面向医学科技文献分类的语义特征增强研究*

宫小翠 安新颖

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔摘要〕 **目的/意义** 构建大批量医学科技文献自动分类方法，以应对医学科技文献快速增长给文献分类和利用带来的新挑战。**方法/过程** 以医学论文为研究对象，利用《医学主题词表》同义词和语义层级结构，增强概念信息的语义特征，采用双向编码器表征模型进行微调训练和测试评估，并与随机森林算法的分类结果进行对比。**结果/结论** 十折交叉验证结果显示，该分类方法精确率、召回率、*F*1 值分别达到 95.42%、93.61%、94.47%，优于随机森林算法及其他未进行特征增强的方法，其准确、有效，具有可应用性。

〔关键词〕 医学科技文献；《医学主题词表》；双向编码器表征；自动分类

〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2025.03.007

Study on Semantic Feature Enhancement for Medical Literature Classification

GONG Xiaocui, AN Xinying

Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔**Abstract**〕 **Purpose/Significance** To build an effective automatic classification method for a large number of medical literatures, so as to cope with the new challenges brought by the rapid growth of medical literatures for their classification and utilization. **Method/Process** Taking medical literatures as data source, the study utilizes the synonyms and hierarchical structure of the medical subject headings (MeSH) to enhance the semantic features of concept information, uses bidirectional encoder representations from transformers (BERT) for fine-tuning and testing, and compares the classification results with random forest (RF). **Result/Conclusion** The results of the ten-fold cross-validation method show that the precision, recall and *F*1 score of this medical literature classification method are 95.42%, 93.61%, 94.47%, which are better than the classification results of RF and other methods without feature enhancement, and show accuracy, effectiveness and applicability.

〔**Keywords**〕 medical literatures; medical subject headings (MeSH); bidirectional encoder representations from transformers (BERT); automatic classification

1 引言

科技创新大潮下，科技文献成为学术成果的重要体现，数量逐年增多。科技论文涵盖领域广泛且多样，对不同领域学者或机构进行科技产出评估需要将论文细分到不同领域。探索有效精准的科技论文分类方法有助于快速识别相关领域科技论文，节省人力。医学科技文献蕴含医学专业术语，标题和

〔修回日期〕 2024-10-25

〔作者简介〕 宫小翠，助理研究员，发表论文 10 余篇；通信作者：安新颖，博士，研究员。

〔基金项目〕 中国医学科学院/北京协和医学院医学信息研究所/图书馆青年人才培养专项（项目编号：2024YT14）。

关键词中的医学术语往往是区分类别的关键特征。但标题和关键词文本较短，表达领域知识不够充分，影响分类效果。增强这些特征，是提高医学文献分类精准度的关键，而相关研究尚不充分。本研究利用《医学主题词表》（medical subject headings，MeSH）[1]的同义词和上下位等级关系，对各领域医学术语进行语义增强，并运用双向编码器表征（bidirectional encoder representations from transformers，BERT）[2]模型进行微调训练和测试评估。

2 BERT 模型

随着人工智能[3]技术的高速发展，机器学习在医疗领域已有广泛应用[4]。机器学习算法包括支持向量机、决策树、K－最近邻、随机森林、朴素贝叶斯等方法[5]，无法捕捉文本的语义关联。深度学习[6]能够从原始文本中自动提取隐含语义信息，相关研究正逐年增多。尤其是 Transformer[7]神经网络结构提出后，文本分类领域不断尝试新方法，如 Chang W C 等[8]提出 X－Transformer 模型，在 4 个基准数据集上取得最佳分类效果；Bello A 等[9]将 BERT 与卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）及双向长短期记忆（bidirectional long short－term memory，BiLSTM）组合，用于微博用户情感分类，实验结果表明融合 BERT 模型效果更佳；吴雪华等[10]采用支持向量机、逻辑回归、文本 CNN、BERT 以及 BERT＋TextCNN 的组合模型对应急行动支撑信息分类，结果显示 BERT 和 BERT＋TextCNN 模型优于其他模型。

BERT 是 2018 年提出的预训练语言模型，以 Transformer 双向编码器为基础，设计大量多头注意力机制，利用大规模训练数据学习通用知识，再用少量领域内数据微调，在文本分类领域取得较好效果[11]。BERT 是自编码语言模型，主要采用两个任务预训练模型。一是遮蔽语言模型（masked language model，MLM）：在训练数据中随机遮蔽掉一些单词，让模型预测这些被遮蔽的单词，学习每个单词在上下文中的相对位置和语义信息。二是下一句预测（next sentence prediction，NSP）：给模型提供两个句子，让其预测，学习单词在上下文中的位置和语义。BERT 模型的核心思想是通过双向上下文建模来理解单词的含义，与传统的单向语言模型相比，BERT 同时考虑上下文信息，使模型能够更好地捕捉单词之间的关系。

3 医学文献分类实验设计

3.1 实验流程

选取 SinoMed 数据库 2020—2022 年发布的 53 万条数据作为数据源，基于既往研究[12]，通过构建医学分词词典，对中文论文标题和关键词进行分词、去停用词处理，人工标注分类信息并分组交叉审核，暂不考虑摘要信息。利用 MeSH 对数据集进行特征增强，以提高分类精度，利用其入口词和树状结构等信息，构建特定类别下的特征词表。本研究提出基于语义的特征遴选方法，充分考虑特征词的语义信息，用 BERT 模型进行微调训练，并与特征增强前结果对比，选用随机森林算法作为对比模型，实验流程，见图 1。

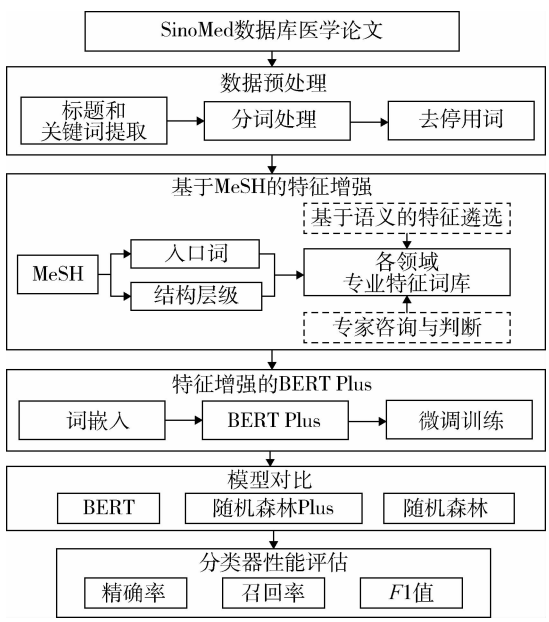


图 1 医学文献分类实验流程

基于 BERT－Base－Chinese 模型进行微调。BERT－Base－Chinese 模型在简体和繁体中文文本上训练，能学习中文文本的语义和语法信息。该模型含有 12 层 Transformer 编码器，每层有 12 个自注意力

头。使用 Python – 3.6 和 Tensorflow – gpu – 2.3.0 构建神经网络模型，NVIDIA RTX 4070 显卡加速训练和推理，CUDA 和 cuDNN 版本分别为 10.1 和 7.6。优化器使用 tf. keras. optimizers. Adam，损失函数使用 tf. keras. losses. SparseCategoricalCrossentropy。

随机森林算法分类精度高、训练速度快，抗拟合能力强，能分析大规模样本，评价特征重要程度[13]。已有研究[14]证实其对医疗短文本分类效果较好，因此选用随机森林算法作为对比模型。为提高模型精确度，对随机森林中的子树数量（n_estimators）、树的高度（max_depth）和最大特征数（max_features）3 个参数进行随机搜索，选择最佳参数。

3.2 基于 MeSH 的特征增强

MeSH[15]作为医学领域最权威的词表之一，涵盖大量医学专业术语和词汇，其入口词提供同义词信息，树状结构信息给出上下位等级关系，见表 1。利用 MeSH 在各领域的同义词、树状上下位概念词及领域专家知识，初步构建各学科特征词表，英文词汇翻译为中文，特征词汇用 Word2Vec[16]进行向量化表示，再与文献特征词汇计算余弦相似度，遴选保留大于某阈值的特征词汇，见图 2。

表 1 MeSH 检索结果（以干眼症（dry eye syndromes）为例）

主题词	入口词	树状结构信息
Dry Eye Syndromes	Dry Eye	Eye Diseases [C11]
	Dry Eye Disease	Lacrimal Apparatus Diseases [C11. 496]
	Evaporative Dry Eye	Dacryocystitis [C11. 496. 221]
	Evaporative Dry Eye Disease	Dry Eye Syndromes [C11. 496. 260]
	Evaporative Dry Eye Syndrome	Keratoconjunctivitis Sicca [C11. 496. 260. 394]
		Sjogren's Syndrome [C11. 496. 260. 719]
		Xerophthalmia [C11. 496. 260. 892]

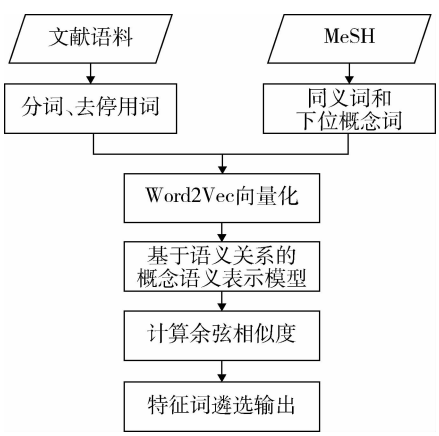


图 2 基于 MeSH 的特征增强

利用 Word2Vec 对特征词汇进行向量化表示。Word2Vec 无法解决一词多义问题，同时同义现象也使训练所得词向量不能完整实现概念的语义表示。因此借助 MeSH 同义词和下位概念词，设计基于语义关系的概念语义表示模型，实现对特征词的准确语义表示。在已训练得到词向量的基础上，以概念为单位将表示相同概念的所有同义词向量整合用于语义表示，得到具有所有同义词向量特征的概念向量。为避免消耗额外计算资源，采取平均池化[17]方法对同概念下所有同义词向量进行特征聚合，见公式（1）。式中，V_{term_i}为该概念同义词对应的词向量。另外，将下位概念的特征信息与概念本身特征信息融合，可以实现基于等级关系的概念语义扩展，将下位概念的特征信息与概念本身特征信息融合，形成关联向量 $V\_\text{relation}_{m,n}$，见公式（2）。式中，$V_{m,n}$代表节点 $\text{Node}_{m,n}$对应概念的向量表示 V_{concept}；ω 代表融合权重；$C_{m,n}$代表节点 $\text{Node}_{m,n}$对应概念的所有下位概念对应向量集合；$C_{m,n}(i)$ 代表节点 $\text{Node}_{m,n}$对应概念第 i 个下位概念的对应向量；Q 代表节点 $\text{Node}_{m,n}$对应下位概念的个数。

$$V_{\text{concept}} = \frac{1}{N}\sum_{i=1}^{N} V_{\text{term}_i} \quad (1)$$

$$V\_\text{relation}_{m,n} = \omega V_{m,n} + (1-\omega)\frac{1}{Q}\sum_{i=1}^{Q} C_{m,n}(i) \quad (2)$$

3.3 分类性能评价指标

利用分类结果的精确率（precision，P）、召回率（recall，R）和 $F1$ 值对分类模型进行综合评价。

本研究为多分类问题，因此需进一步计算整体的综合评价指标，使用“宏_(macro_)”的计算方法[18]，见表 2。其中 TP 表示实际为正且被确认为正的个数，FP 表示实际为负但被确认为正的个数，FN 表示实际为正但被确认为负的个数。

表 2 分类性能评价指标

指标名称	二分类计算方法	多分类计算方法
精确率	$P=\frac{TP}{TP+FP}$	$macro\_P=\frac{1}{n}\sum_{i=1}^{n}P_i$
召回率	$R=\frac{TP}{TP+FN}$	$macro\_R=\frac{1}{n}\sum_{i=1}^{n}R_i$
$F1$ 值	$F1=\frac{2\times P\times R}{P+R}$	$macro\_F1=\frac{1}{n}\sum_{i=1}^{n}F1_i$

4 实验过程与结果

4.1 实验过程

SinoMed 数据库中文论文数据含 R 开头分类号 cn_auo 字段，存在数量不一致、标注不准确等问题，需人工校对并标注为国标分类号 gb_code 和分类名 gb_name，作为文献标准分类结果，见表 3。考虑模型受数据量影响，选取妇产科学、皮肤病学、眼科学、神经病学、骨外科学、口腔医学、呼吸病学、消化病学、心血管病学、耳鼻咽喉科学共 10 个学科 21 万条数据进行实验。

表 3 SinoMed 数据库中文论文分类结果示例

编号	标题	关键词	cn_auo	gb_code	gb_name
2023185492	带蒂瓣修复桥体下龈瘤切除后组织缺损 1 例	龈瘤，带蒂转位瓣，软组织缺损	R781.05	32044	口腔医学
2023187473	经髂骨钉钉道骨盆外固定治疗骨盆骨折患者的疗效	骨盆损伤，骨折，骨钉，骨折固定术，治疗	R683.3	3202745	骨外科学

针对每个类别构建特征词表，利用 MeSH 同义词和下位概念词，如皮肤病学，检索皮肤病树状结构所有词汇及入口词，翻译为中文后，由领域专家判断并补充临床专业词汇。随后，将特征词进行语义向量化表示，与文献特征词计算相似度，探究融合权重 ω 和相似度阈值对分类结果的影响，当 ω 为 0.7、相似度权重为 0.75 时，分类效果最佳，准确率最高。论文特征增强结果，见表 4，如脑水肿增加“脑肿胀”“颅内水肿”等近义词，语义更明确，模型更易区分。每个学科领域保留的特征词汇，见表 5，词汇涵盖学科领域的疾病、病因、症状、治疗等信息，学科特征明显。

表 4 特征增强结果示例

学科	分词、去停用词结果	特征增强后结果
呼吸病学	慢阻肺、气道	慢阻肺、气道、慢性阻塞肺疾病、慢性气道阻塞
消化病学	胆囊炎	胆囊炎、胆囊积脓、胆囊炎症
神经病学	脑水肿	脑水肿、脑肿胀、颅内水肿、脑部水肿

表 5 10 个学科领域特征词示例

学科	特征词
妇产科学	产妇、促排卵、妊娠、宫颈炎、宫腔、刮宫等
皮肤病学	皮疹、痤疮、白癜风、尖锐湿疣、皮炎、毛囊炎等
眼科学	白内障、青光眼、角膜、结膜炎、晶状体、干眼病等
神经病学	帕金森、神经修复、脑科学、脑瘫、脑出血、脑水肿等
骨外科学	骨骼、脊柱、颈椎、骨肉瘤、半月板、肩关节等
口腔医学	牙周、义齿、牙髓、牙槽、龋病、拔牙等
呼吸病学	百日咳、肺纤维化、呼吸衰竭、肺癌、肺损伤、慢阻肺等
消化病学	丙型肝炎、肠穿孔、胆管炎、胆囊癌、肝衰竭、结肠炎等
心血管病学	肺心病、冠心病、心包囊肿、心肌病、心绞痛、心脏猝死等
耳鼻咽喉科学	鼻道、鼻畸形、鼻咽炎、鼻黏膜、耳聋、耳鸣等

使用 BERT-Base-Chinese 模型微调训练，参数设 max_length = 128，batch_size = 32，learning_rate = 2e-5，使用 Hugging Face 提供的 Transformers 库中 BertTokenizer 和 TFBertForSequenceClassification 进行分词和训练。使用随机森林模型计算特征重要度，基于袋外数据分类准确率，去除重要性差的特征，最后利用随机搜索法优化超参数，取 n_estimators = 55，max_depth = 12，max_features = 400，cri-

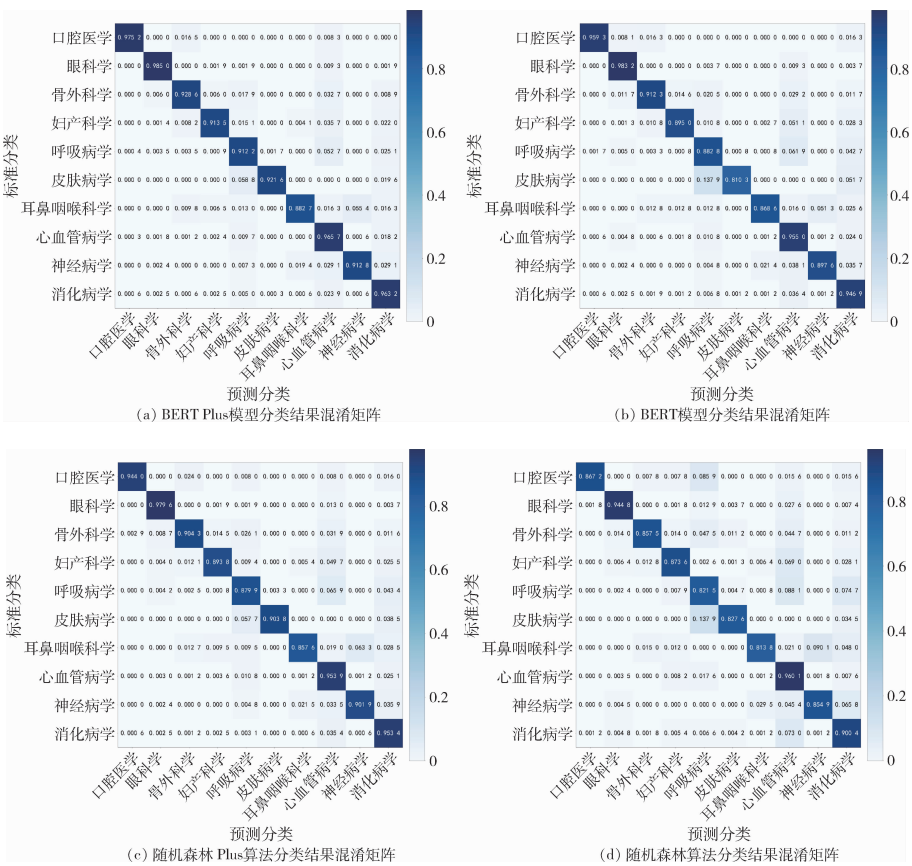
terion = gini，并评估分类精度。

4.2　实验结果

定义特征增强微调训练的 BERT 模型为 BERT Plus，特征增强的随机森林算法为随机森林 Plus。十折交叉验证方法测度 BERT Plus 分类结果的精确率、召回率、*F*1 值分别为 95.42%、93.61%、94.47%，见表 6，与未进行特征增强的 BERT 分类结果相比，平均 *F*1 值提高 2 个百分点，且在各学科领域 *F*1 值均较高。随机森林 Plus 相比随机森林模型，效果同样更优。实验表明，领域知识特征增强能更显著表征语义信息，提升分类效果。分类结果混淆矩阵，见图 3。

表 6　BERT 模型、随机森林算法分类结果

学科	BERT Plus			BERT			随机森林 Plus			随机森林		
	精确率	召回率	*F*1	精确率	召回率	*F*1	精确率	召回率	*F*1	精确率	召回率	*F*1
口腔医学	0.983 3	0.975 2	0.979 3	0.967 2	0.959 3	0.963 3	0.983 3	0.944 0	0.963 3	0.973 7	0.867 2	0.917 4
眼科学	0.972 3	0.985 0	0.978 6	0.954 7	0.983 2	0.968 8	0.961 7	0.979 6	0.970 5	0.946 5	0.944 8	0.945 6
骨外科学	0.945 5	0.928 6	0.936 9	0.934 1	0.912 3	0.923 1	0.931 3	0.904 3	0.917 6	0.941 7	0.857 5	0.897 7
妇产科学	0.979 4	0.913 5	0.945 3	0.977 9	0.895 0	0.934 6	0.970 8	0.893 8	0.930 7	0.939 6	0.873 6	0.905 4
呼吸病学	0.953 0	0.912 2	0.932 2	0.946 2	0.882 8	0.913 4	0.955 6	0.879 9	0.916 2	0.924 0	0.821 5	0.869 7
皮肤病学	0.949 5	0.921 6	0.935 3	0.940 0	0.810 3	0.870 4	0.921 6	0.903 8	0.912 6	0.705 9	0.827 6	0.761 9
耳鼻咽喉科学	0.957 6	0.882 7	0.918 6	0.944 3	0.868 6	0.904 8	0.944 3	0.857 6	0.898 8	0.921 8	0.813 8	0.864 4
心血管病学	0.909 2	0.965 7	0.936 6	0.885 0	0.955 0	0.918 7	0.882 5	0.953 9	0.916 8	0.825 1	0.960 1	0.887 5
神经病学	0.952 0	0.912 8	0.932 0	0.949 6	0.897 6	0.922 9	0.942 5	0.901 9	0.921 8	0.915 0	0.854 9	0.883 9
消化病学	0.940 5	0.963 2	0.951 7	0.913 0	0.946 9	0.929 7	0.912 5	0.953 4	0.932 5	0.888 6	0.900 4	0.894 5
宏平均	0.954 2	0.936 1	0.944 7	0.941 2	0.911 1	0.925 0	0.940 6	0.917 2	0.928 1	0.898 2	0.872 1	0.882 8



(a) BERT Plus模型分类结果混淆矩阵　(b) BERT模型分类结果混淆矩阵

(c) 随机森林 Plus算法分类结果混淆矩阵　(d) 随机森林算法分类结果混淆矩阵

图 3　各模型 10 个学科领域分类结果混淆矩阵

BERT 模型和随机森林算法通过特征增强后对于不同领域分类结果的召回率均有所提升，如呼吸病学两个模型分别提升 3 个百分点和 5 个百分点，消化病学两个模型分别提升 2 个百分点和 5 个百分点，证明经过特征增强后，文本表达语义信息更加清晰，误判为其他学科的概率降低。

另外，通过对比发现，BERT Plus 模型较随机森林 Plus 分类结果更好，*F*1 值提高了近 2 个百分点，见图 4，体现了 BERT 能够通过通用知识与领域知识相融合，更明显地表征语义信息，提升分类效果。

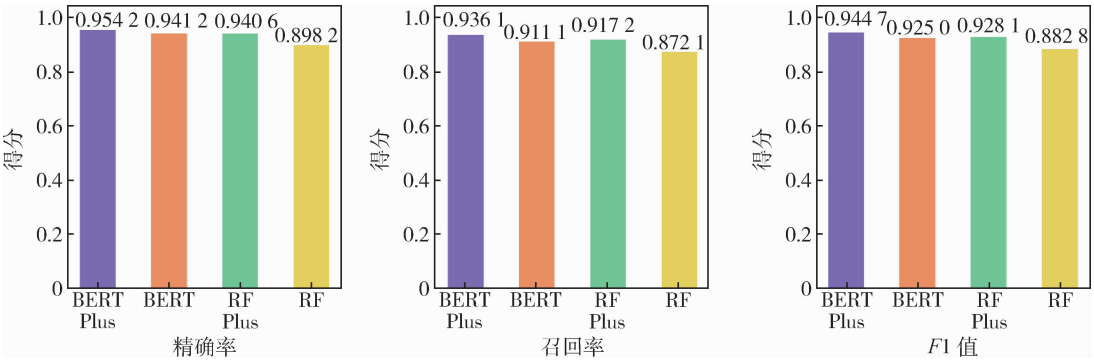


图 4 各模型分类总体结果对比

5 结语

本研究针对医学文献分类特征不足，设计特征增强方法，以医学科技文献为研究对象，利用 MeSH 同义词和上下位关系构建特征词表，进行向量化并与语义向量融合，计算与文献特征词的相似度，保留阈值内特征词。对比本研究方法与未进行特征增强的 BERT 分类结果，以及随机森林算法进行特征增强前后的分类结果。两组实验结果表明，特征增强后分类指标较好，精确率、召回率和 *F*1 值均有所提高，证明领域知识特征增强能更准确地表征语义信息，提升分类效果。本研究旨在构建更精准的文本分类方法，以加速医学文献组织，助力医学科技评价工作，减轻分类困难带来的人力、财力及时间浪费。受标注语料限制，未进行更细粒度分类，需进一步人工标注审核及实验。未来将继续探索利用深度学习、大语言模型优化文本分类方法。

作者贡献：宫小翠负责研究设计、数据分析、论文撰写；安新颖负责提供指导。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 吴霞，曾建勋，吴雯娜. 汉语主题词表生物、医学、农业领域顶层 MeSH 语义类型框架研究［J］. 情报科学，2022，40（1）：94－101.

2 DEVLIN J，CHANG M W，LEE K，et al. BERT：pre－training of deep bidirectional transformers for language understanding［C］. Minneapolis：The 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2019.

3 苏尤丽，胡宣宇，马世杰，等. 人工智能在中医诊疗领域的研究综述［J］. 计算机工程与应用，2024，60（16）：1－18.

4 HAUG C J，DRAZEN J M. Artificial intelligence and machine learning in clinical medicine，2023［J］. The New England journal of medicine，2023，388（13）：1201－1208.

5 刘晓明，李丞正旭，吴少聪，等. 文本分类算法及其应用场景研究综述［J］. 计算机学报，2024，47（6）：1244－1287.

6 LECUN Y，BENGIO Y，HINTON G. Deep learning［J］. Nature，2015，521（7553）：436－444.

7 VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need［C］. Long Beach：The 31st International Conference on Neural Information Processing Systems，2017.

8 CHANG W C，YU H F，ZHONG K，et al. Taming pretrained transformers for extreme multi－label text classification［C］. New York：The 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2020.

（下转第 67 页）

[J]. 数字图书馆论坛，2020（8）：7－14.
8 Neo4j [EB/OL]. [2024－09－01]. http://neo4j.com.
9 黄贺瑄，王晓燕，顾正位，等. 医学知识图谱构建技术及发展现状研究 [J]. 计算机工程与应用，2023，59（13）：33－48.
10 张璐璐，杨晟，史涪仁，等. 本体支持的生物医学领域元数据异质性与可兼容性研究 [J]. 中国生物医学工程学报，2019，38（3）：324－331.
11 张庆，吕少妮，轩扬. 本体在生物医学领域中应用研究热点分析 [J]. 医学信息学杂志，2019，40（1）：63－67.
12 于彤，崔蒙，李海燕，等. 中医药学语言系统的语义网络框架：一个面向中医药领域的规范化顶层本体 [J]. 中国数字医学，2014，9（1）：44－47.
13 贾李蓉，李海燕，刘静，等. 中医药学术语系统研究概述 [J]. 中国中医药图书情报杂志，2015，39（5）：7－10.
14 李思思，燕燕，夏书剑，等. 中医汤剂知识图谱的构建与样例查询方法研究 [J]. 中华中医药学刊，2024，42（8）：31－36.
15 叶其松. 汉语术语学术语体系构建的“不变”与“变”——从国际术语标准 ISO1087 说起 [J]. 中国科技术语，2024，26（1）：54－62.
16 唐晓波，王琼赋，牟昊. 基于词共现与词向量的概念层次关系自动抽取模型——以学术论文评价领域为例 [J]. 情报科学，2022，40（10）：3－11.
17 许冬冬，林惠，段会龙，等. 健康行为改变干预本体的构建与应用 [J]. 中国生物医学工程学报，2023，42（1）：74－81.
18 高峰，刘晴，靳英辉，等. 基于知识微调和信息融合的医学指南知识抽取 [J/OL]. 武汉大学学报（理学版），1－11 [2024－06－19]. https://doi.org/10.14188/j.1671－8836.2024.0032.
19 张宇，张舒琪，佟琳，等. 基于中医药学语言系统的中医药优势病种术语集构建研究 [J]. 中国中医药信息杂志，2024，31（5）：36－41.
20 祝振媛，李广建. 多视角下的情报分析模型研究综述 [J]. 图书情报工作，2019，63（19）：136－147.
21 范瑾研. 中医小儿哮喘诊疗知识本体构建研究 [D]. 长春：长春中医药大学，2024.
22 崔一迪. 寻常型银屑病中医诊疗本体知识库的构建与应用 [D]. 北京：中国中医科学院，2020.
23 张萌. 中医古籍疫病方剂知识库构建研究 [D]. 长春：吉林大学，2024.
24 李贺，祝琳琳，刘嘉宇，等. 基于本体的简帛医药知识组织研究 [J]. 图书情报工作，2022，66（22）：16－27.
25 林睿凡. 基于本体方法构建唐本《伤寒论》知识图谱 [D]. 北京：中国中医科学院，2022.
26 孙华君，李海燕，贾李蓉，等. 基于国际标准 ISO/TS16843－6：2022 的针刺效应本体构建研究 [J]. 中国针灸，2023，43（1）：73－78.

（上接第 41 页）

9 BELLO A, NG S C, LEUNG M F. A BERT framework to sentiment analysis of tweets [J]. Sensors, 2023, 23 (1): 506.
10 吴雪华，毛进，陈思菁，等. 突发事件应急行动支撑信息的自动识别与分类研究 [J]. 情报学报，2021，40（8）：817－830.
11 胡昊天，邓三鸿，王东波，等. 情报学视角下的预训练语言模型研究进展 [J]. 图书情报工作，2024，68（3）：130－150.
12 宫小翠，安新颖，单连慧. 基于 Labeled LDA 主题模型的医学文献自动分类法 [J]. 中华医学图书情报杂志，2018，27（10）：53－58.
13 YAO D, YANG J, ZHANG X. Feature selection algorithm based on random forest [J]. Journal of Jilin university (engineering and technology edition), 2014, 44 (1): 137－141.
14 张梦芸，丁敬达. 面向短文本分类的语义增强研究 [J]. 图书情报工作，2023，67（9）：4－11.
15 MeSH [EB/OL]. [2024－05－20]. https://meshb.nlm.nih.gov/.
16 唐焕玲，卫红敏，王育林，等. 结合 LDA 与 Word2vec 的文本语义增强方法 [J]. 计算机工程与应用，2022，58（13）：135－145.
17 罗宏宇，刘伟. 基于语义层级细粒度的海量文献标引研究 [J]. 情报理论与实践，2024，47（5）：194－203，193.
18 徐璐，卢小宾，杨冠灿. 金融科技专利识别与分类方法构建及应用 [J]. 图书情报工作，2020，64（11）：87－95.