基于深度学习的评论文本情感分析

顾昕健[1] 陈 涛[1]
GU Xinjian CHEN Tao

摘 要 在互联网高度发达、高度普及的今天，大量用户更倾向于在互联网平台上发表自己的意见与评论。如何更快、更准确地分析海量评论文本中所蕴含的情感倾向，已是自然语言处理领域的热点问题。针对该问题，提出一种基于改进 TextCNN 网络的 STCNN，使用 Mish 函数替代了原始 TextCNN 模型中的 ReLU 激活函数，规避了负输入的梯度损失问题，另在池化层混合了最大池化层和平均池化层的输出，弥补了上下文信息提取不充分的不足，并融入了注意力机制。基于 RoBERTa 和 STCNN 模型对微博疫情评论文本进行情感分析，利用 RoBERTa 模型提取文本特征向量，输入 STCNN 中获得更丰富的语义信息。实验结果表明，所提出的 RB-STCNN 模型在评论集情感分类中的准确率、F_1 评分等指标均有良好结果，较优于对比实验中的其他模型。

关键词 情感分析；注意力机制；激活函数；RoBERTa；TextCNN

doi：10.3969/j.issn.1672-9528.2024.07.008

0 引言

如今，互联网发展势头迅猛，在人们的日常生活中愈发重要。由于在互联网发布和获取信息的成本低、速度快，信息交流模式逐渐由互联网占据主导地位。2024 年 3 月 22 日，中国互联网络信息中心（CNNIC）发布了第 53 次《中国互联网络发展状况统计报告》[1]。报告显示，截至 2023 年 12 月，中国境内的互联网用户数量已攀升至 10.92 亿人，相较于前一年同期用户基数净增 2480 万人。以此计算，全国互联网普及率达到了 77.5%，相较于 2022 年同月，实现了 1.1 个百分点的提升。

在此基础上，人们更倾向于在互联网平台发表和交流自己的想法和意见。以微博为例，作为一款以用户关系为核心的社交媒体平台，它集成了信息多样性、操作简易性、裂变式传播效率以及高效的用户交互等特性 [2]。截至 2023 年第四季度末，其月活跃用户数达到了 5.98 亿人次，相较于上年同期实现净增长约 1100 万人。在如此大的用户基数下，所产生的评论文本数据更是浩如烟海。这些评论文本有着短小精炼、数量庞大、情感蕴含丰富等特点，包含了很多值得挖掘的信息。对这些用户生成的数据进行自动分析，可以有效监控公共舆论并帮助决策，这也是近年来情感分析在研究界更受欢迎的原因之一 [3]。通过情感分析算法挖掘出海量文本中所蕴含的情感极性，这在舆情监控、虚假识别、产品推荐、决策预测等多个应用场景中有着举足轻重的意义。

1. 南京审计大学计算机学院 江苏南京 211800

1 研究背景

情感分析，也称为意见挖掘，旨在研究人们对某些实体的情绪 [4]。简单来说，它是对带有情绪倾向的文本进行自动分析处理，总结规律，得出结论，其本质上是一种文本分类。文本分类一直受到学术界的广泛关注，其关键技术也在不断演变中 [5]。从最初的基于规则的文本分类（如情感分析领域中基于情感词典的分析方法），到传统的基于机器学习的方法，再到近年来流行起来的基于深度学习的方法，之后随着 Transformer[6] 模型的兴盛，又出现了一系列预训练的大规模模型。相比于较为古早的机器学习等算法，深度学习模型缩减了人工设立特征这一工作量庞杂的步骤，自动挖掘文本特征，且准确率往往优于传统机器学习模型，而其中预训练大模型则在海量数据中学习到了通用的语言特征，针对跨领域、数据量小等特殊情况有着一马当先的性能表现。

2003 年，Nasukawa[7] 首次提出不再简单地将文本分为正负极，而是从文档中提取积极和消极的情绪。自此以后，情感分析的研究受到了学术界的广泛关注，其关键技术也随着文本分类技术的演变而不断迭代更新。在其发展过程中，主要经过了三个阶段，即基于规则、机器学习、深度学习三种方案。

基于规则的情感分析方法是一种传统的方案，其原理是基于文本分类中的词匹配方法 [8]，具体到情感分析领域，便是使用人为编写的情感词典与文本比对，从而判定其情感倾向。比如“快乐”属于积极极性，“悲伤”属于消极极性。情感词典配合不同领域的专业知识词典，可以在一定程度上

判定文本的情感分类，但是同时它也存在着明显的弊端。首先，在构建情感词典的过程中需要用到大量的人力物力，情感词典的质量高低直接影响了分类的准确率。再者，在不同领域中，情感词典的复用率很低，不同的领域对应着不同的专业名词，对于新的领域便需要构建新的情感词典。最后，情感词典无法对文本的上下文信息进行识别，同一个词语在不同的语境下会有不同的情感倾向，如果仅仅是按照其期望情感分数计算，或者对不同情况的情感分数进行加权操作，那么情感分析的结果必然会出现不尽如人意的情况。

于是，人们尝试利用计算机模拟人类学习的过程，对大量的同一类别文本数据进行特征学习，从而得到情感分类的结果，这一方法被称为基于机器学习的情感分析方法。它通常分为三个步骤：文本的预处理、特征提取、分类计算。特征提取方法比较常见的有词袋模型、TF-IDF 等，而分类计算较为常用的则包括朴素贝叶斯、K 近邻算法、支持向量机等。Pang 等人 [9] 在 2002 年使用了基于朴素贝叶斯的算法，对电影评论数据做了情感分类工作，各项指标证明，其分类效果相较基于情感词典的情感分析方法，要更加优秀，这被认为是早期将机器学习应用在情感分析之中的案例。Asep 等人 [10] 同时利用支持向量机和朴素贝叶斯等方法，对有关新旧金融技术的公共舆情进行分辨，其中支持向量机方法优于朴素贝叶斯方法 5.7 个百分点，证明了新的金融技术更加受公众欢迎。然而，基于机器学习的情感分析仍然不够关注文本数据的上下文信息。与此同时，其特征构造过程也成本高昂，耗时耗力。利用机器学习所构造的模型，对于不同领域的泛化能力也偏弱。因此，一些研究人员开始对神经网络展开深入的研究。

对比传统的机器学习方法，基于深度学习的情感分析不再依赖工程量浩大的手工提取特征工程，而是直接从原始数据中自动学习情感特征，大大降低了资源消耗成本。基础的神经网络通常指前馈神经网络、循环神经网络和卷积神经网络。其中，循环神经网络在应对提取文本上下文信息方面表现优异，代表有长短期记忆网络、门控循环记忆网络等。而卷积神经网络更擅长提取文本的局部特征，主要包括文本卷积神经网络、深度卷积神经网络等。此后，研究人员在基础网络的基础上，发展出了混合神经网络、胶囊神经网络、图神经网络和预训练模型等各种架构，成果丰富而又杂乱。近年来，大数据预训练模型更是成为研究热点，基于 Tranformer 架构的 Bert 系列、Gpt 系列都拥有着强大的性能。许多研究人员利用深度学习在情感分类获得显著成效，Kim 等人 [11] 较早提出了将卷积神经网络应用于情感分析领域中，利用卷积核提取文本局部特征，并在多个数据集上获得了良好的成果。Chatterjee 等人 [12] 提出利用双向长短期记忆网络（BiLstm）来评估能源管理方法的有效性，并取得一定的成效。林伟等人 [13] 则结合了预训练模型 Bert 与双向门控网络、卷积网络三者，利用循环网络捕捉上下文信息，卷积网络捕捉局部特征，在中文微博数据集上有着较好的表现。兰正寅等人 [14] 采取更为先进的 RoBERTa 预训练模型，与融合了注意力机制的 Bilstm 模型结合，建立了针对新闻分类的深度学习模型，在今日头条等新闻数据集中的实验表明，该模型的准确率高于一般模型。

深度学习模型消除了传统机器学习提取文本特征工作量需求大的弊端，对于上下文信息的挖掘更加透彻。部分学者结合了不同的深度学习神经网络的优势、融合注意力机制、预训练模型等等，显著提高了模型的性能和泛化能力。即使基于深度学习的模型有时需要更大数量的训练集提升模型效果，对计算资源需求度也更高，但它仍不可改变地成为当前情感分析领域的一个研究趋势。

2 基于 RB-STCNN 模型的评论文本情感分类

本文基于 RoBERTa 和 STCNN 模型构建复合神经网络 RB-STCNN 模型用于情感分类，模型主要分为文本预处理模块、预训练模型 RoBERTa 模块和 STCNN 特征学习模块。模型结构如图 1 所示。

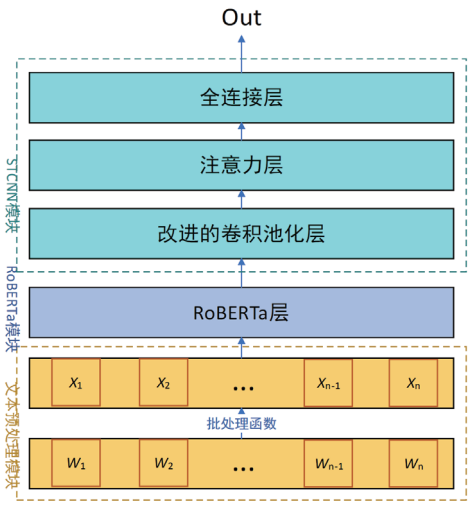

图 1 模型结构

2.1 文本预处理模块

本文采取了 Kaggle 网站竞赛的微博评论文本数据集，针对该数据集，分别进行了数据获取、数据清洗两个步骤。

（1）数据获取。数据集包含微博 ID、发布时间、评论内容、情感倾向等多维数据，提取数据的评论列数据和对应列的情感倾向数据，分别命名为 text、label。然后根据 6:2:2 的比例对处理好的数据集进行分割，分别得到训练集、验证集和测试集。

（2）数据清洗。数据集来自微博评论，这些数据是由大量的用户随意生成的，含有大量噪声。本文对数据集进行数据清洗。检测并删除异常数据，如 label 值异常。利用正则等方法检测空数据、重复数据、特殊符号等数据噪声并删除。经过数据清洗，去除无用噪声，尽可能地保证了数据集的可靠性，方便下一步建模分析。

2.2 预训练模型 RoBERTa 模块

RoBERTa 预训练模型的本质是一种对于 Bert 模型的优化方法，是一种优化的高性能 Bert 预训练模型。类似于 Bert，RoBERTa 同样是基于 Transformer 架构，由多层双向编码器堆叠而成，每一层都包含了多头注意力机制和前馈神经网络层，且含有残差连接与层归一化。这种结构可以更好地捕捉文本中的长距离依赖关系，更深层次地学习文本特征，动态获取文本蕴含的语义信息。其结构如图 2 所示。

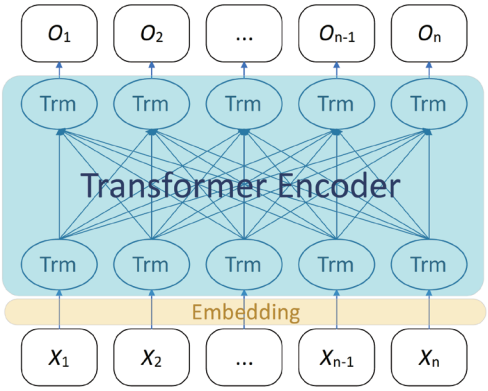

图 2 RoBERTa 结构图

区别于原生 Bert 预训练模型，RoBERTa 模型在一些方面做了改进。首先，RoBERTa 采取了动态掩码的方式，即随机遮蔽句子中的部分词汇，按照上下文预测词汇内容，并且只有在训练时，会动态地将句子中的随机词汇替换成<Mask>，规避了原生 Bert 无意义的复制所导致的内存占用问题。其次，RoBERTa 取消了 Bert 模型中原有的 NSP 任务，即 Next Sentence Prediction（下句预测任务）。下句预测任务的作用一直饱受争议，原生 Bert 模型做此任务的目的是让模型能够学习到语言所含有的连贯性。RoBERTa 为了研究 NSP 任务的必要性，进行了多种实验，证明了取消 NSP 任务对于下游任务有着积极影响。最后，RoBERTa 使用了更大的训练数据，并采用了更大的训练批次，延长了训练时间。增加训练数据无疑是对于提升模型鲁棒性最有效的方法之一，实验证明，大量的语料库显著提升了模型的泛化能力和对语句的理解能力。

本文根据使用的数据集形状，设置了合适的句长 200，利用 RoBERTa 提供的标准将原始文字分词并编码，获取其句子嵌入向量、位置嵌入向量、计算注意力时需要忽略的位置等。将处理后的文本序列输入 RoBERTa 预训练模型中，提取高级语义特征。由于 RoBERTa 最后一层隐藏状态的默认大小为 768，所以最后获取到批次大小数量的 [200,768] 形状的三维张量。将这部分语义特征再进一步送到 STCNN 模块进行特征学习。

2.3 STCNN 特征学习模块

卷积神经网络（CNN）最初是应用在图像领域中，通过对图像进行卷积操作，可以有效提取图像中的局部特征，从而达到监测、识别等应用效果。而 TextCNN 是将卷积网络应用在文本信息中的早期尝试，它利用不同大小的卷积核在经过处理的文本序列上滑动来提取局部的 N-gram 特征。传统的 TextCNN 方法使用 ReLU 作为激活函数，存在负输入坍缩为 0 的梯度损失问题和近零输入梯度变化不平缓的梯度消失问题。且传统方法在池化层选择最大池化操作，这一方法在提取图像特征时确实可以有效提取到图像的特征信息，抑制背景信息，但是应用在自然语言处理领域中时则会丢失部分上下文信息，某些情况下是不合适的。因此，本文对于传统的文本卷积网络做出改进，使用 Mish 函数作为激活函数，在池化层选择混合最大池化和平均池化的输出，之后融入注意力机制，加强捕捉远距离依赖的能力。改进后的模型包括输入层、卷积层、池化层、注意力层和输出层。模型结构如图 3 所示。

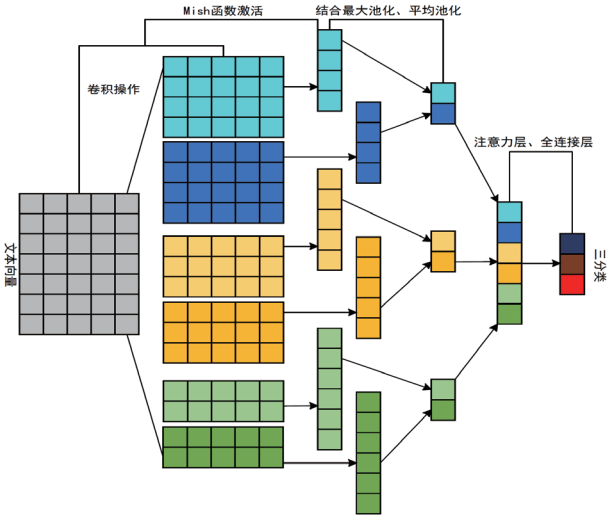

图 3 STCNN 模型结构

（1）输入层

文本经过 RoBERTa 预训练模型处理得到每一句的向量表示，获得 $S\times N\times M$ 的三维张量，其中 S 为批次大小，N 为最大句长，M 为隐藏层特征维度。由于输入通道为 1，对该张量进行维度扩展，获得 $S\times 1\times N\times M$ 的四维张量，输入 STCNN 模型中。

（2）卷积层

卷积层使用不同宽度的卷积核在词向量上移动，不同宽度对应着不同的 N-gram 窗口，通过对每个窗口上的词向量进行卷积操作，提取该窗口的语义特征。本文为了充分提取文本情感信息，设置了 2、3、4 三个大小的卷积核。对于卷积结果，传统 TextCNN 采取 ReLU 作为激活函数，其计算公式为：

$$f(x)=\max(0,x) \tag{1}$$

这种激活函数虽然简单有效，但是它在 $x<0$ 时梯度为 0，负的梯度会直接置 0，导致特征屏蔽太多，减少了模型的有效容量，因此本文优化了激活函数，将传统的 ReLU 激活函

数替换为 Mish 函数，其计算公式为：

$$f(x)=x\tanh\left(\ln\left(1+e^{x}\right)\right) \tag{2}$$

Mish 函数在所有点都是连续且可微的，它规避了在 $x<0$ 时的梯度损失问题，并且在 x 接近 0 时梯度变化也更为平缓，这有助于提升模型的稳定性。它还拥有非单调的特性，意味着在模型中可以捕捉更加复杂的语义特征。因此，它的表现往往比 ReLU 要更好。

（3）池化层

池化层的目的是将卷积所得到的特征进行压缩提取，获得具有代表性的特征。传统的 TextCNN 一般采用最大池化策略，可以捕捉文本中的局部特征，丢弃弱特征。这一策略在图像领域显然是非常有效的，可以忽略图像的背景特征，捕捉到关键特征。然而在自然语言处理领域，这一策略会丢失部分上下文信息。因此，本文采用混合最大池化和平均池化的策略，将两者的输出加权融合，既关注了文本的显著特征，又结合了对全局信息的考虑，从而更好地保留语义特征。

（4）注意力层

相较于传统的 TextCNN，本文在卷积层后添加了注意力层，将卷积输出 o_i 输入到注意力层，得到注意力分数 l_i，借助 softmax 函数对注意力分数进行归一化得到需要的注意力权重 α_i，结合权重对输出 o_i 进行加权求和，得到结合注意力机制的输出 t，使模型能够动态地强调对分类任务贡献较大的特征，抑制无关或噪声特征，其公式为：

$$l_i=\tanh(W_a o_i+b) \tag{3}$$

$$\alpha_i=\frac{e^{l_i}}{\sum_{i=1}^{j}e^{l_i}} \tag{4}$$

$$t=\sum_{i=1}^{j}\alpha_i o_i \tag{5}$$

式中：W_a 是系数矩阵，b 是偏置项。

（5）输出层

在输出层，本文对注意力层输出的特征向量应用 Dropout 正则化以防止过拟合，然后输入到全连接层得到最终的向量输出，维度为三维，对应着正极、负极和中性情感三个分类。

3 实验分析

3.1 数据集

实验使用的数据集是 Kaggle 官网的竞赛疫情期间网民情绪识别所提供的 Weibo nCoV Data。本文从中抽取部分微博评论及其对应情感倾向数据，按照 6:2:2 的比例裁分成训练集、验证集和测试集。本文对所使用数据集做了预处理工作，提高了该数据集的可靠性。

3.2 实验环境配置

本文所使用的环境配置如表 1 所示。

表 1 实验环境参数

实验环境	配置信息
操作系统	Windows 11
内存	16 GB
中央处理器型号	Intel i7-12700H
显卡型号	GeForce RTX 3060
开发语言	Python 3.10
开发框架	PyTorch 2.2.2
开发环境	PyCharm

3.3 模型参数与评价指标

本文所使用的模型参数如表 2 所示。

表 2 模型参数设置

参数名称	参数值
预训练模型	RoBERTa_Chinese
批处理大小	128
迭代轮次	15
随机失活比例	0.4
学习率	0.001
卷积核尺寸	[2,3,4]
卷积核数量	128

本文引入混淆矩阵，用以计算模型分类结果的准确率（acc），每一类的精准率（pr）、召回率（re）和 F_1 分数（F_1），其中精准率、召回率和 F_1 分数在各个类别分别计算后取加权平均。这些指标可以衡量模型在不同任务上的性能和泛化能力。相关公式如下：

$$\text{acc}=\frac{T_P+T_N}{T_P+F_P+T_N+F_N}\times 100\% \tag{6}$$

$$\text{pr}=\frac{T_P}{T_P+F_P}\times 100\% \tag{7}$$

$$\text{re}=\frac{T_P}{T_P+F_N}\times 100\% \tag{8}$$

$$F_1=\frac{2\text{pr}\cdot\text{re}}{\text{pr}+\text{re}}\times 100\% \tag{9}$$

式中：T_P 为正类别中被正确预测的样本数量，T_N 为负类别中被正确预测的样本数量，F_P 为正类别中被错误预测的样本数量，F_N 为负类别中被错误预测的样本数量。

3.4 实验结果与分析

为了验证 RB-STCNN 模型的分类效果，本文分别引入 BiLstm、TextCNN、RB-BiLstm、RB-TextCNN、RB-BiLstm-Att、RB-TextCNN-Att 与本文的模型进行比对。实验环境与模型参数保持不变，评价指标采用准确率、精确率、召回率和 F_1 分数。实验结果如表 3 所示。其中，RB 代表与预训练模型 RoBERTa 结合，A 代表融入了注意力机制，BL 代表 BiLstm 模型，TC 代表 TextCNN 模型。由实验结果可以看出，融合预训练模型 RoBERTa 会具有较高的提升，可见该模块是显著有效的模块。而融合注意力机制，在 F_1 分数等指标方面会有一定提升，其他指标仅有些微提升或者持平。综合来看，注意力机制也是较为有效的模块。

表 3 实验结果

分类模型	准确率 /%	F_1 分数 /%	精确率 /%	召回率 /%
BL	62.52	57.69	61.73	62.52
TC	64.62	60.91	63.32	64.62
RB-BL	70.47	70.25	70.79	70.47
RB-TC	71.93	70.41	73.24	71.93
RB-BL-A	70.18	70.52	71.51	70.18
RB-TC-A	72.47	72.07	72.51	72.47
OURS	73.15	72.87	73.90	73.15

而本文的改进模型 RB-STCNN，在各指标上都优于未改进的 RB-TC 模型，也优于仅融合了注意力机制的 RB-TC-A 模型。可见，改良了 TextCNN 的激活函数与池化层策略，对词向量的处理能力有着一定程度的提高，同时增强了模型提取上下文特征的能力，性能提升显著。

本文模型相对于对比实验中的其他模型，各项指标都有明显提升，其结果在数据集所对应的比赛排行榜上的排名也位于第一梯队，距离第一名的分数仅仅相差 1.2 个百分点左右。可见，本文提出的模型 RB-STCNN 的性能优异，效果良好。

4 结论

本文分析了情感分析的研究现状，并针对评论文本集数量庞大、情感蕴含丰富的问题，构建了 RB-STCNN 的复合情感分析模型。该模型结合了 RoBERTa 预训练模型和改进后的 STCNN 网络结构。RB-STCNN 模型通过 RoBERTa 模块捕获文本的高级语义特征，克服了传统特征提取方法的局限性，确保模型能够理解复杂的文本情境和潜在情感含义。而在 STCNN 特征学习模块中，通过引入 Mish 激活函数解决了 ReLU 存在的梯度消失问题，同时运用混合最大池化和平均池化策略兼顾了局部显著特征和全局上下文信息的提取。此外，模型还在卷积层后添加了注意力机制，使得模型能够根据任务需求动态突出重要特征，从而提升了情感分类的准确性。实验结果表明，RB-STCNN 模型在微博疫情评论文本情感分析任务上表现出色，各项评价指标如准确率、F_1 分数、精确率和召回率均优于对比实验中的其他变体模型。不过，本文更偏向将卷积神经网络与预训练模型结合，从而提升情感分类效果。未来可以通过融合循环神经网络，进一步提升模型对上下文信息的提取能力，以期在情感分析任务上实现更高的性能突破。

参考文献：

[1] 第 53 次《中国互联网络发展状况统计报告》发布 [J]. 新闻论坛 ,2024,38(2):17.

[2] 刘桂海 , 崔福龙 , 卢彩菡 , 等 . 公众对假房源的关注点和态度 : 基于微博评论的文本挖掘研究 [J]. 管理评论 , 2023, 35(11): 153-165.

[3]WANKHADE M, ANNAVARAPU C S, KULKARNI C, et al.A survey on sentiment analysis methods,applications,and challenges[J].Artificial intelligence review,2022,55(7):1-50.

[4]FANG X, ZHAN J.Sentiment analysis using product review data[J].Journal of big data,2015,2:5.

[5] 刘晓明 , 李丞正旭 , 吴少聪 , 等 . 文本分类算法及其应用场景研究综述 [J/OL]. 计算机学报 :1-44[2024-03-07].http://kns.cnki.net/kcms/detail/11.1826.TP.20240229.1608.002.html.

[6]VASWANI A, SHAZEER N, PARMAR N, et al.Attention is all you need[EB/OL].(2017-06-12)[2024-03-12].https://doi.org/10.48550/arXiv.1706.03762.

[7]NASUKAWA T, YI J.Sentiment analysis:capturing favorability using natural language processing[C]//Proceedings of the Second International Conference on Knowledge Capture,October 23-26,2003,Florida,USA.New York:The Association for Computing Machinery Inc., 2003:70-77.

[8]APTÉ C, DAMERAU F, WEISS S M.Automated learning of decision rules for text categorization[J].ACM transactions on information systems,1994,12(3):233-251.

[9]PANG B, LEE L, VAITHYANATHAN S.Thumbs up? Sentiment classification using machine learning techniques[C]//2002 Conference On Empirical Methods In Natural Language Processing.New Brunswick,NJ:Association for Computational Linguistics,2002:79-86.

[10]ASEP T N, BENNY M A, WIDYA S, et al.Sentiment analysis of user preference for old vs new fintech technology using SVM and NB algorithms[J].Management systems in production engineering,2023,31(4):373-380.

[11]KIM Y.Convolutional neural networks for sentence classification[C]//Conference on Empirical Methods in Natural Language Processing,vol.3.Stroudsburg,PA:Association for Computational Linguistics,2014:1746-1751.

[12]CHATTERJEE D, BISWAS K P, SAIN C, et al.Bi-LSTM predictive control-based efficient energy management system for a fuel cell hybrid electric vehicle[J].Sustainable energy,grids and networks,2024,38:101348.

[13] 林伟 , 陈雁 . 融合 BERT-BiGRU 和多尺度 CNN 的中文微博情感分析 [J]. 中国电子科学研究院学报 ,2023,18(10):939-945.

[14] 兰正寅 , 周艳玲 , 张龑 , 等 . 基于 RoBERTa-ATTLSTM 新闻分类方法研究 [J]. 计算机与数字工程 ,2023,51(11):2620-2626.

【作者简介】

顾昕健（1999—），男，江苏南京人，硕士，研究方向：深度学习。

陈涛（1970—），男，浙江淳安人，博士，研究方向：人工智能审计、大数据分析、模式识别等。

（收稿日期：2024-04-30）