基于 K-BERT 的测井文本分类方法研究

曹茂俊,肖 阳
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:在石油勘探与开发领域,测井文本数据的处理和分类是提高测井数据解读效率和准确性的关键环节。然而,测井文本中包含大量专业术语和复杂的数据结构,使得传统的文本分类技术在面对专业领域数据时效果有限,难以满足实际应用需求。为了解决这一问题,该文提出了一种改进的 K-BERT 文本分类方法。该方法结合了 K-BERT 模型和 TextCNN 的文本特征提取能力。K-BERT 通过引入测井领域的知识图谱,将领域知识嵌入模型中,增强了模型对专业术语和复杂语义的理解能力,从而提升了模型在专业领域文本分类中的语义捕捉效果。而 TextCNN 利用卷积神经网络的特性,能够有效提取文本的局部特征,捕捉文本细节信息,进一步提升分类的精度与鲁棒性。两者的结合为测井文本的分类提供了一种创新的解决方案。通过实验对比分析,该方法在宏精确率、宏召回率及宏 F1 值等指标上均优于传统文本分类模型,验证了其在专业领域文本分类中的有效性和优越性。

关键词:K-BERT;TextCNN;测井文本;文本分类;测井知识图谱
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2025)05-0197-08
doi:10.20165/j.cnki.ISSN1673-629X.2024.0390

Research on Logging Text Classification Method Based on K-BERT

CAO Mao-jun,XIAO Yang
(School of Computer and Information Technology,Northeast Petroleum University,Daqing 163318,China)

Abstract:In the field of petroleum exploration and development,the processing and classification of well logging text data are crucial steps for enhancing the efficiency and accuracy of well logging data interpretation. However,well logging texts contain a plethora of professional terminology and complex data structures,which limit the effectiveness of traditional text classification techniques when dealing with domain-specific data,thus failing to meet practical application requirements. To address this issue,we propose an improved K-BERT text classification method. This method integrates the text feature extraction capabilities of the K-BERT model and TextCNN. By incorporating a knowledge graph specific to the well logging domain,K-BERT embeds domain knowledge into the model,enhancing its understanding of professional terminology and complex semantics,and improving the model's semantic capture performance in domain-specific text classification. On the other hand,TextCNN leverages the characteristics of convolutional neural networks to effectively extract local features of texts and capture detailed textual information,further enhancing classification accuracy and robustness. The combination of these two techniques provides an innovative solution for the classification of well logging texts. Experimental comparisons demonstrate that the proposed method outperforms traditional text classification models in terms of macro precision,macro recall,and macro F1 score,validating its effectiveness and superiority in domain-specific text classification.

Key words:K-BERT;TextCNN;logging text;text classification;logging knowledge graph

0 引 言

测井技术是石油勘探开发过程中不可或缺的重要技术手段。该技术通过准确记录地下地质层的物理化学性质,可以提供与石油天然气潜力密切相关的岩石类型、孔隙度、含水量等详细信息。这些信息对于储层评价至关重要,并为后续钻井和生产决策提供了科学依据。随着全球能源需求持续增长和勘探技术不断进步,勘探活动范围和深度也显著扩大。相应地,测井过程产生的数据量也急剧增加。这些数据不仅庞大复杂,还包含众多地质、物理和化学参数。处理分析这些海量数据对于提高石油天然气勘探效率和准确性至关重要,但也带来了巨大挑战。在这种情况下,文本分类技术的应用尤为重要。文本分类技术是一种先进信息处理方法,它通过自动分类技术快速有效地将大量文

收稿日期:2024-09-06 **修回日期**:2025-01-08
基金项目:中石油科技术开发项目(2021DJ4001);黑龙江省建设项目(YJSKCSZ_202309);黑龙江省高等教育教学改革项目(SJGY20220253)
作者简介:曹茂俊(1978-),男,副教授,博士,研究方向为深度学习、智能计算在测井中的应用;通信作者:肖 阳(1998-),男,硕士生,研究方向为自然语言处理、知识图谱。

本数据划分到预设类别中。文本分类技术有助于测井专家快速提取有用信息,并帮助他们识别复杂数据中关键模式趋势以做出更准确的地质解释决策。

文本分类的发展经历了专家规则、传统机器学习和深度学习三个主要阶段。文本分类最初是通过专家规则(Pattern)[1]进行,利用知识工程建立专家系统,这样做的好处是比较直观地解决了问题,但费时费力,覆盖的范围和精确率都有限。在这之后,人们试图使用机器学习方法来自动化进行文本分类[2],支持向量机(Support Vector Machines, SVM)[3]和随机森林(Random Forest)[4]等机器学习方法已被广泛应用于文本分类任务。近年来,深度学习技术已经成为大数据与人工智能领域的研究热点[5]。与传统机器学习模型相比,深度学习模型能够深入对文本进行分析和理解,从而实现文本分类。Kim 等人[6]提出一种用于文本分类的卷积神经网络模型 TextCNN,可以在一定程度上避免梯度消失的问题,而且在处理短文本和固定长度文本时表现良好。Devlin 等人[7]提出了 BERT 模型,该模型是一种基于 Transformer 网络的预训练模型,可用于自然语言处理任务,如文本分类、语言推断等。然而,当处理专业领域文本时,这些模型表现出了一定的局限性,尤其是在理解领域内专业知识方面的不足,导致其分类效果有限。为了解决这一问题,Liu 等人[8]提出了 K-BERT 模型,该模型通过引入知识图谱(Knowledge Graph,KG),能够在理解文本时获得额外的上下文信息,这对于处理专业术语和领域特定概念尤其有效,增强了模型对文本含义的理解。

近年来,也有一些研究尝试将 BERT 这类基于 Transformer 的预训练语言模型与 TextCNN 等传统卷积神经网络模型进行结合,以期获得性能上的突破,进一步提升文本分类的精度和效率。例如,张淦等人[9]提在智能制造成熟度评估任务中,通过融合 BERT 与 TextCNN 模型,显著提高了分类的准确性,展示了这种结合策略的有效性。类似地,杨忠霖等人[10]在电信网络诈骗案情文本分类问题上,也设计了一种 BERT 与 TextCNN 相结合的模型,该模型首先对 BERT 进行微调以适应特定任务,再与 TextCNN 结合,从而提高了分类的精度和鲁棒性。此外,万铮等人[11]和鲍彤等人[12]分别针对中文新闻分类和农业问句分类任务,提出了各自的 BERT 与 TextCNN 结合模型,这些研究不仅验证了融合模型在不同应用场景下的有效性,也进一步丰富了 BERT 与 TextCNN 结合策略的研究内涵。同时,谢佩君等人[13]提出了 BERT-TextCNN-Highway 模型,该模型在提升分类性能的同时,还通过引入 Highway 网络增强了系统的稳定性,为 BERT 与 TextCNN 的结合提供了新的思路和方法。这些研究成果充分展示了 BERT 这类基于 Transformer 的预训练语言模型与 TextCNN 相结合在文本分类任务中的广泛适用性和优异表现。

测井属于专业领域,测井文本分类研究十分有限,通用的分类技术可以直接应用于分类,但存在以下问题:一是领域特定语言和术语问题。测井领域有丰富的领域特定语言和术语,这些语言和术语可能难以被通用模型理解,从而降低了文本分类精度。二是背景知识问题。测井领域涉及测井技术、测井解释、测井分析和历史背景等方面的知识。这些知识可能对于模型是未知的,需要特殊的处理来识别和理解。三是文本复杂性的问题。测井领域的文本非常复杂,包含大量专业术语,这需要模型识别和理解的能力。

该文在上述工作的基础上,针对以上方法存在的问题,通过构建了一个专门针对测井领域的知识图谱与 K-BERT 模型相结合,并利用 TextCNN 对关键词和局部特征更加敏感的优势,设计了一种改进 K-BERT 的分类模型(K-BERT-TextCNN)。主要贡献和创新点如下:

(1)引入了一个针对测井领域的知识图谱,通过使用知识表达的 K-BERT 模型生成语料的词向量,解决了在测井文本数据上多样化词向量编码空间不一致和语句偏离核心语义的问题。由于融合了专业领域的知识图谱,使得该模型可以更准确地进行测井领域的文本分类。

(2)提出了一个结合 K-BERT 和 TextCNN 的测井文本分类模型,该模型能够有效地提取并利用文本中的特征。它通过融合 K-BERT 模型的深层语义信息和 TextCNN 的局部特征,提高模型对测井数据的理解和分析能力。模型在保持文本特征细粒度的同时,还避免了传统深度学习模型可能会遇到的梯度消失和特征损失问题。

(3)通过一系列实验验证了该方法在测井文本分类任务上的有效性和优越性。

1 相关工作

1.1 K-BERT 模型

K-BERT 模型是一种融合知识图谱的模型,旨在提高机器理解自然语言的能力。其模型的体系结构主要由四个模块组成,即知识模块、嵌入模块、视图模块和掩码变换器模块,模型总体框架如图 1 所示。

1.1.1 知识模块

知识模块是 K-BERT 的核心组成部分,负责引入和整合外部知识。在这一层中,知识图谱被用作信息的来源,以增强模型对文本中隐含知识点的理解,该层的任务是将知识图谱的实体和关系有效地映射到文本

的词汇中,形成一个包含背景知识的句子树(Sentence Tree)。这个过程可以分为知识查询(K-Query)和知识注入(K-Inject)。知识查询针对句子中的所有实体,都查询一遍知识图谱,而知识注入是将查询到的知识图谱三元组嵌入到句子中的合适位置上形成句子树。

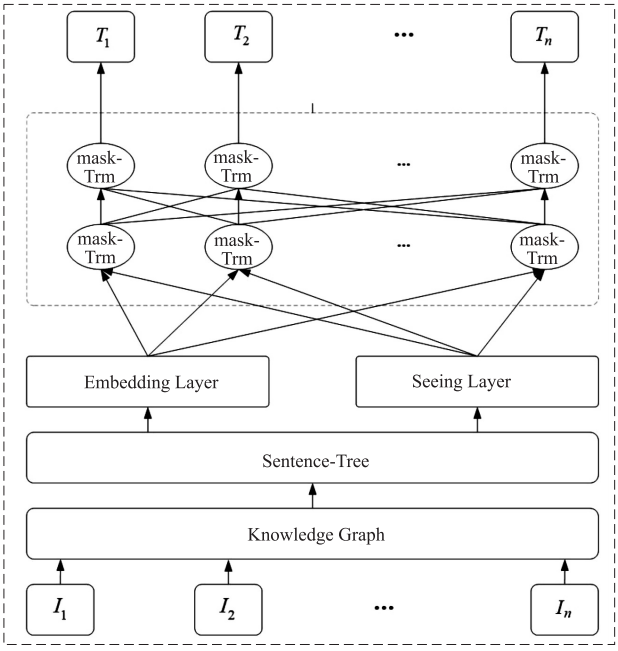

图 1　K-BERT 模型框架

1.1.2　嵌入模块

嵌入模块包括标记嵌入(token embedding)、位置嵌入(position embedding)和段嵌入(segment embedding)三部分,其作用是将文本中的词汇转换为向量形式,这些向量能够捕捉单词的语义信息。嵌入模块不仅包括传统的词嵌入,还包括从知识模块得到的实体嵌入。每个词汇不仅有其原始的语义表示,还包含了与之相关的知识信息。

1.1.3　视图模块

视图模块允许模型在不违反原始 BERT 架构的情况下,观察并处理插入的知识实体。该层通过生成一个可见矩阵 $\boldsymbol{M}$,利用"软位置关系"技术,使得模型能够在不改变原始文本顺序的前提下,处理知识增强的文本序列。$\boldsymbol{M}$ 定义如式 1 所示。

$$\boldsymbol{M}_{ij}=\begin{cases}0, & w_i, w_j \text{ 相互可见}\\ -\infty, & w_i, w_j \text{ 相互不可见}\end{cases} \tag{1}$$

其中,$\boldsymbol{M}_{ij}$ 表示一个矩阵,其元素关联到权重 w_i 和 w_j;w_i 和 w_j 分别代表与硬位置索引 i 和 j 相关的权重;当 $\boldsymbol{M}_{ij}=0$ 时,表明 w_i 和 w_j 相互可见;当 $\boldsymbol{M}_{ij}=-\infty$ 时,表明 w_i 和 w_j 相互不可见。

1.1.4　掩码变换器模块

掩码变换器模块基于 Transformer Encoder,是 BERT 模型的关键组成部分。在 K-BERT 中,掩码变换器进行了特殊的调整来处理额外加入的知识实体。通过引入掩码机制,模型能在训练时忽略知识实体,从而在不泄露答案的情况下进行自监督学习。其核心思想是让一个词的词嵌入只来源于其同一个枝干的上下文,而不同枝干的词之间相互不影响。掩码变换器模块由 12 层掩码自注意力模块堆叠而成,自注意力的定义如式 2~4 所示。

$$Q^{i+1}, K^{i+1}, V^{i+1}=h^i\boldsymbol{W}_q, h^i\boldsymbol{W}_k, h^i\boldsymbol{W}_v \tag{2}$$

$$S^{i+1}=\mathrm{softmax}\left(\frac{Q^{i+1}K^{i+1}+\boldsymbol{M}}{\sqrt{d_k}}\right) \tag{3}$$

$$h^{i+1}=S^{i+1}V^{i+1} \tag{4}$$

其中,$\boldsymbol{W}_q$,$\boldsymbol{W}_k$,$\boldsymbol{W}_v$ 是模型需要学习的矩阵向量参数;h^i 是隐状态的第 i 个自注意力模块;d_k 是缩放因子,用于控制训练过程中的梯度稳定性;$\boldsymbol{M}$ 为可见矩阵。

句子树、嵌入模块、视图模块和可见矩阵是 K-BERT 处理的关键技术,四者之间的关系如图 2 所示。从知识模块得到句子树后,对句子树同时构建可视化矩阵和送入嵌入模块编码,这两个过程得到的信息归并后输入到掩码自注意力中进行计算。

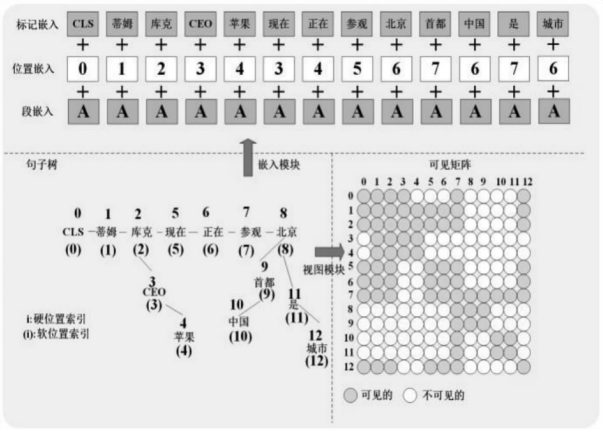

图 2　K-BERT 处理的关键技术

1.2　卷积神经网络

卷积神经网络[14](Convolutional Neural Networks,CNN)是一种在计算机视觉领域取得了巨大成功的深度学习模型,设计灵感来自于生物学中的视觉系统,旨在模拟人类视觉处理的方式。在过去的几年中,CNN 也运用在文本分析任务上,包括情感分析、文本分类等。

在传统的神经网络架构中,输入层的每个神经元均与隐藏层中的神经元全连接。相比之下,卷积神经网络(CNN)采用的是稀疏连接策略,其中输入层神经元的一个小子集仅与隐藏层的相应神经元相连,构成所谓的"局部感受野"。此设计基于图像处理领域的观察,即空间邻近的像素点在视觉上往往具有更高的相关性,此现象在文本处理中亦有相似的表现。因此,CNN 的初级阶段专注于局部信息的提取,随后将这些

局部特征集成以形成全局的特征表示。

与传统神经网络类似,CNN 亦涵盖权重与偏置参数,这些参数在学习过程中通过训练样本进行适应性调整。不同之处在于,CNN 的隐藏层在不同区域采用相同的权重和偏置,即所谓的“参数共享”机制。此机制使网络具备对输入数据的平移不变性。

在神经元的激活阶段,激活函数用于转换每个神经元的输出。修正线性单元(ReLU)是常用的激活函数[15],其作用是将神经元的输出映射到非负值,即正值保持不变,而负值归零,这一过程称为“单侧抑制”。有了这单侧抑制才使得神经网络中的神经元具备了稀疏激活性[16]。

TextCNN 模型是一种将 CNN 应用到文本分类任务的分类模型,与传统图像的 CNN 网络相比,TextCNN 在网络结构上没有任何变化,该模型最大优势在于其参数数量少、训练速度快、网络结构简单、计算量适中,其模型结构如图 3 所示。

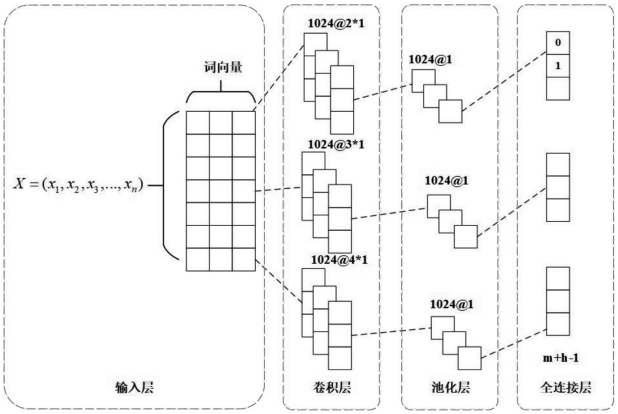

图 3 TextCNN 模型结构

1.3 测井知识图谱

测井技术在油气勘探和生产中具有重要的作用,它通过测量井孔中的地层参数,为地质学家和工程师提供了有关地层岩石类型、含油气性质和储层特征的信息。随着测井数据的爆炸性增长和复杂性增加,如何有效地组织、管理和利用这些数据成为一个重要的挑战。

近年来,知识图谱技术在信息管理和知识表示领域引起了广泛关注。知识图谱是一种以图形形式表示知识的结构化数据模型,它包含实体、属性和实体之间的关系。测井知识图谱是基于测井领域的专业知识构建的一种知识表示模型,旨在整合和表达测井领域中的知识。

测井知识图谱涉及三个核心组成部分:实体、关系和属性。实体代表测井领域中的关键概念和对象,例如测井曲线、岩石类型、流体性质等。这些实体是测井知识图谱的基本单元,用于表示领域内的具体事物或概念。关系描述了实体之间的关联,如测井曲线与岩石类型之间的关系、测井参数与流体性质之间的关系等。这些关系反映了实体之间的相互作用和依赖,是测井知识图谱中信息流动和推理的基础。属性为实体和关系提供了更详细的描述和特征信息。例如,测井曲线的测量单位、岩石类型的物理特征等。属性进一步丰富了测井知识图谱的内容,使其能够更准确地表示领域知识。

本研究所使用的测井知识图谱主要依托大庆测试服务分公司、中国石油勘探开发研究院等权威机构提供的测井解释产生的多源异构数据进行构建。该图谱涵盖了测井领域的多个关键方面,包括测井方法、岩石类型、流体性质、测井参数等。通过构建测井知识图谱,将测井领域的知识以结构化的方式组织起来,便于进行查询、推理和分析。

在测井知识图谱的架构中,每个实体均被精准地赋予了一个独一无二的标识符,确保了其在整个知识网络中的唯一可识别性。例如,“自然伽马测井”是一个实体,它与“伽马射线强度”属性相关联,并通过“反映”关系与“岩石类型”实体相连。

测井知识图谱的这种结构化表示方法,不仅极大地提升了信息的组织效率和表达的清晰度,还为处理复杂查询、执行高效推理以及进行深入分析提供了坚实的基础。借助图谱的查询功能,用户可以轻松追踪特定实体及其相关的属性和关系,迅速获取所需信息。而图谱的推理机制则能够基于现有知识推导出新的结论,为测井领域的决策制定和问题解决提供有力支持。

1.4 模型微调

模型微调是指在预训练阶段基础上,将模型进一步训练以适应具体任务的过程。该文将测井领域的多标签文本数据集作为训练集和测试集进行输入,并引入测井知识图谱,根据损失函数和评价指标对模型进行训练和调优。模型微调时需要用到的交叉熵损失函数如式 5 所示。

$$L_{\log}(Y,P) = -\log \Pr(Y \mid P) = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k}\log p_{i,k} \tag{5}$$

其中,$y_{i,k}$ 为样本 i 的真实标签;$p_{i,k}$ 是第 i 个样本预测为第 k 个标签值的概率。

2 基于 K-BERT 的测井文本分类模型

基于 K-BERT 对核心语义的良好获取能力,以及 TextCNN 对信息的序列特征提取能力,该文提出了测井文本分类模型 K-BERT-TextCNN。

K-BERT-TextCNN 模型如图 4 所示,分为 K-BERT 词嵌入表示层、TextCNN 特征提取层和 Softmax 分类层。K-BERT 词嵌入表示层将背景知识添加到

语料库中并生成词向量;TextCNN 特征提取层对 K-BERT 提取的词向量进行深层、浅层和上下文特征提取;Softmax 作为分类层,用于获得句子的分类。

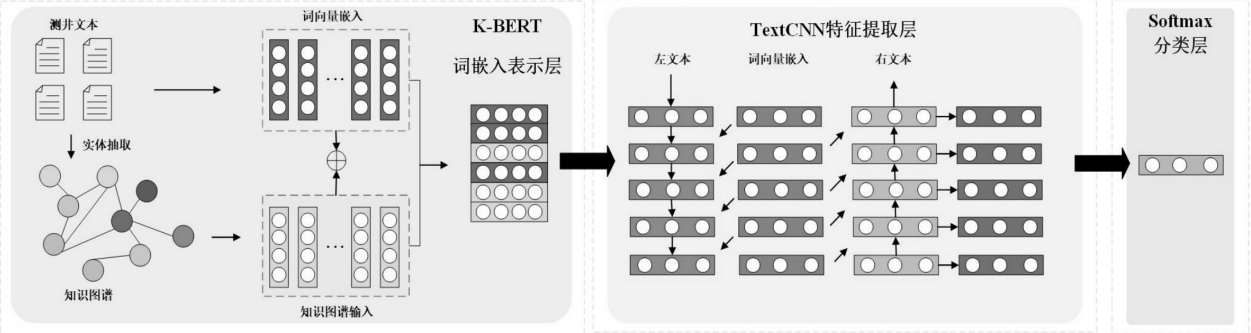

图 4　K-BERT-TextCNN 模型框架

2.1　K-BERT 词嵌入表示层

K-BERT 词嵌入表示层的主要功能是将自然语言文本转化为神经网络可处理的特征向量。传统词嵌入方法,如 Word2Vec[17] 和 GloVe[18],为每个单词分配一个静态的向量表示,但这些方法未能有效应对多义词的语义多样性。相比之下,K-BERT 模型采取了一种上下文感知的策略,通过结合单词的语境以及相关的背景知识,提供了一种增强的语义表示。这种方法在处理输入的文本时,能够引入额外的背景知识,最终生成的词向量更为丰富,包含了特定于测井领域的信息,从而能更准确地捕捉到词义的细微差别。

2.2　TextCNN 特征提取层

TextCNN 特征提取层旨在全面提取词向量中的序列信息,解决 K-BERT 模型生成的动态词向量在序列信息方面存在不足等问题。模型架构分为四层:首层为输入层,负责接收文本序列并将其转化为词嵌入表示,其中每个词语对应一个独特向量,并依照原文本序列顺序排列成矩阵形式。次层为卷积层,通过不同尺寸的卷积核处理前述矩阵,以提取文本的局部特征。第三层是池化层,主要功能是降低特征图的维数并捕获关键特征,通过最大池化策略从滑动窗口中选取最显著特征,并将这些特征拼接成一维向量。最后一层为全连接层,它将池化层的输出连接到一个或多个密集层,这对于学习不同特征之间的相互关系和执行最终分类任务至关重要。

2.3　Softmax 分类层

经过词嵌入和特征提取处理后的测井文本信息使用 Softmax 层进行测井文本分类。Softmax 函数为每个输出分类的结果均赋值一个概率,表示每个类别的可能性,Softmax 函数的定义如式 6 所示。

$$\mathrm{Softmax}(Z_i)=\frac{\mathrm{e}^{Z_i}}{\sum_{c=1}^{C}\mathrm{e}^{Z_c}} \tag{6}$$

其中,$\mathrm{Softmax}(Z_i)$ 是当前元素与所有元素的比值,即当前元素 i 的概率;Z_i 为第 i 个节点的输出值;C 为输出节点的个数。

3　实验结果及分析

3.1　数据集

实验数据主要采用大庆油田的真实测井文本数据集以及测井知识图谱,一共包括 10 000 条测井文本,一共 7 个分类类别:地球物理测井技术、油田地质与勘探、油田类型与特点、储集层分析与评价、中国石油工业发展历史、技术应用与案例分析和非测井知识,标签设置为 0 ~ 6。通过文本长度统计分析可以看出文本较短,最长文本长度为 120。同时,按 8 ∶ 2 的比例将标注数据集划分为训练集和测试集。

3.2　实验环境与超参数设置

使用阿里云平台(www. aliyun. com)进行模型训练以及测试,框架基于 Pytorch[19],具体实验环境设置如表 1 所示。

表 1　实验环境设置

相关名称	配置
操作系统	Ubuntu-20. 04. 1
CPU	Hygon C86 3285 8-core Processor
内存	32 GiB
GPU	NVIDIA V100
语言	Python3. 8
框架	Pytorch2. 0. 1-CUDA11. 7. 1

经过多次实验之后,选取最优的超参数配置,如表 2 所示。

表 2　模型超参数设置

参数	参数含义	值
pad_size	序列长度	128
learning_rate	学习率	2e-5
batch_size	批量大小	128
hidden_size	隐藏层神经元个数	768
filter_sizes	卷积核大小	3, 4, 5
dropout	dropout 参数	0. 5

3.3 评价指标

实验的评价指标分别是宏精确率(macro-P)、宏召回率(macro-R)和宏 F1 值(macro-F1),表示对精确率、召回率和 F1 值的算术平均数,各指标的计算公式如下:

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

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

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R} \tag{9}$$

$$P_{\mathrm{macro}} = \frac{1}{k}\sum_{i=1}^{k} P_i \tag{10}$$

$$R_{\mathrm{macro}} = \frac{1}{k}\sum_{i=1}^{k} R_i \tag{11}$$

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{F1}_i \tag{12}$$

其中,TP 表示被预测为正类的正样本;FP 表示被预测为正类的负样本;FN 表示被预测为负类的正样本;P_i、R_i 和 $\mathrm{F1}_i$ 分表示第 i 类的精确率、召回率和 F1 值。

3.4 对比实验设计

为了研究文中模型的有效性,选择了现在主流的文本分类模型:CNN、RCNN(RNN-CNN)[20]、MacBERT[21]和 RoBERTa[22],以及其他知识增强型预训练语言模型:KG-BERT[23]和 DKPLM[24]进行对比实验。

3.5 实验结果与分析

将文中模型与其他模型在测井文本数据集上进行对比,采用宏精确率、宏召回率和宏 F1 值作为评判指标,实验结果如表 3 所示。表 3 列出了多种文本分类模型的宏精确率、宏召回率和宏 F1 值。这些指标反映了模型在不同类别上的整体性能,从而提供了对比不同模型在处理测井文本分类任务时的效果的清晰概览。从表中可以看出,各种模型在处理测井文本分类任务时的表现存在显著差异。

表 3 实验结果

Model	macro-P	macro-R	macro-F1
CNN	0.798	0.792	0.795
RCNN(RNN-CNN)	0.800	0.784	0.792
MacBERT	0.917	0.824	0.868
RoBERTa	0.916	0.823	0.867
KG-BERT	0.859	0.773	0.814
DKPLM	0.912	0.823	0.865
K-BERT-TextCNN	0.929	0.829	0.876

表 4 展示了模型在测井文本数据集中针对 7 个具体类别的分类性能评估结果,采用宏精确率、宏召回率和宏 F1 值作为评判指标。这些详细的指标反映了模型在每个特定类别上的表现,从而提供了对模型分类能力的深入洞察。通过对比不同类别的指标,可以进一步了解模型在处理具体测井文本分类任务时的优势和局限性。从表中可以看出,不同类别之间的分类性能存在差异。

根据表 3 所呈现的实验结果,该文对多种文本分类算法在测井文本数据集上的综合性能进行了严格对比,采用的评价标准包括宏精确率、宏召回率以及宏 F1 值。在此框架下,该文创新性构建的 K-BERT-TextCNN 模型展现出了显著优越的分类效能,具体而言,其宏精确率高达 0.929,宏召回率为 0.829,宏 F1 值更是达到了 0.876,这一系列数据均明显优于其他参与比较的模型。

为了更深入地剖析 K-BERT-TextCNN 模型在测井文本分类任务中的具体表现,进一步对该模型在测井文本数据集中针对 7 个具体类别上的分类性能进行了详尽的评估与分析(如表 4 所示)。评估结果表明,K-BERT-TextCNN 模型在各个类别上的宏精确率、宏召回率和宏 F1 值均呈现出高度的一致性和稳定性。然而,值得注意的是,不同类别间的分类性能仍存在一定的差异性。特别是在“地球物理测井技术”和“中国石油工业发展历史”等专业性较强的类别上,模型展现出了极高的分类准确性和稳定性,宏精确率、宏召回率和宏 F1 值均达到了较高水平。相比之下,在“非测井知识”这一相对宽泛且内容多样的类别上,尽管模型的宏精确率依然保持较高水平,但宏召回率却略显逊色,这可能是由于该类别文本的复杂性和多样性给模型的识别带来了更大的挑战。

表 4 实验结果(按类别)

类别编号	类别名称	macro-P	macro-R	macro-F1
0	地球物理测井技术	0.941	0.862	0.904
1	油田地质与勘探	0.925	0.813	0.860
2	油田类型与特点	0.915	0.833	0.879
3	储集层分析与评价	0.930	0.801	0.866
4	中国石油工业发展历史	0.952	0.852	0.861
5	技术应用与案例分析	0.907	0.820	0.861
6	非测井知识	0.901	0.782	0.842

3.6 消融实验

为了深入探究并验证 K-BERT-TextCNN 模型中各个关键组成部分的有效性以及它们对整体分类性能的贡献程度,该文构建了以下四种模型变体,并进行了全面、系统的对比实验。

(1)K-BERT 模型:此变体仅保留了 K-BERT 部

分,移除了 TextCNN 组件,旨在单独评估 K-BERT 的性能及其测井知识图谱的增强效果。

(2)TextCNN 模型:此变体仅保留了 TextCNN 部分,去除了 K-BERT 组件,目的是评估 TextCNN 在测井文本分类任务中的独立表现能力。

(3)BERT 模型:此变体为原始的 BERT 模型,既去除了 TextCNN 组件,也未融入测井知识图谱,用以评估基础的 BERT 模型在测井文本分类任务中的基准性能。

(4)BERT-TextCNN 模型:此变体将 BERT 模型与 TextCNN 模型相结合,但去除了测井知识图谱,旨在评估这种组合的性能是否接近或达到 K-BERT-TextCNN 的水平。

在实验实施过程中,该文严格控制了除模型结构变化之外的所有其他实验条件,以确保实验结果的准确性和可比较性。

表 5 详细列出了各模型的宏精确率、宏召回率和宏 F1 值。

表 5　消融实验结果

Model	macro-P	macro-R	macro-F1
K-BERT-TextCNN	0.929	0.829	0.876
K-BERT	0.919	0.826	0.870
TextCNN	0.820	0.810	0.815
BERT	0.916	0.823	0.867
BERT-TextCNN	0.920	0.825	0.871

实验结果表明,K-BERT-TextCNN 模型在性能上显著优于其他所有模型变体,验证了模型中各个组件的有效性,进一步证明了这些组件之间的协同作用对于提升模型性能的重要性。测井知识图谱的融入有效增强了 BERT 模型的语义理解能力,而 TextCNN 的加入则显著提升了模型对文本特征的捕捉能力。两者的有机结合,使得 K-BERT-TextCNN 模型在测井文本分类任务中展现出了卓越的性能。

综上所述,提出的 K-BERT-TextCNN 模型在测井文本分类领域取得了显著成效,其分类精度达到了最优水平,充分展现了该模型在实际应用场景中的广阔前景。通过创新性地将测井领域知识图谱与 BERT 预训练模型相结合,并融合 TextCNN 的多尺度卷积特征提取技术,有效地利用了测井知识图谱的丰富语义信息和深度学习模型的强大特征学习能力,为测井文本分类任务构建了一个高效且精准的解决方案。未来的研究可进一步深化测井知识图谱的构建与优化,提升其知识覆盖度和准确性,同时探索更为先进的文本特征表示方法和分类算法,以进一步提升测井文本分类的性能。

4　结束语

针对中文测井文本数据,该文旨在提高文本分类模型在这一特定领域的性能。测井文本通常包含丰富的专业术语和复杂的上下文语义,传统的文本分类方法难以充分捕捉这些特征。因此,提出了一种融合 K-BERT 与 TextCNN 模型的创新型分类方法,即 K-BERT-TextCNN 模型。该模型利用 K-BERT 预训练模型整合测井知识图谱的信息,以增强模型对测井相关内容的上下文语义表示能力,进一步通过 TextCNN 提取上下文相关特征,提升对文本语义的理解和分类能力。

实验结果表明,该模型在多个指标上均优于现有的文本分类方法,展示了其在测井文本分类任务中的显著优势。未来的研究可以在知识图谱扩展、多模态数据融合、模型可解释性和在线学习等方面继续探索,以进一步提升模型性能和应用价值。通过不断优化和创新,K-BERT-TextCNN 模型有望在测井领域以及其他垂直领域的文本分类任务中发挥重要作用。尽管 K-BERT-TextCNN 模型在本研究中取得了显著成果,但仍有一些值得进一步探索的研究方向:

(1)测井知识图谱的扩展与优化:进一步丰富和优化测井领域的知识图谱,提升其覆盖范围和精确性,从而增强模型的语义表示能力。

(2)多模态数据融合:在测井作业中,文本数据往往与其他类型的数据(如图像、传感器数据等)密切相关。未来可以探索将多模态数据融合到模型中,以提升整体分类性能。

(3)模型的可解释性研究:提高模型的可解释性,使得分类结果更加透明和易于理解,帮助工程师更好地利用分类结果进行决策。

(4)在线学习与自适应调整:开发能够在线学习和自适应调整的模型,使其能够实时适应新数据和新环境,保持分类性能的持续提升。

参考文献:

[1] SCOTT S,MATWIN S. Feature engineering for text classification[C]//Proceedings of the sixteenth international conference on machine learning (ICML´99). San Francisco: Morgan Kaufmann Publishers Inc.,1999:379-388.

[2] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM Comput. Surv.,2002,34(1):1-47.

[3] CORTES C,VAPNIK V. Support-vector networks[J]. Machine Learning,1995,20(3):273-297.

[4] BREIMAN L. Random forests[J]. Machine Learning,2001,45:5-32.

[5] SILVER D,HUANG A,MADDISON C J,et al. Mastering the game of Go with deep neural networks and tree search

[J]. Nature,2016,529:484-489.
[6] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha: Association for Computational Linguistics,2014:1746-1751.
[7] DEVLIN J,CHANG M W,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:human language technologies, volume 1. Minneapolis: Association for Computational Linguistics,2019:4171-4186.
[8] LIU W,ZHOU P,ZHAO Z,et al. K-BERT:enabling language representation with knowledge graph[C]//Proceedings of the AAAI conference on artificial intelligence. [s. l.]:AAAI,2020:2901-2908.
[9] 张 淦,袁堂晓,汪惠芬,等. 基于 BERT 和 TextCNN 的智能制造成熟度评估方法[J]. 计算机集成制造系统,2024,30(3):852-863.
[10] 杨忠霖,顾益军. 一种基于 BERT 微调-TextCNN 的电信网络诈骗案情文本分类设计[J]. 电子测试,2023(3):47-53.
[11] 万 铮,王 芳,黄树成. 基于权重词向量与改进 TextCNN 的中文新闻分类[J]. 软件导刊,2023,22(9):59-64.
[12] 鲍 彤,罗 瑞,郭 婷,等. 基于 BERT 字向量和 TextCNN 的农业问句分类模型分析[J]. 南方农业学报,2022,53(7):2068-2076.
[13] 谢佩君,迟呈英,战学刚. 基于改进的 TextCNN 模型的中文文本分类系统[J]. It 经理世界,2021(10):125-126.
[14] HADJI I,WILDES R P. What do we understand about convolutional networks? [J]. arXiv. 1803. 08834,2018.
[15] 张有健,陈 晨,王再见. 深度学习算法的激活函数研究[J]. 无线电通信技术,2021,47(1):115-120.
[16] 何 静,程 涛,黄良辉,等. 深度可分离卷积神经网络在自动分拣中的应用[J]. 包装学报,2018,10(6):33-40.
[17] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301. 3781,2013.
[18] PENNINGTON J,SOCHER R,MANNING C. GloVe:global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha:Association for Computational Linguistics,2014:1532-1543.
[19] PASZKE A,GROSS S,CHINTALA S,et al. Automatic differentiation in PyTorch[J/OL]. 2017[2024-08-24]. https://api. semanticscholar. org/CorpusID:40027675.
[20] ZHONG Z,SUN L,HUO Q. An anchor-free region proposal network for Faster R-CNN-based text detection approaches [J]. International Journal on Document Analysis and Recognition,2018,22:315-327.
[21] CUI Yiming,CHE Wanxiang,LIU Ting,et al. Revisiting pre-trained models for Chinese natural language processing [C]//Findings of the association for computational linguistics: EMNLP 2020. Online: Association for Computational Linguistics,2020:657-668.
[22] LIU Y,OTT M,GOYAL N,et al. RoBERTa:a robustly optimized BERT pretraining approach[J]. arXiv:1907. 11692, 2019.
[23] YAO L,MAO C,LUO Y. KG-BERT:BERT for knowledge graph completion[J]. arXiv:1909. 03193,2019.
[24] ZHANG T,WANG C,HU N,et al. DKPLM:decomposable knowledge-enhanced pre-trained language model for natural language understanding[C]//Proceedings of the AAAI conference on artificial intelligence. [s. l.]:AAAI,2022:11703-11711.