.891 9
	20	0.905 3	0.902 7		20	0.900 2	0.898 1
	30	0.903 1	0.901 4		30	0.896 8	0.896 3
	40	0.901 2	0.896 7		40	0.882 1	0.873 3

对比4类集在不同词特征规模下的召回率和$F1$值,在相同样本集条件下,两者对词特征规模的变化并不敏感,这说明RNN-AM算法对4类集的情感分类适用度高,而在3个分类指标中均表现出MicroblogPCU集最优,而SMP2014集最差的情况,可能与样本类别数和本身的样本分类难度有关系.

3.2 不同注意力窗口大小的分类性能

采用注意力机制将词特征加权求和组建句子特征的过程中,注意力窗口大小决定了组建句子特征的词特征数量,根据组建词特征占总词向量的比例α,确定RNN-AM算法的分类准备率和分类时间.

表3 不同注意力窗口对应的分类性能

Table 3 Classification performance corresponding to different attention windows

数据集	α/%	准确率	分类时间/s	数据集	α/%	准确率	分类时间/s
MicroblogPCU	20	0.8372	1.337	NLPCC2013	20	0.8264	1.315
	40	0.8961	2.235		40	0.8897	2.469
	60	0.9043	3.482		60	0.9001	3.622
	80	0.9218	4.239		80	0.9198	4.831
	100	0.9297	6.538		100	0.9258	6.796
COAE2014	20	0.8276	1.403	SMP2014	20	0.8091	1.662
	40	0.8933	2.560		40	0.8324	2.724
	60	0.9024	3.588		60	0.8652	3.679
	80	0.9205	4.715		80	0.8995	4.912
	100	0.9267	6.593		100	0.9071	6.816

从表3知,选择不同的注意力窗口尺寸,RNN-AM算法的微博情感分类准确率和效率影响较大. 对于相同的样本集,在$\alpha=100\%$时,也就是所有的词特征均参与加权至句子向量时,采用RNN算法求解的分类准确率最高,α值越小,分类准确率越低,这也说明了微博情感分类的准确度与预料分词的完整性存在着密切联系,对比发现,当$\alpha=80\%$时,RNN-AM的分类准确率虽没有达到最高,但是和$\alpha=100\%$时所对应的准确率已非常接近. 而从分类时间来看,当α值越大,采用注意力机制被选择参与运算的词特征向量数越多,分类所需时间越长. 因此,为了平衡分类准确率和时间,在注意力窗口尺寸设置上应合理,后续仿真实验中选择均设置$\alpha=80\%$.

3.3 AM对RNN情感分类影响

为了验证AM对RNN微博情感分类的优化性能,分别采用RNN和RNN-AM算法进行微博情感分类仿真.

从图4可知,RNN-AM算法相比于RNN对于4类微博集的分类准确率提升明显,RNN-AN均在0.9以上,而RNN最高值仅为0.84,所以对分词进行AM策略后,其对RNN微博情感分类准确率提升效果明显.

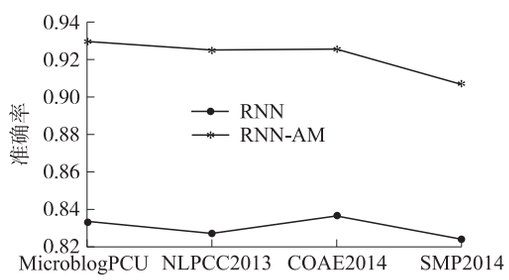

图4 RNN和RNN-AM算法的分类准确率

Fig. 4 Classification accuracy of RNN and RNN-AM algorithms

表4 RNN与RNN-AM的召回率及$F1$

Table 4 Recall rate of RNN and RNN-AM and $F1$

数据集	算法	召回率	$F1$值	数据集	算法	召回率	$F1$值
MicroblogPCU	RNN	0.8179	0.8042	NLPCC2013	RNN	0.8134	0.7953
	RNN-AM	0.9172	0.9023		RNN-AM	0.9125	0.9008
COAE2014	RNN	0.8155	0.8009	SMP2014	RNN	0.8166	0.7855
	RNN-AM	0.9143	0.9011		RNN-AM	0.9075	0.8812

由表4得,在召回率与$F1$性能方面,RNN-AM相比于RNN优势明显,这也说明AM策略对RNN的微博情感分类作用明显,这可能是因为微博文本较长且表述模糊,采用AM后更能够实现词特征到句子特征的精准拼接,从而为分类提供帮助.

3.4 不同算法分类性能

为了对比RNN-AM算法和常用微博情感分类算法的性能,分别采用SVM[14]、胶囊网络(capsule

network,CN)[15]、LSTM分类[16]和RNN-AM进行实例仿真,仿真结果如表5所示.

表5 4种算法的分类微博情感分类性能

Table 5 Classification performance of four algorithms for Weibo emotion classification

MicroblogPCU数据集				NLPCC2013数据集			
算法	准确率	召回率	$F1$	算法	准确率	召回率	$F1$
SVM	0.8471	0.8223	0.8165	SVM	0.8193	0.7966	0.7847
CN	0.9019	0.8973	0.8905	CN	0.8958	0.8873	0.8758
PSO-LSTM	0.9105	0.9053	0.8926	PSO-LSTM	0.8973	0.8892	0.8743
RNN-AM	0.9297	0.9172	0.9023	RNN-AM	0.9258	0.9125	0.9008
COAE2014数据集				SMP2014数据集			
算法	准确率	召回率	$F1$	算法	准确率	召回率	$F1$
SVM	0.8249	0.8046	0.7927	SVM	0.7925	0.7813	0.7655
CN	0.9020	0.8858	0.8795	CN	0.8873	0.8725	0.8439
PSO-LSTM	0.9093	0.8933	0.8864	PSO-LSTM	0.8902	0.8799	0.8466
RNN-AM	0.9267	0.9143	0.9011	RNN-AM	0.9071	0.9075	0.8812

从表5知,对于相同样本,4类微博情感分类算法的分类性能差异较大,其中RNN-AM的分类性能最好,PSO-LSTM算法较好,SVM最差,前3者都是深度学习算法在微博情感分类中的运用,后者未采用深度学习算法,这表明深度学习算法在微博情感分类的场景适用性更好,而RNN-AM对比PSO-LSTM和CN,采用了注意力机制进行词特征选择,其得到的分类效果更优,这也说明在深度学习分类训练中,采用注意力机制有主次选择性的词特征分类更能获得最佳情感分类性能.

4 结论

本文在RNN微博情感分类训练中,采用注意力机制用于微博词特征加权求和,然后构建句子特征,通过句子特征的分类实现微博情感分类. 这种通过注意力机制对重点词特征进行筛选后再进行分类的方法,有效提高了RNN的分类准确度. 后续研究将对RNN的参数求解进行优化,考虑采用仿生算法进行求解,以进一步提高RNN-AM算法在微博情感分类中的适用度.

[参考文献]

[1] 马晓慧,马尚才,闫俊伢,等. 基于距离感知的目标情感分类模型[J]. 南京师大学报(自然科学版),2021,44(4):111-116.

[2] 段吉东,刘双荣,马坤,等. 基于集成学习的文本情感分类方法[J]. 济南大学学报(自然科学版),2019,33(6):483-488.

[3] 朱亚军,次曲,拥措. 基于SVM算法的藏文微博情感分析研究[J]. 计算机仿真,2022,39(8):226-229.

[4] 冯媛媛,刘克剑,李伟豪. 基于BiLSTM+Self-Attention的多性格微博情感分类[J]. 西华大学学报(自然科学版),2022,41(1):67-76.

[5] NIU Z,ZHONG G,YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing,2021,452:48-62.

[6] GUO M H,XU T X,LIU J J,et al. Attention mechanisms in computer vision:a survey[J]. Computational visual media,2022,8(3):331-368.

[7] 张旭辉,张郴,李雅南,等. 城市旅游餐饮体验的注意力机制模型建构——基于机器学习的网络文本深度挖掘[J]. 南京师大学报(自然科学版),2022,45(1):32-39.

[8] LI J,JIN K,ZHOU D,et al. Attention mechanism-based CNN for facial expression recognition[J]. Neurocomputing,2020,411:340-350.

[9] WANG Y,WU H,ZHANG J,et al. Predrnn:a recurrent neural network for spatiotemporal predictive learning[J]. IEEE transactions on pattern analysis and machine intelligence,2022,45(2):2208-2225.

[10] HEWAMALAGE H,BERGMEIR C,BANDARA K. Recurrent neural networks for time series forecasting:current status and future directions[J]. International journal of forecasting,2021,37(1):388-427.

[11] YIN B,CORRADI F,BOHTÉ S M. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks[J]. Nature machine intelligence,2021,3(10):905-913.

[12] KISVARI A,LIN Z,LIU X. Wind power forecasting—a data-driven method along with gated recurrent neural network[J]. Renewable energy,2021,163:1895-1909.
[13] WEI X,ZHANG L,YANG H Q,et al. Machine learning for pore-water pressure time-series prediction:application of recurrent neural networks[J]. Geoscience frontiers,2021,12(1):453-467.
[14] 邓君,孙绍丹,王阮,等. 基于 Word2Vec 和 SVM 的微博舆情情感演化分析[J]. 情报理论与实践,2020,43(8):112-119.
[15] 吴仁彪,乔晗,贾云飞,等. 基于胶囊网络的中长微博情感分析[J]. 信号处理,2022,38(8):1632-1641.
[16] 林伟. 基于 PSO-LSTM 的中文微博情感分类研究[J]. 中国人民公安大学学报(自然科学版),2022,28(1):95-101.

[责任编辑:陆炳新]