基于 RoBERTa-wwm 与 BiLSTM 模型的铁路数据分类分级方法研究

齐晨虹[1] 朱 明[1] 欧阳慎[1] 赵红涛[1] 许丹亚[1]
QI Chenhong ZHU Ming OUYANG Shen ZHAO Hongtao XU Danya

摘 要 《中华人民共和国数据安全法》明确要求建立数据分类分级保护制度，根据数据在经济社会发展中的重要程度实行分类分级保护。数据分类分级的实施对于推动数据经济的健康发展、保护数据安全、优化数据治理结构以及贯彻国家相关政策具有至关重要的作用。目前，铁路行业的数据分类分级主要依赖人工操作，存在着效率低和成本高等问题。针对上述问题，文章提出基于 RoBERTa-wwm 与 BiLSTM 的铁路数据分类分级模型。实验结果表明，该模型在准确率、召回率和 F_1 值等指标上表现出较高的准确性，有效地提高铁路数据分类分级工作效率，具有广泛的应用前景。

关键词 数据安全；铁路数据；分类分级；RoBERTa-wwm；BiLSTM

doi：10.3969/j.issn.1672-9528.2024.12.027

0 引言

在数字化进程持续推进且网络环境愈发错综复杂的态势下，铁路行业正处于数据安全风险中，面临着诸如数据泄露风险的严峻考验、非法访问行为的频繁侵扰以及恶意篡改隐患的潜在威胁等一系列复杂且棘手的数据安全挑战。因此，实行数据分类分级是保障数据安全的前提[1]。数据分类是指依据数据的属性或特征，按照特定的原则和方法进行区分和归类，建立相应的分类体系和排列顺序，从而更有效地管理和使用数据的过程。数据分级是指依据数据的敏感程度，以及其遭到篡改、破坏、泄露或非法利用后对国家安全、企业利益和个人隐私所造成的影响程度，按照特定的原则和方法进行分级的过程。

1 研究背景

国家及国铁集团积极推进相关工作，相继制定并颁布了一系列法律政策与标准规范，涵盖《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律层面的纲领性文件，以及《数据安全技术 数据分类分级规则》（GB/T 43697—2024）《网络数据安全管理条例（征求意见稿）》《铁路数据分类分级指南（暂行）》等技术标准与行业规范。这些法律政策与标准规范相互配合、协同作用，为铁路数据分类分级工作的有序开展构建起坚实的法律依据框架，并提供了极具针对性与可操作性的实践指导准则，有力地推动铁路数据管理工作朝着规范化、科学化、法治化方向迈进，在保障铁路数据安全、促进数据合理开发利用等多方面发挥着极为关键的引领与支撑作用。

目前，国内许多行业在数据分类分级的研究中主要依赖人工判断，辅助预定义规则的方法[2]。随着数据量的爆炸式增长，此类方法逐渐暴露出效率低下、成本高昂的问题。人工分类分级不仅需要业务和数据专业的人员投入大量精力，还容易受到主观因素的影响；预定义规则的方法则存在规则制定复杂、维护成本高、难以应对异常情况等问题，进而影响数据分类分级的效率和精度。

文献研究显示，深度学习技术在数据分类分级方面的应用日益增多。顾荣杰等人[3]提出了基于 TFR 模型的公安云平台数据分级分类安全访问控制模型，通过对数据、人员、权限的分级分类，实现精准的访问控制，用于处理公安大数据中的敏感信息管理问题；陈美等人[4]分析了政府数据分类分级政策的合理性和执行效果，通过采用描述性分析、共现词分析和聚类分析方法对政策文本进行研究，结合 PMC-NMF 模型对政策进行量化评估，该框架能够有效评价政策的科学性和适用性。此外，王继晔等人[5]通过分析交通运输行业中的数据分级需求，采用结合卷积神经网络（CNN）、双向门控循环单元（BiGRU）和胶囊网络（CapsNet）的深度学习方法，有效地推进了数据自动分级处理。

深度学习技术在数据分类分级方面取得了一定成效，但现有方法往往未考虑铁路数据的业务特性。这可能导致模型对铁路运营中的关键信息误解或遗漏，从而影响其在铁路系统中的应用效果。基于此，本文以现行的国家和铁路行业政策法规为参考依据，结合中国铁路郑州局集团公司的数据登记工作，提出了一种新型的数据分类分级方法。该方法基于

1. 中国铁路郑州局集团有限公司信息技术所 河南郑州 450052
[基金项目] 国铁集团青年科研专项课题（Q2023W002）

RoBERTa-wwm（robustly optimized BERT pretraining approach with whole word masking）预训练语言模型和 BiLSTM（bidirectional long short-term memory）网络模型，采用注意力机制，增强对上下文相关性的处理能力，提高分类准确性。

2 铁路数据分类分级模型

本文提出了一种结合 RoBERTa-wwm 和 BiLSTM 的铁路数据分类分级模型。使用 RoBERTa-wwm 模型提取文本的语义特征，并将这些特征输入改进的 BiLSTM 模型中，以获得更全面的上下文信息，采用注意力机制层处理最后一个时序输出的特征向量，最终通过全连接层神经网络完成类别和级别的预测。模型的结构如图 1 所示。

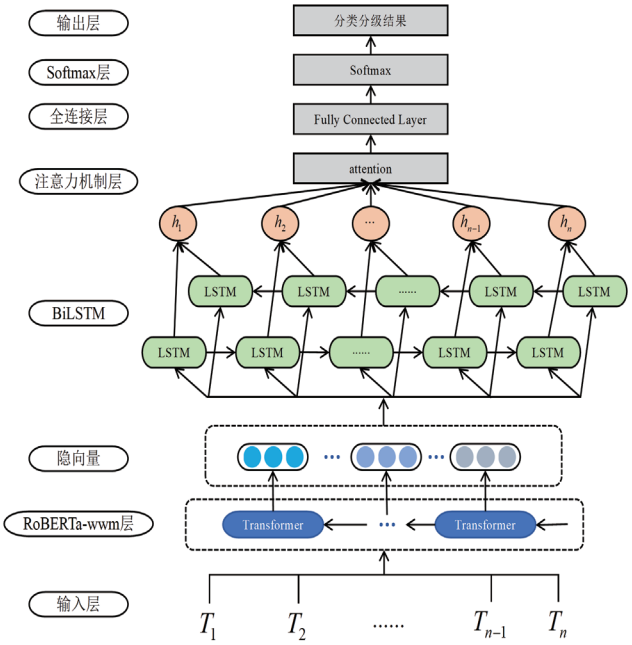

图 1 铁路数据分类分级模型结构

2.1 基于 RoBERTa-wwm 的特征提取模块

BERT 模型基于双向 Transformer 网络架构，通过分析大量语料库中的文本来学习其深层次特征，实现对文本的深度语义理解 [6]。BERT 模型结构如图 2 所示，其中 $\boldsymbol{T}_i$ 是模型的输入，通过 Transformer 层构建的双向编码器译码后，输出动态词向量 $\boldsymbol{H}_i$。RoBERTa 模型是对 BERT 预训练语言模型的一种改进，通过使用更多训练数据、延长训练周期、采用更严格的掩码策略以及更深的网络结构，显著提高了模型的性能 [7]。基于上述改进，RoBERTa-wwm 模型采用全词遮蔽（whole word masking，WWM）策略 [8]，该策略是对 BERT 和 RoBERTa 中使用的掩码语言模型（masked language model，MLM）的进一步优化，两者区别如表 1 所示。在中文文本处理中，WWM 策略能更有效地识别中文的词边界，相较于传统的字符级遮蔽方法，它能更精确地把握词汇的含义和语义结构。得益于这一优化的 WWM 策略，RoBERTa-wwm 在特征提取方面表现出更高的效率和准确性，尤其是在处理复杂语境的中文文本时，其性能超越了原始的 BERT 模型。

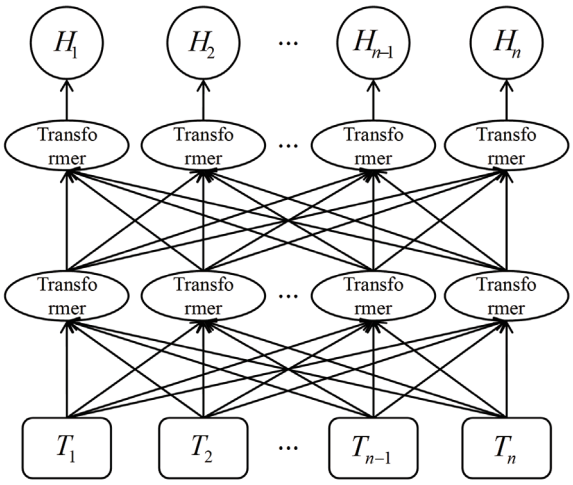

图 2 BERT 模型结构

在铁路数据分类分级的场景中，文本数据不仅包含大量专业术语，如“编组”“股道”等，还包含需要上下文语境理解的词语，如“作业”“线路”等。RoBERTa-wwm 的掩码策略与编码能力能够更准确地理解这些信息之间的语义关联和细微差别，这对于提高铁路数据分类分级的准确性和可靠性至关重要。

表 1 不同掩码策略区别

掩码策略	例子
原始句子	使用语言模型来预测下一个词的 probability
中文分词	使用 语言 模型 来 预测 下 一个 词 的 probability
MLM	使用语言 [MASK] 型来 [MASK] 测下一个词的 pro[MASK]lity
WWM	使用语言 [MASK][MASK] 来 [MASK][MASK] 下一个词的 [MASK][MASK][MASK]

2.2 基于改进 BiLSTM 的特征增强模块

长短期记忆网络（long short-term memory，LSTM）是一种特殊的循环神经网络（recurrent neural network，RNN），旨在解决传统 RNN 在处理长序列数据时出现的梯度消失或梯度爆炸问题，使网络能够学习到长期依赖关系 [9]。LSTM 结构如图 3 所示，其中 $\boldsymbol{x}_t$ 是当前时间步的输入；$\boldsymbol{h}_{t-1}$ 是前一时间步的隐藏状态；$\boldsymbol{C}_{t-1}$ 是前一时间步的单元状态；$\boldsymbol{f}_t$ 是遗忘门；$\boldsymbol{i}_t$ 是输入门；$\boldsymbol{o}_t$ 是输出门；$\boldsymbol{h}_t$ 是当前时间步的隐藏状态；$\boldsymbol{C}_t$ 是当前时间步的单元状态。LSTM 单元每个时间步 t 的更新用以下公式描述：

$$\boldsymbol{f}_t=\sigma\left(\boldsymbol{W}_f\cdot\left[\boldsymbol{h}_{t-1},\boldsymbol{x}_t\right]+\boldsymbol{b}_f\right) \tag{1}$$

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_i \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_i\right) \tag{2}$$

$$\widetilde{\boldsymbol{C}}_t = \tanh\left(\boldsymbol{W}_C \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_C\right) \tag{3}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t * \boldsymbol{C}_{t-1} + \boldsymbol{i}_t * \widetilde{\boldsymbol{C}}_t \tag{4}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_o \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_o\right) \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh\left(\boldsymbol{C}_t\right) \tag{6}$$

BiLSTM 网络由两个独立的 LSTM 组成，分别处理数据的正向和反向序列。这种结构使网络能够从两个时间方向学习，捕捉过去和未来的上下文信息，从而提升模型对信息的整体理解能力。每个方向的 LSTM 单独提取时间序列的特征，将两个方向的特征合并，形成丰富的数据表示 [10]。但在面对特别长或信息密集的文本时，模型可能无法有效区分信息的重要性。

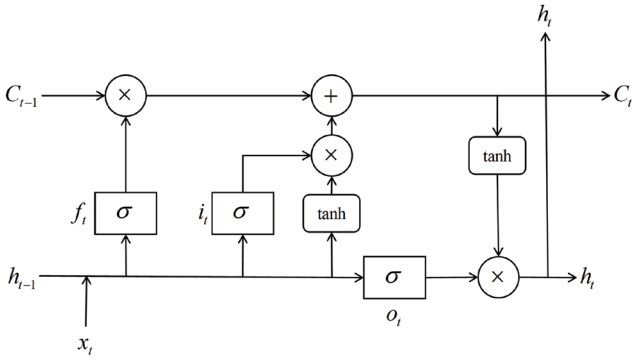

图 3 LSTM 结构图

在本文中使用的铁路数据分类分级数据集中涉及关键实体，如“列车运行”“安全设备”和“数据”。这些实体在铁路运输领域各自具有显著特征，但这些特征并不能单独决定实体是否属于“重要数据”级别。判断实体级别时，需要详细了解这些实体之间的相对位置，并据此评估实体组合的特征贡献度。例如，当 3 个实体按“列车运行安全设备数据”的顺序组合时，此组合被定义为“重要数据”级别。反之，若组合顺序为“列车运行数据安全设备”，则定义为“一般数据”级别。这种差异主要由于“数据”这一关键词在不同组合中的含义和重要性的变化。

为了适应铁路领域数据的特定需求，本文引入注意力机制对标准 BiLSTM 模型进行改进，旨在提高模型在铁路数据分类分级任务中对关键词的理解能力。

2.2.1 注意力机制

在铁路数据分类分级数据集中，每条数据包含多维信息，其中各维度的重要性也不同。深入分析数据集发现，关键词往往是决定数据归属类别和级别的重要因素，如表 2 所示。为了增强关键信息在分类时的作用，本文引入注意力机制，通过突出数据中的关键特征，提高分类的准确性。

表 2 关键词示例

原文	关键词	类别
日期 单位代码 运输总收入 旅客票价收入 货物运费收入 建设基金 其它收入 客其他 货其他 装车数 发送吨 发送人 集装箱收入 集装箱	运输、货物、集装箱	运输生产
推送批次号 录入单位编号 录入单位名称 发料/供应地区 库位编码 库位名称 上级库位 应税用途编号 应税用途名称 存货属性 是否实体库 是否车间库	发料、库位、存货、车间	资源管理
主键 项目 包件 响应承包人 承包类型 承包金额 工期	项目、承包、工期	建设管理
日期 路局码 车站代码 车种车型 车号 车辆标识 车辆日 转入转变日期 转入转变时分 转入命令号码 转出命令日期 转出转变时分 转出命令号码 备注	车站、命令	战略决策
联络人邮箱 值班电话 机构编码 出口地址 网信主管科室 职务 姓名 电话 排序 路电 类型	邮箱、电话、姓名、路电	综合协同
运单号码 货票票种 发局代码 发站电报码 到局代码 到站电报码 发货人代码 收货人代码 发专用线代码 到专用线代码 发送上门装车 到达上门卸车 到达送货里程 发送取货里程 运到期限 总重量 费用合计 制票日期 保价运输	货票、发货、收货、费用、保价	经营开发

注意力机制的核心思想是模拟人类的注意力聚焦行为，使模型在处理大量数据时优先关注对当前任务最重要的信息 [11]。具体到铁路数据的分类与分级，是指模型能够根据不同的上下文，识别并聚焦关键特征，如在列车调度数据中可能更加关注列车运行时刻、线路安排等敏感数据内容。

注意力机制的基本原理是通过为序列中的每个元素分配不同的权重，使权重较大的元素对最终结果的影响更显著。模型首先计算一个称为“注意力分数”的量，该分数表示了在给定查询（Query）的情况下，序列中每个键（Key）的重要性。将计算得到的注意力分数通过 Softmax 函数进行归一化，使用这些归一化后的注意力分数作为权重，对所有的值（Value）进行加权平均。这一步的输出是聚合了所有输入信息的向量，其中包含了模型认为最关键的信息，其计算公式为：

$$\boldsymbol{e}_t = \tanh\left(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b}\right) \tag{7}$$

$$\alpha_t = \frac{\exp\left(\boldsymbol{e}_t\right)}{\sum_{t'=1}^{T} \exp\left(\boldsymbol{e}_{t'}\right)} \tag{8}$$

$$\boldsymbol{c} = \sum_{t=1}^{T} \alpha_t \boldsymbol{h}_t \tag{9}$$

式中：$\boldsymbol{h}_t$ 是时间 t 的 BiLSTM 输出；$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是可学习的权重矩阵和偏置参数，用于计算 $\boldsymbol{h}_t$ 的注意力分数 $\boldsymbol{e}_t$；T 是序列总长度；α_t 是归一化后的注意力权重；$\boldsymbol{c}$ 是最终输出的上下文向量。

通过引入注意力机制，基于改进 BiLSTM 的特征增强模块能更有效地理解和处理铁路数据中的时序和上下文信息，

而且能更准确地反映出数据中的细微差异和复杂关系，为后续的分类模块提供精确的特征信息。

2.3 分类模块

在铁路数据分类分级模型的分类模块中，全连接层（fully connected layer，FC）和 Softmax 层是核心组件，用于实现最终数据类别和级别的标签预测。

全连接层将改进的 BiLSTM 输出的特征向量进行整合，为最终的分类决策提供输入。Softmax 层是分类模块的最后一层，它的作用是将全连接层的输出转化为概率分布，表明每一个类别的可能性。

3 实验与分析

3.1 数据集描述

本文按照国铁集团印发的《铁路数据分类分级指南（暂行）》文件要求，从业务运营角度对铁路数据进行了分类，具体分为战略决策、经营开发、运输生产、资源管理、建设管理、综合协同和其他七类；从数据安全保护角度对数据进行了级别划分，具体分为一般 -S1、一般 -S2、一般 -S3 和一般 -S4，共四级。

初始数据集包含 1611 条记录，主要来源于前期的数据登记工作。考虑到初始数据集规模较小，本课题组依据《“十四五”铁路网络安全和信息化规划》文件，通过人工收集各大网站的公开信息进行标注，将数据集规模扩充至 3754 条。

3.2 实验设置

实验采用 Linux 操作系统，硬件配置为：Intel(R) Core(TM) i7-12700KF CPU、NVIDIA GeForce RTX 3060 GPU、32 GB RAM、256 GB SSD 以及 1TB HDD 等。实验所使用的编程语言为 Python 3.8，深度学习框架选用 PyTorch 1.7。

分类分级模型的特征提取模块使用了哈工大讯飞联合实验室开源的中文 RoBERTa-wwm-ext-large 预训练模型权重。特征增强模块的 BiLSTM 层数设置为 2，隐藏层神经单元个数设置为 320，Dropout 设置为 0.3。为提高模型训练效率和性能，在全连接层后增加了批量归一化。

分类分级模型训练参数如下：

输入文本长度设置为 128，超长会被截断。Batch Size 设置为 10，epochs 设置为 100，为避免模型过拟合设置 Early Stopping 的耐心等待参数为 10。采用 Adam 优化模型训练，初始学习率设定为 2e-5，训练过程中通过使用循环学习率 CyclicLR 策略动态调整学习率来提高模型性能和训练效率。

为了验证提出模型的有效性，选择 BiLSTM、TextCNN、BERT+BiLSTM 以及 RoBERTa+BiLSTM 模型与本文模型进行对比实验。

3.3 结果分析

本研究在构建的数据集上分别进行数据类别预测和数据级别预测的实验，得到本文方法和对比模型的实验结果。为了评估模型的性能，采用混淆矩阵计算了 4 种指标进行比较，分别是准确率（accuracy，A）、精确率（precision，P）、召回率（recall，R）和 F_1 值。混淆矩阵是机器学习中常用于评估分类模型性能的工具，展示了分类结果与实际情况的对比，具体如表 3 所示。

表 3 混淆矩阵

Confusion Matrix		预测值	
		Postive	Negative
实际值	Postive	TP	FN
	Negative	FP	TN

4 种评价指标的计算公式分别为：

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

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

分类和分级预测的消融实验结果详见表 4 和表 5。由表中数据可知，本文提出的基于 RoBERTa-wwm 与 BiLSTM 的铁路数据分类分级模型在铁路数据分类分级任务上相较于其他基线模型，在各项指标上均有提升。具体而言，本文方法与其他方法相比，分类准确率至少提升了 0.94%，分级准确率至少提升了 1.20%。尽管本文方法相比同样使用预训练模型的 BERT+BiLSTM 和 RoBERTa+BiLSTM 仅有小幅度的性能提升，但本文方法使用较少的 epoch 就能达到更好的效果。

表 4 分类实验结果

单位：%

指标	Accuracy	Precision	Recall	F_1 值
BiLSTM	67.43	64.20	64.31	63.67
TextCNN	81.35	77.54	78.51	77.60
BERT+BiLSTM	86.76	83.88	85.26	84.40
RoBERTa+BiLSTM	87.30	84.97	84.14	84.40
本文方法	88.24	85.44	85.38	85.37

表 5 分级实验结果

单位：%

指标	Accuracy	Precision	Recall	F_1 值
BiLSTM	69.19	66.08	66.59	66.33
TextCNN	80.22	77.03	77.48	77.25
BERT+BiLSTM	85.93	82.54	82.12	82.33
RoBERTa+BiLSTM	86.08	83.25	84.63	83.93
本文方法	87.28	84.31	83.85	84.08

根据表 4、表 5 数据显示，BiLSTM 模型的准确率得到了显著的提升，从几乎为零增长到大约 60%，随后趋于稳定。

这表明传统的 BiLSTM 模型能够在一定程度上捕捉到铁路数据的业务特性，但其性能受限于模型的容量和对复杂数据模式的处理能力。

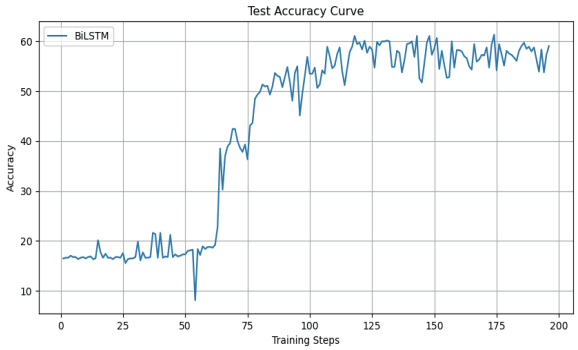

图 4 BiLSTM 准确率曲线

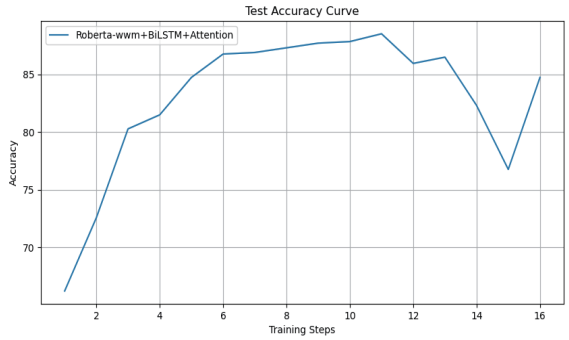

图 5 本文方法准确率曲线

综上所述，本文提出的组合模型在初始阶段就表现出高达 75% 的准确率，并在训练的早期阶段迅速提升至约 85%，证明了其应用在铁路复杂场景中的优越性。

4 结论

为提升铁路数据安全管理效率，减轻人工负担，本研究提出了一种基于 RoBERTa-wwm 与 BiLSTM 模型的铁路数据分类分级方法。在构建的铁路数据分类分级数据集上进行的对比实验表明，该方法在分类任务中的准确率达到 88.24%，在分级任务中的准确率为 87.28%，优于其他几种主流模型，验证了其在实际应用中的有效性和可靠性。由于铁路数据的多样性和复杂性，下一步需扩展数据集规模和优化模型结构，研究更准确更泛化的铁路数据分类分级方法。

参考文献：

[1] 饶伟，李碧秋，任宸莹，等．铁路数据分类分级保护路径研究 [J]. 铁道通信信号 ,2023,59(11):49-54.

[2] 董智华．企业数据分类分级的研究与思考 [J]. 软件和集成电路，2023(12):74-80.

[3] 顾荣杰，吴治平，石焕．基于 TFR 模型的公安云平台数据分级分类安全访问控制模型研究 [J]. 计算机科学，2020, 47(z1): 400-403.

[4] 陈美，何祺．基于特征分析的政府数据分类分级政策量化评价 [J]. 情报资料工作 ,2024,45(1):78-88.

[5] 王继晔，张少博，叶润泽，等．基于深度学习的交通运输行业数据自动分级方法研究 [J]. 应用科技 ,2024,51(2):145-150.

[6] DEVLIN J, CHANG M W, LEE K, et al. BERT: pretraining of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational . Stroudsburg, PA: Association for Computational Linguistics , 2019:4171-4186.

[7] LIU Y H, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach[DB/OL]. (2019-07-26) [2024-07-16].https://doi.org/10.48550/arXiv.1907.11692.

[8] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[9] SCHMIDHUBER J. Deep learning in neural networks: anoverview[J]. Neural networks, 2015, 61:85-117.

[10] ZHOU P, QI Z Y, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of The Association for Computational Linguistics (volume 2: Short papers). Berlin:Springer, 2016: 207-212.

[11] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[C]//NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. Cambridge, MA : MIT Press, 2014:2204-2212.

【作者简介】

齐晨虹（1989—），女，河南商丘人，硕士，高级工程师，研究方向：数据治理及大数据技术。

朱明（1997—），男，河南郑州人，硕士，助理工程师，研究方向：数据安全。

欧阳慎（1969—），男，上海人，本科，高级工程师，研究方向：网络安全。

赵红涛（1976—），男，河南洛阳人，本科，高级工程师，研究方向：信息技术与大数据应用。

许丹亚（1996—），女，河南濮阳人，硕士，工程师，研究方向：数据安全。

（收稿日期：2024-08-30）