基于DistilBERT的领域智库文本自动分类方法研究*

王小丫[1,2,3]　马建玲[1,2,3]
（1.中国科学院西北生态环境资源研究院，兰州 730000；2.甘肃省知识计算与决策智能重点实验室，兰州 730000；3.中国科学院大学经济与管理学院信息资源管理系，北京 100190）

摘要：随着国外智库成果数量的激增及其在政策研究与情报分析等应用场景中的参考价值日益凸显，高效处理与深度分析海量智库文本成为相关研究与信息分析工作的核心问题，其中实现智库文本的自动分类是提升信息检索效率、支撑后续深度分析的重要基础。本研究旨在构建面向领域的国外智库文本自动分类方法，为情报研究与分析提供技术支撑。针对目前领域智库文本标注数据稀缺、分类标准差异较大的问题，提出基于无监督主题建模与轻量化预训练语言模型微调相结合的混合自动分类技术框架：首先通过BERTopic实现无监督主题发现，构建数据驱动的分类标准；其次利用知识蒸馏模型DistilBERT进行领域自适应微调，实现对领域智库文本的高效准确分类。基于生物医药领域智库文本的分类实验结果表明，该方法在F1值（0.746 5）和召回率（0.750 0）上均优于多种基线模型，验证了其在领域智库文本分类任务中的有效性。

关键词：智库；智库文本；文本自动分类；主题建模；DistilBERT；生物医药
中图分类号：G250.7　DOI：10.3772/j.issn.1673-2286.2026.02.001
引文格式：王小丫，马建玲. 基于DistilBERT的领域智库文本自动分类方法研究[J]. 数字图书馆论坛，2026，22（2）：1-10.

近年来，全球智库数量快速增长。据《2020年全球智库指数报告》（2020 Global Go To Think Tank Index Report）统计，智库与公民社会项目（Think Tanks and Civil Societies Program，TTCSP）全球智库数据库中共收录11 175家智库，其中超过6 500家智库被纳入评估范围[1]。智库数量的激增伴随着影响力的持续攀升，其产出的海量报告、简报等成果已成为政策制定与决策支持不可或缺的知识来源[2]。智库成果形式多样，涵盖文本、音频、视频等多种媒介，而文本是各智库普遍采用、更易于系统化获取与分析的核心成果形态。随着智库成果数量的快速增长，高效处理海量文本，进而实现快速、准确的信息检索与分析，成为相关研究与信息分析工作的关键问题。因此，本研究以国外智库公开文本为研究对象，重点研究如何通过自动分类技术，有效解决智库文本的快速归类问题，从而提升政策分析与知识服务的效率。

在实际应用中，智库文本的自动分类面临多重挑战：①不同智库和数据库采用的分类方法与标准存在差异，为数据整合与分析带来困难；②智库文本具有较强的政策导向性、专业性与跨学科特征，增加了分类难度；③公开可用的标注语料相对有限，制约了深度学习方法的应用。为应对上述挑战，本研究提出一种无监督主题建模与轻量化预训练语言模型相结合的领域智库文本自动分类方法：首先利用无监督主题建模从语料中

收稿日期：2025-10-15
*本研究得到中国科学院文献情报中心项目“智库资源保障平台建设及资源采集服务”（编号：E340090701）资助。

归纳主题结构，为缺乏统一标准的领域智库文本构建可用的分类标准；然后基于轻量化预训练语言模型在少量标注样本上进行微调，实现小样本条件下的高效自动分类。本研究以生物医药领域智库文本为案例进行实证研究，以验证所提方法在小样本场景下的分类效能。

1 相关研究

作为服务于政策研究与战略决策分析的特殊文本类型，智库文本在内容、结构与应用场景上的独特性对文本自动分类方法构成了显著挑战。在内容专业性方面，智库文本的语言风格通常更为精练，政策指导性较强，且使用大量的领域专业性词汇，跨领域词汇交叉现象也较为常见[3]，这些内容上的独特性使得通用分类模型难以在智库文本上直接套用，构成了语义理解和分类建模的主要挑战。在结构复杂性方面，智库成果形式多样，既包括短小精炼的政策简报（Policy Brief）、摘要和声明（Abstract & Press），也涵盖篇幅较长、内容深入的研究报告（Research Report）、工作文件（Working Paper）等，且同类成果在不同智库之间也会存在明显的篇幅差异与风格差异。这些文本通常融合事实陈述、数据论证与政策建议等，呈现出多重逻辑层次。在应用场景方面，智库成果的根本目的在于支撑决策，具有强烈的政策导向性。由于智库成果通常属于灰色文献，缺乏统一的出版规范和元数据描述，其获取与组织方式更加复杂，进一步加剧了自动化处理的难度。

在分类标准方面，目前尚无统一的智库文本主题分类标准。不同智库数据库和检索平台通常根据自身研究重点和机构背景，采用差异较大的分类标准。标准的不一致性在跨机构或跨领域的数据整合、对比分析中造成明显障碍，并且目前智库文本分类技术多采用人工方式，分类效率低，无法实现大量智库文本的快速主题识别与分类。现有面向智库文本的自动分类研究较少，朱金箫等[4]针对智库文本主要为长文本的特点，提出的基于自注意力的层次结构文本分类模型SHTC（Hierarchical structure model based on self-attention）在智库文本分类任务中准确率达到82.43%，表现出更优的分类效果。但该研究主要针对国内智库文本，对于国外智库（主要是以英文为官方语言的智库）的文本分类效果尚未有明确评估。

文本分类作为自然语言处理领域的基础任务，其核心目标是将文本数据分配至预定义类别体系中，该技术在信息检索和内容推荐等领域发挥着重要作用。深度学习模型不依赖人工获取的文本特征，可以直接对文本内容进行学习、建模[5]，与传统机器学习相比在多种文本分类任务中表现更为出众[6-7]。在此基础上，以BERT（Bidirectional Encoder Representations from Transformers）、ERNIE（Enhanced Representation through Knowledge Integration）等为代表的预训练语言模型在文本分类任务中的应用为处理上下文关系和语义理解提供了新的思路。例如：祝思炬[8]结合ERNIE、注意力机制和深度金字塔卷积网络，优化了新闻文本的分类效果；施国良等[9]使用融合RoBERTa和文本增强技术的方法，有效提升了网络问政留言分类模型的F1值，并通过对比证实了该方法的有效性。

针对特定领域和特定语言特性的文本分类，研究者们结合领域知识和数据特征提出针对性方法有效地解决了实际问题：张典等[10]针对农业领域文本特征提出了基于MacBERT（MLM as correction-BERT）、深度金字塔卷积网络和注意力机制的农业文本分类模型，在农业领域文本分类任务中表现优越；武同心等[11]提出了结合文本卷积网络和注意力机制的电力文本分类模型，提高了电力文本分类的效率；李江涛等[12]针对医疗信息敏感程度的不同，提出了大小模型融合分类的方法来完成医疗数据分类及加密工作。

综上，尽管通用与领域化的文本分类方法不断进步，但面对智库文本的自动分类，仍存在明显的研究空白：如何为缺乏统一分类标准的领域智库文本构建分类标准，以及如何在标注资源稀缺的条件下实现高效的文本分类。本研究提出了一种结合BERTopic和DistilBERT的智库文本分类方法，该方法能够通过无监督主题建模构建不同智库领域的独特分类标准，并借助轻量化预训练语言模型的迁移学习能力，通过在少量标注数据上微调以实现高效的文本分类，为实现小样本场景下的领域文本自动分类提供一种可行的解决方案。

2 领域智库文本自动分类技术框架

2.1 自动分类技术框架设计

智库文本通常具有专业术语集中、政策导向鲜明、跨学科交叉的特点，对分类方法的语义理解能力提出了较高的要求。本研究提出结合无监督主题建

模与轻量化预训练语言模型的两段式技术框架（见图1），既强调对文本语义的准确把握，又兼顾资源受限环境下的适用性与可扩展性。文本数据预处理包括去噪、分词、统一大小写和去停用词等步骤，为后续分类任务提供了基础文本。第一阶段：通过BERTopic模型对文本语料进行无监督主题建模，并结合主题间距离图与层次聚类图结果，对初步生成的主题类别进行人工判读与主题合并，同时参考权威数据分类标准，最终形成明确且可用于监督学习的领域智库文本分类标准。第二阶段：采用DistilBERT模型，完成文本预处理、模型构建与微调训练，并通过多项评估指标验证模型在智库文本分类任务中的有效性。

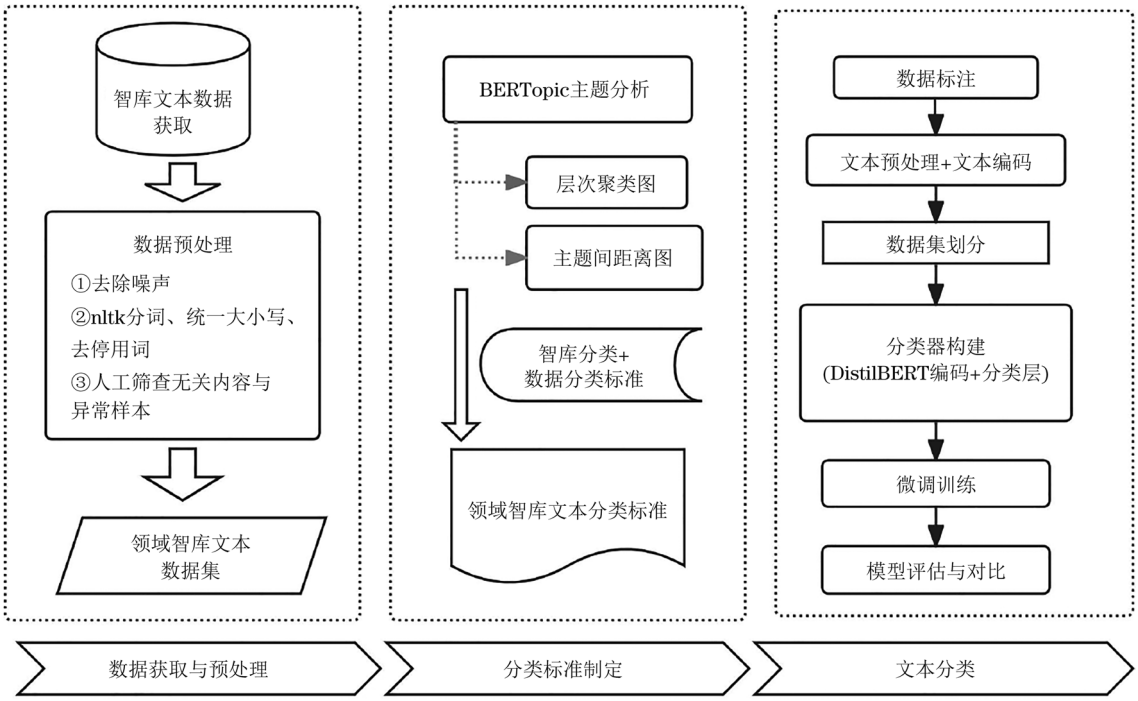

图1　技术框架

2.2　构建领域智库文本分类标准

BERTopic[13]通过三个阶段技术流程实现客观分类：首先基于Sentence-BERT获取具有上下文语义的文本嵌入，然后结合UMAP降维与HDBSCAN聚类实现主题划分，最后使用c-TF-IDF从每个聚类簇中提取关键词生成可解释性的主题表示。这一流程不仅适配智库文本的内容特征，也能在长文本（如研究报告）与短文本（如政策简报）的主题识别中保持良好的性能[14]。本研究在制定特定领域智库文本的分类标准过程中采用了BERTopic主题建模方法，使分类标准能够更客观地反映智库文本的内在结构特征。

为确保分类标准的科学性，本研究在BERTopic主题分析结果的基础上对主题进行整理。一方面，基于主题向量相似度计算，采用AGNES算法迭代合并相近主题，生成层次聚类图实现冗余主题压缩；另一方面，基于UMAP降维将主题向量投射到二维空间，生成主题间距离图，用于快速定位核心主题或离群主题，直观呈现议题聚合与分化趋势。综合两种可视化结果，确保分类标准既贴合数据特性，又具备多维度分析视角。主题合并与命名参考了领域智库分类体系、智库对不同主题内容的定义及相关政策规范。

2.3　智库文本分类模型与微调

DistilBERT是由Sanh等[15]提出的一种基于知识蒸馏（Knowledge Distillation）技术的轻量化预训练语言模型，其核心目标是在保留BERT[16]97%语义理解能力的前提下，显著降低模型的计算成本。DistilBERT用于完成文本分类任务的效果得到了多方验证：该模型在GLUE基准测试中表现显著优于ELMo基线[15]；Silva等[17]和Akpatsa等[18]的研究进一步证实，DistilBERT在处理双语种文本分类和新闻文本分类任务中，不仅准确率表现优异，还具备运行时间短、存储成本低的优势。

本研究选用DistilBERT作为分类器，主要出于对任务特性与模型适配性的双重考量：一方面，智库文本通常具有专业术语密集、标注数据稀缺以及跨领域语义复杂等特征，这要求分类模型能够在有限样本下实现高效的语义理解，因此借助预训练语言模型从大规模语料库中学习到的通用语言表征，使用微调方法来适配垂直领域任务，同时轻量化模型更适用于小样本场景；另一方面，尽管生成式大语言模型在诸多任务上表现突出，但其核心目标为文本生成而非文本分类这类判别式任务，在直接用于分类任务时可能出现稳定性和可靠性不足的问题。基于上述考虑，本研究回归深度学习中的轻量化预训练语言模型，在分类效果与计算效率之间寻求平衡。

本研究采用的文本分类流程如图2所示。①构建标注语料库，对采集的数据实施人工标注、文本预处理与文本编码、数据集划分；②加载预训练语言模型，选用Hugging Face提供的DistilBertForSequenceClassification模型执行分类任务，该模型架构由特征提取层及分类层构成；③使用标注语料库对模型实施微调（Fine-tuning），通过训练更新模型参数；④生成适用于目标领域的分类模型。此方法直接利用预训练DistilBERT模型已习得的通用语言知识，通过微调使其适配特定分类任务，在显著降低计算资源消耗的同时，保障了分类性能。

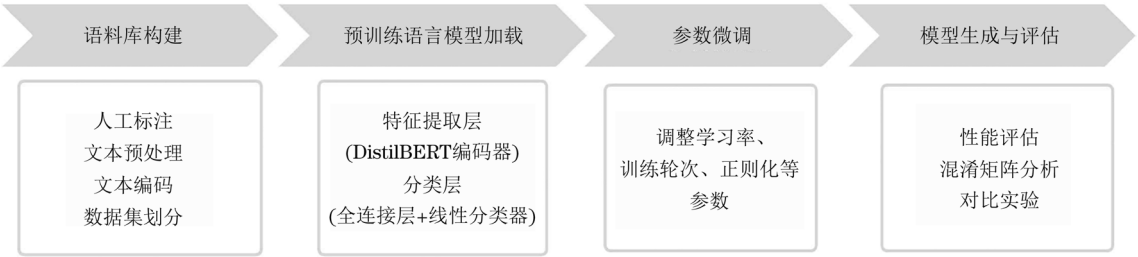

图2　文本分类流程图

3　生物医药领域智库文本自动分类

3.1　数据来源

本研究以生物医药领域智库文本为例进行研究，数据来源于TTCSP发布的《2020年全球智库指数报告》中Global Health与Domestic Health两个榜单。为确保数据集的质量与数据采集的可行性，本研究以机构官网稳定可访问、研究成果持续更新，且在公共卫生或者生物医药领域具有代表性为主要筛选标准，最终选取了14家智库作为研究对象。这些智库既包括美国的兰德公司（RAND Corporation）、全球发展中心（Center for Global Development），英国的海外发展研究所（Overseas Development Institute）等具有全球影响力的综合性智库，也涵盖了澳大利亚格拉坦研究所（Grattan Institute）、亚太卫生系统和政策观察站（Asia Pacific Observatory on Health Systems and Policies）等区域特色鲜明的智库，为研究提供了主流且高质量的文本基础。为确保时效性，数据选取时间范围限定在2021—2025年，涵盖各机构官网公开发布的报告（Report）、政策简报（Policy Brief）、学术论文（Paper）及智库活动记录（Event）等多类型成果。文本内容包括智库成果的标题、摘要、关键词和全文。在数据筛选过程中，结合各智库官网的检索分类功能初步定位生物医药相关文本，并通过人工审核对文本主题相关性作进一步确认以提高数据集的准确性，最终构建包含1 182条非均匀长度文本的专题语料库，该数据集覆盖了从标题到全文的不同信息粒度，能够较好地反映智库成果在体裁、篇幅和语义特征上的差异性。

3.2　文本分类标准构建

本研究使用BERTopic进行主题建模，核心参数设置如下：采用all-MiniLM-L6-v2模型生成文本嵌入；通过UMAP算法将语义向量降至8维；使用HDBSCAN进行聚类，最小聚类簇大小设为10；主题数量控制参数设为None（不自动合并主题），以确保初步发现的语义结构得以保留；最终通过c-TF-IDF算法提取各主题的关键词。本研究设置上述参数，旨在优先发现数据中潜在的多样主题，为后续基于层次聚类图与主题间距离图的可视化结果开展语义辨析和主题合并工作奠定

基础。

应用BERTopic主题建模方法[13]对上述数据集进行主题分析后共识别出19个主题，主题及其对应关键词概率分布如图3所示。

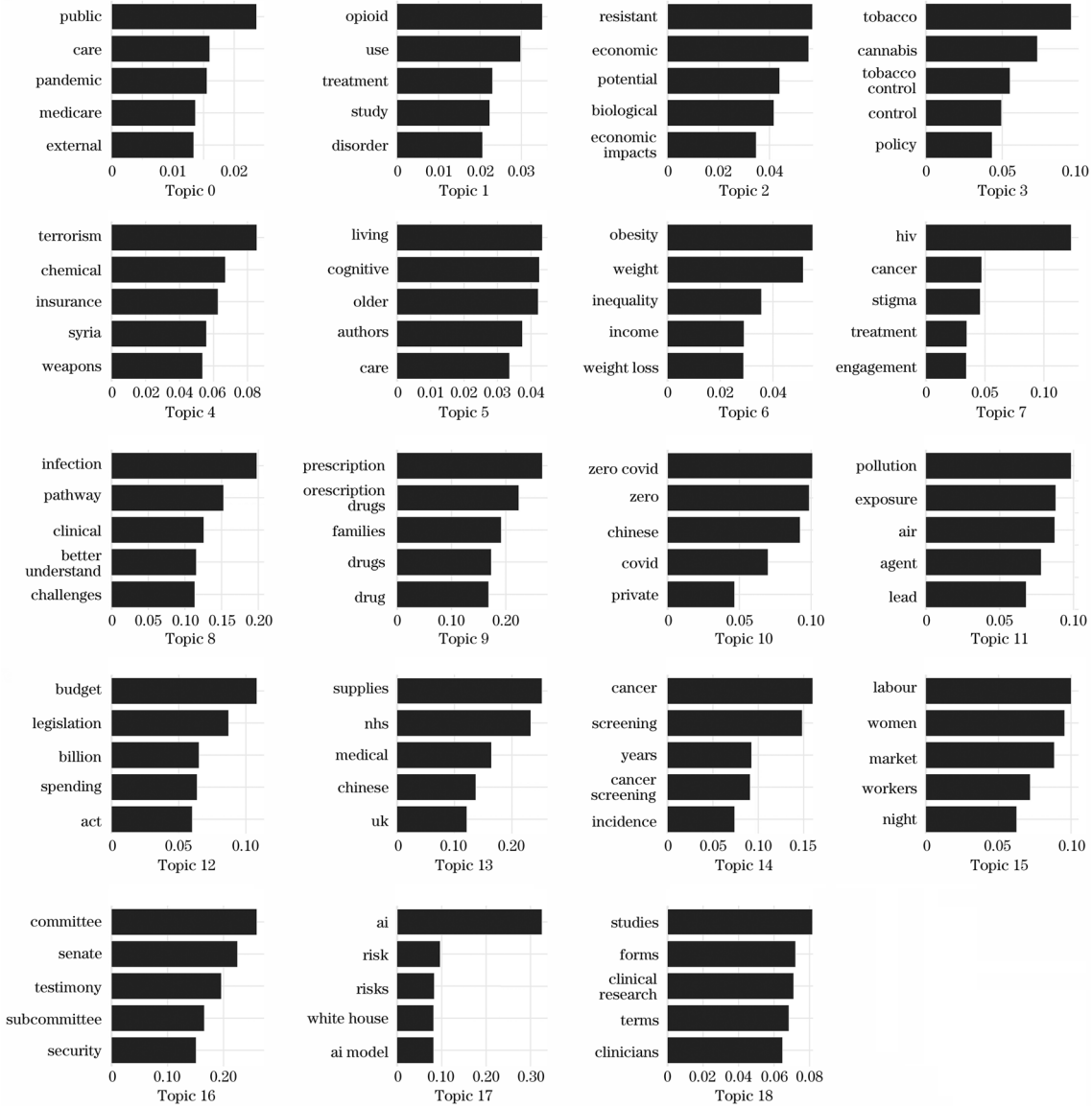

图3　主题-关键词概率分布图

结果显示，不同主题在语义空间中呈现出一定的聚集性，但同时存在部分主题边界模糊、语义交叉的情况。这与智库文本跨学科、政策导向性强的属性有关，不同主题反复使用相同的领域高频词汇从而造成语义边界模糊。如果完全依赖自动聚类，可能出现语义差异较大的文本被错误聚合，或语义接近的文本被过度分散的情况。因此，本研究在数据驱动的基础上加入人工矫正，对初始主题进行合并：首先依据BERTopic的层次聚类图（见图4）和主题间距离图（见图5）识别自然成组的主题（如Topic16、Topic17）；其次对于聚类有异议的主题回归文本本身进行语境判读（如Topic9、Topic12）；最后参考权威分类标准，对合并后的主题进行统一命名与规范化界定。

最终，19个主题被整合为8个主要类别，由这8个类别确立的分类标准如表1所示。类别命名主要参考宾夕法尼亚大学《2020年全球智库指数报告》中的智库分类标准、生物医药领域智库对部分主题的界定以及美国《信息类型和信息系统与安全类别映射指南》（Guide for Mapping Types of Information and Systems to Security Categories）[19]。由于多数智库以政策研究和

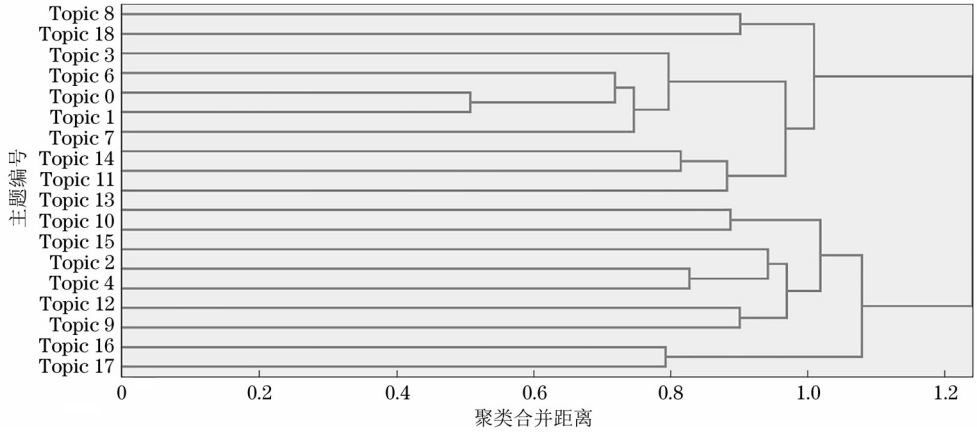

图4　层次聚类图

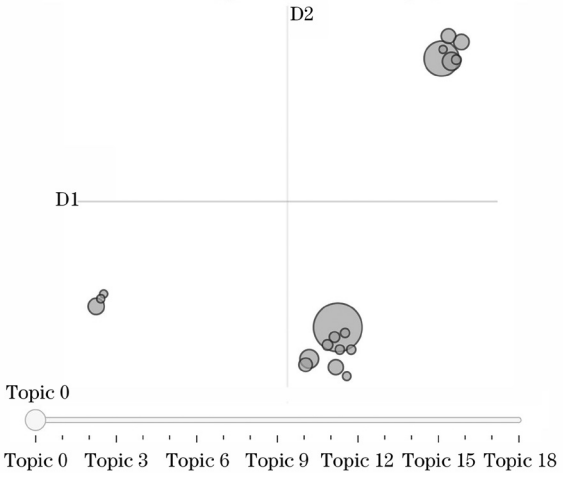

图5　主题间距离图

政策建议为核心职能，具有明显的政策导向性，因此即使只有Topic 12与政策相关，本研究仍然保留了这一类别。同时为应对不同智库分类及检索系统不统一可能导致的噪声数据干扰问题，以及考虑潜在的、新兴小众的主题文本出现的情况，在本研究的分类标准中加入了“其他”类别进行补充。

上述分类方法与两种常见分类思路存在明显差异。第一，区别于依赖智库地域、类型等聚焦外部特征的划分方式，本研究提供了一种跨机构的、统一的内容分类方法，使不同智库的成果能在同一语义框架下进行比较与整合；第二，区别于医学主题词表（MeSH）等追求学科知识全覆盖的通用体系，本研究提出的是一种问题导向与政策聚焦的方法，直接基于智库文本的语义表示，通过聚类分析生成分类类别，旨在刻画智库文本本身所回应的现实政策关切与复合型社会议题；第三，该方法具有动态可扩展的潜力，数据驱动的方法既有助于削弱单纯依赖专家经验可能带来的主观偏差，也能通过“其他”类观察窗口和重新聚类流程识别并吸纳新兴议题，从而适应研究热点的动态演变。

3.3　基于DistilBERT的文本分类实现

高质量的文本数据是进行文本分类的基础，因此在训练文本分类模型之前需要对数据进行预处理。

（1）数据标注。考虑到智库文本语义的复杂性，数据标注采用人工标注进行。在标注过程中对各类别

表1　分类标准

类别	包含主题	依据
疫苗与传染病防控	Topic 8、Topic 18	文献[20-21]
生物安全与风险管理	Topic 4、Topic 16、Topic 17	文献[22]
经济与贸易（与生物医药领域相关）	Topic 2、Topic 10、Topic 13	基于主题关键词与文本侧重点
公共卫生及干预措施	Topic 0、Topic 6、Topic 9	文献[23]
卫生服务与管理	Topic 1、Topic 3、Topic 5、Topic 7、Topic 14	文献[19]
生物医学基础研究	Topic 18	基于主题关键词与文本侧重点
政策分析与评论	Topic 12	基于主题关键词与文本侧重点
其他	Topic 11、Topic 15	基于主题关键词与文本侧重点

数据集的边界进行规范，数据集分类及文本示例如表2所示。标注后发现疫苗与传染病防控、生物安全与风险管理、经济与贸易、生物医学基础研究这4个类别的数据数量较少，所以采用了数据增强策略对这些类别的数据进行了小批量的补充，以提升文本分类性能。具体采用了两种方法：一是同义词替换，即替换原文中部分非关键术语（主要是形容词、副词等）；二是同义改写，即在保持关键术语与文本原意的前提下调整句式与表达。研究对采用数据增强策略的原始文本严格控制，每条生成1~2条文本，在确保数据多样性的同时最大程度维持文本的原始语义与质量。此策略在文本分类任务中，特别是在较小数据集上的性能已得到验证[24]。最终各个类别的数据都保持在130~180条，共得到1 280条数据用于模型微调。

表2　数据集分类及文本示例

类别	分类说明	文本示例原文	文本示例译文
疫苗与传染病防控	疫苗（研发、接种、资源分配等），新发传染病与疫情的监测与应急响应	Lately, stories about outbreaks seem to be spreading faster than the diseases themselves. An outbreak of measles in Ohio is just part of an 18-year high of U.S. cases. Meanwhile, polio continues to circulate in Pakistan, Afghanistan, and Nigeria, while spreading to other countries, like Cameroon, Equatorial Guinea, and Syria	近来，关于疫情暴发的消息似乎比疾病本身传播得还要迅速。俄亥俄州暴发的麻疹疫情，仅是美国病例数创下18年来新高这一整体趋势中的一部分。与此同时，脊髓灰质炎（小儿麻痹症）仍在巴基斯坦、阿富汗和尼日利亚持续传播，并正向喀麦隆、赤道几内亚和叙利亚等其他国家蔓延
生物安全与风险管理	生物实验室安全标准，生化武器的防控，基因编辑技术等技术的伦理问题或风险	Ensuring robust biosafety protocols in biological laboratories has become a priority as the potential for accidental pathogen release remains a constant threat. Strengthening laboratory containment systems, as well as increasing regulatory oversight, is essential to mitigate risks associated with high-risk biological experiments	确保生物实验室拥有健全的生物安全规程已成为当务之急，因为意外病原体释放的风险始终存在。加强实验室防护系统以及加大监管力度，对于降低高风险生物实验的相关风险至关重要
经济与贸易（与生物医药领域相关）	生物医药产业链，生物医药及相关领域的出口管制与经济制裁，医疗设备关税，生物技术资金投入与产业化应用等	This paper studies the impact of the first joint licensing platform for patented drugs, the Medicines Patent Pool, on global drug diffusion and innovation. The pool allows generic firms worldwide to license drug bundles cheaply and conveniently for sales in a set of developing countries	本文研究了首个专利药品联合许可平台——药品专利池对全球药品扩散和创新的影响。该专利池允许全球仿制药企业以低成本、便捷的方式获得药品组合许可，并在特定的一组发展中国家销售
公共卫生及干预措施	烟草与酒精控制政策，药物滥用防治，慢性病预防与管理制度	Patterns and consequences of opioid use are changing dramatically. Researchers provide a nuanced assessment of America's opioid ecosystem, highlighting how leveraging system interactions can reduce addiction, overdose, suffering, and other harms	阿片类药物的使用模式和后果正在发生巨大变化。研究人员对美国的阿片类药物生态系统进行了细致入微的评估，重点阐述了如何利用系统互动来减少成瘾、过量用药、痛苦和其他危害
卫生服务与管理	医疗保健服务，人口健康管理，特殊人群的医疗保健需求；安全有效的医疗产品及设备的研发与测试	We exploit a quasi-experiment in the province of Alberta, Canada, to identify how changes in the schedule of medical benefits affected the provision of primary care services to patients with multiple co-morbidities	本文借助加拿大阿尔伯塔省开展的一项准自然实验，分析医疗福利目录的调整对多重共病患者基层医疗服务供给情况产生的影响
生物医学基础研究	包含实验室研究、临床前药物研发、动物实验、生物信息分析等技术研发或实验报告	Antimicrobial drugs form the backbone of modern medicine. Yet over time, use of these drugs selects for mutations that survive exposure to those same drugs, driving "antimicrobial resistance" or AMR	抗菌药物是现代医学的基石。然而，随着时间的推移，这些药物的使用会筛选出能够耐受这些药物的突变菌株，从而导致"抗菌素耐药性"（AMR）
政策分析与评论	跨领域政策评估（如技术-经济-伦理平衡）、智库观点报告、战略规划、政策法律动态	Epidemics and pandemics pose a sporadic and sometimes severe threat to human health. How should policymakers prioritize preventing and preparing for such events, relative to other needs	流行病与大流行对人类健康构成间歇性威胁，且有时危害严重。在各类社会需求中，政策制定者应如何优先安排此类事件的预防与应对工作
其他	涉及多主题或新兴小众议题等无法明确归类的文本	Susan Chemerynski, ScD, MPH, joined the Department of Health Policy and Management in August as an associate professor and director of the Risk Sciences and Public Policy Institute	苏珊·切梅林斯基博士，公共卫生硕士，于8月加入卫生政策与管理系，担任副教授兼风险科学与公共政策研究所所长

（2）文本预处理与编码。本研究采用Hugging Face提供的DistilBertTokenizer对文本进行分词与编码，该分词基于子词级的字节对编码算法，将句子切分为子词单元并转换为模型可接受的数字序列表示。最大序列长度设定为512个词元（token），以兼顾长文本信息与计算效率。

（3）数据集划分。数据集按8∶1∶1划分为训练集（80%用于模型训练）、验证集（10%用于超参数调优和模型选择）和测试集（10%用于最终性能评估）。

（4）智库文本分类。本研究采用DistilBERT-base-uncased模型完成智库文本分类任务，该模型在保持BERT核心性能的同时显著降低了计算成本。实验利用Hugging Face的Trainer API实现模型训练，经多轮实验确定最优训练参数：训练轮次（num_train_epochs）为5，学习率（learning_rate）为5e-5，权重衰减系数（weight_decay）为0.01，批次大小（per_device_train_batch_size）为8。

3.4 实验结果与分析

本研究使用精确率（Precision）、召回率（Recall）、F1值3个指标来评价模型效果。为了验证本研究所提DistilBERT模型用于文本分类的效果，选择了5个模型作为对照组，包括传统机器学习方法TF-IDF+SVM，深度神经网络模型TextCNN、RCNN，以及基于预训练语言模型的BERT-base、DistilRoBERTa，其中预训练语言模型均采用微调加分类范式。为保证不同预训练语言模型的可比性，实验在相同数据划分策略（训练集、验证集与测试集按 8∶1∶1划分）下进行，并在文本最大长度、训练轮数、学习率、权重衰减等关键训练参数上保持一致，仅替换预训练语言模型以考察模型差异对分类效果的影响。各模型分类效果评估结果如表3所示。

表3 各模型分类效果评估结果

模型	精确率	召回率	F1
TF-IDF+SVM	0.682 5	0.584 1	0.601 7
TextCNN	0.557 0	0.376 9	0.381 2
RCNN	0.648 6	0.595 8	0.602 1
BERT-base	0.702 8	0.686 0	0.670 8
DistilRoBERTa	0.722 5	0.698 7	0.704 9
DistilBERT	0.761 7	0.750 0	0.746 5

从整体表现来看，基于预训练语言模型的方法在各项指标上优于传统机器学习方法和深度神经网络模型。RCNN相比TF-IDF+SVM在召回率和F1值上有所提升，而TextCNN在本实验中的表现明显低于TF-IDF+SVM，表明其在本数据集上的适应性相对有限，这可能与TextCNN依赖固定卷积窗口、在长文本语义建模能力方面存在一定局限有关；RCNN虽引入循环结构，但整体表现仍不及预训练语言模型，反映了其在语义复杂文本处理任务中存在一定局限。

在预训练语言模型中，DistilBERT模型在精确率、召回率与F1值上均取得了最优效果，相较BERT-base的F1值（0.670 8）提升约11.29%，这可能与DistilBERT通过知识蒸馏在减少模型参数的同时保持较高性能有关[15]；DistilRoBERTa的整体表现低于DistilBERT、略高于BERT-base。DistilRoBERTa各项结果未能超过DistilBERT，这可能与其预训练语料领域分布与本研究使用的数据（生物医药领域智库文本）存在差异有关。整体而言，蒸馏模型DistilBERT在本研究场景下展现了更优的综合性能，这得益于知识蒸馏技术在小样本条件下的泛化能力以及对过拟合风险的有效抑制。

本研究在此基础上进行了消融实验，以探究数据集中文本长度差异对分类性能的影响。在DistilBERT模型设置保持不变的前提下，将测试集样本按长度划分为短文本（1~64词）、中等长度文本（65~256词）和长文本（大于256词），并分别验证不同长度区间的文本分类效果。为了避免样本数量较多的类别对整体结果造成影响，实验使用了宏平均F1（macro-averaged F1-score）作为主要评估指标。

不同长度文本分类效果对比如表4所示。其中短文本和中等长度文本的分类性能明显优于长文本，其宏平均F1值分别为0.666 2、0.660 2，而长文本仅为0.568 2。这一结果表明在本数据集与模型设置下，文本长度差异对分类效果产生了较为明显的影响，尤其是长文本的处理更加复杂，可能导致模型在此类文本上的特征提取与判别能力较差；短文本和中等长度文本的语义结构相对简洁清晰，更有利于模型完成有效分类，而长文本往往因包含信息层次更复杂多元，从而增加了分类的难度。

表4 不同长度文本分类效果对比

	Overall	短文本	中等长度文本	长文本
Macro-averaged Precision	0.667 5	0.695 8	0.715 2	0.576 0
Macro-averaged Recall	0.648 0	0.668 7	0.756 9	0.616 1
Macro-averaged F1-score	0.645 4	0.666 2	0.660 2	0.568 2

4　结论与展望

本研究针对智库文本的特点提出一种融合BERTopic与DistilBERT的混合方法实现领域智库文本自动分类，并开展实验研究从方法流程层面证实了该框架的可行性与有效性。基于1 280条生物医药领域智库文本的分类实验结果表明，DistilBERT模型的F1值达到0.746 5，与BERT-base模型的F1值（0.670 8）相比提升约11.29%。数据标注及文本分类实验结果验证了分类标准的有效性，也证明了知识蒸馏模型DistilBERT在小样本场景下的优势：压缩模型参数量（减少40%）、保留核心语义理解能力、降低过拟合风险，为领域智库文本自动分类提供了一套有效的技术方法。此外，本研究虽以生物医药领域智库文本为实证研究对象，但研究重点在于验证所提出文本分类方法的可行性，并非针对单一领域构建专用模型。

本研究仍存在一定局限性：①如实验结果显示，数据集中文本长度差异对分类效果有较为明显的影响；②在分析混淆矩阵结果时发现，部分类别的分类表现相对其他类别偏低，在一定程度上拉低了模型整体F1值；③在人工数据标注过程中发现，部分文本的分类表现出较强的主题交叉性和语义复杂性，智库文本鲜明的政策导向性使得许多文本在阐述具体内容时会自然融入政策视角的分析，因此根据其文本侧重点将其归入单一类别可能会影响实验结果的客观性和准确性。

在此基础上，未来研究可从以下方面展开：①探索面向长文本的分段编码策略或引入注意力机制等方法，以提升模型对不同长度文本的建模能力；②尝试构建多标签分类方法，例如修改模型输出层并设定置信度阈值、与多标签分类适配的神经网络模型结合进行模型改进等，以更准确捕捉智库文本的多主题交叉特征，避免单标签分类造成的信息损失；③随着大语言模型的飞速发展，未来可结合更多、更高质量的垂直领域语料开展针对性微调，并与轻量化预训练语言模型进行对比，以进一步探讨不同模型在小样本场景下的适用性。

参考文献

[1] MCGANN J G. 2020 Global Go To Think Tank Index Report [EB/OL]. [2025-10-14]. https://repository.upenn.edu/entities/publication/9f1730fa-da55-40bd-a1f4-1c2b2346b753.

[2] ABELSON D E. Old world, new world: the evolution and influence of foreign affairs think-tanks[J]. International Affairs，2014，90（1）：125-142.

[3] 张誉曜. 关键词视角下的美国智库文本研究：以人工智能领域为例[D]. 乌鲁木齐：新疆师范大学，2021.

[4] 朱金箫，陈媛媛. 基于深度学习的智库文本分类研究[J]. 情报杂志，2022，41（8）：179-184.

[5] 汪家伟，余晓. 基于深度学习的文本分类研究综述[J]. 电子科技，2024，37（1）：81-86.

[6] LI Q，PENG H，LI J，et al. A survey on text classification: from traditional to deep learning[J]. ACM Transactions on Intelligent Systems and Technology，2022，13（2）：1-41.

[7] MINAEE S，KALCHBRENNER N，CAMBRIA E，et al. Deep learning-based text classification: a comprehensive review[J]. ACM Computing Surveys，2021，54（3）：1-40.

[8] 祝思炬. 基于ERNIE与注意力特征融合的新闻文本分类研究[D]. 武汉：武汉工程大学，2024.

[9] 施国良，陈宇奇. 文本增强与预训练语言模型在网络问政留言分类中的集成对比研究[J]. 图书情报工作，2021，65（13）：96-107.

[10] 张典，刘畅，陈雯柏，等. 融合注意力机制的MacBERT—DPCNN农业文本分类模型[J]. 中国农机化学报，2025，46（8）：83-89.

[11] 武同心，纪鑫，杨成月，等. 基于双层注意力机制的电力文本分类模型[J]. 中国电力，2025，58（11）：156-163.

[12] 李江涛，马礼，李阳. 基于大小模型融合的医疗数据分类方法[J/OL]. 计算机工程：1-11[2025-09-24]. https://doi.org/10.19678/j.issn.1000-3428.0070408.

[13] GROOTENDORST M. BERTopic: neural topic modeling with a class-based TF-IDF procedure[EB/OL]. [2025-05-11]. https://arxiv.org/abs/2203.05794.

[14] 卢文昊，邹杭纳，汤颖. BTVis：基于BERTopic的交互式层次主题建模可视分析系统[J]. 计算机辅助设计与图形学学报，2025，37（12）：2112-2130.

[15] SANH V，DEBUT L，CHAUMOND J，et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [EB/OL]. [2025-05-04]. https://arxiv.org/abs/1910.01108.

[16] 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. Minneapolis：Association for Computational Linguistics，2019：4171-4186.

[17] SILVA B R，AKABANE A T. Towards transfer learning techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for automatic text classification from different languages: a case study[J]. Sensors，2022，22（21）：81-84.

[18] AKPATSA S K，LEI H，LI X，et al. Online news sentiment classification using DistilBERT[J]. Journal of Quantum Computing，2022，4（1）：1.

[19] Guide for Mapping Types of Information and Systems to Security Categories[EB/OL]. [2025-10-14]. https://csrc.nist.gov/publications/sp800.

[20] Brown University School of Public Health. Infectious Disease[EB/OL]. [2025-10-14]. https://epidemiology.sph.brown.edu/research/areas/infectious-disease.

[21] Gates Foundation. Vaccine Development and Surveillance [EB/OL]. [2025-10-14]. https://www.gatesfoundation.org/our-work/programs/global-health/vaccine-development-and-surveillance.

[22] What we do[EB/OL]. [2025-10-14]. https://internationalbiosafety.org/who-we-are/what-we-do/.

[23] Bloomberg Philanthropies. Public Health[EB/OL]. [2025-10-14]. https://www.bloomberg.org/public-health/.

[24] WEI J，ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong：Association for Computational Linguistics，2019：6382-6388.

作者简介

王小丫，女，硕士研究生，研究方向：知识发现与知识系统。
马建玲，女，硕士，硕士生导师，研究馆员，通信作者，研究方向：知识发现与知识系统，E-mail：majl@lzb.ac.cn。

Research on the Automatic Classification Method of Domain Think Tank Texts Based on DistilBERT

WANG XiaoYa[1,2,3] MA JianLing[1,2,3]
(1.Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, P.R. China; 2.Key Laboratory of Knowledge Computing and Intelligent Decision, Gansu Province Lanzhou 730000, P.R. China; 3.Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, P.R. China)

Abstract: With the surge in the number of foreign think tank results and their reference value in application scenarios such as policy research and intelligence analysis, efficient processing and in-depth analysis of massive think tank texts have become core issues in related research and information analysis. Among them, automatic classification of think tank texts is an important foundation for improving information retrieval efficiency and supporting subsequent in-depth analysis. This research aims to construct a field-oriented automatic classification method for foreign think tank texts to provide technical support for intelligence research and analysis. In view of the current problems of scarcity of text annotation data and large differences in classification standards in domain think tanks, a hybrid automatic classification technology framework based on a combination of unsupervised topic modeling and lightweight pre-trained language model fine-tuning is proposed: First, unsupervised topic discovery is realized through BERTopic to build a data-driven classification standard; then the knowledge distillation model DistilBERT is used for domain adaptive fine-tuning to achieve efficient and accurate classification of domain think tank texts. Experimental results on think tank texts in the field of biomedicine show that this method is superior to various baseline models in both F1 value (0.746 5) and recall rate (0.750 0), verifying its effectiveness in text classification tasks in domain think tanks.

Keywords: Think Tank; Think Tank Text; Automatic Text Classification; Topic Modeling; DistilBERT; Biomedicine

（责任编辑：王玮）