DOI:10.12158/j.2096-3203.2024.06.015

基于字词混用集成模型的电力变压器缺陷记录文本挖掘方法

李元[1]，李睿[1]，林金山[1]，金凌峰[2]，邵先军[2]，张冠军[1]
（1. 西安交通大学电气工程学院，陕西 西安 710049；
2. 国网浙江省电力有限公司电力科学研究院，浙江 杭州 310014）

摘 要：变压器运维管理中积累了海量以文本形式记录的非结构化缺陷数据，但缺乏有效挖掘手段导致其利用率极低。文中提出一种基于字词混用集成模型的变压器缺陷记录文本挖掘方法，首先对变压器缺陷文本进行文本分词、去除停用词、文本增强、文本特征表示等预处理，以文本数学向量形式为输入，集成多个词汇级和字符级分类模型，通过元学习器对各基学习器性能的协同互补作用，实现变压器缺陷类型的准确识别和分类。与单一文本分类算法相比，该方法能够更全面地获得文本的语义特征，分类精确率达 91%，模型准确率和召回率的综合评价分数 $F_1=0.9$。将自然语言处理技术应用于电力设备缺陷记录文本，可以实现精准高效分类和故障识别，唤醒数据资源，显著提升电力变压器智能化管理水平。

关键词：电力变压器；自然语言处理；文本挖掘；故障诊断；集成学习；人工智能
中图分类号：TM8 **文献标志码**：A **文章编号**：2096-3203(2024)06-0153-10

0 引言

随着电网规模不断增大，智能化与数字化水平快速发展，积累的电网设备运行和状态监测数据呈几何式增长，数据量已超数百 TB[1]。在变压器运维管理过程中，设备信息、故障描述、故障部位等非结构化数据多以文字、数字及其混合文本形式记录，在电网设备大数据中占 80%，体量远超结构化数据[2-3]。但由于中文文本语义和结构的复杂性，非结构化数据结构往往不规则、不完整，难以采用关系型数据库表示，潜在信息无法直接挖掘，导致目前对电力设备运检数据利用率极低[4-5]。亟须发展专业领域自然语言处理技术，唤醒海量电力设备数据资源，挖掘电力设备缺陷文本信息，实现设备缺陷类型和健康状态自动化识别与评估。

针对文本挖掘问题，国内外已开展了大量研究，主要聚焦于文本分类、语义对比、观点提取。文献[6]基于卷积神经网络(convolutional neural network，CNN)模型建立文本分类卷积神经网络(text convolutional neural network，TextCNN)，奠定了利用深度学习算法提取词汇语义、实现文本挖掘的应用基础。近年来，研究者不断更新、优化深度学习算法，通过提取文本中的专业词汇语义特征进行句义识别。文献[7]将 CNN 应用于变压器设备缺陷文本分类，但仅对文本进行一次串行浅层特征提取，不能很好地挖掘长文本深层语义信息。文献[8]将循环神经网络(recurrent neural network，RNN)和 CNN 相结合，充分发挥 2 种算法的优势，更好地融合了上下文信息，但对长文本特征和重要语义信息的提取能力仍不足。文献[9]利用双向长短时记忆(bi-directional long short term memory，Bi-LSTM)神经网络解决了长文本特征记忆问题，文献[10]在此基础上引入注意力机制，使设备缺陷程度的分类准确率提高了 2.4%。

事实上，对于中文文本，字符是构成词汇的基本语义单位，字符和词汇的语义分析对中文文本的分类具有同样重要的意义[11]。仅注重词汇级特征而忽略字符级特征提取，可能会得到大相径庭的语义分析结果。若词汇特征提取中忽略“不”“否”“差”“未”等字符信息，则获得的状态语义甚至与真实语义相反。

为了解决上述问题，已有研究结合字符级和词汇级的文本特征对中文文本分类方法进行优化。文献[12]将基于位置和聚类特征的字符向量组合形成句子向量，优化文本特征表示方法，提高文本语义情感分析准确率。文献[13]综合与类别相关的重要词汇和字符注意力模型，并以一定权重合并为具有代表性的文本特征向量，实现信息语义单元的有效选择。文献[11]提高了歧义性字符的选择能力，以保证所选字符具有高度代表性，降低特征表示的冗余性，增强对词汇词义的补充和纠错能力。但由于电力领域文本的专业性和特殊性，目前仍缺乏对于专业性字符级语义特征的研究，亟须针对电力设备缺陷记录文本特点，实现字符级与词汇

收稿日期：2024-04-07；修回日期：2024-06-29
基金项目：国家自然科学基金资助项目(52107165)

级集成式特征提取，提高电力设备缺陷记录文本语义识别和故障分类准确率。

文中提出基于字词混用集成模型的电力变压器缺陷记录文本挖掘方法，以解决电力领域专业性字符级语义特征难提取、字符与词汇特征结合不足的问题。首先，以变压器状态评价导则[14]为指导，分析缺陷文本特点，并结合专家经验构建电力变压器缺陷记录文本类别；然后，对文本信息进行标注，通过文本分词、去除停用词、文本增强和文本特征表示等环节建立电力变压器缺陷记录文本数据集；最后，使用 Stacking 集成方法融合管理多个词汇级与字符级集成的文本分类模型，实现变压器缺陷记录文本的自动诊断和分类功能。在 Stacking 集成框架下，考虑多种词汇级与字符级文本分类模型的信息观测空间，提出多模型融合的变压器缺陷记录文本自动诊断和分类方法。

1 电力变压器缺陷记录文本特点及类别

1.1 缺陷记录文本特点

电力变压器缺陷记录文本是现场作业人员在设备检修系统中输入的实际数据。基于管理便捷性考虑，通常以文本形式详细记录变压器的缺陷情况，其中一般包含缺陷设备的名称、部件、现象及缺陷程度等。根据缺陷部件可分为本体、套管、冷却器系统、有载分接开关、无载分接开关和非电量保护及在线监测装置。不同缺陷部件的电力变压器缺陷记录文本实例如表 1 所示。

表 1 变压器缺陷记录文本实例

Table 1 Examples of transformer defect recording texts

缺陷部件	实例 1	实例 2
本体	#2 主变本体呼吸器硅胶大部分变色	4 号主变 A 相硅胶变色超 2/3
套管	#1 主变 110 kV 套管接头 A 相 46.9 度，B 相 107 度，C 相 49.2 度	#1 主变 110 kV 套管柱头发热，C 相 60 ℃，AB 相 44 ℃，负荷 220 A，环境温度 32 ℃
冷却器系统	#2 主变#5 风扇声响较大	#2 主变冷却器#11 风扇有异常卡涩声响
分接开关	#1 主变有载油位偏低	#2 主变有载呼吸器油位低于最低油位
非电量保护及在线监测装置	4 号主变 A 相绝缘油温度后台显示与现场不一致（后台-34 摄氏度，现场 41 摄氏度）	监控后台上#1 主变油温显示不正确，为负数

由实例可见，电力变压器缺陷记录文本具有以下特点。

（1）语义结构不固定。由于缺陷文本没有统一的书写标准，这些缺陷的描述通常具有个人语言习惯，难以利用固定的语义结构直接提取缺陷信息。表 1 中，2 个缺陷文本实例均反映套管的过热问题，实例 1 仅记录了“套管接头”三相温度差异显著，而实例 2 包含“套管柱头发热”字段。2 个实例对同一部位的名称不统一，语义省略和关键字也不同。

（2）不同缺陷类别的文本数量极不平衡。由于缺陷文本主要来源于现场人工巡视、调度监控等，多反映可观察的缺陷类别（如呼吸器硅胶变色、渗漏油等），而针对检测诊断性缺陷类别的信息很少（如绕组变形、局部放电等），不同缺陷类型的文本数量严重失衡。

（3）文本长度不统一。根据缺陷的复杂程度，现场检修人员对缺陷的描述详略不一。常见缺陷文本长度约二三十字，而复杂文本长度可达上百字。

（4）无效信息多。缺陷文本中普遍含有对缺陷文本诊断和分类没有意义的信息，需要进行清洗，如电站或线路名称、设备编号、班组名称等。

1.2 缺陷记录文本类别

由于变压器缺陷文本专业性较强，文中基于电力变压器状态评价导则[14]，并汲取电力专家经验，对缺陷文本语料进行深入分析，提炼文本中的缺陷类型关键信息，得到 102 种缺陷类别，如图 1 所示。

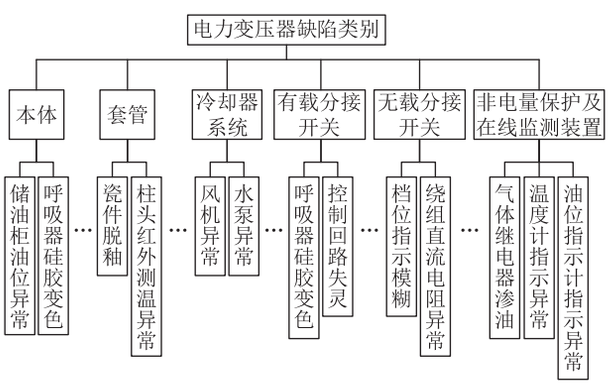

图 1 电力变压器缺陷记录文本类别

Fig.1 Power transformer defect recording text types

2 电力变压器缺陷记录文本预处理

电力变压器缺陷记录文本信息是典型的非结构化数据，文本数据预处理阶段需要去除语料中的无效信息，并通过文本增强来平衡缺陷类型数量，最终提取文本特征实现文本语料的向量化计算[15]。文本预处理包括文本分词、去除停用词、文本增强和文本特征表示 4 个步骤。

2.1 文本分词

中文文本词与词之间没有空格作为间隔，因此切分中文文本是文本挖掘的首要步骤[16]。建立完善的电力变压器缺陷文本词典是实现文本分词的

基础。考虑到变压器缺陷记录文本中含有较多电力领域的专业词汇,文中基于半监督学习建立变压器缺陷文本词典。首先基于自然语言公开通用词典进行初始分词,使用词频-逆文档词频[17](term frequency-inverse document frequency,TF-IDF)提取分词结果中的高频词汇,从而进行复查形成专业基础词库;再对文本语料进行重复分词和人工核查,直到实现较好的分词效果,得到优化的变压器缺陷记录文本词典。该文本词典包含厂站名称、缺陷定位、缺陷描述等关键信息,部分内容见表2。

表2 电力变压器缺陷记录文本词典(部分)

Table 2 Dictionary of power transformer defect recording text (partial)

类别	子类	词典内容
厂站名称	变电站名称	110 kV 城中变、220 kV 柯岩变、220 kV 渡东变等
	部件生产厂家	常州东芝变压器有限公司、浙江电力变压器有限公司、山东电工电气集团有限公司等
缺陷定位	缺陷部件	本体、开关、套管、冷却器等
	缺陷位置	硅胶、呼吸器、油位、电源等
缺陷描述	缺陷现象	变色、渗油、受潮等
	缺陷程度	正常、异常、告警等

2.2 去除停用词

需要过滤和去除电力变压器缺陷文本语料中存在的连接词、语气词、地(人)名等对于缺陷类型区分无意义的词和标点符号。基于分词结果,建立由无意义词构成的停用词表,遍历每条变压器缺陷文本,清洗文本语料中的无效信息。经过文本分词和去除停用词后,文本数据预处理示例效果如表3所示。

表3 电力变压器缺陷文本预处理示例

Table 3 Examples of power transformer defect recording text preprocessing

项目	内容
原文本	港头变主变#1 散热片与油箱连接处有渗油痕迹,地上有油迹
文本分词	港头变|主变|#1|散热片|与|油箱|连接处|有|渗油|痕迹|地|上|有|油迹
去除停用词	主变|散热片|油箱|连接处|渗油|痕迹|地|上|油迹

2.3 文本增强

针对变压器缺陷记录因缺陷类别不均衡导致分类模型结果出现过拟合的问题[18-19],文中采用easy data augmentation(EDA)[20]协同RoFormer-Sim[21]方法对变压器缺陷记录语料进行文本增强处理。EDA方法定义同义词为有最高相似度的词向量,并通过同义词随机替换或插入等操作扩充相似语义文本库,因此更适用于含较少关键词的短文本。而RoFomer-Sim方法更适合长文本语料,其通过对语料的关键词位置信息进行编码并训练,然后隐藏部分关键词,使模型自行扩充语句,实现相似句扩增。2种增强方法协同使用可以有效抑制模型过拟合,提升小样本数据集训练结果正确率,示例如表4所示。

表4 电力变压器缺陷文本增强示例

Table 4 Examples of power transformer defect recording text enhancement

示例	项目	内容
示例一	原文本	#2 主变#11 风扇声音异常
	EDA 增强	#2 主变#11 风扇装置声音故障
示例二	原文本	#1 主变 10 kV 套管渗漏油,平均一分钟 2 滴,现场油温油位正常,待备品到后消缺
	RoFormer-Sim 增强	#1 主变 10 kV 套管连接处渗漏故障,平均每分钟 2 滴,现场油温油位正常,待备品到后消缺

2.4 文本特征表示

文本特征表示是把语料的文本特征表示为特定的向量形式,即将非结构化数据转化为计算机可处理的结构化数据[22]。文中基于分布式表示方法,采用连续词袋(continuous bag-of-word,CBOW)模型,将中心词 x_i 的上下文 a 个邻近词汇 $C(x_i)=\{x_{i-a},x_{i-a+1},\cdots,x_{i+(a-1)},x_{i+a}\}$ 进行模型训练,得到 x_i 的词向量[23]。

假设 x_i 为该语句中的第 i 个词汇,上下文邻近词包含 C 个词汇,以语料"主变|油箱|渗油|痕迹"为例,如图2所示,选取 x_i 为"渗油",定义邻近词距离 a 为1,则上下邻近词 $C(x_i)$ 为{"油箱","痕迹"}。图2中,N 为训练样本个数;V 为输入词编码的维度;$\boldsymbol{W}'_{N\times V}$ 为权重矩阵。以每个邻近词的独热编码为输入,即 $[x_1\ x_2\ x_3\ x_4]=[0\ 1\ 0\ 1]$。首先构建前向网络中输入层 $\boldsymbol{x}$ 到隐藏层 $\boldsymbol{h}$ 的映射关系,如式(1)所示。

$$\boldsymbol{h}=\frac{1}{C}\boldsymbol{W}^{\mathrm{T}}(x_1+x_2+\cdots+x_C)=\frac{1}{C}(\boldsymbol{v}_{\omega 1}+\boldsymbol{v}_{\omega 2}+\cdots+\boldsymbol{v}_{\omega C})^{\mathrm{T}} \tag{1}$$

式中:$\boldsymbol{W}$ 为输入层到隐藏层的权值矩阵;$\boldsymbol{v}_{\omega i}$ 为第 ω 句话第 i 个词汇 x_i 的词向量;C 为上下文词汇数量。

然后通过最小化CBOW模型的损失函数,并遍历 x_i 从而实现权值向量 $\boldsymbol{W}$ 的迭代更新,即以"油箱""痕迹"为输入时,输出概率序列中"渗油"对

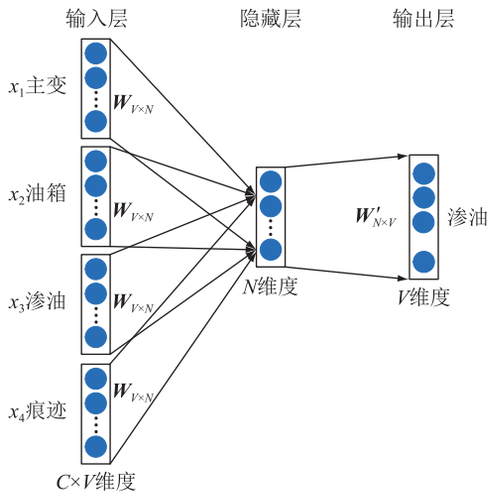

图 2 文本特征表示图解

Fig.2 Illustration of text feature representation

应概率的最大化，以最终权值向量 $\boldsymbol{v}_{\omega i}^{(\mathrm{new})}$ 为 x_i 的词向量表示。损失函数和权值更新分别如式(2)、式(3)所示。

$$E=-u_{j^*}+\lg\left(\sum_{j'=1}^{V}\exp(u_{j'})\right) \tag{2}$$

$$\boldsymbol{v}_{\omega i}^{(\mathrm{new})}=\boldsymbol{v}_{\omega i}^{(\mathrm{old})}-\frac{1}{C}\eta\boldsymbol{M}^{\mathrm{T}} \tag{3}$$

式中：E 为损失函数；u_{j^*}、$u_{j'}$ 分别为输出层第 j^* 个、第 j' 个输出；$\boldsymbol{v}_{\omega i}^{(\mathrm{old})}$ 为第 ω 句话第 i 个词汇 x_i 对应的输入层到隐藏层的权值向量，即 x_i 的词向量；η 为学习率；$\boldsymbol{M}$ 为输出节点对隐藏节点的导数。

经过以上文本预处理后，文本数据集可以转化为样本均衡且包含关键文本特征的向量化数学格式，方便被文本分类任务直接使用。

3 字词混用集成的文本分类模型

文中借鉴集成学习的思想，提出一种基于字词混用集成模型的电力变压器缺陷记录文本挖掘方法，将词汇级和字符级的多种文本分类算法通过 Stacking 集成学习模型结合在一起，以此获得比单一分类算法更优越的性能[24]。采用 5 种分类算法作为字词混用集成模型并行的第 1 层基学习器，即选取 TextCNN、文本循环神经网络（text recurrent neural network，TextRNN）、深度金字塔卷积神经网络（deep pyramid convolutional neural networks，DPCNN）、TextRNN+Attention 模型作为词汇级文本的分类算法，选取 Bi-directional Encoder Representations from Transformers（BERT）模型作为字符级文本的分类算法；在第 2 层选择泛化能力较强的随机森林（random forest，RF）算法作为元学习器，通过归纳并纠正基学习器对于训练集的偏置情况，得到最终的分类结果，模型架构如图 3 所示。

3.1 词汇级文本分类模型

由于变压器缺陷记录文本长度不一、关键信息数量多、专业词汇相关性强等特点，词汇级文本分类算法选取 TextCNN、TextRNN、DPCNN 和 TextRNN+Attention 模型。以下简述 4 种词汇级文本分类算法的原理。

3.1.1 TextCNN 文本分类

TextCNN 的结构简单、计算速度快，适用于数据特征维度高的短句文本分类问题[6]。对一组输入文本向量 $\boldsymbol{x}_{1:n}=\boldsymbol{x}_1\oplus\boldsymbol{x}_2\oplus\cdots\oplus\boldsymbol{x}_n$，其中⊕为拼接操作符，利用 k 个窗口、大小为 h 的卷积核 $\boldsymbol{w}\in\mathbb{R}^{h\times k}$ 对输

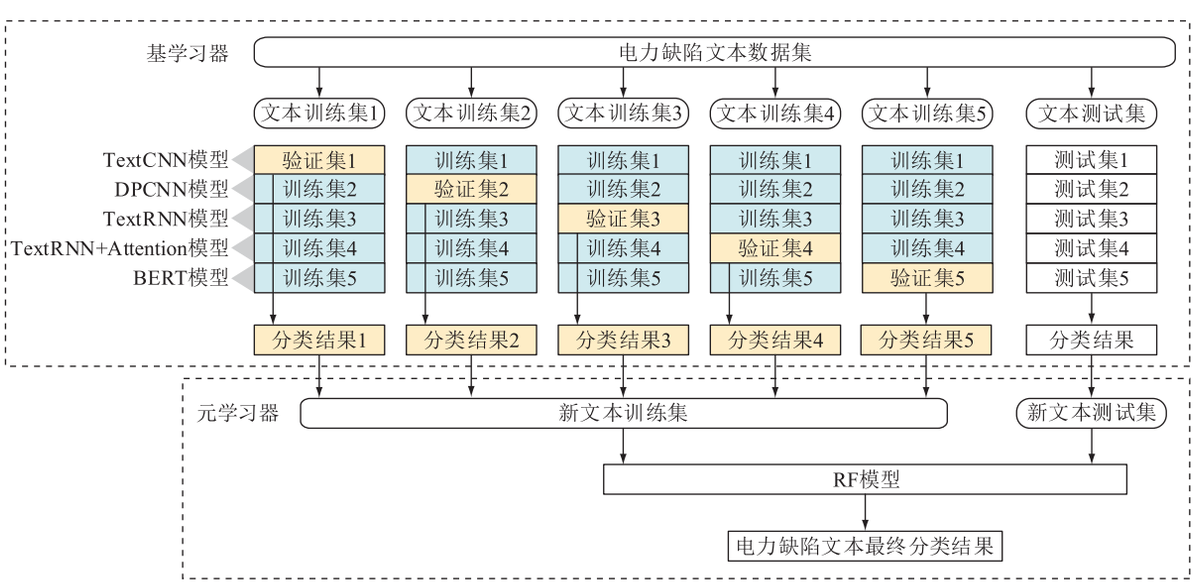

图 3 字词混用集成模型的电力变压器缺陷记录文本挖掘方法

Fig.3 Power transformer defect recording text mining using character-word ensemble learning model

入文本向同一方向依次进行卷积计算,即对 $\boldsymbol{x}_{1:h}$、$\boldsymbol{x}_{2:h+1}$、…、$\boldsymbol{x}_{n-h+1:n}$分别计算特征值 $c_i(i=1,2,\cdots,n)$,从而得到特征图 $\boldsymbol{c}$,如式(4)、式(5)所示。

$$c_i = f(\boldsymbol{w}\cdot\boldsymbol{x}_{i:i+h-1} + b) \tag{4}$$

$$\boldsymbol{c} = [c_1 \quad c_2 \quad \cdots \quad c_{n-h+1}] \tag{5}$$

式中:$f(\cdot)$为非线性函数;$b\in\mathbb{R}$,为偏置项。

将特征图 $\boldsymbol{c}$ 输入最大池化层,将每个卷积核的最大特征图标量 $\hat{c}=\max c_i$ 作为句子特征,再将多个特征拼接输入全连接层进行类别预测。

3.1.2　TextRNN 文本分类

TextRNN 因其循环结构能够捕获更长的序列信息,可以被用于处理较长文本[25]。将输入文本词向量逐一送入 TextRNN 单元,最后的输出可以表示整个序列,再将其输入非线性层即可得到分类结果。训练该 TextRNN 模型的参数使预测分布和实际结果分布的交叉熵最小化误差值 L,如式(6)所示。

$$L(\hat{y}_{\omega j}, y_{\omega j}) = -\sum_{\omega=1}^{NG}\sum_{j=1}^{G} y_{\omega j}\lg(\hat{y}_{\omega j}) \tag{6}$$

式中:$y_{\omega j}$ 为第 ω 句话第 j 个先验知识标签;$\hat{y}_{\omega j}$ 为第 ω 句话第 j 个标签预测概率;G 为类别数量。

3.1.3　DPCNN 文本分类

DPCNN 根据最大句子长度增加神经网络深度,能在不过多增加计算成本的前提下提高分类准确率,可处理更长文本[26]。该算法的本质是在 TextCNN 的基础上对输入的文本向量重复 2 次等长卷积计算,如式(7)所示。DPCNN 采用残差网络向下采样,每次循环可在保留文本信息特征的同时使得卷积核维度减半,从而克服了重复卷积计算引起的词级数量倍增问题。最终所有卷积核数据聚合为一个向量,输入到全连接层实现类别预测。

$$\boldsymbol{R} = f_{\mathrm{Relu}}(\boldsymbol{W}_{\boldsymbol{x}}) + \boldsymbol{W}_{\mathrm{b}} \tag{7}$$

式中:$\boldsymbol{R}$ 为输出;$f_{\mathrm{Relu}}(\cdot)$为 Relu 激活函数;$\boldsymbol{W}_{\boldsymbol{x}}$ 为本层权重向量;$\boldsymbol{W}_{\mathrm{b}}$ 为调节项。

3.1.4　TextRNN+Attention 文本分类

TextRNN+Attention 是在 TextRNN 的基础上引入注意力机制,可直观反映每个关键词对分类结果的贡献,能在提升模型效果的同时增强分类结果的可解释性[27]。具体地,通过计算每个词汇 $\boldsymbol{u}_{\omega i}$ 与其上下文词汇 $\boldsymbol{u}_{\omega}$ 的相似度,通过 Softmax 函数获得该词归一化的重要性权重 $\boldsymbol{\alpha}_{\omega i}$。然后利用词向量及其 $\boldsymbol{\alpha}_{\omega i}$,可合并为句向量 $\boldsymbol{s}_{\omega}$,并输入全连接层实现类别预测,如式(8)、式(9)所示。

$$\boldsymbol{\alpha}_{\omega i} = \frac{\exp(\boldsymbol{u}_{\omega i}^{\mathrm{T}}\boldsymbol{u}_{\omega})}{\sum_{i}\exp(\boldsymbol{u}_{\omega i}^{\mathrm{T}}\boldsymbol{u}_{\omega})} \tag{8}$$

$$\boldsymbol{s}_{\omega} = \sum_{i}\boldsymbol{\alpha}_{\omega i}\boldsymbol{h}_{\omega i} \tag{9}$$

式中:$\boldsymbol{u}_{\omega i}$ 为第 ω 句话的第 i 个词汇的隐含观测量;$\boldsymbol{u}_{\omega}$ 为上下文词汇隐含观测量;$\boldsymbol{h}_{\omega i}$ 为隐藏层权重。

3.2　字符级文本分类模型

由于单个字符(如“未”“不”等)可能对文本语义的理解结果产生重要影响,而词汇级模型对单字的分类能力不足,所以文中还采用了以字为单位的字符级文本分类模型 BERT。相比于传统单向语言模型,BERT 模型能够融合左右双向的上下文信息,实现深层语言表征,从而更全面地理解输入文本的语义信息[28-29],其算法结构如图 4 所示。其中,$\boldsymbol{E}_{\mathrm{CLS}}$ 为句首标记符;$\boldsymbol{E}_{\mathrm{SEP}}$ 为句末标记符;$\boldsymbol{T}_{\mathrm{SEP}}$ 为特殊分割符[SEP]在 BERT 模型中的初态;$\boldsymbol{T}_i$ 为特殊标记符[CLS]在 BERT 模型中的第 i 态。

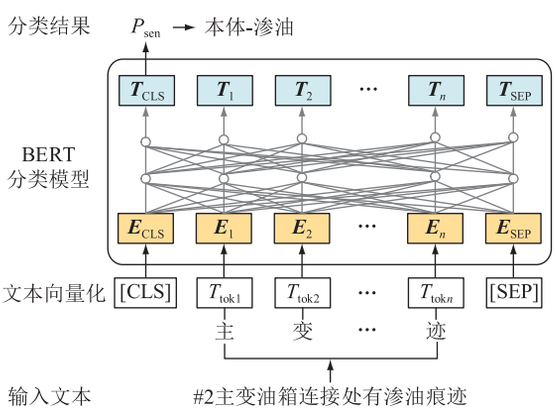

图 4　BERT 字符级分类模型原理

Fig.4　Principle of BERT character-level classification

该模型以单条语句 $T_{\mathrm{tok}i}(i=1,2,\cdots,n)$ 作为输入,在句首添加特殊标记符[CLS],在句尾添加特殊分割符[SEP];将输入文本进行向量化表示为 $\boldsymbol{E}_i$;将固定长度的字符串输入到编码层,自下而上传递计算,特殊标记符[CLS]对应位置输出向量 $\boldsymbol{T}_{\mathrm{CLS}}$;输入该向量到全连接层,按式(10)计算输出 P_{sen},即模型最终类别预测结果。

$$P_{\mathrm{sen}} = \mathrm{Softmax}(\boldsymbol{T}_{\mathrm{CLS}}\boldsymbol{W}^{\mathrm{T}}) \tag{10}$$

3.3　字词混用集成的文本挖掘方法

字词混用集成的文本挖掘方法由两级学习器构成,如图 3 所示,第一级包含 4 个词汇级文本分类基学习器和 1 个字符级文本分类基学习器,即 $\{\Theta_1,\Theta_2,\Theta_3,\Theta_4,\Theta_5\}$,第二级由 1 个元学习器 Θ 构成。文本的准确分类/预测需要建立在字词混用集成模型充分训练的基础上。文中提出的字词混用集成模型的训练过程如下:

(1) 将变压器缺陷文本数据集随机分为包含 A 个样本的训练集 S_{train} 和 B 个样本的测试集 S_{test};为

提高分类模型的泛化能力，将原始的训练集 S_{train} 进行 K 折划分，得到 K 个子集 $\{S_1, S_2, \cdots, S_K\}$。

（2）对第一级中的各分类器，将每个子集 S_p（$p=1,2,\cdots,K$）作为一次验证集，其余子集作为训练集，得到 K 个分类子集 L_p，L_p 为 S_p 中样本分类后的故障类型编码；对测试集 S_{test} 进行类别预测，得到测试子集 T_m（$m=1,2,\cdots,5$），T_m 为 S_{test} 中样本分类后的故障类型编码。

（3）将 K 个分类子集合并成一列，得到第 t 个基学习器对训练集 S_{train} 中所有样本的分类结果集合 E_t（$t=1,2,\cdots,5$）。

（4）针对每一个基分类器 Θ_m 分别进行上述操作，得到 5 个基学习器的分类结果集合 $E=\{E_1, E_2, E_3, E_4, E_5\}$；将测试集分类结果 T_m 组合，得到测试集分类结果集合 $T=\{T_1, T_2, T_3, T_4, T_5\}$。

（5）将新的训练集 E 和新的测试集 T 输入第二级元学习器 Θ 进行训练，得到最终的分类结果。

这种训练方式可以保证每个基学习器测试的数据集均未参加该学习器的训练，使得所有数据在训练或验证模型时仅使用一次，可有效避免过度拟合。核心算法原理流程如图 5 所示。

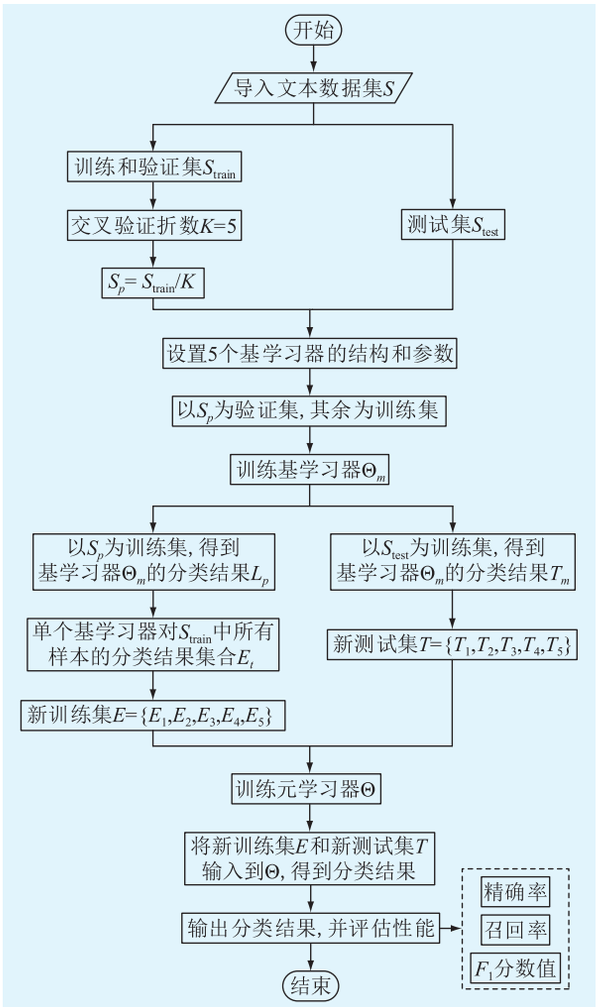

图 5　字词混用集成模型算法原理流程

Fig.5　Process of algorithmic principle of character-word level ensemble integrated model

4　模型实测验证与分析

文中选取某电网公司 2000 年 8 月—2021 年 9 月期间变压器检修产生的 26 237 条变压器缺陷文本作为研究对象，以 10∶1的比例划分为训练集和测试集，测试和验证了提出的字词混用集成文本分类模型，并与常用单一文本分类模型进行性能比较。

4.1　数据集特征分析

4.1.1　词汇级与字符级分词特征

词汇级分词是根据词典将文本语料拆分为单个词汇，拆分后的词汇可反映设备缺陷特征。字符级分词是将文本语料拆分为单个字符。如“主变终端非电量无异常信号”可基于词汇分词为“非电量”“异常”等，基于字符分词为“终”“端”“无”等。字符级分词是对词汇级分词的有益补充，避免部分特殊字符缺失（如“无”）带来的歧义。

4.1.2　词向量的二维映射

电力变压器缺陷文本经预处理后，得到的分词会形成一组高维的词向量，且词汇在文本中距离越近，词向量的相似度越高[30]。为了更直观地观察词向量特征，利用 t-随机邻域嵌入（t-distributed stochastic neighbor embedding，t-SNE）算法对获得的维度为 500 的词向量进行降维，绘制部分典型词向量，如图 6 所示。

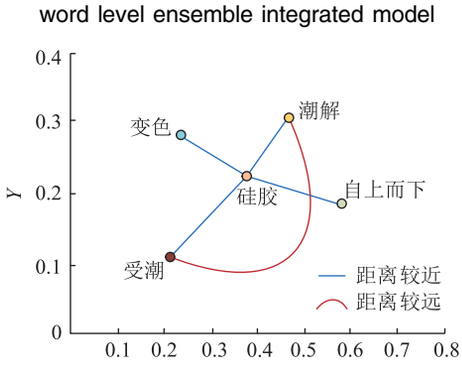

图 6　词向量降维示例

Fig.6　Illustration of dimension-reduction word vectors

可以看到，降维后的“变色”“潮解”“受潮”等词与“硅胶”在二维坐标中的距离较近，说明这些词在缺陷文本里常在“硅胶”的上下文中同时出现。例如，缺陷文本存在大量“硅胶变色”的描述，因此“硅胶”和“变色”应具有较高关联性，而“受潮”和“潮解”由于含义相近，很难会出现在同一文本中，但又因为与“硅胶”存在一定关联，因此 2 个词向量

在二维坐标中与“硅胶”距离稍远。

4.1.3　高频词分析

电力变压器缺陷文本是现场作业人员对缺陷现象的描述,符合实际工程经验,包含缺陷部位、缺陷现象等关键词。将根据模型分词结果得到的变压器缺陷文本的关键词及词频统计列于表 5,变压器缺陷文本高频词示意如图 7 所示。

表 5　电力变压器缺陷文本高频词统计

Table 5　Statistics of high-frequency words in the power transformer defect text

关键词	词频	关键词	词频
硅胶	6 301	正常	2 413
变色	5 576	冷却器	2 364
呼吸器	5 532	故障	2 225
渗油	3 702	装置	2 209
油温	2 806	油位	2 107
后台	2 789	在线	1 913
显示	2 710	异常	1 910
本体	2 638	温度	1 744
检查	2 610	瓦斯	1 729
开关	2 535	电源	1 690

图 7　电力变压器缺陷文本高频词示意

Fig.7　Schematic diagram of high-frequency words in the power transformer defect text

高频词统计在一定程度上反映了变压器易发缺陷类型和缺陷文本特点:① 受限于现场试验条件,具有明显外观特征的缺陷类型记录较多,如呼吸器硅胶变色、渗油等;② 关于易检测和检查频率较高的缺陷类型的记录也较多,如红外测温超标、监测装置故障等;③ 对于设备状态的描述不一致,但常用词含义交叠,如异常、故障、告警等。

4.2　字词混用的深度学习集成模型性能分析

4.2.1　模型性能分析

为了验证字词混用集成模型文本分类的性能,采用 Adam 优化器,设置批处理样本数为 512,自适应调节学习率,并将该模型算法与各单一文本分类算法进行对比,所有算法均使用相同的训练集与测试集,文本特征均采用 Word2Vec 分布式词向量表示,算法性能测试结果见表 6。

表 6　不同算法的性能比较

Table 6　Performance comparison of different algorithms

算法	精确率 P/%	召回率 R/%	F_1 分数值
DPCNN	65	67	0.660
TextCNN	75	74	0.745
TextRNN	66	65	0.655
TextRNN+Attention	71	72	0.715
BERT	85	87	0.860
文中模型	91	89	0.900

采用的模型评价指标有精确率 P、召回率 R 和 F_1 分数值,其中 F_1 分数为模型的整体性能表征[31]。评价指标的计算公式为:

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{11}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{12}$$

$$F_1 = \frac{2PR}{P + R} \tag{13}$$

式中:T_P 为模型正确预测的正例数量;F_P 为真实值为反例但被错误预测为正例的数量;F_N 为真实值为正例但被错误预测为反例的数量。

可以看出,基于字词混用深度集成模型的测试集准确率为 91%,F_1 分数值为 0.9,各个单分类模型中分类效果最好的模型是字符级的 BERT 模型,$F_1 = 0.86$。基于字词混用深度集成模型的评价结果要高于其他单一文本分类模型。这是由于字词混用集成模型不仅可以综合不同算法模型的输出结果,更好地捕捉词汇级和字符级的文本特征,而且有效避免了单个文本分类模型可能出现的过拟合问题。

需要指出,文中提出的分类模型对计算机环境和性能要求很低,算法可以根据损失值自适应调节学习率,模型的收敛速度较快(训练时长小于 2 h,应用响应小于 30 s)。

4.2.2　实例分析

文中提出的字词混用集成模型能够有效提取文本中的词汇与字符信息。分别采用词汇级文本分类模型、字符级文本分类模型和字词混用集成文本分类模型,对同一例变压器缺陷记录文本进行分类,将其分类过程及结果可视化,如表 7 所示。下划线词汇和加点字符分别表示词汇定位和字符选择的可视化结果。

表 7　不同算法的实例分析

Table 7　Example analysis of different algorithms

算法模型	文本分类过程可视化	文本分类结果
词汇级文本分类模型	凤鸣变|#2|主变|本体|绕组|温度计|整定|不正确	非电量保护装置-温度计-指示异常
字符级文本分类模型	凤鸣变|#2|主变|本体|绕组|温度计整定|不正确	本体-红外测温-异常
字词混用文本分类模型（文中模型）	凤鸣变|#2|主变|本体|绕组|温度计整定|不正确	非电量保护装置-温度计-指示异常

观察表 7 可知，词汇级文本分类模型可以捕捉到与缺陷文本类型密切相关、包含关键信息的词汇，如“本体”“温度计”等；字符级文本分类模型可以选择出有区别的字符特征，通过细微的局部差异来区分文本类型，例如“不”具有否定含义，与“正确”搭配则会指向完全相反的结果；2 个不同层次的分类模型有不同的关注点，并且具有一定的互补性，可以提高变压器缺陷文本分类的准确率。该实例联合使用词汇级文本分类模型的关键词“温度计”“正确”和字符级文本分类模型的关键字“不”，可以得到“温度计指示异常”的正确分类结果。

将文中模型应用于 26 237 条变压器缺陷文本批量分类处理，并与词汇级文本分类模型、字符级文本分类模型的分类结果进行比较，得到部分高频缺陷类型分类分布结果，如图 8 所示。

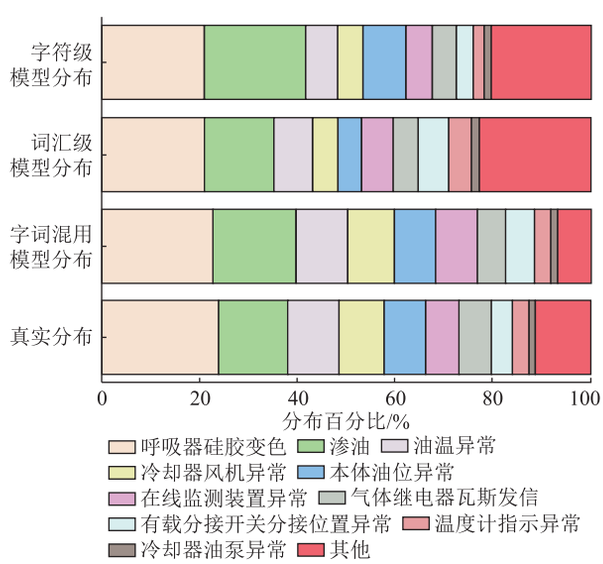

图 8　高频缺陷类型分类分布

Fig.8　Distribution of high-frequency defect type classification

相较于词汇级和字符级文本分类模型，文中模型在保证处理速度的同时（小于 1 h），分类结果的类型分布上与真实值更接近，为设备缺陷差异化处理提供算法和模型支持。

5　结论

针对电力设备海量运检缺陷文本信息无法被有效利用，现有记录文本挖掘技术多关注词汇级语义而忽视字符级语义对文本语义的关键影响等问题，文中提出一种基于字词混用集成模型的电力变压器缺陷记录文本挖掘方法，主要结论如下：

（1）基于字词混用集成模型的电力变压器缺陷记录文本挖掘方法的基学习器由 4 个词汇级和 1 个字符级文本分类模型构成，通过元学习器对基学习器进行集成，充分利用不同算法对文本数据空间与特征的提取能力，获得最优的类别预测结果；

（2）字词混用集成的文本挖掘方法实现了变压器故障类型的精准分类，分类结果的准确率为 91%，模型整体 F_1 分数值高达 0.9，显著优于对比试验中的单一文本分类模型；

（3）字词混用集成的文本挖掘方法提取了字符级和词汇级语义特征，在语义表达上协同互补，可对指向的变压器缺陷类型实现快速准确分类，显著提高非结构化数据利用率，挖掘海量文本信息价值。

文中提出的方法可在设备定检、临检中发挥数据先行、服务保障作用，实现缺陷类型快速准确分批分类处理，较传统文本检索方法节省了大量人力，还可推广到调度运行、客服记录等文本语料的智能化应用。

参考文献：

[1] 江秀臣，盛戈皞. 电力设备状态大数据分析的研究和应用[J]. 高电压技术，2018，44(4)：1041-1050.
JIANG Xiuchen，SHENG Gehao. Research and application of big data analysis of power equipment condition[J]. High Voltage Engineering，2018，44(4)：1041-1050.

[2] 张中文，吐松江·卡日，张紫薇，等. 基于双分支特征融合的电力设备缺陷文本挖掘方法[J]. 高压电器，2024，60(6)：188-196.
ZHANG Zhongwen，TUSONGJIANG Kari，ZHANG Ziwei，et al. Text mining method for power equipment defects based on two-branch feature fusion[J]. High Voltage Apparatus，2024，60(6)：188-196.

[3] 张磐，郑悦，李海龙，等. 配电网电力设备缺陷文本智能辨识运维综述[J]. 电力建设，2022，43(5)：90-99.
ZHANG Pan，ZHENG Yue，LI Hailong，et al. Overview on intelligent text identification and maintenance of power equipment defects in distribution network[J]. Electric Power Construction，2022，43(5)：90-99.

[4] 王宏刚，纪鑫，武同心，等. 基于预训练语言模型的电力领域设备缺陷检测[J]. 电测与仪表，2022，59(5)：180-186.
WANG Honggang，JI Xin，WU Tongxin，et al. Device defect detection in power field based on pre-trained language model[J].

Electrical Measurement & Instrumentation, 2022, 59(5): 180-186.

[5] 李铁成,任江波,刘清泉,等. 基于深度学习的智能录波器配置数据自动化映射方法[J]. 电测与仪表,2022,59(9):76-83.
LI Tiecheng, REN Jiangbo, LIU Qingquan, et al. Automatic mapping method of intelligent recorder configuration datasets based on deep learning[J]. Electrical Measurement & Instrumentation,2022,59(9):76-83.

[6] KIM Y. Convolutional neural networks for sentence classification [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics,2014:1746-1751.

[7] 刘梓权. 电力设备非结构化数据挖掘的应用研究[D]. 杭州:浙江大学,2020.
LIU Ziquan. Research on applications of unstructured data mining in power equipment [D]. Hangzhou: Zhejiang University,2020.

[8] 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 method of power transformer operation and maintenance texts based on deep semantic learning[J]. Proceedings of the CSEE, 2019,39(14):4162-4172.

[9] 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.

[10] 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.

[11] QIAO X, PENG C, LIU Z, et al. Word-character attention model for Chinese text classification[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(12): 3521-3537.

[12] YANG J X, LYU Q J, GAO S, et al. Review aspect extraction based on character-enhanced embedding models [C]//2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC). Beijing, China. IEEE, 2016: 219-223.

[13] ZHOU Y J, XU J M, CAO J, et al. Hybrid attention networks for Chinese short text classification[J]. Computación y Sistemas, 2018, 21(4).

[14] 国家能源局. 油浸式变压器(电抗器)状态评价导则: DL/T 1685—2017[S]. 北京:中国电力出版社,2017.
National Energy Administration. Guide for condition evaluation of oil-immmersed power transformers (reactors): DL/T 1685-2017[S]. Beijing: China Electric Power Press, 2017.

[15] 王宣军,于虹,祁兵,等. 基于注意力机制的混合神经网络电力设备缺陷文本挖掘方法[J]. 电力信息与通信技术,2023,21(9):44-51.
WANG Xuanjun, YU Hong, QI Bing, et al. Hybrid neural network text mining method for power equipment defects based on attention mechanism[J]. Electric Power Information and Communication Technology, 2023, 21(9): 44-51.

[16] LI X Y, MENG Y X, SUN X F, et al. Is word segmentation necessary for deep learning of Chinese representations[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 3242-3252.

[17] 叶雪梅,毛雪岷,夏锦春,等. 文本分类 TF-IDF 算法的改进研究[J]. 计算机工程与应用,2019,55(2):104-109,161.
YE Xuemei, MAO Xuemin, XIA Jinchun, et al. Improved approach to TF-IDF algorithm in text classification[J]. Computer Engineering and Applications, 2019, 55(2): 104-109, 161.

[18] 王绪亮,顾媛丽,张鸿儒,等. 基于知识集成流形的电力设备缺陷文本数据增强方法与应用研究[J]. 电网技术,2024,48(4):1690-1702.
WANG Xuliang, GU Yuanli, ZHANG Hongru, et al. Data augmentation and application of defect texts for power equipment based on knowledge integration manifold [J]. Power System Technology, 2024, 48(4): 1690-1702.

[19] 邓雯瀚,苗宇,许逵,等. 电力变压器状态评估方法的应用及展望[J]. 智慧电力,2023,51(10):93-102.
DENG Wenhan, MIAO Yu, XU Kui, et al. Application and prospect of condition assessment method of power transformer[J]. Smart Power, 2023, 51(10): 93-102.

[20] WEI J, ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks [EB/OL]. [2023-12-29]. https://arxiv.org/abs/1901.11196v2.

[21] SU J L, LU Y, PAN S F, et al. RoFormer: enhanced transformer with rotary position embedding[EB/OL]. [2023-12-29]. https://arxiv.org/abs/2104.09864v5.

[22] 曹靖. 基于文本挖掘技术的电力设备缺陷分析[D]. 杭州:浙江大学,2020.
CAO Jing. Defect analysis of power equipment based on text mining technology [D]. Hangzhou: Zhejiang University, 2020.

[23] MIKOLOVT, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[EB/OL]. [2023-12-29]. https://arxiv.org/abs/1310.4546v1.

[24] 史佳琪,张建华. 基于多模型融合 Stacking 集成学习方式的负荷预测方法[J]. 中国电机工程学报,2019,39(14):4032-4042.
SHI Jiaqi, ZHANG Jianhua. Load forecasting based on multi-model by Stacking ensemble learning[J]. Proceedings of the CSEE, 2019, 39(14): 4032-4042.

[25] LIUP F, QIU X P, HUANG X J, et al. Recurrent neural net-

work for text classification with multi-task learning[EB/OL]. [2023-12-29]. https://arxiv.org/abs/1605.05101v1.

[26] JOHNSONR, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017:562-570.

[27] YANG Z C, YANG D Y, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016:1480-1489.

[28] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2023-12-30]. https://arxiv.org/abs/1810.048-05v2.

[29] 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.

[30] LAI S W, LIU K, HE S Z, et al. How to generate a good word embedding[J]. IEEE Intelligent Systems, 2016, 31(6): 5-14.

[31] 谢庆,蔡扬,谢军,等. 基于 ALBERT 的电力变压器运维知识图谱构建方法与应用研究[J]. 电工技术学报, 2023, 38(1):95-106.
XIE Qing, CAI Yang, XIE Jun, et al. Research on construction method and application of knowledge graph for power transformer operation and maintenance based on ALBERT [J]. Transactions of China Electrotechnical Society, 2023, 38(1): 95-106.

作者简介:


李元

李元(1984),男,博士,副教授,研究方向为高电压试验技术、电力设备状态感知与智能诊断(E-mail:liyuan8490@xjtu.edu.cn);

李睿(1998),女,硕士在读,研究方向为电力设备智能感知与运维;

林金山(1998),男,硕士在读,研究方向为电力装备故障诊断与状态评估。

Character-word level ensemble integrated model for power transformer defect recording text mining method

LI Yuan[1], LI Rui[1], LIN Jinshan[1], JIN Lingfeng[2], SHAO Xianjun[2], ZHANG Guanjun[1]

(1. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. State Grid Zhejiang Electric Power Co., Ltd. Research Institute, Hangzhou 310014, China)

Abstract: The operation and maintenance management of transformers has accumulated a large amount of unstructured defect recording data in the form of text. However, the lack of effective mining method has led to an extremely low utilization rate. A text mining method for transformer defect recording text based on a character-word level ensemble integrated model is proposed in this paper. Firstly, the transformer defect recording texts are preprocessed with text segmentation, stop word removal, text augmentation, and text feature representation to convert the data into mathematical vectors for input. By integrating multiple word- and character-level classification models, the method can realize accurate identification and classification of transformer defect types through the synergistic and complementary effects of meta-learners on the individual base learners. Compared to single-text classification algorithms, this method can obtain the semantic features of the text more comprehensively, achieving a classification precision of 91% and F_1 score of 0.9, which is the comprehensive evaluation score for model precision and recall. By applying natural language processing technology to precise power equipment defect recoding text classification and efficient fault recognition, data resources are awakened, and the intelligent management level of power transformers is significantly improved.

Keywords: power transformer; natural language processing; text mining; fault diagnosis; ensemble learning; artificial intelligence

(编辑 陆海霞)