文飞. 传统与大模型并举:中文文本分类技术对比研究[J]. 智能计算机与应用,2024,14(6):88-94. DOI:10.20169/j.issn.2095-2163.240612

传统与大模型并举:中文文本分类技术对比研究

文 飞
(中卓信(北京)科技有限公司, 北京 100085)

摘 要:本文专注于探索与实践中文文本分类技术的演进,通过严谨的实证对比研究,检验了传统技术方法与基于大模型的先进算法在各类文本分类任务中的表现差异。研究在涵盖情感分析的基础数据集和富含复杂专业信息的多类别文本数据集上展开了深入探索,系统性地对比了传统统计学习方法、经典深度学习算法与当前极具影响力的预训练大模型(如 BERT、LLM 等)。研究核心围绕提升分类准确性这一关键目标,同时审视各模型在资源效率及训练时效性方面的能力。针对预训练大模型,利用了提示工程技术和模型微调手段,以期优化其性能表现。实验结果揭示了大模型在理解和利用语言上下文、提高泛化性能方面的显著优势,在不同数据集、验证集上普遍能降低 10%以上的错误率,同时证实了在特定情境下传统技术依然具备独特且有效的应用价值。通过系统化的对比分析,本文旨在为中文文本分类技术的科学选型及未来发展方向提供有力依据与导向。

关键词:文本分类; BERT; 预训练大语言模型; 提示工程; 微调; 小样本学习
中图分类号:TP391.1 **文献标志码**:A **文章编号**:2095-2163(2024)06-0088-07

Comparative study on traditional and large model-based techniques for Chinese text classification: Leveraging both paradigms

WEN Fei
(ZhongZhuoxin (Beijing) Technology Co., Ltd., Beijing 100085, China)

Abstract: This paper focuses on exploring and practicing the evolution of Chinese text performance differences between traditional methods and advanced algorithms based on large models across various text classification tasks. The paper delves into extensive investigations across foundational datasets for sentiment analysis and multi-class text datasets laden with intricate professional information, systematically comparing traditional statistical learning approaches, classical deep learning algorithms, and the currently influential pre-trained large models such as BERT and LLMs. Central to the proposed research is the enhancement of classification accuracy, while concurrently assessing the resource efficiency and training time effectiveness of each model. With respect to pre-trained large models, the paper employs prompt engineering techniques and model fine-tuning strategies to optimize their performance. The proposed experimental outcomes vividly demonstrate the substantial advantages of large models in understanding and leveraging linguistic context, thereby boosting generalization capabilities, universally reduces the error rate by more than 10% across diverse datasets and validation sets. Meanwhile, the proposed findings confirm the unique and effective application value of conventional techniques under specific scenarios. Through systematic comparative analyses, this study aims to provide strong evidence and direction for the scientific selection and future development path of Chinese text classification technologies.

Key words: text classification; BERT; pre-trained large language models; prompt engineering; fine-tuning; few-shot learning

0 引 言

文本分类在信息检索、智能对话系统构建、任务自动化分配等诸多现实应用场景中起着不可或缺的作用,是现代自然语言处理技术(NLP)的核心研发领域之一。工业生产的需求持续驱动着研究者们不断追求更高性能、更快训练速度以及更具成本效益的文本分类技术革新。本文以探究中文文本分类技术的进步与实践为核心,通过对传统技术手段与新兴大模型方法在文本分类任务上的详实对比实证研究,深度剖析其各自特性与优势。

研究将审视经典的文本特征提取方法与传统的

作者简介:文 飞(1981-),男,本科,主要研究方向:自然语言处理。Email:21645988@qq.com
收稿日期:2024-05-12

机器学习分类算法在文本分类任务中的应用现状。同时，聚焦近年来引领技术潮流的大规模预训练模型，如 BERT、LLM 等，深入探讨如何实现的变革并优化中文文本分类的表现。在此基础上，进行了深入的研究实验，试图从不同角度呈现、评估各类方法的性能。

通过这些对比与分析，本文致力于为中文文本分类技术的未来发展提供坚实的理论依据与实用指导，以期为科研工作者和实践者在选择与优化文本分类技术时提供有价值的方向性参考。

1　相关知识

1.1　文本分类

文本分类是将文本数据自动划分为预定义的一组类别，属于监督学习的一种。文本分类广泛应用于诸多场景，如情感分析、主题分类等。文本分类技术的发展历程见证了从传统方法到现代人工智能技术的转变。早期，基于规则的方法占主导地位，这一时期的特点是手工设计特征和规则，依赖于专家知识，泛化能力有限。随后，随着数据量的增长和计算能力的提升，统计学习方法逐渐兴起，这些方法开始利用数据驱动的方式来自动学习特征，提高了模型的泛化性能，但在处理高维稀疏数据和复杂语言结构方面仍面临挑战。进入 21 世纪第 2 个 10 年，基于深度学习的方法、特别是神经网络的广泛应用彻底改变了文本分类的方式。循环神经网络、长短时记忆网络以及卷积神经网络等模型的引入，极大地提升了对文本序列理解和上下文捕捉的能力，使得模型能够自动学习到更深层次的语义特征。近年来，随着预训练大模型如 BERT、GPT 系列等模型的出现，文本分类技术达到了新的高度。这种方法不仅大幅提高了分类任务的性能，还显著减少了对标注数据的依赖，展示了极强的迁移学习能力，成为了当前文本分类技术的主流趋势。

1.1.1　基于规则的方法

（1）词语匹配：基于关键词或短语的匹配是最早的文本分类方法之一，根据文本中是否存在特定词语或短语来决定其类别。这种方法简单直观，但容易受词汇多样性、同义词和多义词，及反讽、多重否定等修辞方法影响，准确率和泛化能力比较有限。

（2）规则引擎：基于领域专家经验根据文本关键词、结构或者模式编写规则构建规则库，设计规则处理系统，利用规则引擎的逻辑处理能力来分类文本。这种方法需要大量的人工劳动和领域知识。

1.1.2　统计学习方法

（1）朴素贝叶斯分类器（Naive Bayes Classifier，NB）：这种分类算法通过逐一计算每个特征对目标类别的影响程度，独立计算各个特征对结果类别的条件概率，依据各类别条件概率的最大化原则，预测文本分类目标类别。

（2）支持向量机（Support Vector Machines，SVM）：SVM 通过建立一个最优的超平面结构来分割各种类别的训练文本信息，间隔最大化不同类别的数据点，具有优良的泛化性能。

（3）逻辑回归（Logistic Regression，LR）：这是一种广义线性模型，将文本通过特征提取后的文本向量作为输入特征，通过一个线性函数进行计算，得到一个线性预测值，然后通过逻辑函数将其转换为概率值，优势在于模型结构简单、易于解释，在数据量不大时也能取得较好的效果。

（4）随机森林（Random Forest，RF）：这是一个高度集成的模型，利用对决策树模型的大量集成，实现了对数据的有效分析与回归估计，在现实应用中取得良好效果。

（5）XGBoost（eXtreme Gradient Boosting）：这是一个先进的梯度提升式决策树计算的实现，是最主要的机器学习方法之一。

在利用统计学习方法进行文本分类前，需要先对文本进行预处理和特征提取工作，对文本进行向量化。常用的特征提取方法有 Bag-of-Words 和 TF-IDF。其中，Bag-of-Words 模型将文本表示为一个词频向量，向量每一个维度对应字典中的一个词，数值则为词在文本中出现的次数。TF-IDF 由词频（TF）和逆文档频率（IDF）两部分组成，综合体现了词在文档中的局部重要性和全局重要性，其计算公式为：

$$TF(t,d)=\frac{n_{t,d}}{\sum_{w\in d} n_{w,d}} \tag{1}$$

$$IDF(t,D)=\log\left(\frac{|D|}{|\{d\in D:t\in d\}|}+1\right) \tag{2}$$

$$TF-IDF(t,d,D)=TF(t,d)\times IDF(t,D) \tag{3}$$

其中，$n_{t,d}$ 表示词 t 在文档 d 中出现的次数；$\sum_{w\in d} n_{w,d}$ 表示文档 d 中所有词出现的次数之和；$|D|$ 表示文档集中所有文档的总数；$|\{d\in D:t\in d\}|$ 表示包含词 t 的文档的数量。

1.1.3　基于深度学习和神经网络的分类方法

（1）神经网络（Neural Networks，NN）：前馈神

经网络被应用于文本分类的早期阶段,然而,单层或多层神经网络在捕捉文本的序列性和上下文依赖性时能力有限。是基础灵活的模型,适用于简单场景的分类任务,如基于关键词的分类。

(2)卷积神经网络(Convolutional Neural Networks, CNN):通过对文本向量的卷积操作,CNN 能够有效学习到文本的局部特征表示,通过池化层转变为固定长度的全局特征向量,最后再利用一个全连接层进行分类。能够捕捉文本局部特征,计算效率高,但不擅长充分捕捉文本的序列性和长距离依赖关系。适合短文本分类和需要识别文本局部特征模式的任务。

(3)循环神经网络(Recurrent Neural Networks, RNN):RNN 具有循环的结构,能够保留并利用前面序列的状态信息,捕捉数据中的序列模式,是一种设计用来处理序列数据的神经网络模型。特别适合序列数据的分类任务,缺点是训练速度相对较慢,且可能发生梯度消失或者爆炸问题。

(4)长短时记忆网络(Long Short - Term Memory, LSTM):在 RNN 基础上为解决长距离依赖问题而发展出来的变体,创造性地引入门控机制,有效地提升模型性能[1]。在提高了模型复杂度和计算量的同时,也对模型调试和理解带来了挑战,但解决了 RNN 的梯度消失问题,在处理长序列数据时表现优异,适用于复杂、高级文本分类任务。

1.1.4 基于 Transformer 的大规模预训练模型分类方法

Transformer 架构是一种深度学习模型架构,通过对自注意力机制的成功应用,解决了长距离依赖的问题,并由于允许大规模并行计算,极大地提高了模型训练效率。自注意力机制的计算公式为:

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{Q},\ \boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}^{K},\ \boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}^{V} \tag{4}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

其中,$\boldsymbol{X}$ 表示输入数据;$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别表示查询(Query)、键(Key)和值(Value)向量;$\boldsymbol{W}^{Q},\boldsymbol{W}^{K},\boldsymbol{W}^{V}$ 分别表示其对应的权重矩阵;$\sqrt{d_k}$ 表示缩放因子[2]。Transformer 架构在机器翻译、文本分类、问答系统、文本生成、语音识别等多种自然语言处理任务中都取得了显著效果,并发展出 BERT、RoBERTa、GPT 系列等大规模预训练模型。

BERT (Bidirectional Encoder Representations from Transformers)是在 Transformer 架构上发展出来的一种双向编码器架构,主要特点包括预训练与微调、双向 Transformer 编码器结构等。在执行文本分类任务时,在预处理阶段,对输入文本添加[CLS]作为序列开头标识。在预训练过程中,其独特的双向 Transformer 编码器结构,能够同时考虑词语的所有上下文信息。在输出层,计算出[CLS]标记对应的隐藏状态向量,作为输入文本的特征向量,并利用全连接层进行分类输出,或者进一步构建神经网络模型进行分类输出。

GPT 模型的核心是依托于 Transformer 架构,通过自我监督学习策略,在大规模无标签文本数据集上进行训练,从而深刻理解和捕捉到复杂的语言结构与模式。目前,GPT 系列模型已在广泛的自然语言处理任务上展现出了卓越性能。尽管原始设计初衷并非专门针对文本分类任务,但鉴于其对自然语言深度理解的能力,GPT 模型可以通过创新的应用策略、如提示工程技术(Prompt Engineering)以及模型微调(fine-tuning)手段,成功地拓展到文本分类应用场景中,并取得良好的分类效果。这意味着 GPT 模型经过适当的调整和配置,同样能够成为一种有力的文本分类工具。

1.2 提示工程技术

提示工程技术(Prompt Engineering)是与预训练大语言模型进行沟通的艺术,通过技术性精巧设计的提示词,规范、引导大语言模型的行为和输出,从而获得更好的性能,更具有实用性和准确性的输出产物[3]。在与像 GPT 这样的大语言模型打交道时,提示词不仅仅是简单的输入指令,而是通过构造具有一定上下文、结构和目标导向性的文本序列,来激发模型更好地理解用户意图并生成符合期望的答案或内容。通常包括提示词设计、参数调整、实验与优化等环节。通过提示工程技术,不仅可以提高模型的实用性和准确性,还可以探索模型潜在能力的边界,推动自然语言处理技术的实际应用价值。

思维链(Chain-of-Thought, CoT)是提示工程技术的一种形式,要求预训练大语言模型在处理复杂推理任务时,按照用户的要求,执行并显式地输出思考推理过程,而不是按照大语言模型的“直觉”直接输出结果,通过推理过程的思考,可以显著提高大语言模型解决特定问题的能力,提高输出的准确性[4]。

1.3 Fine-tuning

Fine-tuning 是机器学习和深度学习领域中的一种常见技术。为了让预训练模型在具体的任务上有更强的表现力,往往需要对其进行微调(Fine-

tuning)，即在预训练模型的基础上，利用目标任务的特定数据集重新训练模型的部分或全部参数，以进一步优化模型在特定任务上的表现。微调过程中，模型会学习到更具体的模式和特征，从而更好地适应新任务的需求。

LoRA(Low-Rank Adaptation) 是 Fine-tuning 技术的一种，通过向模型中的指定权重矩阵添加低秩矩阵进行修正优化，而不需要修改所有的权重矩阵，通过非常少量的参数调整，大量减少训练优化时间和资源成本，并提升模型的性能。这意味着 LoRA 只微调一小部分额外参数(低秩矩阵分解出的因子)，而不是直接调整整个模型的大量参数。

具体来说，LoRA 在保持预训练模型原有大部分权重不变的同时，通过引入 2 个较小的低秩矩阵(通常维度较低)相乘，来更新原始权重矩阵，这种方式有效地限制了模型在微调阶段所需的计算资源和存储空间，同时减少了对预训练知识的潜在破坏。就是强调在不显著改变预训练模型参数的前提下，通过低秩近似来进行高效且有效的微调操作。经过 LoRA 微调后，原权重矩阵的操作变为：

$$\boldsymbol{h} = \boldsymbol{W}_0\boldsymbol{x} + \boldsymbol{BAx} \tag{6}$$

其中，$\boldsymbol{W}_0$ 表示原权重矩阵；$\boldsymbol{A}$、$\boldsymbol{B}$ 表示 LoRA 引入的低秩矩阵；$\boldsymbol{x}$ 表示输入向量；$\boldsymbol{h}$ 表示输出矩阵[5]。LoRA 可将可训练参数减少上万倍，资源需求降低 3 倍以上，同时保持甚至提升模型性能。

随着预训练大模型技术的突飞猛进的发展，利用预训练大模型进行文本分类的研究也越来越受到关注和重视。Hegselmann 等学者[6](2023)探讨了预训练大语言模型在零样本和少量样本情况下的性能，显示其能与传统方法相媲美；Sun 等学者[7](2023)提出了名为“线索与推理提示”(Clue And Reasoning Prompting，CARP)的新方法用于提升预训练大模型的分类性能，刷新了多项基准测试的纪录；Zhang 等学者[8](2024)提出了 RGPT 的自适应提升框架，在基准测试中胜过多个最先进的预训练模型和最先进的 LLM，并超过了人类分类的表现。

2　实验

2.1　情感分析文本分类实验

本研究旨在对各类分类方法在相对清晰明了的情感分类数据集上的性能进行基础验证。此次实验拟将初步考察多种分类技术的基准性能，尚未进行全面的参数调优，故所得结果仅反映了各方法在未经精细优化条件下的基本效能，而非其潜在的最佳性能表现。

在本研究中，选用 ModelScope 平台上提供的商品评论情感预测数据集“jd”作为实验数据源。训练集总共有45 367条数据，在实验中，使用了包括传统统计学习方法、BERT 和基于大模型的 Prompt 及微调分类方法进行了测试验证和效果评估。

对于 BERT 分类任务，采用了 Bert For Sequence Classification 模型架构，并对 2 种不同预训练模型—“bert-base-chinese”及“bge-large-zh-v1.5”进行了测试。

在运用大模型进行 prompt 分类时，采取了直接询问大模型并获取分类结果的方式，暂未采用提示词工程技术进行优化。评估的大模型包括但不限于 qwen-turbo、qwen-7b-chat、qwen-max 和 chatglm3-6b。

在设计 prompt 时，遵循了角色明确、指示清晰的原则，并结合实例引导，构造了由角色描述、指示信息、实例展示以及输入和输出共同构成的提示词模板(见表 1)，以适应不同模型的特点并提高其在文本分类任务中的表现。

表 1　提示词模板

Table 1　Prompt template

角色描述	你是一个情感分类模型
指示信息	请根据用户评价进行分类，1 表示正面评价，0 表示负面评价，只需要输出分类结果(0 或者 1)，不需要输出任何其它内容
实例展示	用户评价：一百多和三十的也看不出什么区别，包装精美，质量应该不错 输出结果：1
输入	用户评价：{user_input}
输出	输出结果

此处，运用了 one-shot learning 策略，通过固定示范样本来指导和规范模型的输出格式，以期在保持输出一致性的同时，提高分类结果的准确性。

针对大模型微调分类任务，鉴于硬件资源配置和训练时间成本的考量，选择了开源的 chatglm3-6b 模型进行精细化微调。为了适应有限资源条件，训练过程中采用了小规模训练集方案，具体为从训练集中分别抽取了前 100 条和前 500 条样本进行训练。经实验验证，所得测试结果见表 2。

表 2 jd 数据集情感分类结果比较

Table 2 Comparison of sentiment classification for jd dataset

Method	Memory usage/GB	GPU usage/GB	Training time/s	*Accuracy*
Naive Bayes Classifier	0.235	–	0.013	0.880
Logistic Regression	0.240	–	0.347	0.900
Random Forest	0.371	–	90.009	0.880
Support Vector Machines	0.243	–	509.200	0.880
XGBoost	0.281	–	477.931	0.790
TextCNN	2.523	0.813	164.876	0.885
BERT (based on the bert-base-Chinese model)	3.473	5.429	5.335×10^3	0.935
BERT (based on the bge-large-zh-v1.5 model)	4.501	13.041	17.359×10^3	0.920
Prompt(based on the chatglm3-6b model)	–	–	–	0.870
Prompt(based on the qwen-7b-chat model)	–	–	–	0.855
Prompt(based on the qwen-max model)	–	–	–	0.900
Prompt(based on the qwen-turbo model)	–	–	–	0.880
Fine-tuning(based on the chatglm3-6b model,100 samples)	2.512	22.766	528.854	0.910
Fine-tuning(based on the chatglm3-6b model,500 samples)	2.505	22.766	2.648×10^3	0.930

通过上述实验数据分析可见,BERT 模型在执行文本分类任务时的性能表现极为出色。在情感分类场景中,即便未经专门训练与细致调优,大型语言模型(LLM)依旧能够展现出令人满意的性能水平,这有力证明了此类大模型对于自然语言的语义内涵及情感色彩已有深刻理解,并且随着模型参数规模的扩大,其性能也相应得到了显著提升。

通过分析实验数据,不难发现,大模型经过训练微调后,可显著提升其性能。即使在一个参数规模相对不是很大的模型基础上,通过对一个小规模的训练集进行微调训练,微调后的模型性能也能超越参数规模更大的未经微调的预训练模型,说明微调对模型性能作用较大,可显著降低大语言模型的训练及应用成本,有着重大的实践意义。这意味着,在有限的训练数据条件下,通过微调大模型同样可以达到较高的分类性能。也意味着,小参数量大模型在适当微调后,也能达到媲美、甚至超越大参数量模型的优秀效果。鉴于这些小参数量大模型的开源属性、训练和运行成本较低的优势,在实际应用中具有广泛的实用性与巨大的发展潜力。

值得注意的是,不论是机器学习模型、还是预训练模型,往往伴随着较高的内存占用和 GPU 资源需求。这一需求随着模型参数量的指数级增长而急剧上升,直接导致训练时间和维护成本的大幅增加。尤其是预训练大语言模型,其训练成本往往不是一般的项目可以接受的。因此,在生产环境的选型决策中,模型的资源效率、训练时间成本及其后续维护开销成为了不可忽视的关键考量要素。

另外,考虑到数据集中存在评论内容与标签不符或矛盾的现象,通过人工筛查并剔除了一部分问题数据,实验结果见表 3。这一举措对训练集质量的改善产生了积极影响,各模型的分类准确率均有显著提升。在排除这些干扰因素后,特别是对于超大规模参数的大模型、例如 qwen-max,即使在未经任何微调优化的状态下,在情感分类任务中也已然展现出了极为优越的性能。

表 3 清洗后数据集情感分类结果比较

Table 3 Comparison of sentiment classification for filtered dataset

Method	*Accuracy*
Naive Bayes Classifier	0.920
Logistic Regression	0.940
Random Forest	0.930
Support Vector Machines	0.940
XGBoost	0.840
TextCNN	0.925
BERT (based on the bert-base-Chinese model)	0.980
BERT (based on the bge-large-zh-v1.5 model)	0.970
Prompt(based on the chatglm3-6b model)	0.905
Prompt(based on the qwen-7b-chat model)	0.905
Prompt(based on the qwen-turbo model)	0.945
Prompt(based on the qwen-max model)	0.985
Fine-tuning(based on the chatglm3-6b model,100 samples)	0.975
Fine-tuning(based on the chatglm3-6b model,500 samples)	0.975

通过实验数据分析,总结了各种文本分类方案的特点见表 4。

表 4　文本分类方案比较

Table 4　Comparison of text classification scheme

方案	准确率	优点	缺点
统计学习方法/TextCNN	中	模型小，训练快，资源要求低，成本低	性能和泛化能力相对较弱
BERT	高	模型中等	训练时间长，资源要求中等
大模型 Prompt	中/高	不需要训练，可直接利用大模型的能力	资源要求高，成本高 高性能的大语言模型往往不开源
大模型微调	高	可以对小参数量大模型进行小数据量微调，达到媲美大参数量模型的优秀效果，有诸多开源模型可以选用	资源要求相对较高

2.2　复杂文本多分类实验

2.1 节的实验虽然突显了大模型在某些文本分类任务中的卓越性能，但仍不能据此断定大模型能完美应对所有文本分类挑战。事实上，在面对复杂文本的多类别分类问题时，大模型的性能表现并未臻至理想，这意味着还要探寻更加有效的策略，有助于大模型适应此类复杂任务。

接下来的实验采用了 NCAA2024-中文糖尿病问题分类评测数据集，该数据集源自糖尿病医学专业领域，其内含问题的复杂性较高。部分问题涉及到医学专业知识，分类的界限也比较精妙，需要充分掌握相关的领域专业知识，并仔细研究分类规律斟酌分类界限，即便如此，人工标注也可能存在误差。

与情感分类实验一样，使用相同的方法，对这个数据集进行了测试评估，得到的实验结果见表 5。

表 5　NCAA2024 数据集文本分类结果比较

Table 5　Comparison of text classification for NCAA2024 dataset

Method	*Accuracy*
Naive Bayes Classifier	0.690
Logistic Regression	0.710
Random Forest	0.670
Support Vector Machines	0.710
XGBoost	0.670
TextCNN	0.695
BERT (based on the bert-base-Chinese model)	0.769
BERT (based on the bge-large-zh-v1.5 model)	0.770
Prompt(based on the chatglm3-6b model)	0.520
Prompt(based on the qwen-7b-chat model)	0.520
Prompt(based on the qwen-turbo model)	0.530
Prompt(based on the qwen-max model)	0.520
Fine-tuning(based on the chatglm3-6b model,500 samples)	0.737
Fine-tuning(based on the chatglm3-6b model,9 000 samples)	0.783

从实验数据中能够看出，BERT 和经过微调的预训练大语言模型继续保持了优秀的性能表现。与此相反，单纯依赖大语言模型 Prompt 直接进行专业领域文本分类的准确率并不理想。究其原因可知，仅凭一个分类名去精准划分专业文本类别，即使是人类专家也可能面临很大挑战。

然而，进一步分析得到，是否可以通过对分类依据进行详细的描述，让大模型掌握基本分类规则，从而提升分类准确率呢？研究可知，这个问题并不容易解决。一方面，往往难以用简洁的语言来完整表述分类标准；另一方面，即使设法将分类规则进行了精炼表达，大模型对这些规则的理解和运用也无法与人类智慧相提并论。此外，尤其在多类别间的边界模糊地带，大模型的表现则常常不尽如人意。

详细描述分类规则后采用大语言模型 Prompt 方法分类的实验结果见表 6。尽管研究尝试优化分类提示以期提升准确率，但从实验结果来看，此类改进并未带来显著的效果提升。进一步引入思维链(Chain of Thought)后，虽然分类性能略有增长，但总体提升幅度依然有限，表明当前技术下，在复杂专业文本分类问题上，大模型的表现仍有很大的提升空间。

表 6　大语言模型提示工程分类结果比较

Table 6　Comparison of prompt engineering results for LLMs

Method	*Accuracy*
chatglm3-6b	0.525
qwen-7b-chat	0.525
qwen-turbo	0.530
qwen-max	0.525
qwen-max(employing COT)	0.590

为提升大语言模型 Prompt 分类方法的准确性，尝试采用 knn-few-shot learning 策略，即在分类请求的 Prompt 中融入相关示例。具体来说，通过从训练集文本向量中选出最近似的 10 个样本作为参照案例，随后将这些示例与待分类文本一起提交至大模型以获取分类结果。从实验结果(见表 7)分析不难发现，这种方法显著提升了分类准确率，而且大语言模型的参数量越大，提升的效果越明显。

表 7 大语言模型提示工程仿真结果比较

Table 7 Comparison of prompt engineering results for LLMs

Method	*Accuracy*
chatglm3-6b	0.565
qwen-7b-chat	0.655
qwen-turbo	0.740
qwen-max	0.775

3 结束语

通过本次实验探究,得出以下关键发现:传统的统计学习和机器学习分类技术,在处理较为简洁、要求适度准确率的文本分类任务时,仍保持可观的性能表现,并在对运行效率和资源要求严苛的环境中展现出广泛应用价值。而对于复杂度较低的任务场景,这类方法因其实现简易、资源需求较小而广受欢迎。

BERT 模型在实验中展现了稳定优秀的性能,且模型尺寸适中,运行维护成本不高,这都有利于生产实践中的应用。训练时则需要一定的硬件资源投入,尤其是面对大规模训练数据集时训练耗时较长,成本较高。

基于预训练大语言模型的分类方案在处理通用性场景下有不错的性能,在资源消耗和成本方面相对较高,同时也带来了效果突出的性能优点。这些大语言模型可以称作智能底座,具有零样本学习和小样本学习的能力,只需要少量样本即可通过微调等技术手段获得出色的性能。其灵活性使其能在不同生产环境下适应各种应用场景,特别是在文本分类乃至整个自然语言处理领域的广阔潜力尚待深度挖掘,预期将在未来带来更多的创新成果,创造更大实际应用价值。

参考文献

[1] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. Long Beach, USA: NIPS Foundation, 2017, 30: 5998-6008.

[3] WHITE J, FU Q, HAYS S, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt [J]. arXiv preprint arXiv:2302.11382, 2023.

[4] WEI J, WANG Xuezhi, 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.

[5] HU E J, SHEN Yelong, WALLIS P, et al. Lora: Low-rank adaptation of large language models [J]. arXiv preprint arXiv: 2106.09685, 2021.

[6] HEGSELMANN S, BUENDIA A, LANG H, et al. Tabllm: Few-shot classification of tabular data with large language models [C]// International Conference on Artificial Intelligence and Statistics. Valencia, Spain: PMLR, 2023: 5549-5581.

[7] SUN Xiaofei, LI Xiaoyu, LI Jiwei, et al. Text classification via large language models [J]. arXiv preprint arXiv: 2305.08377, 2023.

[8] ZHANG Yazhou, WANG Mengyao, REN Chenyu, et al. Pushing the limit of LLM capacity for text classification [J]. arXiv preprint arXiv:2402.07470, 2024.

[9] CHEN Y. Convolutional neural network for sentence classification [D]. Waterloo, Canada: University of Waterloo, 2015.

[10] LIU N F, LIN K, HEWITT J, et al. Lost in the middle: How language models use long contexts [J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173.

[11] BAHARAT S M, MYRZAKHAN A, SHEN Z. Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4 [J]. arXiv preprint arXiv:2312.16171, 2023.

[12] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

[13] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [J]. arXiv preprint arXiv: 1806.05266, 2020.

[14] KOWSARI K, JAFARI M K, HEIDARYSAFA M, et al. Text classification algorithms: A survey [J]. Information, 2019, 10 (4): 150.

[15] ABBURI H, SUESSERMAN M, PUDOTA N, et al. Generative ai text classification using ensemble llm approaches [J]. arXiv preprint arXiv:2309.07755, 2023.

[16] CHEN Siyuan, WU Mengyue, ZHU K Q, et al. Llm empowered chatbots for psychiatrist and patient simulation: Application and evaluation [J]. arXiv preprint arXiv:2305.13614, 2023.

[17] WEI Fusheng, KEELING R, HUBER-FLIFLET N, et al. Empirical study of LLM fine-tuning for text classification in legal document review [C]//2023 IEEE International Conference on Big Data (BigData). Sorrento, Italy: IEEE, 2023: 2786-2792.

[18] HOWARD J, RUDER S. Universal language model fine-tuning for text classification [J]. arXiv preprint arXiv:1801.06146, 2018.

[19] MENG Yu, ZHANG Yunyi, HUANG Jiaxin, et al. Text classification using label names only: A language model self-training approach [J]. arXiv preprint arXiv:2010.07245, 2020.

[20] ZHAO Biao, JIN Weiqiang, DEL S J, et al. ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification [J]. Neurocomputing, 2023, 557: 126708.

[21] MAYER C W F, LUDWIG S, BRANDT S. Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models [J]. Journal of Research on Technology in Education, 2023, 55(1): 125-141.

[22] GUO Yuting, OVADJE A, AL-GARADI M A, et al. Evaluating large language models for health-related text classification tasks with public social media data [J]. arXiv preprint arXiv: 2403.19031, 2024.

[23] 王淼,丁德锐. SmBERT(SmallerBert):一种更小更快的文本分类模型[J]. 智能计算机与应用,2023,13(1):129-135.

[24] 张铭泉,周辉,曹锦纲. 基于注意力机制的双 BERT 有向情感文本分类研究[J]. 智能系统学报,2022,17(6):1220-1227.

[25] 李可悦,陈轶,牛少彰. 基于 BERT 的社交电商文本分类算法[J]. 计算机科学,2021,48(2):87-92.

[26] 严佩敏,唐婉琪. 基于改进 BERT 的中文文本分类[J]. 工业控制计算机,2020,33(7):108-110,112.