基于ERNIE预训练的电力设备缺陷文本挖掘方法

杨帆[1],黄河[1],李海英[2],李东东[1],赵耀[1]
(1.上海电力大学 电气工程学院,上海 200090;
2.天津电气科学研究院有限公司,天津 300180)

摘要:电力设备缺陷文本中蕴含了大量的设备运行规律,可采用自然语言处理技术对其进行挖掘,为设备运维提供指导。针对该现状构建了一种基于知识增强的语义表示模型(ERNIE)预训练的电力设备缺陷文本挖掘方法。首先,分析了传统词向量模型的局限性,并总结了电力设备缺陷文本的特点,针对以上问题采用一种基于ERNIE进行预训练,得到包含缺陷文本上下文语境信息的词向量;接着,采用双向长短时记忆网络(BiLSTM)对缺陷文本双向学习,进一步提取缺陷文本语义特征并分类;最后,通过实例分析表明ERNIE-BiLSTM模型具有更良好的分类性能,为今后电力设备缺陷文本挖掘提供了新的思路。

关键词:电力设备;缺陷文本挖掘;预训练模型;知识增强;双向长短时记忆网络
中图分类号:TM28 **文献标识码**:A **DOI**:10.19457/j.1001-2095.dqcd22951

Power Equipment Defect Text Mining Method Based on ERNIE Pre-trained

YANG Fan[1],HUANG He[1],LI Haiying[2],LI Dongdong[1],ZHAO Yao[1]
(1. *School of Electrical Engineering*,*Shanghai University of Electrical Power*,*Shanghai* 200090,*China*;
2. *Tianjin Research Institute of Electric Science Co.*,*Ltd.*,*Tianjin* 300180,*China*)

Abstract: There are a lot of equipment operation rules in power equipment defect text,which can be mined by natural language processing technology to provide guidance for equipment maintenance. Aiming at this situation,a defect text mining method based on enhanced representation through knowledge integration(ERNIE)pre-trained was constructed. Firstly,the limitations of traditional word vector model were analyzed,and the characteristics of power equipment defect text were summarized. Aiming at the above problems,ERNIE was used for pre training to get the word vector containing the context information of the defect text. Then,bidirectional long short term memory(BiLSTM) was used to learn the defect text in two directions,and the semantic features of the defect text were further extracted and classified. Finally,an example shows that ERNIE-BiLSTM has better classification performance,which provides a new idea for future power equipment defect text management.

Key words: power equipment;defect text mining;pre training model;knowledge enhancement;bidirectional long short term memory(BiLSTM)

电力系统在日常运行中积攒了大量的缺陷文本数据,这些文本数据中蕴含了设备的潜在运行规律[1]。在当前“电力大数据”的时代背景下,如何利用基于数据驱动的人工智能方法从数据中挖掘出相关运行规律,对提升电力系统的运维检修智能化具有重要意义[2]。

近年来,在电力文本挖掘领域,主要应用研究有:设备缺陷等级分类、历史文本经验指导和健康状态评价等。文献[3]利用One-hot编码实现缺陷文本的向量化,并采用K最近邻算法(K-nearest neighbor,KNN)进行分类,提升断路器状态评价合理性,但One-hot模型添加新词困难,面对电力设备缺陷文本这种高维数据集容易导致维数灾难等问题,无法衡量词语之间的语义关联度;文献[4]利用隐马尔可夫法(hidden Markov model,HMM)对断路器缺陷文本进行向量化,并

基金项目:国家自然科学基金(51977127)
作者简介:杨帆(1981—),女,博士,副教授,Email:smartgridcontrol@163.com

采用比率型状态信息融合模型对断路器全寿命健康状态进行展示，但HMM这种词库型自然语言处理技术无法很好地处理电力缺陷文本这种专业性强的语料；文献[5]通过BiLSTM深度提取缺陷文本句级别的深层语义特征，提高分类效果；文献[6]提出一种基于深度语义学习的变压器运维文本挖掘方法，研究给出的健康指数计算方法有助于运维人员更直观理解变压器缺陷程度；文献[7]在文献[5]的基础上提出一种基于注意力机制的BiLSTM缺陷文本分类方法，提出的缺陷处理优先指数具有实际应用价值。文献[5-7]的共性在于均采用了Word2vec词嵌入模型[8]对缺陷文本进行向量化，以无监督学习的方式学习语义信息，将分词后的缺陷文本映射为K维向量，这种向量被称为静态词向量，词向量无法根据上下文语境动态调整。但文本信息特征表示的好坏对缺陷文本分类的准确率影响重大，好的词向量能够有效提升模型的学习效果[9]。电力设备缺陷文本本身即存在着特征稀疏等问题，导致Word2vec并不能完整地表示缺陷文本的语义信息。另一方面，不同电力设备的缺陷类型有所不同，且电力文本还包括了调度文本等，Word2vec需要根据应用场景进行重新训练，词向量效果无法保证，模型整体的鲁棒性和泛化能力较差。

为了更好地利用缺陷文本上下文语义信息，保留句子层面的语义特征，学界目前的研究重点已经转向以句为单位的预训练语言模型[10]，使词向量能够动态地根据上下文进行调整，解决一词多义问题的同时充分提升了模型的泛化能力。为了提高预训练语言模型的分类准确率，Matthew E. Peters等人[11]在2018年提出的ELMo模型通过深度双向语言模型进行预训练，使产生的词向量能够包含丰富的句法和语义信息，并且能对多义词进行建模，这种能够根据上下文语境进行调整的词向量被称为动态词向量。然而ELMo的双向学习机制在时间上是不同步的，为了解决这一问题，Google于2018年提出了真正意义上的基于双向Transformers编码表示（bidrectional encoder representations from transformers，BERT）的预训练语言模型[12]，各项研究均表明预训练语言模型在自然语言处理领域的多项任务中都表现出优异的性能。在电力文本挖掘领域，近年来预训练语言模型也开始展现出了优势，文献[13]提出了一种基于PBERTBiLC的电力实体信息识别方法，利用大量电力领域相关文本语料进行学习，生成具有电力特色的BERT模型，并结合BiLSTM和条件随机场（conditional random field，CRF）共同构建了电力实体信息识别模型，提高了电力实体信息识别准确率；文献[14]针对电网调度语言的特点对BERT的输入特征进行改进，提高了模型对调度语言的适应性，并验证了其在困惑度和语音识别效果中的优势。

为了提高电力设备缺陷文本挖掘模型的泛化能力和准确率，本文首先总结了电力设备缺陷文本的特点，在此基础上采用一种基于知识增强的预训练语言模型ERNIE对缺陷文本进行向量化，训练出包含上下文语义信息的缺陷文本词向量，增强其语义表示能力；其次，为了进一步获取缺陷文本的双向语义信息，本文在ERNIE模型的基础上增加了BiLSTM单元，将ERNIE训练后得到的缺陷文本词向量作为BiLSTM的输入进行双向学习，并采用Softmax分类器进行分类；最后，以准确率、精确度及召回率作为评价指标，与传统词向量模型及分类模型进行比较，所提出的ERNIE-BiLSTM模型具有更强的语义学习能力和泛化能力。

1　电力设备缺陷文本特点

电力设备缺陷文本内容通常包括了缺陷所在单位、缺陷编号、缺陷发现日期、缺陷设备、缺陷内容、缺陷等级、技术原因、处理详情等。缺陷内容记录了电力设备的具体缺陷情况，如设备产生缺陷的具体部件、缺陷的具体现象等，运维人员会将具体的缺陷信息以文本的形式详细记录在缺陷内容中，缺陷内容对于缺陷等级的评估具有重要的价值。因此，缺陷文本蕴含了大量的健康状态既往史和同类设备运行规律，如果能正确地通过缺陷内容判断缺陷的严重程度，避免人为主观因素的影响，便可为设备的检修周期提供决策参考，在设备生命将至时提前对设备进行维修或更换。图1给出了缺陷文本的数据结构。

从图1中可以看出，相较于普通的中文文本，电力设备缺陷文本具有以下特点：

1）缺陷文本专业性非常强，包含了大量的电气专业术语；

2）缺陷文本中掺杂着大量的数字、单位及特殊符号等，例如“主变本体渗油，大约20 s 1滴，目前油位在1/2处”。包含的数字对于缺陷文本的等级判断起着重要的作用，传统的文本挖掘方法在

学习缺陷文本时容易忽略数字部分的量化信息；

3）缺陷文本往往包含了大量缩写，长短差异也较大。缺陷文本通常为人工输入，由于经验及习惯的差异，存在着记录方式的差异，例如“呼吸器”、“吸湿器”和“吸潮器”均表示同一个含义，以及“主变有载油位过高”与“主变有载油位高于正常油位的上限”，两条缺陷文本表示的含义相同，但文本的长度和记录形式均有所差异。

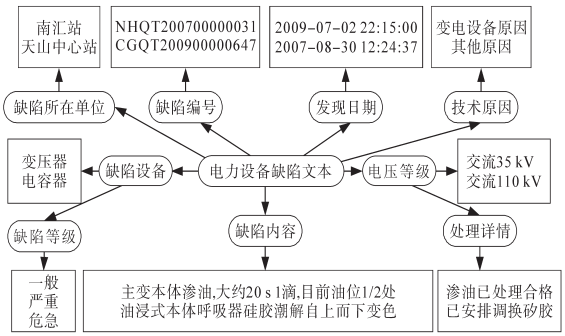

图1　电力设备缺陷文本数据结构图

Fig.1　Data structure of power equipment defects text

2　基于ERNIE-BiLSTM的电力设备缺陷文本分类模型

预训练语言模型本质上采用了迁移学习的思想。预训练语言模型能将电力相关领域数据集中学习到的先验语言知识迁移到缺陷文本领域，从而有效地提升了模型泛化能力。因此，为了解决缺陷文本在向量化时容易丢失上下文语义信息、泛化能力较差等问题，本文采用了一种更适用于中文文本挖掘的的预训练模型ERNIE，相较于BERT和ELMo，ERNIE是一种基于大规模中文语料训练出来的语言模型，多项研究均已证明ERNIE在文本分类、命名实体识别等多个中文文本挖掘任务上都要表现得更为优秀。其次，为了得到更好的缺陷文本分类结果，采用BiLSTM模型对ERNIE产生的词向量进一步进行特征提取，BiLSTM克服了传统循环神经网络的梯度爆炸和梯度消失的问题，能对缺陷文本进行双向编码提取特征。

ERNIE-BiLSTM的结构如图2所示，模型包括了3个部分。假设预处理后的电力设备缺陷文本为$\boldsymbol{E}=[E_1,E_2,\cdots,E_N]$，利用ERNIE进行预训练求取其词向量表示$\boldsymbol{T}=[T_1,T_2,\cdots,T_N]$；然后将ERNIE输出的词向量$\boldsymbol{T}$作为BiLSTM层的输入进行特征提取，进一步提取电力设备缺陷文本中每个词语的上下文语义信息，经过双向学习后的缺陷文本词向量表示为$\boldsymbol{H}=[H_1,H_2,\cdots,H_N]$；最后采用Softmax层进行分类，计算输入的缺陷文本的分类标签的概率分布，判断其缺陷等级。

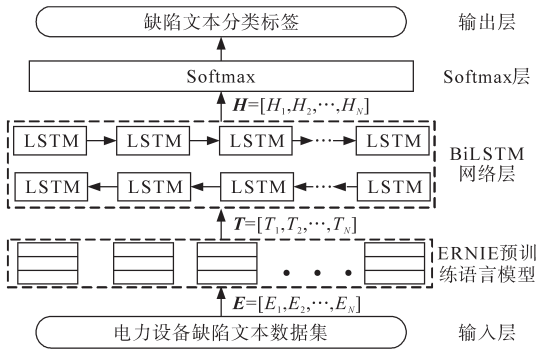

图2　ERNIE-BiLSTM结构图

Fig.2　ERNIE-BiLSTM structure

2.1　ERNIE预训练语言模型

ERNIE是一种知识增强的语义表示模型，通过对海量数据中的词、实体及实体关系等先验语义知识建模，进而学习到完整概念的语义表示[15]。ERNIE与BERT都是以多层双向Transformer编码器为基础单元构建的预训练语言模型，结构如图3所示[16]。

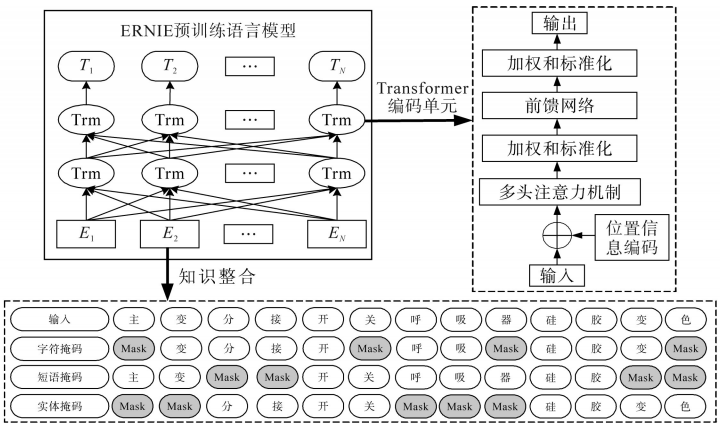

图3　ERNIE结构示意图

Fig.3　Structure of ERNIE

ERNIE包括编码和知识整合两个部分。编码部分利用Transformer编码器生成融合了上下文语义信息的缺陷文本词向量，从图中可以看出Transformer编码器基于注意力机制进行建模，注意力机制能够反映缺陷文本中不同词汇之间的重要程度，提高有助于缺陷文本分类信息的权重，进一步提高模型特征抽取的准确性；在知识整合部分，ERNIE提出了一种多阶段知识遮蔽策略，融合了字符、短语以及实体级别的缺陷文本语义知识。知识遮蔽指的是通过随机掩码一些字符，训练模型预测被掩码的部分，进而有效学习到被掩码部分的上下文信息。相较于BERT模型的单一字掩码策略，ERNIE引入了字符、短语和实体三种级别的掩码方式，通过多阶段知识遮蔽策略，ERNIE能够生成一个包含了丰富缺陷文本语义信息的词向量，有效保留了缺陷文本中各成分之间的关联性，从而保证重要的语义信息不会丢失。

2.2　BiLSTM模型

传统的机器学习方法依赖人工预先进行特征提取，再进行文本分类，容易导致前后文信息丢失。长短时记忆网络（long short term memory，LSTM）为了解决这一问题，利用门控单元来控制长期信息的传递过程，增强上下文的关联性。LSTM具有三个门控结构，包含了输入门（i_t）、遗忘门（f_t）、输出门（o_t）及记忆单元（c_t），能够有效克服一般神经网络所存在的梯度消失问题。LSTM结构如图4所示。

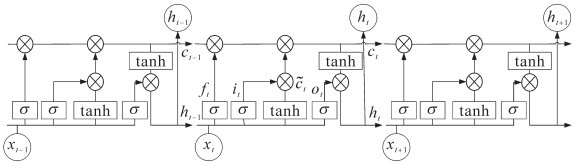

图4　LSTM结构图

Fig.4　LSTM structure

图4中，x_t为t时刻的输入；f_t为t时刻对t+1时刻的输出；o_t为t时刻的输出；h_t为隐藏层，表示t时刻的输出；σ为Sigmoid函数；c_t为t时刻的单元状态。LSTM中门控单元计算公式如下所示：

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, x_t] + b_i) \tag{1}$$

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, x_t] + b_o) \tag{3}$$

$$h_t = o_t \tanh c_t \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{5}$$

式中：$\boldsymbol{W}_i$，$\boldsymbol{W}_f$，$\boldsymbol{W}_o$为门控单元连接的权重矩阵；b_i，b_f，b_o为偏移值。

然而，LSTM只能学习电力设备缺陷文本的上文语义。BiLSTM能对缺陷文本进行双向编码，结构如图5所示。其中，x_t为输入；$\vec{h}_t$为正向输出；$\overleftarrow{h}_t$为反向输出；h_t为t时刻BiLSTM的输出。

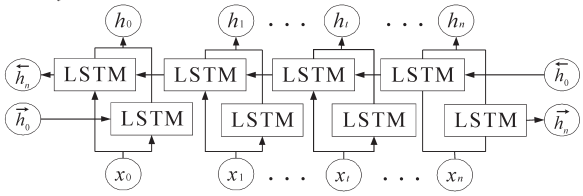

图5　BiLSTM结构图

Fig.5　BiLSTM structure

例如，某变压器缺陷文本数据“主变本体矽胶全部变色”，通过正向编码能学习到“主变”等语义信息，而反向编码则能学习到“变色”等语义信息。BiLSTM能考虑缺陷文本上下文信息，对有价值信息的语序含义进行保留，提高缺陷文本分类准确率。

3　实例分析

3.1　实验环境及实验数据集

具体实验环境如下：选择Python语言作为开发语言，CPU型号为Intel Core i7-8750H，开发工具选择JetBrains PyCharm。

为了研究本文所构建的ERNIE-BiLSTM模型对于电力设备缺陷文本的分类效果，本文经过筛选、去除无效文本，选取某地电网公司2010—2016年的变压器缺陷文本共计13 095条，每条缺陷文本均已人工标注了缺陷等级，缺陷等级分为“一般”、“严重”和“危急”3种。将所有变压器缺陷文本随机以4∶1的比例划分为训练集与测试集，进行5倍交叉验证，变压器缺陷文本数目统计如表1所示。

表1　变压器缺陷文本数目统计

Tab.1　Statistics of transformer defect text number

缺陷等级	一般	严重	危急	共计
训练集	6 076	4 204	196	10 476
测试集	1 519	1 051	49	2 619
共计	7 595	5 255	245	13 095

3.2　评价指标

为了验证缺陷文本分类模型的效果，在完成训练和测试后，需要对测试结果进行评价，因此采用准确率Acc、精确度P及召回率R作为评价指标，定义如下：

$$Acc = \frac{TP + TN}{TP + FP + FP + FN} \tag{6}$$

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

式中：TP为缺陷文本正类正确的样本数目；FP为缺陷文本负类正确的样本数目；TN，FN为与缺陷文本实际分类预测相反的样本数目。

由表1可知，缺陷文本分类是多分类问题，且数据集分布不平衡，“危急”缺陷样本数目较少。因此为了保证评价结果的合理性，本文还采用宏平均综合指标F_1对模型的全局性能进行评估，宏平均综合指标将多分类问题拆分成多个二分类问题来综合每个类别的分类准确情况，定义如下：

$$F_1 = \frac{2PR}{P + R} \tag{9}$$

3.3 实验结果对比及分析

为了最大限度保留缺陷文本中电气专业词汇信息，本文根据变压器缺陷文本特点总结了相关领域词典，以Jieba分词工具为基础实现缺陷文本的分词。此外，为了降低缺陷文本冗余度，结合哈工大停用词表设立了停用词，并补充了地址等对判断缺陷等级无用的词汇，进一步提升特征提取的准确率。缺陷文本预处理流程示意图如图6所示。

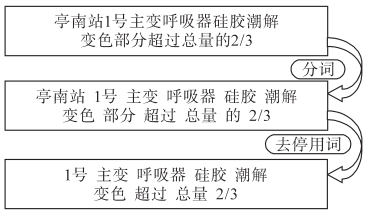

图6　缺陷文本预处理示意图

Fig.6　Schematic diagram of defect text preprocessing

3.3.1 预训练对缺陷文本分类结果的影响

传统的静态词向量模型虽然有效解决了维数灾难等问题，但缺乏对上下文语境信息的考虑。为了验证ERNIE对缺陷文本分类准确率的提升，在文本表示阶段，选取了Word2Vec，GloVe两种静态词向量模型和BERT预训练语言模型与其进行对比。图7给出各模型在训练过程中的准确率变化曲线。

从图7中可以看出，随着迭代次数的增加，所有模型均呈现整体上升趋势，预训练语言模型均能在迭代次数较少的时候取得不错的效果，并且由于ERNIE基于中文语料进行建模，能以最少的迭代次数达到最优。而传统的静态词向量模型准确率在迭代过程中波动幅度较大，且准确率始终低于预训练语言模型。最终实验结果如图8所示。

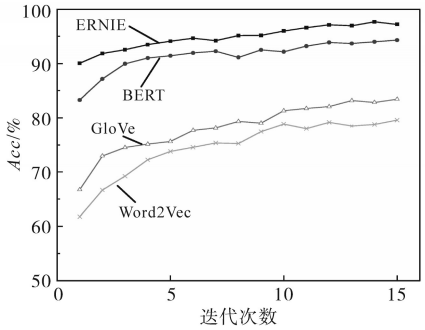

图7　训练集准确率变化曲线

Fig.7　Training set accuracy curves

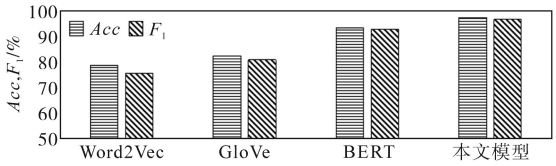

图8　词向量模型对分类结果的影响

Fig.8　Influence of word vector model on classification results

从图8中可以看出，Word2Vec的分类准确率最低，主要原因是缺陷文本记录内容表达多样化，自行编撰的领域词典难以概括全部内容；GloVe由于引入了缺陷文本的全局统计信息，故性能要略优于Word2Vec。传统的静态词向量无法解决变压器缺陷文本中错字等情况，例如“变压气”等。

预训练语言模型对于分类性能的提升非常明显。相较于Word2Vec只能在单个字符或词汇上提取缺陷文本特征信息，BERT结合了缺陷文本的全部内容，对每个字符都进行概率预测，提取句子层面语义信息。而ERNIE对变压器缺陷文本进行动态语义编码优于BERT的原因主要有两方面，ERNIE通过知识增强的方法引入了实体、短语级别的多阶段掩码策略，学习到真实的缺陷文本语义关系，而BERT只能对字符进行掩码学习，两者掩码策略上的区别如图9所示。

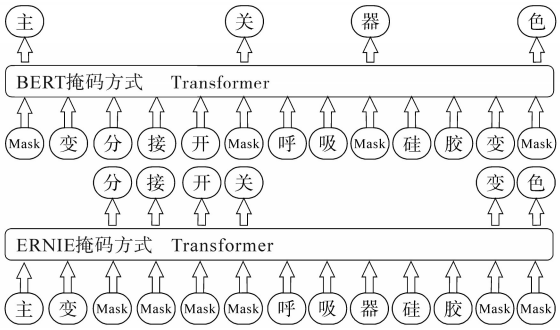

图9　ERNIE和BERT掩码策略的差异

Fig.9　The different masking strategy between ERNIE and BERT

从图9中可以看出，BERT通过学习局部的“变”和“开”字，能够成功预测被掩码的“主”和“关”字，但单一基于字符的掩码策略缺乏对实体单元及语义关系的建模学习，例如BERT无法建模“变色”的到底是“硅胶”还是“呼吸器”；ERNIE由于引入了短语和实体级别的语义知识，能够学习完整的实体单元，例如“分接开关”，同时能够建模“硅胶”和“变色”之间的语义关系，提高模型的可解释性。并且ERNIE是一种基于多种中文语料库训练出来的语言模型，训练语料的扩展不仅提高了ERNIE的语义表示能力，也显著提高了其处理缺陷文本时的泛化能力，使其在面对电力文本挖掘的不同应用场景时具备更好的鲁棒性。

3.3.2　不同缺陷文本分类模型性能对比

为了验证BiLSTM在缺陷文本分类问题上的优势，在文本分类阶段，选取了支持向量机（support vector machine，SVM）和KNN两种机器学习算法及卷积神经网络（convolutional neural networks，CNN）、循环神经网络（recurrent neural network，RNN）以及LSTM作为对比，实验结果如表2所示。

表2　不同文本分类模型结果统计

Tab.2　Statistical results of different text classification models

模型	*Acc*/%	*P*/%	*R*/%	F_1/%
SVM	54.13	46.52	44.87	45.68
KNN	59.24	57.27	60.11	58.66
CNN	87.27	83.16	86.39	84.74
RNN	88.39	85.43	89.29	87.32
LSTM	90.53	90.74	89.46	90.11
BiLSTM	94.33	93.98	94.25	94.09

从表2中可以看出，SVM和KNN两种机器学习算法分类性能要显著低于深度学习。传统机器学习方法依赖词向量的质量，而深度学习方法具有更强的泛化能力，能够自适应提取缺陷文本特征，提高分类准确率。

由于CNN使用固定窗口作为缺陷文本的上下文信息，因此只能学习到局部特征，卷积核大小对分类性能影响较大；RNN无法解决长距离依赖问题，在面对长文本时容易丢失重要语义信息。LSTM虽然克服了CNN和RNN的缺点，但忽略了未来信息和局部语义特征对缺陷文本的影响。BiLSTM通过双向编码学习融合了全局上下文的特征信息，为缺陷文本分类准确率提供双重保证。并且BiLSTM在解决长文本分类问题上优势会更为明显，具有良好的泛化性能，能应用于不同的电力文本挖掘问题，例如操作票、调度文本等长文本内容。

4　结论

针对缺陷文本的特点，本文提出一种基于ERNIE和BiLSTM融合的电力设备缺陷文本分类方法。ERNIE预训练语言模型通过多阶段知识遮蔽策略生成包含缺陷文本上下文信息的词向量，BiLSTM进一步提取缺陷文本的句义特征。所提模型能有效提升缺陷文本分类准确率，减少运维人员的工作量。此外，采用人工智能的技术能改善因经验差异导致的部分亚健康缺陷难以判断缺陷等级的情况，促进电网的智能化发展。

参考文献

[1] 王慧芳，曹靖，罗麟. 电力文本数据挖掘现状及挑战[J]. 浙江电力，2019，38（3）：1-7.
Wang Huifang，Cao Jing，Luo Lin. Current status and challenges of power text data mining[J]. Zhejiang Electric Power，2019，38（3）：1-7.

[2] 杨帆，邓一帆，李东东，等. 基于最优等级数的多维特征量开关柜健康状态评价方法[J]. 高电压技术，2020，46（11）：3934-3942.
Yang Fan，Deng Yifan，Li Dongdong，*et al*. Evaluation for switchgear health states based on multi-dimensional feature with optimal rank[J]. High Voltage Engineering，2020，46（11）：3934-3942.

[3] 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.

[4] 邱剑，王慧芳，应高亮，等. 文本信息挖掘技术及其在断路器全寿命状态评价中的应用[J]. 电力系统自动化，2016，40（6）：107-112，118.
Qiu Jian，Wang Huifang，Ying Gaoliang，*et al*. An access control protocol for intelligent electricity consumption information acquisition terminals[J]. Automation of Electric Power Systems，2016，40（6）：107-112，118.

[5] 杜修明，秦佳峰，郭诗瑶，等. 电力设备典型故障案例的文本挖掘[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.

[6] 蒋逸雯，李黎，李智威，等. 基于深度语义学习的电力变压器运维文本信息挖掘方法[J]. 中国电机工程学报，2019，39（14）：4162-4172.
Jiang Yiwen，Li Li，Li Zhiwei，*et al*. An information mining