基于 LLMs 的危化品典型事故文本分类研究

张又元[1,2],马新春[1,2],赵 军[3]
(1. 新疆大学 软件学院,新疆 乌鲁木齐 830046;
2. 新疆电子研究所股份有限公司,新疆 乌鲁木齐 830013;
3. 新疆维吾尔自治区安全科学技术研究院,新疆 乌鲁木齐 830000)

摘 要:为了实现对危化品事故案例的有效管理,首先需要对事故案例文本进行精确分类。尽管当前的大语言模型(LLMs)在经过简单微调后能够在特定领域的中长文本分类任务中表现良好,但却忽视了生成式大语言模型强大的推理能力对于此类任务的重要促进作用。生成式大语言模型不仅具有生成分类结果的能力,还具有生成推理过程的能力。基于此,该文提出了一种基于推理模式的大语言文本分类模型,具体构建流程如下:首先,利用大型 LLMs“通义千问 2.5”的推理能力,模拟生成连接案例文本和真实标签的中间推理过程;然后,将生成的推理过程编码为结构化的提示信息,并嵌入到用于分类任务的提示模板中;最后,在低资源条件下,选择小型 LLMs“Qwen1.5-4B”作为该文采用的文本分类器,利用“通义千问 2.5”LLMs 构建的提示模板进行微调。实验证明,该方法在危化品事故案例的小样本数据集中表现优异,其 F_1 值达到了 90.22%。此外,在公开新闻数据集上验证了该方法的泛化性能,其 F_1 值也达到了 88.04%。

关键词:危化品事故;大语言模型;中长文本;中间推理过程;提示模板;微调
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2025)07-0133-07
doi:10.20165/j.cnki.ISSN1673-629X.2025.0054

Research on Text Classification of Typical Hazardous Chemical Accidents Based on Large Language Models

ZHANG You-yuan[1,2], MA Xin-chun[1,2], ZHAO Jun[3]
(1. School of Software Engineering, Xinjiang University, Urumqi 830046, China;
2. Xinjiang Electronics Research Institute Co., Ltd., Urumqi 830013, China;
3. Xinjiang Uyghur Autonomous Region Institute of Safety Science and Technology, Urumqi 830000, China)

Abstract: To effectively manage hazardous chemical accident cases, precise classification of the accident case texts is essential. Although current large language models (LLMs) perform well in specific domain's medium and long text classification tasks after simple fine-tuning, they overlook the significant role of generative LLMs' powerful reasoning capabilities in such tasks. Generative LLMs not only have the ability to generate classification results but also to generate the reasoning process. Based on this, we propose a large language text classification model based on reasoning patterns. The specific construction process is as follows: Firstly, the reasoning ability of the large LLM "Qwen 2.5" is utilized to simulate and generate the intermediate reasoning process connecting the case text and the real label. Then, the generated reasoning process is encoded as structured prompt information and embedded into the prompt template used for the classification task. Finally, under low-resource conditions, the small LLM "Qwen 1.5-4B" is selected as the text classifier, and the prompt template constructed by "Qwen 2.5" LLMs is used for fine-tuning. Experiments prove that the proposed method performs excellently in the small sample data set of hazardous chemical accident cases, with an F_1 value reaching 90.22%. Additionally, the generalization performance of the proposed method is verified on the public news dataset, with an F_1 value reaching 88.04%.

Key words: hazardous chemical accidents; large language models; medium and long texts; intermediate reasoning process; prompt templates; fine-tuning

收稿日期:2024-12-17 **修回日期**:2025-04-18
基金项目:新疆维吾尔自治区重点研发项目(2022B03004-4)
作者简介:张又元(1996-),男,硕士研究生,通讯作者,研究方向为自然语言处理;马新春(1967-),男(回族),正高级工程师,研究方向为物联网大数据技术。

0 引 言

针对危化品事故文本分类任务,由于事故案例的日益增多,传统人工分类方法已经不适用。为解决这个问题,机器学习相关技术开始应用于此类任务。然而,传统的机器学习模型虽然能够很好地完成一些短文本分类任务,但是面对危化品事故文本复杂多变的文字长度、语义及语境,传统的机器学习文本分类模型如朴素贝叶斯、支持向量机、K 最近邻等已不能满足任务需求。深度学习作为机器学习方法的延伸,为中长文本分类任务提供了理论支撑。随着深度学习的发展,涌现出了一批优秀的自然语言处理模型,如 CNN(卷积神经网络)、RNN(循环神经网络)、LSTM(Long Short - Term Memory)以及 BERT[1](Bidirectional Encoder Representations from Transformers)等模型。之后,预训练模型[2]的出现使自然语言处理领域有了新的发展。这些模型(如适应中文的 BERT - base - Chinese)通过自监督学习在大规模中文文本数据上应用双向 Transformer 架构进行无监督学习,从而能够学习全局的语义表征,并在各种中文自然语言处理任务中展现出良好性能。

然而,上述模型在出现错误的情况下由于隐层的数据无法直接理解,因此引入了注意力机制[3]来提高模型的解释性和准确性。Prabhakar 等人[4]构造出了两种新的深度学习架构来学习文本分类任务。一是,利用 4 个通道构建四通道混合长短期记忆的深度学习模型;二是,开发并实施了集成多头注意力的混合双向门控循环单元深度学习模型。Deng 等人[5]构建出了通过融合多种神经网络模型的特性并集成 Attention 以达到提取文本的全局语义特征目标的 FMNN(Fused with Multiple Neural Network)模型。之后,Yu 等人[6]构建了 BERT4TC 文本分类模型。此模型是以 BERT 模型作为基础,通过构造辅助的文本,利用 BERT 能实现 NSP 任务的能力,不仅通过构造辅助句获取任务,同时解决了小文本数据规模受限的问题。Hu 等人[7]使用 BERT 模型作为分类任务的编码器,通过 BERT 的双向性、预训练能力、通用性和适应性,大大提高了短文本分类任务的分类准确率。

尽管已存在多种文本分类模型,但它们在长文本分类任务中仍存在提升空间。长文本复杂多变的语言环境对模型的理解与特征提取能力提出了更高要求。随着近年来以 ChatGPT[8]、Llama[9]、Qwen[10]以及 ChatGLM[11]为代表的生成式大语言模型在文本理解和生成方面展现出很强的能力,尤其在处理长文本复杂任务时表现出强大的泛化能力和深度理解能力,越来越多的中文中长文本分类任务开始应用大语言模型。

虽然,现阶段的大语言模型仅通过简单的微调方式便可使得大语言模型快速、良好地适应于各类下游任务。但是,这种简单的微调方式完全忽略大语言模型强大的推理能力,大语言模型不仅能够生成任务执行结果,还能够生成任务执行的推理过程,因此,大语言模型在各类下游任务上还有更大的潜力。基于此,为了充分利用大语言模型强大的推理能力,该文设计了一种融入 CoT(Chain of Thought)[12]的大语言模型微调模式,应用于危化品案例长文本分类任务中。

1 相关工作

1.1 实验数据集介绍

本实验使用的数据集有两个:一个是该文的应用实例—危化品事故案例数据集,一个是为验证该文提出的提示模版构建策略泛化性能的公开新闻长文本数据集[13]。两者的数据类别原始分布如表 1 所示。

表 1 实验数据集数据类别分布

新闻文本数据集			危化品事故案例数据集		
新闻类别	训练集数量	测试集数量	事故类别	训练集数量	测试集数量
科技	9 908	1 052	生产	215	71
财经	13 040	873	存储	296	98
游戏	7 874	833	运输	109	36
体育	9 330	1 022	使用	26(50%)	26
社会	8 501	996	其他	83	27
房产	9 618	859	--	--	--
娱乐	8 015	922	--	--	--
教育	8 601	953	--	--	--
家居	6 876	1 139	--	--	--
时政	6 999	1 351	--	--	--

危化品事故案例来源于国内各大应急管理局,该文的研究内容专注于化工企业发生的与危化品相关的事故,对于日常生活中发生的事故需要选择性忽略,对于摔伤、撞伤类似的事故也要排除在外。在经过筛选、

去重、清洗后获得了 987 篇高质量的典型事故案例文章,案例文本字数集中在[300,500]这个区间,属于小样本中长文本数据集。

危化品事故的分类标准[14]一般有两大类型:一是基于事故的伤害类型分类,主要将危险化学品事故分为六类;二是基于危化品所处的过程对危险化学品事故案例共分五类,包括危险化学品生产事故、储存事故、运输事故、使用事故和其它事故。由于危化品事故在发生时,爆炸、火灾等伤害类型会一起出现,因此伤害类型不适合作为危化品事故精准分类的标准,而过程类型则不会产生混淆,便于对事故的精确定位。各过程类型数量对如表 1 所示。观察表 1 中各类型所占比例可知,运输过程的事故案例占比较低,或可能影响对这一类型的识别效果。因此,为了验证大语言模型少样本甚至无样本微调的能力,在划分训练测试时,运输过程的事故案例有 50% 被划分为训练集,其他类型则按照 3 ∶ 1 的比例划分训练集与测试集。

另外,由表 1 可知,新闻文本数据集规模远超事故案例数据集,数量极其庞大,这就带来了两个问题:一是大语言模型的训练难度随着样本数量的增加而增加,按照原数据集的规模在现有资源的条件下进行微调所花费的时间成本难以估算;二是该文是建立在实际应用基础上的大语言模型微调任务,现实应用中无法达到公开数据集的数量规模,该文的危化品事故案例领域就是一个典型的例子。考虑到这两点,本实验在每个新闻类别的训练集中随机抽取 300 篇新闻报道构成新的训练集,在测试集中随机抽取 100 篇新闻报道构成新的测试集。新的训练集包括 10 种新闻类别,共 3 000 篇文章;新的测试集包括 10 种新闻类别,共 1 000篇文章。注意的是,在重新组织数据集时,随机抽取的每篇文章的文本长度与事故案例数据集的文本长度分布保持一致,都集中在[300,500]区间内。

1.2　基础架构

该文以主线任务危化品事故案例中长文本分类任务为例,讨论构建融入 CoT 数据的提示模版构建策略对于充分发挥生成式大语言模型在长文本分类任务中的重要作用。以此为出发点,该文将这种推理思维具象化为思维链,以此来提升事故案例所属类型划分的各项指标。所设计的基于 LLMs 的中长文本识别算法的建模包含 CoT 数据构建和大语言模型微调两个核心步骤,其整体训练流程架构如图 1 所示。

将整个训练策略分为两个主要阶段,首先是 CoT 数据构建,其次是大语言模型的微调。

在 Step1 中,利用“通义千问 2.5”强大的推理能力,将其作为该文构建思维链的工具。推理模板的构建包括指令、案例文本、具体类型三部分,尤其是具体类型的引入,形成了一种案例输入与模型输出反推分析过程的巧妙模式,这种模式可以充分激活“通义千问 2.5”的强大推理能力,确保生成的分析过程的逻辑性和准确性。这种方法的优势在于,通过输入和输出反推过程,可以更好地挖掘出输入与输出之间的关联关系和隐含语义。通过这种策略,该文可以在危化品事故数据集上,完成各个案例划分为某个过程类型的中间分析过程的构建。

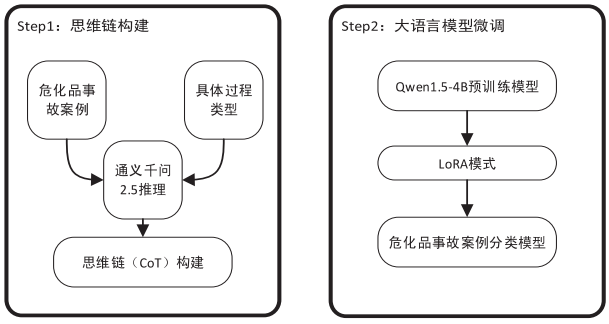

图 1　整体训练流程架构

在 Step2 中,基于 Qwen 系列模型庞大参数规模覆盖范围以及优秀的性能,该文选定 Qwen1.5-4B 作为预训练模型。预训练模型的输入包括两个部分:一部分是指令,暗示系统进入“专家分析模式”;另一部分则是原始文本。通过对文本信息进行逐层深入解析,并将中间推理步骤作为输出的一部分,通过推理分析过程获取最终的分类结果。这种方法使得支持中文推理的预训练大语言模型能够更准确地进行文本分类,从而构建出更加精准且具备强大逻辑推理能力的模型,显著提升了大语言模型在文本分类任务中的各项性能。

1.3　CoT 数据构建

构建 CoT 数据的过程,本质上是一个将事故案例与特定过程类型相匹配的中间推理步骤转化为文字描述的过程。为了更高质量地构建出中间推理过程,该文提出了一种创新方法,即利用大型 LLMs 实现这一过程。“通义千问 2.5”作为 Qwen 系列模型现阶段最先进的大语言模型,其强大的推理能力能够很好地完成这一任务。该文利用“通义千问 2.5”的高级推理功能,实现事故案例长文本分类任务中的中间推理过程的自动化生成。鉴于 CoT 数据对于提升生成式大语言模型(LLMs)能力的重要性,该文提出的方法通过巧妙构思的上下文模板与实例,引导“通义千问 2.5”大语言模型进行深度文本解析。具体实施过程中,通过搜索引擎调用“通义千问 2.5”的 API(应用程序接口)进行推理,针对每个样本及其对应的真实标签,发送请求并接收模型生成的响应,进而自动构建出一系列中间思维链条。CoT 数据的构造流程如图 2 所示。

首先,给“通义千问 2.5”大语言模型下达一个指

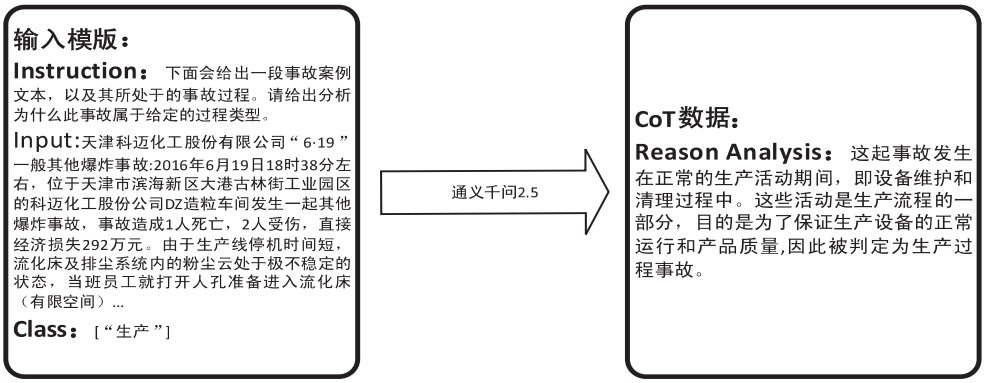

图 2 CoT 数据的构造流程

令“下面会给出一段事故案例文本,以及其所处的事故过程。请给出分析为什么此事故属于给定的过程类型”,暗示大语言模型进入“专家模式”,然后大语言模型通过自身的推理能力依据事故案例文本的上下文语境与真实分类标签,模仿人类的思维分析上下文与真实标签之间的因果关系,进而生成中间推理过程,图 2 中的“Reason Analysis”便是 CoT 数据的示例展示。

1.4 LLMs 微调流程

微调 Qwen1.5-4B 大语言模型的整体流程如图 3 所示,任务的核心为提示模板的构造与 LoRA(Low-Rank Adaptation)[15]模式下的低参微调。

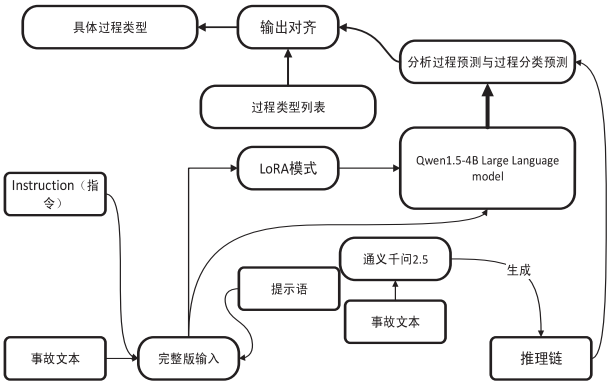

图 3 Qwen1.5-4B 模型微调架构

1.4.1 完整提示模板构造

完整提示模板包括两大部分:输入模板和输出模板。其中输入模板包括两部分,第一部分是指令,第二部分为案例文本;输出模板也包括两部分,分别是中间推理流程与最终结果分类。示例如图 4 所示,图 4 中左侧部分为提示模板的输入样例构成,右侧部分为模型的输出样例构成。

1.4.2 LoRA 训练模式

大语言模型的参数量是非常庞大的,以 10 亿为基本单位,而常规的模型,例如 BERT、LSTM 等参数量远低于 LLMs,它们的参数量对比如图 5 所示。可以看出,训练 LLMs 所需要的资源是训练常规模型的数十上百倍,尤其是在长文本上,这种资源消耗一般的 GPU 无法满足。

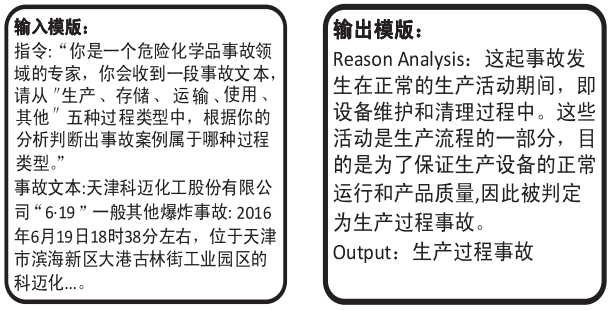

图 4 提示模板示意图

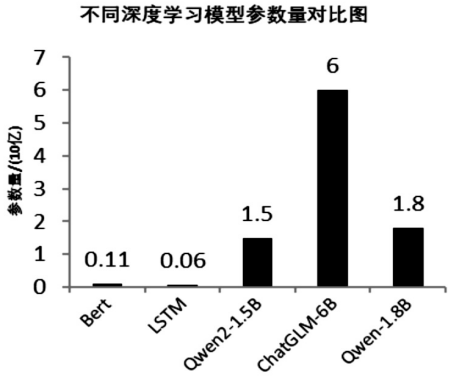

图 5 分类任务模型参数规格(10 亿)对比

同时,如果是在小样本数据集上微调大语言模型,全参微调会导致所有的权重参数向样本数据集倾斜,会出现过拟合的问题,大大减弱 LLMs 的泛化能力。同时,权重的改变,导致 LLMs 本身的推理能力受到影响,而且全参微调大语言模型所要的经济成本和时间成本也是很难被满足的。因此,需要调整微调大语言模型的策略。

LoRA 算法是一种用于高效微调大型预训练语言模型的技术,通过在模型的特定层引入两个低秩—r 矩阵 $\boldsymbol{A}$ 、$\boldsymbol{B}$,通过控制 r 的大小,控制参与计算的参数的规模。具体结构如图 6 所示,其中“ * ”代表参数冻结不参与训练。

这种方法显著减少了需要调整的参数数量,从而降低了计算和存储成本,同时保持或提升了模型的性能。LoRA 的创新结构使得在资源有限的环境中快速微调 Qwen1.5-4B 模型的任务得以实现。

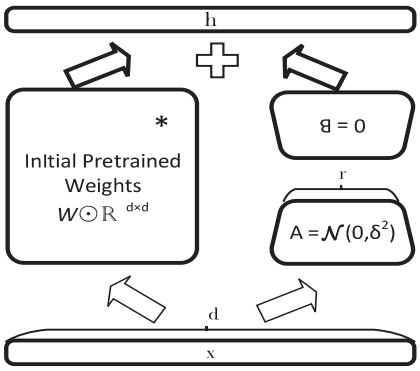

图 6　LoRA 训练架构

预设初始的预训练模型的权重为 W_0,在微调阶段更新权重,如公式 1 所示。

$$W = W_0 + \Delta W \tag{1}$$

其中, ΔW 是实验人员在输入模板上重新训练的权重,计算过程如公式 2 所示。

$$\Delta W = \boldsymbol{BA} \tag{2}$$

其中, $\Delta W \in R^{d\times k}$, $\boldsymbol{B} \in R^{d\times r}$, $\boldsymbol{A} \in R^{d\times r}$ 。其中 $r << \min(d,k)$, $\boldsymbol{B}$ 初始化为 0 矩阵。

在分类模型中,模板输入定义为 Input = {I,T} ,其中“I”为微调指令,“T”为原始文本。在输入预训练之后模型之前会对输入进行文本编码,初始层输入 Z_0 如公式 3 所示。

$$Z_0 = \boldsymbol{W}_e + \boldsymbol{W}_p \tag{3}$$

其中, $\boldsymbol{W}_e$ 是字词嵌入矩阵, $\boldsymbol{W}_p$ 是位置嵌入矩阵。

在获取到初始编码输入之后,假设定义预训练 LLMs 中第 $t+1$ 层的隐层输出为 Z_{t+1} ,如公式 4 所示。

$$Z_{t+1} = \mathrm{Trans}(Z_t), t \in [0,N] \tag{4}$$

其中,Trans 代表 transformer 转换。

transformer 每一层的核心模块包括带跳层连接和前置标准化操作的多头注意力机制 MHA 和前馈网络 FFN 两部分。多头注意力机制的运算过程中,假设第 $t+1$ 层的注意力投影矩阵分别为 Query($\boldsymbol{Q}$)、Key($\boldsymbol{K}$)和 Value($\boldsymbol{V}$),三个矩阵经过一个全连接层,分别得到权重矩阵 $\boldsymbol{W}^{\boldsymbol{Q}}_{t+1}$ 、$\boldsymbol{W}^{\boldsymbol{K}}_{t+1}$ 、$\boldsymbol{W}^{\boldsymbol{V}}_{t+1}$ 。具体计算过程如公式 5 所示。

$$\boldsymbol{Q},\boldsymbol{K},V = \boldsymbol{W}^{\boldsymbol{Q}}_{t+1}(Z_t), \boldsymbol{W}^{\boldsymbol{K}}_{t+1}(Z_t), \boldsymbol{W}^{\boldsymbol{V}}_{t+1}(Z_t) \tag{5}$$

因此,最终第 $t+1$ 层的隐层输出可以定义为公式 6 的形式。

$$Z^{\mathrm{MHA}}_{t+1} = \mathrm{SoftMax}\left(\frac{\boldsymbol{Q} * \boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\boldsymbol{V}\right) \tag{6}$$

2　实验与分析

2.1　实验设置

设置的实验参数如下:batch_size(每批次数据量)为 4, learning _ rate(学习率)为 1e-4,优化器为 AdamW,Epoch(训练轮数)为 50,真实标签数量为(5,10),可视化工具为 SwanLab,dropout 为 0.1, r (训练权重的秩)为 2,LoRA_alpha(缩放系数)为 16, LoRA_dropout(随机失活概率)为 0.1。软硬件实验环境如表 2 所示。

表 2　实验环境详情

序号	名称	配置及参数
1	操作系统	Windows10
2	CPU	i7-9750H
3	GPU	NVIDIA GTX 1650
4	显存容量/G	8
5	内存容量/G	40
6	Cuda 版本	12.4
7	开发工具	pycharm
8	Python 版本	3.10.5
9	Pytorch 版本	2.4.0+cu124

实验将宏平均准确率(Precision, P)、宏平均召回率(Recall, R)以及宏 F_1 值作为评估标准。真假标准判定为当且仅当分类结果与真实标签相同时,才认定该预测数据的结论为真,否则即为假。P 、R 以及 F_1 值的计算流程如式 7 ~9 所示。

$$P = \frac{1}{C}\sum_{i=1}^{C}\frac{N_{\mathrm{TP}_i}}{N_{\mathrm{TP}_i} + N_{\mathrm{FP}_i}} \tag{7}$$

$$R = \frac{1}{C}\sum_{i=1}^{C}\frac{N_{\mathrm{TP}_i}}{N_{\mathrm{TP}_i} + N_{\mathrm{FN}_i}} \tag{8}$$

$$F_1 = \frac{2(P * R)}{P + R} \tag{9}$$

式中, C 为类别规模, N_{TP_i} 为第 i 类正例的样本被正确地预测为正例的数量, N_{FP_i} 为第 i 类负例的样本被错误地预测为正例的数量, N_{FN_i} 为第 i 类别正例的样本被错误地预测为负例的数量。

2.2　对比实验

该文将生成式大语言模型与传统意义上的标注式模型进行实验对比,以此来验证生成式大语言模型对于传统标注式文本分类模型的优势。对比实验模型包括 RNN、TextCNN[16]、BERT-LSTM-Self_Attention(BLS)[17]、LSTM、BERT[1]共 5 个标注式文本分类模型。具体对比实验结果如表 3 所示。

RNN 是递归神经网络,能够处理序列数据并保持记忆,能学习文本时序特征,有一定捕获长距离依赖关系的能力,但在处理长序列时容易出现梯度消失或梯度爆炸的问题。

TextCNN 通过卷积层和池化层学习文本空间特征,相较于 RNN,TextCNN 具有更高的并行计算速度,但无法捕获长距离依赖关系。

LSTM 是一种特殊的 RNN,能够更好地解决 RNN 中的梯度消失和梯度爆炸问题,有着更好地捕获长距

离依赖关系的能力。

BERT 则是基于 Transformer 架构的预训练模型，利用大规模文本数据进行预训练，可用于各种自然语言处理任务。每种模型都有其独特的优势和适用场景，在选择模型时需考虑具体的任务需求、数据特点以及计算资源等因素。

BLS 模型，集成 BERT 的语义理解、LSTM 的序列建模与远距离依赖捕获，并以 Self-Attention 对这两方面进一步增强，在长文本分类任务中已取得更加全面和精确的结果。可以说 BLS 模型是目前依靠单源特征进行长文本分类任务的 SOTA 模型。

表 3 分类任务对比实验结果

对比模型	危化品事故案例数据集			长文本新闻数据集		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%
RNN	80.26	81.79	81.02	82.26	84.26	83.25
TextCNN	76.92	78.76	77.83	80.86	82.14	81.50
LSTM	81.14	82.59	81.86	84.16	84.69	84.42
BERT	83.16	85.20	84.17	85.05	85.45	85.25
BLS	84.74	86.74	85.73	85.88	86.18	86.03
New-CoT-Qwen1.5-4B	**89.69**	**90.75**	**90.22**	**87.85**	**88.23**	**88.04**

从表 3 危化品事故案例数据集的实验结果中可以观察到：经过微调后的 Qwen1.5-4B 模型相较于从单方面考虑特征的 TextCNN、RNN、LSTM、BERT 模型，在 F_1 值的表现上，对比最优的 BERT 模型有约 6 百分点的提升；与 BLS 模型相对比，该文微调后的大语言模型，在 F_1 分数上有着约 4.5 百分点的提升。这证明了经过微调后的 New-CoT-Qwen1.5-4B 模型凭借其强大的学习能力，对于长文本复杂的语境和上下语义有着更强的理解能力。

从表 3 的长文本新闻数据集实验结果中可以观察到，该文的策略在精确率 P、召回率 R、F_1 值三项核心指标上分别达到了 87.85%、88.23%、88.04%。对比较为先进的 BLS 模型，在这三项指标上分别约有 2 百分点的提升。这再一次证明了，提出的策略在中长文本分类任务中有着良好的表现。

为了进一步验证该模型在危化品数据集上的优越性，将完整的案例数据集作为验证集，调用保存好的大语言模型模型，统计在各个分类中正例的数量与初始数量进行对比，验证微调后的大语言模型对于各个分类的识别性能，具体情况如图 7 所示。

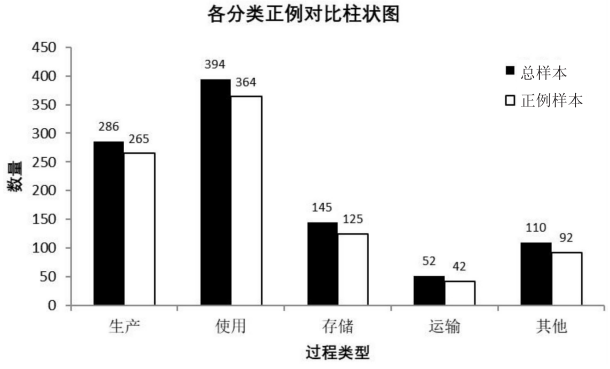

图 7 案例样本正例样本比对柱形图

观察图 7 得出结论，生产过程、使用过程的分类准确率达到了 90% 以上，存储过程与其他过程的分类准确率也达到了 80% 以上，分类性能良好。值得一提的是，运输过程的分类准确率也达到了约 81%。这也再次证明了大语言模型优秀的少样本学习能力。综合来说，整个数据集总体的识别准确率达到了 90%，再一次证明本实验微调后的 LLMs 在危化品事故案例分类任务中具有良好的分类性能。

2.3 消融实验

消融实验是在危化品事故案例数据集上进行的，共有两个。第一个是在简单提示模板的基础上引入 CoT 数据的消融实验，第二个是在引入 CoT 数据后更换同级别或略高、略低级别的大语言预训练模型的消融实验。

第一个消融实验的具体实验结果如表 4 所示，表 4 中的 Qwen1.5-4B 代表基线模式，New-Qwen1.5-4B 代表不采用 CoT 数据构建提示模板的简单微调模式。

表 4 提示模板构建模块消融实验

提示模板构造方式	P/%	R/%	F_1/%
New-CoT-Qwen1.5-4B	89.69	90.75	90.22
New-Qwen1.5-4B	86.54	87.03	86.78
Qwen1.5-4B	82.16	84.37	83.25

观察表 4 的实验结果，可以确定未经过微调的大语言模型 Qwen1.5-4B，对于分类任务也有良好的效果，其 F_1 值达到了 83.25%，证明了大语言模型 Qwen1.5-4B 对于文本分类任务的泛化能力极强。更进一步，虽然简单微调的 New-Qwen1.5-4B 在原始大模型的基础上 F_1 值提升了 3.5 百分点，但是对比 New-CoT-Qwen1.5-4B 在 F_1 值上还是存在 3.5 百分点的

差距,这也证明了基于 CoT 数据构建的推理模板,对比普通模板在中长文本的分类任务中有着明显的优势。

第二个消融实验的具体实验结果如表 5 所示。更换的大语言预训练模型包括 Qwen2-1.5B 大语言模型、Qwen-4B 大语言模型和 ChatGLM-6B 大语言模型。

表 5　预训练模块更换消融实验

预训练模块	P/%	R/%	F_1/%
Qwen1.5-4B	89.69	90.75	90.22
Qwen2-1.5B	87.26	87.95	87.60
Qwen-4B	88.16	89.47	88.81
ChatGLM-6B	89.23	89.84	89.53

从表 5 中的实验结果可以观察到,相比于文中的 Qwen1.5-4B 预训练模型,参数级别较低的 Qwen2-1.5B 预训练模型在 F_1 值上降低了 2.6 百分点,同级别的 New-Qwen-4B 预训练模型在 F_1 值上也降低了 1.4 百分点,而级别略高的 ChatGLM-6B 在 F_1 值上也是极为相近,仅仅是差距了 0.7 百分点。这再一次证明了在参数规模相近的情况下,该文的基座预训练模型对比其他预训练模型有着更好的性能。

3　结束语

围绕危化品事故案例文本分类任务,该文提出了一种基于推理模式的大语言文本分类模型,并详细阐述了其构建流程与应用效果。研究的核心在于利用大型生成式大语言模型(LLMs)的推理能力,模拟生成连接案例文本和真实标签的中间推理过程,进而构建结构化的提示信息嵌入到提示模板中,以此指导小型 LLMs 进行分类任务。这一方法不仅充分挖掘了 LLMs 的推理潜力,还有效提升了在长文本分类任务中的表现。此外,该模型在公开新闻数据集上也验证了其良好的泛化性能,F_1 值达到 88.04%。相较于传统的文本分类模型,该方法不仅提高了文本分类任务的各项性能,还具备更强的逻辑推理能力。

尽管本研究取得了一定成果,然而在模型构建过程中,虽然利用了大型 LLMs 的推理能力,但是这种推理能力本身具有不确定性,这种不确定性会制约提示模板的构造质量,进而影响 LLMs 的微调效果。因此,如何生成更高质量的中间推理过程,以减少这种不确定性并提高模型的稳定性和准确性,是未来的一个重要研究方向。

参考文献:

[1] DEVLIN J,WEI M,KENTON C,et al. BERT:pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the conference of the North American chapter of the association for computational. linguistics:human language technologies,Minneapolis:ACL,2019:89-97.

[2] QIU X,SUN T,XU Y,et al. Pre-trained models for natural language processing:a survey[J]. Science China Technological Sciences,2020,63(10):1872-1897.

[3] SUBAKAN C,RAVANELLI M,CORNELL S,et al. Attention is all you need in speech separation[C]//ICASSP 2021-2021 IEEE international conference on acoustics,speech and signal processing (ICASSP). Toronto:IEEE,2021:21-25.

[4] PRABHAKAR S K,WON D O. Medical text classification using hybrid deep learning models with multi-head attention [J]. Computational Intelligence and Neuroscience,2021(3):1-16.

[5] DENG Weibin,ZHU Kun,LI Yunbo,et al. FMNN:text classification model fused with multiple neural networks[J]. Computer Science,2022,49(3):281-287.

[6] YU S,SU J,LUO D. Improving BERT-based text classification with auxiliary sentence and domain knowledge[J]. IEEE Access,2019,7:176600-176612.

[7] HU Y,DING J,DOU Z,et al. Short-text classification detector:a BERT-based mental approach[J]. Computational Intelligence and Neuroscience,2022(4):1-11.

[8] 李春涛,闫续文,张学人. GPT 在文本分析中的应用:一个基于 Stata 的集成命令用法介绍[J]. 数量经济技术经济研究,2024,41(5):197-216.

[9] 宋　婧. Meta 公布最新大模型 Llama 2[N]. 中国电子报,2023-07-28(006).

[10] XIONG H,WANG S,ZHU Y,et al. Doctorglm: Fine-tuning your chinese doctor is not a herculean task[J]. arXiv:2304.01097,2023.

[11] 赵　雪,赵志枭,孙凤兰,等. 面向语言文学领域的大语言模型性能评测研究[J]. 外语电化教学,2023(6):57-65.

[12] WEI J,WANG X,SCHUURMANS D,et al. Chain-of-thought prompting elicits reasoning in large language models [J]. Advances in Neural Information Processing Systems,2022,35:24824-24837.

[13] 许楠楠,柯圆圆,胡晓莉. 基于增强语言表示模型的网络新闻长文本分类的研究[J]. 江汉大学学报:自然科学版,2024,52(4):37-44.

[14] 李安康. 基于卷积神经网络的危化品事故案例分类研究[D]. 青岛:中国石油大学(华东),2019.

[15] HU E J,SHEN Y,WALLIS P,et al. Lora:low-rank adaptation of large language models[J]. arXiv:2106.09685,2021.

[16] GONG L,JI R. What does a TextCNN learn? [J]. arXiv:1801.06287,2018.

[17] SHOBANA J,MURALI M. An improved self attention mechanism based on optimized BERT-BiLSTM model for accurate polarity prediction[J]. The Computer Journal,2023,66(5):1279.