

DOI:10.13296/j.1001-1609.hva.2024.06.024

基于双分支特征融合的电力设备缺陷文本挖掘方法

张中文¹, 吐松江·卡日¹, 张紫薇², 崔传世¹, 邵 罗¹

(1. 新疆大学电气工程学院, 乌鲁木齐 830049; 2. 清华四川能源互联网研究院, 成都 610299)

摘要: 针对电力设备缺陷文本信息知识挖掘与分析任务中存在缺陷文本特征信息提取不足、缺陷文本分类精度不够的问题,提出一种基于BERT(bidirectional encoder representations from transformers)的双分支特征融合的电力设备缺陷文本分类模型。首先,对缺陷文本数据进行预处理,删除异常缺陷文本,并归纳了电力设备缺陷文本特点;然后,采用BERT模型作为文本编码器,将文本转化为向量后分别输入至BiLSTM-Attention(attention-based bidirectional long short-term memory)模块和多分支CNN(multi-scale convolutional neural network, MCNN)模块,提取缺陷文本语义信息特征和局部关键信息特征;最后,将所提取出的语义特征和多维关键特征向量进行融合,并通过Softmax层实现对缺陷文本分类。与基准模型BERT-BiLSTM-Attention相比,其准确率、召回率及 F_1 值分别提高了2.76%、3.58%和4.39%,表明所建模型在缺陷文本分类任务中性能的优越性。

关键词: 预训练模型; 多维特征提取; 语义信息特征; 缺陷文本分类

Text Mining Method for Power Equipment Defects Based on Two-branch Feature Fusion

ZHANG Zhongwen¹, TUSONGJIANG Kari¹, ZHANG Ziwei², CUI Chuanshi¹, SHAO Luo¹

(1. School of Electrical Engineering, Xinjiang University, Urumqi 830049, China; 2. Sichuan Energy Internet Research Institute, Chengdu 610299, China)

Abstract: Aiming at the problem of insufficient extraction of defective text feature information and insufficient accuracy of defective text classification in the task of knowledge mining and analysis of defective text information of electric power equipment, a kind of two-branch feature fusion based on BERT (bidirectional encoder representations from transformers) is proposed as a text classification model for electric power equipment defective texts. Firstly, the defective text data is preprocessed to delete abnormal defective text and summarize the characteristics of defective text of power equipment. Then, the BERT model is adopted as a text encoder, and the text is transformed into vectors and input into BiLSTM-Attention (attention-based bidirectional long short-term memory) module and multi-branch CNN (multi-scale convolutional neural network, MCNN) module to extract the semantic information features and key information features of the defective text; Finally, the extracted semantic features and multi-dimensional key feature vectors are fused and the Softmax layer is used to achieve the defective text classification of the defective text. Compared with the benchmark model BERT-BiLSTM-Attention, its accuracy, recall, and F_1 value are improved by 2.76%, 3.58%, and 4.39%, showing the superior performance of the proposed model in defective text categorization task.

Key words: pre-trained models; multi-dimensional feature extraction; semantic information feature; defective text categorization

收稿日期:2023-11-14; 修回日期:2024-01-25

基金项目:新疆维吾尔自治区自然科学基金面上项目资助(2022D01C35);国家自然科学基金项目资助(52067021,52207165)。

Project Supported by Natural Science Foundation of Xinjiang(2022D01C35), National Natural Science Foundation of China (52067021,52207165).

0 引言

电力公司储存的大量电气设备运维数据蕴含着丰富的设备健康信息,其中缺陷文本的重要性尤为突出^[1-2]。目前,缺陷文本的分类工作主要由人工完成,然而这种方式既效率低下,又容易受限于运检人员知识储备和经验水平,导致对某些等级不明确缺陷文本常出现分类错误^[3]。随着人工智能技术在电气领域的广泛应用,运用文本挖掘技术对电力设备缺陷文本进行分析愈发重要。该技术不仅能减少人为主观因素干扰,提高分类准确性,还能减少人工处理缺陷文本的工作量,进一步节约人力成本。

针对电力缺陷文本挖掘的研究中,文[4]采用基于one-hot编码的K近邻(K-nearest neighbor, KNN)文本分类模型,通过引入历史缺陷信息的方法,运用KNN模型对断路器缺陷文本进行挖掘,验证了断路器状态评估结果合理性。但当利用one-hot方法对缺陷文本进行编码时,由于缺陷文本稀疏性,导致编码时出现维度灾难和缺乏语义关联的问题。文[5]使用隐马尔可夫法(hidden Markov model, HMM)对告警信号文本进行编码,并采用支持向量机(support vector machine, SVM)对告警信号文本进行分类,可帮助调度员在短时间内做出正确决策,但HMM利用已经构建好的词库进行编码,无法处理专业性较强的电气缺陷文本语料。文[6]考虑了电力设备缺陷文本特点,提出了基于CNN的缺陷文本分类模型,从分类准确率和效率方面充分地说明了模型处理电力缺陷文本的有效性,但该方法只能提取缺陷文本局部信息,不适用于长文本深层语义特征的提取。文[7]利用BiLSTM模型提取缺陷文本深层语义,抽取了故障问题和原因形成的因果关系,为下一步缺陷文本挖掘奠定了基础。文[8]在BiLSTM的基础上提出了基于注意力机制的BiLSTM缺陷文本分类方法,构建了电力设备缺陷等级优先处理指数,为设备运维提供了可靠指导。上述文献利用BiLSTM模型有效解决了长文本序列的特征记忆问题,但因电力缺陷文本具有稀疏性的特点,导致该模型不能完整地提取缺陷文本语义信息^[9]。

为更好地利用缺陷文本背景信息,文本研究的重点已转向预训练语言模型^[10]。此类模型能够在大规模的电气缺陷文本语料库上进行预训练,学习到专业的语言知识和语义表示能力。这使得预训练模型能够捕捉专业文本中的复杂特征和领域相关的语义信息,显著提升了模型对专业文本的处理能力。Google于2018年提出了基于双向Transformer

编码表示预训练语言模型,并被快速应用于电气领域^[11-12]。文[13]提出基于BERT模型对电网现场作业风险自动评级方法,利用作业风险等级表和文本库电气文本对BERT模型进行训练,从而生成具有电力特色的分类模型,该模型对电网作业风险自动评级具有较高准确性。文[14]针对电气文本特点,利用改进后的BERT模型对电气文本进行分词,提高了电气领域分词质量,并验证了优化后BERT模型性能。上述文献运用预训练BERT模型对电力文本进行挖掘,提取文本语义特征信息,实现了电力文本分类任务,但存在文本语义信息提取不充分的问题。

针对上述文献中模型存在的不足,文中提出了一种基于BERT的BiLSTM-Attention和MCNN的双分支特征融合电力设备缺陷文本分类模型(BERT-BiLSTM-Attention-MCNN, BBAMC)。首先,对缺陷文本进行预处理,筛选出完整数据,并根据缺陷文本特点分析了提取特征信息的难点,为后续特征提取提供方向。然后,在BERT对缺陷文本进行预训练的基础上,结合BiLSTM-Attention模块能提取缺陷文本上下文语义特征信息的优势和MCNN能够提取局部重点信息的优势,融合构建了BBAMC模型。最后,实验结果表明,文中模型从分类的精确率、召回率和 F_1 值上都超过了对比的基准模型,证明了BBAMC模型的有效性与优越性。

1 电气设备缺陷文本

原始电力设备缺陷文本含有大量噪声,导致提取文本特征变得困难,影响缺陷文本分类模型准确率。因此,需对缺陷文本进行预处理,以获取有效且可用的数据。此外,电力缺陷文本也具有较强的专业性,导致分类模型理解电力知识存在困难,影响模型分类效果。因此,需要考虑电力缺陷文本的特点,以提高模型分类准确率。

1.1 缺陷文本预处理

电力设备缺陷文本是非结构化数据,计算机不能直接对文本进行处理,需用相应的向量特征对文本进行表示,而文本特征提取的好坏决定着模型分类效果。在文本预处理过程中,首先,删除不完整和异常数据;然后,对设备缺陷文本进行数据清洗,包括分词和去除停用词;最后,将不同文本描述相同设备的文本进行规范化,便于电气缺陷文本分类。

1.2 电气设备缺陷文本的内容

为了解电力系统设备的运行状态,需要对设备运行状态进行监测与记录^[15]。缺陷信息是评价缺陷等级的主要依据。缺陷记录的内容通常包括缺陷

发现日期、缺陷设备名称、缺陷内容等,从而形成大量的电力缺陷文本。缺陷文本涵盖了设备缺陷的具体部件、缺陷原因、缺陷现象以及其带来的影响等重要要素。这些文本蕴含着设备运行状态的健康状况和设备运行规律。借助缺陷文本内容,能够快速、准确地判定设备缺陷等级,为巡检人员提供辅助性的参考^[16]。

1.3 电气设备缺陷文本的特点

电气设备缺陷文本与普通的文本相比,具有以下的特点:

1)电力缺陷文本具有较强的领域特点,其中包含了大量电气领域内术语。

2)电气设备缺陷文本内容一般包括设备缺陷名称、原因及解决方法,还包括电气设备生产公司等一系列信息。但在记录缺陷时,由于巡检人员知识储备存在差异,且记录人习惯不同,导致记录同一件案例会出现多种不同描述方式。加大了对缺陷文本挖掘的难度,如“指示灯”和“信号灯”、“隔离开关”和“刀闸”等^[17]。

3)由于电气设备种类多,数量大,造成设备发生缺陷的原因众多,且记录缺陷文本没有特定的模式,导致记录同一缺陷所使用的文本长度不同。从现有的数据中可以得到,缺陷文本长度范围从十几字到几百字不等^[18-20]。

4)许多缺陷文本中使用了大量电气领域的特定字符和数字对缺陷进行记录,如“变压器油中H₂体积分数为100 mL/L”,这些字符蕴藏着重要的信息;然而,字符信息在文本处理过程中常常会丢失其原有的含义和价值。

2 BBAMC模型介绍

模型对文本特征信息提取的程度影响文本分类模型的效果。传统文本分类模型存在丢失上下文语义信息的问题,并且只能对电力设备缺陷文本中的字符和数字进行简单表示,无法充分提取其含义,而这些字符和数字对缺陷等级具有重要影响甚至决定性作用。为了解决这些问题,文中提出了基于预训练的BBAMC模型,利用迁移学习将电力缺陷文本先验知识迁移到分类任务中,提取电气领域特殊背景信息。为了捕捉长距离依赖和时序信息,引入了BiLSTM模块和注意力机制,使模型能够更好地关注重要单词特征并提取文本语义信息。为了提取文本的局部特征,运用不同卷积的CNN模块可以有效地提取编码向量的局部特征。将局部特征向量与BiLSTM-Attention的文本语义特征融合,提高了对缺陷文本分类的准确率。

文中结合BERT、BiLSTM、Attention、CNN模块优势,构建BBAMC缺陷文本分类模型,模型具体结构见图1。BBAMC模型主要由7个部分构成,分别为数据预处理、BERT层、BiLSTM、Attention层、CNN层、特征融合层和Softmax层。文中首先对缺陷文本进行数据预处理,过滤掉无效数据;其次使用BERT模型对处理后的文本数据进行编码,获得包含缺陷文本语义特征的向量;然后将文本向量分别输入BiLSTM-Attention网络和多通道CNN网络,进一步获取语义特征信息和关键特征信息,并将经过多维特征提取后的向量进行融合;最终,将融合后的语义向量输入Softmax层,以此来计算分类标签概率。

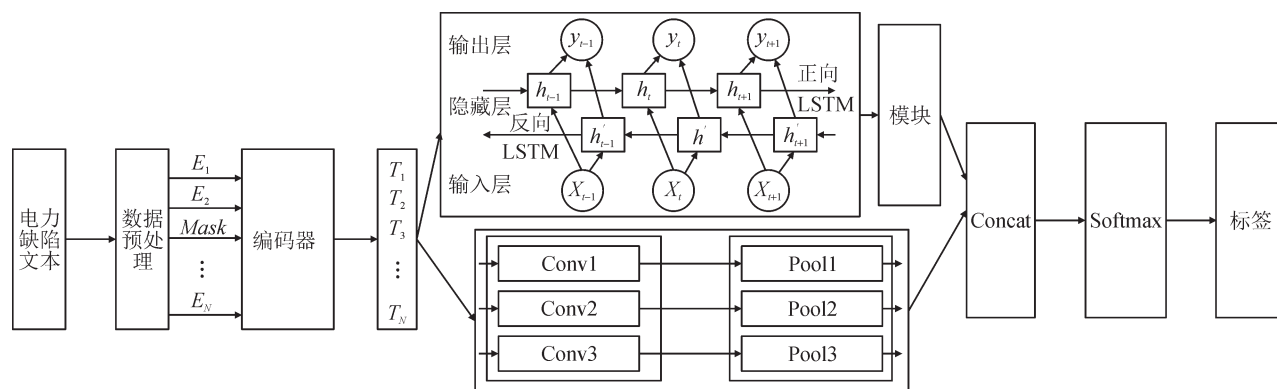


图1 BBAMC模型结构

Fig. 1 BBAMC network structure

2.1 BERT模块

BERT模型是基于Transformer的大规模预训练语言模型,见图2。设备缺陷文本输入多个Transformer结构,进行文本特征提取,其输出为缺陷

文本向量。BERT模型将非结构化数据转化为结构化数据,完成对文本的编码。BERT模型主要有预训练和微调两个步骤。而预训练部分有两个任务:掩码语言模型(MLM)和下一句预测(NSP)^[21],使模型

能理解缺陷文本的上下文信息和句子关系。在电力缺陷文本分类任务中,利用BERT的MLM和NSP任务可以增强模型对上下文信息和句子关系的理解,提高分类准确性。BERT微调部分是在预训练完成后根据具体任务进行微调。例如使用电气文本语料进行微调,使模型能够更好地理解电力系统领域的语言模式和知识,准确的分类电气缺陷文本^[22]。对于传统文本分类模型,使用空间向量模型对文本进行向量化,仅能对文本进行简单的表示,无法提取电力文本字符信息,如“本体瓦斯继电器至变压器器身连接接头漏油 210 mL/min”,传统的文本分类模型在进行向量表示时,只对数字符号进行简单的表示,忽略数字符号部分的量化信息,导致传统文本分类模型准确率低。而BERT模型在大量电气文本语料进行无监督学习,能够理解电力设备缺陷文本相关字符的含义,使得模型能够更好地提取缺陷文本中相关字符信息,从而提高模型分类的准确率。

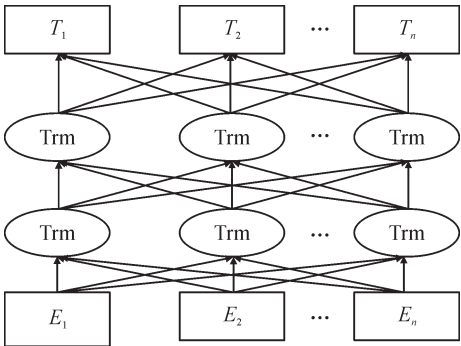


图2 BERT网络结构

Fig. 2 BERT network structure

2.2 BiLSTM模块

LSTM是一种具有特殊记忆和遗忘机制的循环神经网络,能够有效地处理文本语义特征^[23-24]。其基本单元结构包括输入门、输出门和遗忘门。

BiLSTM结构利用正向和反向两个LSTM网络进行提取语义特征,能够更好地捕获文本的前后语义特征,见图3。与单向的LSTM相比,BiLSTM模型对于提取复杂的电力缺陷文本的语义特征更具有优势。

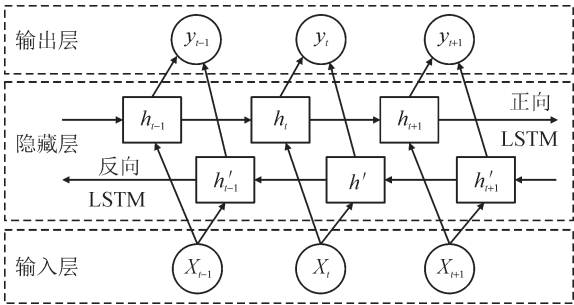


图3 BiLSTM网络结构图

Fig. 3 Structure diagram of BiLSTM

2.3 Attention 模块

传统的神经网络模型在处理文本时,存在所有的文本被赋予相同的权重问题,无法准确区分不同文本部分的重要性,导致关键信息难以被保留,从而降低了模型的性能^[25-26]。为解决这个问题,引入注意力机制打破了传统编码器模型对固定长度向量的依赖。通过计算输入文本隐藏层向量与输出的相关性得分,注意力机制分配权重给隐藏层向量,提取关键信息并减少无关部分的影响。这种灵活的权重使编码器能够有效地提取不同程度的关键信息,解决了BiLSTM网络在处理文本长距离依赖和短期记忆问题上的限制。注意力机制扩展了文本建模范围,模型更好地捕捉到文本的重要特征。

2.4 CNN 模块

Kim^[27]等人将CNN用于文本分类,提出了TextCNN以此来处理文本。TextCNN网络结构见图4。

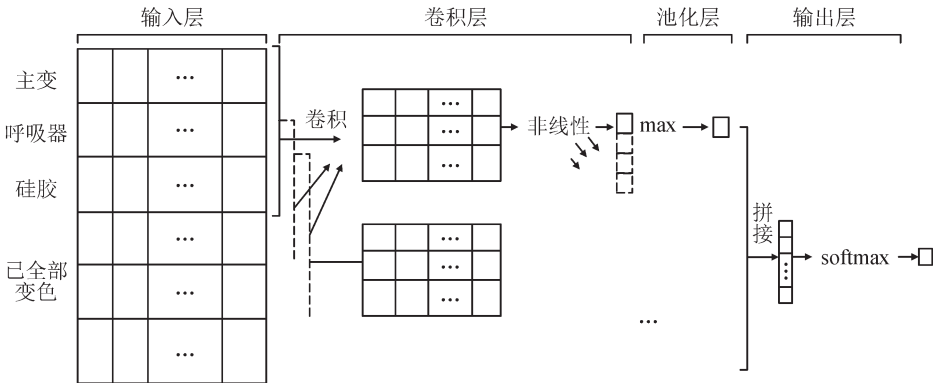


图4 卷积神经网络结构

Fig. 4 Convolutional neural network structure

文本卷积是一种滑动窗口的操作,将窗口内的文本序列与卷积核进行点积运算,得到一个特征

值^[28-29]。因缺陷文本没有固定的格式,且巡检人员的知识储备和习惯的不同,导致在表达相同意思的

时候,可能采用不同的文本进行描述,如缺陷文本中的“全部+变色”与“已+全部+变色”表达的意思相同,但需要不同的卷积窗口来提取文本信息,因此文中提出利用不同卷积来提取局部信息可以解决该问题。

3 实验结果及分析

3.1 实验环境

采用PyTorch框架进行实验,该框架涵盖了许多常用的深度学习模型,其实验环境配置见表1。

表1 实验环境

Table 1 Experimental environment	
实验环境	配置
开发框架	Pytorch
CPU	Intel Xeon Platinum 8255C
GPU	RTX 3080
内存	43 GB

3.2 实验数据与评价指标

为研究BBAMC模型在电气设备缺陷文本中分类效果,将来自南方某电力公司的27 389条电气设备缺陷文本用于实验,其中每条缺陷文本数据均已被标注缺陷等级。根据电力设备缺陷分类标准^[30],可将缺陷文本分为“一般”、“重要”和“紧急”3类。将电气设备缺陷文本随机按照8:1:1划分训练集、验证集和测试集,以10则交叉验证进行实验。具体实验数据统计见表2,实验数据示例见表3。

表2 电力设备缺陷文本数目统计

Table 2 Statistics of the number of defective texts of power equipment

缺陷等级	训练集/条	验证集/条	测试集/条	共计/条
一般	18 902	2 089	2 115	23 106
严重	2 646	350	322	3 318
紧急	787	88	90	965
共计	23 106	2 527	2 527	27 389

该任务为多分类问题,需要采取综合指标对模型进行评估,文中引入精确率、召回率以及 F_1 值对模型效果进行评估,其计算公式为:

$$P = \frac{T_p}{T_p + F_p} \quad (1)$$

$$R = \frac{T_p}{T_p + F_n} \quad (2)$$

$$F_1 = \frac{2PR}{P+R} \quad (3)$$

式(1)~(3)中: P 表示精确率; R 表示召回率; T_p 表示真正例; F_p 表示为伪正例; F_n 表示为伪反例。

表3 数据示例

Table 3 Lots of data examples

标签	缺陷等级	案例文本
0	紧急	35 kV 2M 母线及其附属设备在五防验收时发现:35 kV 2M 母线2号电容器组电容器侧32340接地开关电磁锁不能接通,经继保班检查电磁锁损坏
0	紧急	主变电量保护电源插件烧坏,运行灯灭
1	严重	10 kV 电容器组电抗器C相出线母排螺丝发热,发热相温度达到80.5℃,其他两相分别是40℃和39℃,已向监控申请将10 kV 4号电容转为热备用
1	严重	CT二次接线端子渗油,1 min 3滴
2	一般	3号电容513开关小车位置闭锁电磁铁工作
2	一般	电源整流模块损坏,导致电磁铁无法归位,使小车在摇出过程中卡滞,无法摇出
2	一般	电能表停走

3.3 BBAMC模型性能分析

文中所提出的BBAMC分类模型能充分提取缺陷文本语义特征信息,在缺陷文本自动分类任务中有着准确率高的优点。用缺陷文本数据集对BBAMC模型进行训练,训练过程见图5、6。由图5、6可知,模型在经过大约200个Iter就达到稳定,模型收敛,模型的准确率达到85.25%,此后训练集与测试集的准确率和损失值趋于水平方向稳定,由此证明文中所提出的BBAMC模型在对故障缺陷文本的分类是可行、有效的。为了使模型的分类效果达到最优,所使用反向传播算法训练模型,并结合多次实验结果,最终采取实验模型参数见表4,对比实验模型相同参数采用表4的参数,模型特有的参数使用默认参数。

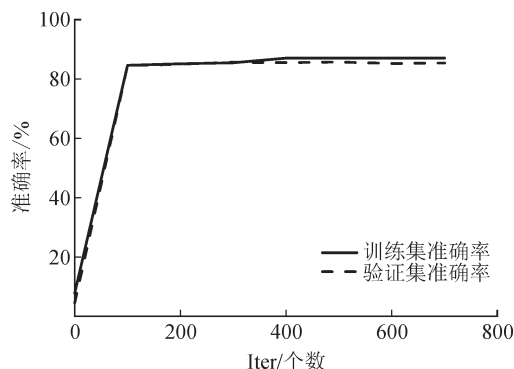


图5 训练准确率和验证准确率变化曲线

Fig. 5 Training accuracy and validation accuracy change curves

表4中:lr为模型训练过程中设置的学习率;epoch为模型训练的迭代次数;bs为每一次迭代训

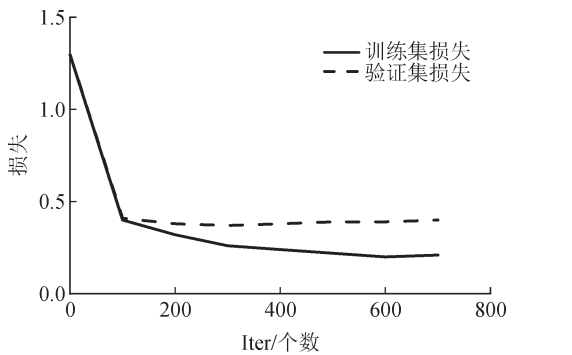


图6 训练损失函数和验证损失函数变化曲线

Fig. 6 Training loss function and validation loss function change curves

表4 BBAMC模型参数

Table 4 BBAMC model parameter

参数	数值	参数	数值
lr	6×10^{-5}	epoch	10
bs	325	hs	768
ps	256	ls	12
dp	0.1	Conv1	$2\times3\times4$
Conv2	$3\times4\times5$	Conv3	$4\times5\times6$

练,批量处理的样本个数;hs为模型隐藏层的大小;ps为每句话被处理的长度;ls为模型层数;dp为加在每层的损失率,防止模型出现过拟合。Conv1-3分别为卷积层1-3的卷积核尺寸。

验证集准确率变化曲线见图7。从图7可以得出不同模型的验证准确率变化情况。TextRNN模型的准确率波动较大,在约500个Iter才开始趋于收敛。CNN-LSTM混合模型在大约400个才Iter开始收敛,但波动较大。与此相比,BBAMC模型的曲线从约200个Iter开始收敛,且波动相对较小且更平缓。总体而言,BBAMC模型的训练过程更为稳定,且在缺陷文本数据集上的精确率达到85.25%,显著优于实验所对比的模型。在后续的训练中,BBAMC模型始终处于优势地位,充分展现了其优越性。

3.4 与传统的机器学习方法的对比

在缺陷文本分类任务中,为验证文中提出模型的优越性,将文中模型与传统的机器学习方法进行实验对比,实验结果见表5。

从表5实验结果可知,文中所提出的模型相较于SVM模型在精确率、召回率和F₁值分别提高了31.25%、29.15%、33.08%。相较于朴素贝叶斯(naive Bayesian, NB)在精确率、召回率和F₁值分别提高了20.25%、20.15%、23.08%。BBAMC模型的性能优于SVM、NB,这是因为SVM、NB的分类效果主要依赖文本特征的质量,文中所对比的SVM、NB均采用词

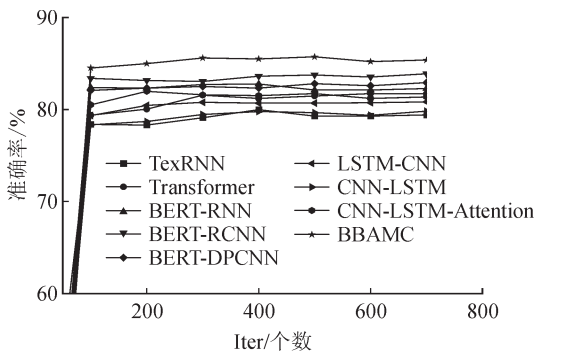


图7 验证集准确率变化曲线

Fig. 7 Validation set accuracy change curve

表5 文中模型与传统的机器学习模型性能对比

Table 5 Performance comparison of the proposed model with the traditional machine learning models %

模型	精确率	召回率	F ₁ 值
NB	65.00	67.00	63.00
SVM	54.00	58.00	53.00
文中方法	85.25	87.15	86.08

频一逆文档频率(term frequency-inverse document frequency, TF-IDF)作为文本特征。TF-IDF用词频作为文本特征,词频只能提取词的特征信息,而不能提取句子的语义信息,且因电气文本结构复杂,使用TF-IDF提取特征会损失部分重要文本特征,使SVM和NB的分类效果不理想。TF-IDF的向量特征表示方式忽略了上下文背景信息的作用,而文中所提出的模型不仅提取了词义特性,也提取了句义的特性信息,进一步提取缺陷文本的深层特性,从而提高了文本分类的准确性。

3.5 文中方法与深度学习模型的对比

为了进一步验证文中提出的模型性能优越,将文中所提出的模型与深度学习模型进行实验对比。深度学习模型的文本特征皆采用Word2vec分布式词向量表示。文中实验选取电气文本挖掘的常用模型,如TextRNN、BiLSTM-Attention等模型,实验结果见表6。由表6可知,文中所提出的模型性能均优于对比实验模型,其中本文模型相较于最优的CNN-LSTM-Attention深度学习模型的准确率、召回率、F₁值分别提高了4.01%、4.59%、4.03%。

从表6实验结果可知,文中方法性能明显优于TextRNN、CNN-LSTM、LSTM-CNN、CNN-LSTM-Attention、Transformer。TextRNN模型无法解决长距离依赖和梯度消失和梯度爆炸的问题,导致在长文本时容易丢失上下文信息。CNN-LSTM、LSTM-CNN、CNN-LSTM-Attention模型虽然都能提取局部特征和语义特征,但单个CNN提取更多的局部关键

表6 文中模型与深度学习模型性能对比
Table 6 Performance comparison of the proposed model with deep learning models %

模型	精确率	召回率	F_1 值
TextRNN	79.52	81.67	80.45
CNN-LSTM	79.40	84.29	81.65
LSTM-CNN	80.56	82.75	81.75
CNN-LSTM-Attention	81.24	82.56	82.05
Transformer	81.66	81.06	81.29
文中方法	85.25	87.15	86.08

信息,并且LSTM只能提取前文语义信息,不能与后面语义信息相联系。Transformer模型对输入数据中的噪声和错误较为敏感,缺陷文本因是人工记录,没有固定的记录格式,导致文本存在噪声,即使是输入中的小错误或噪声,也会导致模型的输出结果出现明显的错误。而BBAMC预训练模型结合了BERT、BiLSTM、Attention、CNN模型的优点,能很好的解决上述模型存在的问题,可以在处理设备缺陷文本时高效地进行特征提取和分类,并且对于缺陷文本错误或语法错误的案例也能准确的识别。在缺陷文本分类实验中,BBAMC对缺陷文本分类效果更好。

3.6 文中方法与预训练模型对比

为进一步验证文中提出模型的优越性,将所提出的模型与组合预训练模型进行实验对比。对比实验模型选取的参数见表4,其实验结果见表7。由表7可知,文中提出的模型相较于预训BERT-RNN模型的精确率、召回率、 F_1 值提高了2.76%、3.58%、3.19%。

表7 文中模型与预训练模型性能对比
Table 7 Performance comparison of the proposed model with the pre-trained model %

模型	精确率	召回率	F_1 值
BERT-RNN	82.49	83.57	82.89
BERT-RCNN	83.76	84.24	84.01
BERT-PCNN	84.21	85.45	84.75
文中方法	85.25	87.15	86.08

根据表7可知,文中提出的模型在电力设备缺陷文本分类任务中表现优于BERT-RNN、BERT-RCNN和BERT-DPCNN。其中,BERT-RNN模型能够提取上下文信息,但存在梯度消失或梯度爆炸问题。BERT-RCNN在利用BERT编码基础上,运用RCNN捕捉局部特征和上下文语义特征,但其卷积窗口难以有效捕捉文本之间的依赖关系,从而无法获取窗口之间的上下文信息。BERT-DPCNN对局

部和全局语义信息的提取相对较好,但其在建模文本序列的长距离依赖关系方面有限。相比之下,文中结合上述模型的优势,提出BBAMC模型,该模型在学习电力领域缺陷文本信息的基础上,使用了一种融合上下文语义特征和多个局部关键信息的方法,能够更全面地捕捉文本中的重要特征和语义信息特征。

3.7 消融实验

为了更好地分析模型的不同模块对电力设备缺陷文本分类的影响,文中设计了消融实验,对比实验模型选取的参数见表4,实验结果见表8。

表8 缺陷文本分类模型的消融实验结果
Table 8 Comparison of ablation experimental results of different text classification models %

模型	精确率	召回率	F_1 值
MCNN	80.77	83.39	82.16
BiLSTM-Attention	81.46	84.21	82.13
BERT	81.26	83.81	80.78
BERT-BiLSTM-Attention	83.32	84.57	83.93
BERT-MCNN	83.37	84.82	84.05
文中方法	85.25	87.15	86.08

从表8实验结果可知,文中提出模型相较基准模块MCNN模型的精确率、召回率、 F_1 值分别提高了4.48%、3.76%和3.92%,相较于BiLSTM-Attention模型的精确率、召回率、 F_1 值分别提高了3.79%、2.94%和3.95%,而相较于BERT模型提高了3.99%、3.34%和5.3%。尽管BERT-MCNN在BERT编码的基础上,重点提取了文本局部特征信息,但对文本上下文语义信息提取依然不足;BERT-BiLSTM-Attention能充分提取上下文语义信息,却忽略了局部关键信息。由于文中所建模型结合了提取局部关键信息和提取上下文语义信息的双重优势,因此文中提出的模型在电力设备缺陷文本中具有更好的自动分类效果。

4 结论

针对当前电气设备缺陷文本主要依靠人工分类的问题,文中提出了基于预训练模型的BBAMC模型,得出如下结论:

1)针对电力设备缺陷文本特点,提出基于预训练模型方法。该方法结合预训练模型迁移学习以及能熟悉专业知识的优势,解决了传统分类模型不能识别特殊字符的问题,提高了模型对缺陷文本分类准确率。

2)针对文本语义特征和局部信息对缺陷等级分类起着关键性作用的特点,文中引入BiLSTM、Attention、多层CNN模块。该模块的融合能同时处理文本序列的正向和逆向信息,能够有效地捕捉文本上下文中的语义关联和局部关键信息特征,提高了模型分类的准确率。

3)为了证明模型的有效性,文中在相同的数据集下进行了对比实验,实验结果表明,BBAMC模型在缺陷文本分类的任务中具有令人满意的分类性能,优于其他常用对比模型,表明了所建模型的有效性与优越性。

文中提出的基于预训练模型、BiLSTM模块、Attention模块和多个CNN相结合的电力设备缺陷文本自动分类方法,为缺陷文本自动分类提供了一种解决方案。但文中只考虑提取缺陷文本的特征未能引入外部知识对缺陷文本进行分析,后期将基于现有工作开展引入外部知识对缺陷文本自动分类任务研究,提升分类模型总体性能。

参考文献:

- [1] 朱彦名,徐潇源,严正,等.面向电力物联网的含可再生能源配电网运行展望[J].电力系统保护与控制,2022,50(2):176-187.
ZHU Yanming, XU Xiaoyuan, YAN Zheng, et al. Prospect of renewable energy integrated distribution network operation in the power internet of things[J]. Power System Protection and Control, 2022, 50(2):176-187.
- [2] 陈敬德,盛戈皞,吴继健等.大数据技术在智能电网中的应用现状及展望[J].高压电器,2018,54(1):35-43.
CHEN Jingde, SHENG Gehao, WU Jijian, et al. Application status and prospect of big data technology in smart grid [J]. High Voltage Apparatus, 2018, 54(1):35-43.
- [3] 郑翔,王慧芳,严娴峥等.用于APP的缺陷文本自动分类与质量保证方法[J].电力系统及其自动化学报,2020,32(7):131-136.
ZHENG Xiang, WANG Huifang, YAN Xianzheng, et al. Defect text automatic classification and quality assurance method for APP[J]. Proceedings of the CSU-EPSA, 2020, 32(7):131-136.
- [4] 马润泽,王龙响,余佳文,等.考虑历史缺陷文本信息的断路器状态评价研究[J].机电工程,2015,32(10):1375-1379.
MA Runze, WANG Longxiang, YU Jiawen, et al. Circuit breakers condition evaluation considering the information in historical defect texts[J]. Journal of Mechanical & Electrical Engineering, 2015, 32(10): 1375-1379.
- [5] 汪崔洋,江全元,唐雅洁,等.基于告警信号文本挖掘的电力调度故障诊断[J].电力自动化设备,2019,39(4):126-132.
WANG Cuiyang, JIANG Quanyuan, TANG Yajie, et al. Fault diagnosis of power dispatching based on alarm signal text mining[J]. Electric Power Automation Equipment, 2019, 39(4):126-132.
- [6] 刘梓权,王慧芳,曹靖,等.基于卷积神经网络的电力设备缺陷文本分类模型研究[J].电网技术,2018,42(2):644-650.
LIU Ziquan, WANG Huifang, CAO Jing, et al. A classification model of power equipment defect texts based on convolutional neural network[J]. Power System Technology, 2018, 42(2):644-650.
- [7] 杜修明,秦佳峰,郭诗瑶,等.电力设备典型缺陷文本的文本挖掘[J].高电压技术,2018,44(4):1078-1084.
DU Xiuming, QIN Jiafeng, GUO Shiyao, et al. Text mining of typical defects in power equipment[J]. High Voltage Engineering, 2018, 44(4):1078-1084.
- [8] 冯斌,张又文,唐昕,等.基于BiLSTM-Attention神经网络的电力设备缺陷文本挖掘[J].中国电机工程学报,2020,40(S1):1-10.
FENG Bin, ZHANG Youwen, TANG Xin, et al. Power equipment defect record text mining based on BiLSTM-Attention neural network[J]. Proceedings of the CSEE, 2020, 40(S1):1-10.
- [9] MANNING C D, SURDEANU M, BAUER J, et al. The stanford core NLP natural language processing toolkit[C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2014:55-60.
- [10] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. Louisiana, USA: Association for Computational Linguistics, 2018:2227-2237.
- [11] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minnesota, USA: Association for Computational Linguistics, 2019:4171-4186.
- [12] 晏鹏,黄晓旭,黄玉辉,等.基于BERT-DSA-CNN和知识库的电网调控在线告警识别[J].电力系统保护与控制,2022,50(4):129-136.
YAN Peng, HUANG Xiaoxu, HUANG Yuhui, et al. Online alarm recognition of power grid dispatching based on BERT-DSA-CNN and a knowledge base[J]. Power System Protection and Control, 2022, 50(4):129-136.
- [13] 虞佳淼,王慧芳,张亦翔,等.基于BERT的电网现场作业风险自动评级方法[J].电网技术,2023,24(1):1-9.
YU Jiamiao, WANG Huifang, ZHANG Yixiang, et al. Automatic risk rating method for power grid field operation based on BERT[J]. Power System Technology, 2023, 24(1):1-9.
- [14] 夏飞,陈帅琦,华珉,等.基于改进BERT的电力领域中文分词方法[J].计算机应用,2023,1(1):1-10.
XIA Fei, CHEN Shuaiqi, HUA Min, et al. Chinese word segmentation method in electric power domain based on improved BERT[J]. Journal of Computer Applications, 2023, 1

- (1):1-10.
- [15] 任晓辉. 电网调控自动化系统运行状态在线监视与智能诊断研究及应用[J]. 电力系统保护与控制, 2018, 46(11): 156-161.
- REN Xiaohui. Research application states of grid dispatching and control system[J]. Power System Protection and Control, 2018, 46(11): 156-161.
- [16] 谢庆, 张焯宇, 王春鑫, 等. 新一代人工智能技术在输变电设备状态评估中的应用现状及展望[J]. 高压电器, 2022, 58(11): 1-16.
- XIE Qing, ZHANG Xuanyu, WANG Chunxin, et al. Application status and prospect of the new generation artificial intelligence technology in the state evaluation of power transmission and transformation equipment[J]. High Voltage Apparatus, 2022, 58(11): 1-16.
- [17] 郝莎, 徐建源, 林莘, 等. 基于隔离开关结构和操作特性的电弧放电模型研究与应用[J]. 高压电器, 2022, 58(4): 131-136.
- HAO Sha, XU Jianyuan, LIN Xin, et al. Research and application of arc discharge model based on the structure and operation characteristics of disconnector[J]. High Voltage Apparatus, 2022, 58(4): 131-136.
- [18] 邓科, 侯晓松, 林湘宁, 等. 基于5G通信的电气设备监测终端性能在线评估方案[J]. 电力系统保护与控制, 2021, 49(7): 39-47.
- DENG Ke, HOU Xiaosong, LIN Xiangning, et al. Online evaluation scheme for the performance of an electrical equipment monitoring terminal based on 5G communication[J]. Power System Protection and Control, 2021, 49(7): 39-47.
- [19] 邹晓峰, 沈冰, 蒋献伟. 5G通信条件下配网差动保护快速动作方案研究[J]. 电力系统保护与控制, 2022, 50(16): 163-169.
- ZOU Xiaofeng, SHEN Bing, JIANG Xianwei. A quick action scheme of differential protection for a distribution network with 5G communication[J]. Power System Protection and Control, 2022, 50(16): 163-169.
- [20] 李文君, 段登伟, 朱雨, 等. 基于5G通信模式下的配电网自愈保护应用[J]. 电力系统保护与控制, 2022, 50(24): 152-159.
- LI Wenjun, DUAN Dengwei, ZHU Yu, et al. Application of distribution network protection based on a 5G end-to-end communication mode[J]. Power System Protection and Control, 2022, 50(24): 152-159.
- [21] 陆晓蕾, 倪斌. 基于预训练语言模型的BERT-CNN多层级专利分类研究[J]. 中文信息学报, 2021, 35(11): 70-79.
- LU Xiaolei, NI Bin. BERT-CNN: A hierarchical patent classifier based on pre-trained language model[J]. Journal of Chinese Information Processing, 2021, 35(11): 70-79.
- [22] 陈玺, 杨雅婷, 董瑞. 面向汉维机器翻译的BERT嵌入研究[J]. 计算机工程, 2021, 47(12): 112-117.
- CHEN Xi, YANG Yating, DONG Rui. Research on BERT embedding for Chinese - uyghur machine translation[J]. Computer Engineering, 2021, 47(12): 112-117.
- [23] 朱晓红, 杨伟荣, 张蓉, 等. 基于RNN-LSTM神经网络的小电流接地故障选线方法[J]. 高压电器, 2023, 59(7): 213-220.
- ZHU Xiaohong, YANG Weirong, ZHANG Rong, et al. Line selection method of low current grounding fault based on RNN-LSTM neural network[J]. High Voltage Apparatus, 2023, 59(7): 213-220.
- [24] 赵建文, 范文璐, 胡雨佳. 基于零序电流包络线拟合斜率的配电网故障选线方法[J]. 电力科学与技术学报, 2022, 37(3): 117-125.
- ZHAO Jianwen, FAN Wenlu, HU Yujia. A distribution network fault line selection method based on zero-sequence current envelope fitting slope[J]. Journal of Electric Power Science and Technology, 2022, 37(3): 117-125.
- [25] 李佳玮, 王小君, 和敬涵, 等. 基于图注意力网络的配电网故障定位方法[J]. 电网技术, 2021, 45(6): 2113-2121.
- LI Jiawei, WANG Xiaojun, HE Jinghan, et al. Distribution network fault location based on graph attention network[J]. Power System Technology, 2021, 45(6): 2113-2121.
- [26] 白浩, 黄升平, 潘姝慧, 等. 配电网中性点电压主动调控的接地故障零序电流多判据选线方法[J]. 电力科学与技术学报, 2022, 37(4): 105-113.
- BAI Hao, HUANG Shengping, PAN Shuhui, et al. Zero sequence current multi-criteria line selection method based on active regulation of neutral voltage in distribution networks[J]. Journal of Electric Power Science and Technology, 2022, 37(4): 105-113.
- [27] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). Doha, Qatar, Association for Computational Linguistics, 2014: 1746-1751.
- [28] 王春雷, 张建林, 李美惠, 等. 结合卷积Transformer的目标跟踪算法[J]. 计算机工程, 2023, 49(4): 281-288.
- WANG Chunlei, ZHANG Jianlin, LI Meihui, et al. Object tracking algorithm combining convolution and Transformer[J]. Computer Engineering, 2023, 49(4): 281-288.
- [29] 王辉东, 姚海燕, 郭强, 等. 基于多尺度卷积神经网络的变压器故障诊断方法[J]. 电力科学与技术学报, 2023, 38(4): 104-112.
- WANG Huidong, YAO Haiyan, GUO Qiang, et al. A transformer fault diagnosis method based on multiscale 1DCNN[J]. Journal of Electric Power Science and Technology, 2023, 38(4): 104-112.
- [30] 浙江省电力公司. 电网变电一次设备缺陷用语规范: Q/ZDJ 44—2005[S]. 2005.
- Zhejiang Electric Power Corporation. Grid substation primary equipment defect terminology specification: Q/ZDJ 44—2005[S]. 2005.

张中文(1998—), 男, 硕士研究生, 研究方向为电力文本处理(E-mail: 107552101446@stu.xju.edu.cn)。

吐松江·卡日(1984—), 男, 博士, 副教授, 博士生导师, 研究方向为电力系统数字化关键技术研究、新能源并网与预测等(通信作者)(E-mail: minyun229@163.com)。